Asymptotically pivotal transforms of residuals sample autocorrelation for weak VARMA models

Maciej Opuchlik, UC3M

November 10, 2013

Abstract

We propose a method for deriving a distribution free-transform of empirical autocorrelations of residuals in case of VARMA models. Statistics being functions of residuals are basic tools for testing for lack of correlation of model errors. The problem being adressed is that asymptotically the residuals obtained after estimating a model are dependent also on the estimation error which affect each covariance of lag h in a different, yet significant manner. We are also weakening the H_0 hypothesis from iid innovations to lack of correlation of the innovations. It is presented how the vectorized autocovariance matrix of lags $1, \ldots, H$ may be standarized by its asymptotic covariance matrix such that proposed pivotal transform causes the resulting vector of autocorrelations to be asymptotically a vector of standard normals. Our method does not specify the number of covariances being transformed and how many of them are going to be applied in Box-Pierce test, it does not require a specific estimator of parameters of VARMA model and the estimation of covariance matrix of parameters estimators is not necessary. The finite sample results are being examined by a Monte Carlo experiment performed for classical Box-Pierce test of lack of correlation for iid and weakly dependent innovations in case of VAR and VARMA generated data. The method proposed is compared to the alternative methods that are currently available to solve the problem of asymptotically significant estimation error in residual autocorrelations.

Keywords: Goodness-of-fit test, higher order serial dependence, model checking, portmanteau Ljung-Box and Box-Pierce, recursive residuals, recursive derivatives of residuals, residual autocorrelation, weak VARMA models.

1 Introduction

In a general parametric time series framework functions of residuals are a key tool for model checking. In case of testing lack of correlation, the standard method is based on using statistics of empirical residuals autocorrelations. Usually portmanteau tests like Ljung-Box (1976) and Box-Pierce (1970) are being utilized. The same applies for vector autoregressive (VAR) and vector autoregressive moving average models (VARMA) that are a standard econometric tool used for macroeconomic data. This class of models is a natural expansion of univariate ARMA models and thus they were extensively studied during the 90's (see *e.g.* Lütkepohl, 1993).

The motivation of VARMA models is based on the fact that they are able to effectively represent multivariate economic data by allowing for empirically evident autocorrelation structures, that can not be achieved by using finite order VAR models. VARMA models may be usually found in financial models, because autocorrelation of shocks have a natural interpretation. As far as macroeconomic data is concerned VAR models are less parsimonious than VARMA models and are not closed to simple subvector transforms like marginalization and aggregation which lead may from VAR to VARMA model (see *e.g* Dufour, Palletier, 2008). Among other fields, vector autoregressive model checking tools, especially goodness-of-fit testing of empirical residuals, may also be used in the functional time series using the Principal Component decomposition of arbitrary functional time series (see *e.g.* Kokoszka, 2011).

VARMA(p, q) in R^d has a very simple model formulation, however still the number of estimated parameters is growing with the square of the dimension considered which may cause statistical problems in estimation even if the order (p, q) is correct. The problem is that usually the orders of the VARMA(p, q) are not known. While estimating the model for too small p or q will lead to inconsistent estimation, estimating it for p or q too large will lead to overfitting that increases the variance of estimated set of parameters, given the number of observations N.

Nevertheless the solution of the problem of order choice of vector autoregressive models leads to validation techniques that are naturally based on assesing goodness-of-fit of the estimated model to the observed multivariate time series. This idea leads exactly to testing if the obtained residuals which depend on the set of parameters $\theta \in \Theta$ for given (p, q) are satisfying the "null" hypothesis: "Innovations are uncorrelated", under which the parameters of VARMA(p, q) were estimated. Thus from the problem of determining the correct order of VARMA(p, q), that has been usually approached with BIC-AIC criterions we may move to more general setup better fitting to weakly dependent innovations formulating the "null": "There exists the VARMA model, given the dataset, that would produce uncorrelated residuals".

It is widely known that given the iid innovations of ARMA generated data, the autocorrelations of lags $1, \ldots, H$ are asymptotically distributed as standard normals. This leads to Ljung-Box (Box-Pierce) family of tests that are using this fact for testing the independence of residuals of estimated ARMA(p,q) which is easily expanded to a multivariate set up. This idea in the multivariate set up has been formulated by Chitturi (1974) and Hosking (1981). However as early as in Box & Pierce (1970) and Durbin (1970) it has been shown that residuals for estimated ARMA models are going to be neither independent nor identically distributed even if the parameters are consistently estimated and the true innovations are iid. The source of this problem is the estimation error of the parameters which affects the asymptotic distributions of the autocorrelations in a differing manner depending on the lags considered. As a result the Ljung-Box (Box-Pierce) test will asymptotically underreject the "null" hypothesis as well as it will have decreased power against natural alternative hypothesis i.e. against sufficiently close VAR representations, unless the number of lags in Q_H is very large.

Another problem connected to using Ljung-Box (Box-Pierce) statistic for goodness-of-fit testing is allowing for non-independence of the true innovations for data generating VARMA process. In this case asymptotic distribution of even serial autocorrelations will not be standard normal and surely they will not be independent. Notable examples of weakly dependent innovations include allowing for serial dependence of innovations modeled as GARCH dependent volatility processes. In this case vectorized autocorrelations even for true innovations will be far from iid

normally distributed (see *e.g.* Franq, Roy, Zakoïan, 2005).
As far as statistical analysis of VARMA models is concerned, the innovations errors ε_n are generally assumed to be iid (see *e.g.* Lüthepohl, New Introduction to Time Series Analysis, Definition 11.3.1, 2005). To understand that it is a restrictive assumption it is sufficient to note that if we allow for non-independence of innovations, for example conditional heteroscedasticity, then linear VARMA is going to be merely the best linear predictor of fundamentally more complicated structure (see *e.g* Dufour, Palletier, 2005 for weak VARMA modelling). Estimation of weak VARMA models has been studied in Francq & Boubacar Mainassara (2009).
Both noted problems have been usually studied separately. Recently Francq, Roy & Zakoïan (2005) have derived the distribution of Box-Pierce statistic for weakly dependent innovations in case of VAR model. The asymptotic distribution in case of VARMA innovations has been derived by Boubacar Mainassara (2009). Both papers proposed a correction of the multivariate Box-Pierce statistic asymptotic distribution due to estimation error and non-independent innovations, using Imhof algorithm for obtaining p-values.
A different approach was presented by Delgado & Velasco (2011) in which both problems are being adressed at the same time. The method is based on transformation of residual autocorrelation vector due to possible serial dependence in order to obtain asymptotically multivariate normal distribution and then a pivotal asymptotic transform of empirical autocorrelations that would orthogonalize the system of m residual serial autocorrelations. The second step rationale is to eliminate the asymptotic effect of estimation error. We are going to concentrate on this approach for VARMA in improving the effectiveness of Box-Pierce test and compare it with other results available (see *e.g.* Francq, Roy, Zakoïan, 2005 and Boubacar Mainassara, 2009).
The article is going to be organized as follows: in Section 2 we are going to introduce the standard goodness-of-fit portmanteau test and estimation of VARMA models under the assumption that innovations are independent and identically distributed. Section 3 will present the algorithm of pivotal asymptotic orthogonalization of vectorized autocorrelations vector that would asymptotically eliminate

the estimation error. In Section 4 we introduce the method of estimation of long run serial dependence covariance that will allow to weight the vector of autocovariances such that they will be asymptotically i.i.d multivariate normal. Section 5 is devoted to Monte Carlo simulation experiment that would compare this method to the available alternatives. All the technical proofs are being relegated to mathematical appendix.

2 Diagnostic checking in VAR(p) and VARMA(p,q) models

Let us consider the d-dimensional VARMA(p,q) model

$$\begin{aligned} X_n &= \sum_{i=1}^{p} A_i X_{n-i} + \sum_{j=1}^{q} B_j \varepsilon_{n-j} + \varepsilon_n, \\ &\{\varepsilon_n\}_{-\infty}^{\infty} \text{ are iid } \mathcal{N}(0, \Sigma_\varepsilon) \text{ for all } n \in \mathbb{Z} \end{aligned} \tag{1}$$

with the assumptions imposed on matrices A_i and B_i that would grant stationarity and invertibility of X_n. We define parameters $\theta \subset \Theta$ as the function

$$\begin{aligned} &\theta \in \Theta : R^s \to R^{(p+q)d^2} \\ &\theta \to \operatorname{vec}[A_1, \ldots, A_p, B_1, \ldots, B_q] \end{aligned} \tag{2}$$

so that we have $A_1 = A_1(\theta), A_2 = A_2(\theta), \ldots, A_p = A_p(\theta)$ and $B_1 = B_1(\theta), B_2 = B_2(\theta), \ldots, B_q = B_q(\theta)$. Now, residuals $\{\varepsilon_n(\theta)\}$ are defined as

$$\varepsilon_n(\theta) = X_{n+1} - \sum_{i=1}^{p} A_i(\theta) X_{n-i} - \sum_{j=1}^{q} B_j(\theta) \varepsilon_{n-j}(\theta) \tag{3}$$

for $n = \max(p,q) + 1, \ldots, N$. Because $\varepsilon_n(\theta)$ are treated as functions of $\theta \in \Theta$ we may write that $\varepsilon_n = \varepsilon_n(\theta_0)$.

It is convenient to define the VAR(p) as a separate case despite the fact that the assumptions will be presented for general VARMA(p,q). In a similar manner we have that for any $\theta \in \Theta$ let $A_1 = A_1(\theta), A_2 = A_2(\theta), \ldots, A_p(\theta) = A_p(\theta)$ and

$$X_n = \sum_{i=1}^{p} A_i X_{n-i} + \varepsilon_n \quad \text{for all } n \in \mathbb{Z} \tag{4}$$

with $\det(I_d - \sum_{i=1}^{p} A_i z^i) \neq 0$ for any $|z| \leq 1$ with $z \in \mathbb{C}$. This condition is necessary for equation (4) to generate stationary series $\{X_n\}_{n=-\infty}^{\infty}$. The residuals $\{\varepsilon_n(\theta)\}$ are given as

$$\varepsilon_n(\theta) = X_{n+1} - \sum_{i=1}^{p} A_i(\theta) X_{n-i} \quad \text{for } n = p+1, \ldots, N. \tag{5}$$

For the time being, we assumed in (1) that innovations ε_n are independent identically distributed, which make a processes generated by (4) and (1) so called strong VAR(p) and VARMA(p,q). Now, $E(\varepsilon_n) = 0$ implies that $E(X_n) = 0$ and allows for omitting the intercept estimation which was not included in (1) and (4). Our setup is general in the sense that if we considered VARMA model with intercept

$$X_n = \mu + \sum_{i=1}^{p} A_i X_{n-i} + \sum_{j=1}^{q} B_j \varepsilon_{n-j} + \varepsilon_n,$$

assuming that $\{X_n\}_{n=-\infty}^{\infty}$ is stationary it is possible to adjust the above VARMA(p,q) model to (1) form using the estimate of EX_n (see *e.g.* Lüthepohl, Chapter 11.3.1, 2005).

The assumptions needed to consistently estimate the strong VARMA model using quasi-maximum log likelihood (QML) were given by Boubacar Mainassara and Francq (2009) and are as follows:

A1: $\{\varepsilon_n\}$ is a sequence of independent and identically distributed random vectors with $E(\varepsilon_n) = 0$ and $\mathrm{Var}(\varepsilon_n) = \Sigma_\varepsilon$.

A2: For all $\theta \in \Theta \subset R^s$ we have

$$\det\left(I - \sum_{i=1}^{p} A_i(\theta) z^i\right) \neq 0 \quad \text{for any } |z| \leq 1 \tag{6a}$$

$$\det\left(I - \sum_{j=1}^{q} B_i(\theta) z^i\right) \neq 0 \quad \text{for any } |z| \leq 1 \tag{6b}$$

It is clear that condition (6a) considers the *stability* of $\{X_n\}$ series given that $\{\varepsilon_n\}$ series is stationary. Condition (6b) assures the *invertibility* condition ie. it allows the model (1) to be represented as a purely innovation based series $X_n = \sum_{i=1}^{\infty} \phi_i \varepsilon_{n-i} + \varepsilon_n$ (see *e.g.* Lüthepohl, Chapter 11.3, 2005).

A3: $\theta_0 \in \text{int}(\Theta)$, the true parameter θ_0 exists in the interior of $\Theta \subset R^s$. The mappings

$$\theta \to \text{vec}(A_i) \text{ for } i = 1, 2, \ldots, p$$
$$\theta \to \text{vec}(B_j) \text{ for } j = 1, 2, \ldots, q$$
$$\theta \to \text{vec}(\Sigma_\varepsilon)$$

admit continuous third order derivatives for $\theta \in \Theta \subset R^s$.

Now, according to formulation of empirical residuals (3) we may define the matrices of empirical residual covariances

$$\hat{\Gamma}_\theta(j) = \frac{1}{N} \sum_{i=1}^{N} (\varepsilon_i(\theta) - \bar{\varepsilon}(\theta))(\varepsilon_{i+j}(\theta) - \bar{\varepsilon}(\theta))' \tag{7}$$

for $\theta \in \Theta$ where $\bar{\varepsilon}(\theta)$ is the average across the $1, \ldots, N$ residuals. Note that $\hat{\Gamma}_{\hat{\theta}}(j)$ for any j may be interpteted as an estimator of $\hat{\Gamma}_{\theta_0}(j)$. For the sake of completeness let us define the residual autocovariance matrices evaluated in θ_0 as

$$\hat{\Gamma}_{\theta_0}(j) = \frac{1}{N} \sum_{i=1}^{N} (\varepsilon_i - \bar{\varepsilon})(\varepsilon_{i+j} - \bar{\varepsilon})'.$$

The ”concept” of goodness-of-fit test for univariate strong VARMA models was introduced by Box and Pierce (1970). Modification of Box-Pierce statistic, so called Ljung-Box portmanteau test has been proposed by Box and Ljung (1978).
These two portmanteau statistics are used to test the null hypothesis

H_0 : $\{X_n\}_{-\infty}^{\infty}$ admits the presumed representation i.e in our case a VAR(p) or VARMA(p, q) representation and the residuals $\{\varepsilon_n\}$ satisfy assumption **A1**.

against the alternative

$H_1 : H_0$ is not true.

Now, in the univariate setting the testing procedure using (7) is based on the following Box-Pierce statistic

$$\hat{Q}_H(\theta) = N \sum_{i=1}^{H} \hat{\Gamma}_\theta(i) \hat{\Gamma}_\theta^{-1}(0),$$

$\theta \in \Theta$, which is going to converge asymptotically under the null to $\chi^2(d^2H)$ given that we would evaluate $\hat{Q}_H$ in θ_0. The reason why $\chi^2(\cdot)$ is a natural benchmark for Box-Pierce statistic is straightforward. Note that under H_0 we have that

$$\hat{\rho}_{\theta_0}(h) = \sum_{i=1}^{N-h} \frac{\varepsilon_n(\theta_0)\varepsilon_{n+h}(\theta_0)}{\hat{\Gamma}_{\theta_0}(0)} \xrightarrow{D} \mathcal{N}(0, \frac{1}{N})$$

which comes from independence of innovations ε_n. This implies that

$$Q_H(\theta_0) = N \sum_{i=1}^{H} \hat{\Gamma}_{\theta_0}(i) \hat{\Gamma}_{\theta_0}^{-1}(0) \xrightarrow{D} \sum_{i=1}^{H} Z_i^2 \sim \chi^2(H)$$

because $Z_i \sim \mathcal{N}(0, 1)$. The same rule applies in multivariate setting, the only difference comes from the more elaborate form of statistic $\hat{Q}_H$. The multivariate version of Box-Pierce statistic, (8) was given by Chitturi (1974)

$$\hat{Q}_H(\hat{\theta}) = N \sum_{h=1}^{H} tr\{\hat{\Gamma}'_{\hat{\theta}}(h)\hat{\Gamma}_{\hat{\theta}}^{-1}(0)\hat{\Gamma}_{\hat{\theta}}(h)\hat{\Gamma}_{\hat{\theta}}^{-1}(0)\} \tag{8}$$

or equivalently by Hosking (1981)

$$\hat{Q}_H(\hat{\theta}) = N \sum_{h=1}^{H} \{\mathrm{vec}(\hat{\Gamma}_{\hat{\theta}}(h))'[\hat{\Gamma}_{\hat{\theta}}(0) \otimes \hat{\Gamma}_{\hat{\theta}}(0)]^{-1}\mathrm{vec}(\hat{\Gamma}_{\hat{\theta}}(h))\} \tag{9}$$

with $\hat{Q}_H(\theta_0)$ being the statistic $\hat{Q}_H$ evaluated in $\theta_0 \in \Theta$.
It is widely known that under H_0 the asymptotic distribution of $\hat{Q}_H(\hat{\theta})$ statistic does not follow $\chi^2_{d^2H}$ due to estimation error. Hosking (1980) has shown that finite sample distribution of $\hat{Q}_H(\hat{\theta})$ is closer to $\chi^2_{d^2(H-p-q)}$ than to the asymptotic distribution of $\hat{Q}_H(\theta_0)$.
It is clear that the (8) and (9) statistics are testing the lack of serial correlation of residuals $\varepsilon_n(\hat{\theta})$ in order to indirectly test the specification of a model that was chosen for an estimation. These statistics are the standard method of checking statistical significance of residual autocorelations.

3 Removing the estimation effect on Portmanteau statistics in VARMA(p, q) setting

The asymptotic effect of estimation error $\hat{\theta} - \theta_0$ on $\hat{Q}_H(\hat{\theta})$ under H_0 is straightforward in nature. The sequence $\{\mathrm{vec}\hat{\Gamma}_{\theta_0}(i)\}_{i=1}^{m}$ is iid while $\{\mathrm{vec}\hat{\Gamma}_{\hat{\theta}}(i)\}_{i=1}^{m}$ is going to be autocorrelated. This implies that the asymptotic distribution of $\mathrm{vec}\hat{\Gamma}_{\hat{\theta}}(i)$ is not going to be the same as the asymptotic distribution of $\mathrm{vec}\hat{\Gamma}_{\theta_0}(i)$ for $i = 1, \ldots, m$ (see Box, Pierce, 1970 and Durbin, 1970 for VARMA). In addition estimation error is going to affect estimated residual autocorrelations in the finite sample stronger for small lag number i.

Let us define the autocovariance vectors

$$\hat{\gamma}_{\theta}^{(m)} = \left[\mathrm{vec}(\hat{\Gamma}_{\theta}(1))', \mathrm{vec}(\hat{\Gamma}_{\theta}(2))', \ldots, \mathrm{vec}(\hat{\Gamma}_{\theta}(m))'\right]' \tag{10}$$

and autocorrelation vectors

$$\hat{\rho}_{\theta}(i) = \mathrm{vec}\left(\hat{\Gamma}_{\theta}(1)\hat{\Gamma}_{\theta}^{-1}(0)\right), \ i = 1, \ldots, m, \tag{11}$$

for any $\theta \in \Theta$.

Now let us assume that the model under consideration follows strong VARMA(p, q) defined by (1) and assumptions **A1-A3**. The available method for taking into account the estimation error in asymptotic distribution of $\hat{Q}_m(\hat{\theta})$ is based on estimation of covariance matrix of joint vector $\sqrt{N}(\hat{\gamma}'^{(m)}_{\theta_0}, (\hat{\theta}-\theta_0)')$, (see *e.g* Boubacar Mainassara, 2009 for VARMA, Francq, Raïssi, 2007 for VAR, Francq, Roy, Zakoïan, 2005 for ARMA). The same method for nonlinear models with iid innovations was analised in Li (1992) and Hwang, Basawa & Reeves (1994). We should note that our solution to the problem of error dependence that we are going to present in this section assuming iid innovations will be applicable for nonlinear dependence assumptions on innovations as well. We are going to consider this case in the following sections.

Our solution of the problem of error dependence in $\hat{\gamma}_{\hat{\theta}}^{(m)}$ follows the route introduced in Delgado & Velasco (2011). Instead of deriving the true asymptotic distribution of $\hat{Q}_H(\hat{\theta})$ under the H_0 the idea is to perform the linear transformation of covariance vector $\hat{\gamma}_{\hat{\theta}}^{(m)}$, defined in (10). The key element of the reasoning is based on the fact that we are not going to directly estimate $\hat{\gamma}_{\theta_0}^{(m)}$, but perform the transform using orthogonal projection operator. In practice we are going to use autocorrelation vectors $\hat{\rho}_{\hat{\theta}}^{(m)}$.

Let us define the matrix of derivatives of $\hat{\Gamma}_\theta$

$$\zeta_\theta^{(m)} = \begin{bmatrix} \nabla \mathrm{vec} \hat{\Gamma}_\theta(1) \\ \nabla \mathrm{vec} \hat{\Gamma}_\theta(2) \\ \vdots \\ \nabla \mathrm{vec} \hat{\Gamma}_\theta(m) \end{bmatrix}_{(\theta)} \tag{12}$$

where

$$\nabla \mathrm{vec} \hat{\Gamma}_\theta(i) = \frac{\partial}{\partial \theta'} \mathrm{vec} \hat{\Gamma}_\theta(i), \quad i = 1, \ldots, m.$$

The derivatives $\nabla \mathrm{vec} \hat{\Gamma}_\theta(i)$ evaluated in $\hat{\theta}$ are going to play the crucial role in our method. According to (17) they may be used in the pivotal transform of $\hat{\gamma}_{\hat{\theta}}^{(m)}$ or $\hat{\rho}_{\hat{\theta}}^{(m)}$ given that they may be consistently estimated.
The transform of sample autocorrelations is based on recursive projections on the space spanned by orthogonal sample autocorrelations of residuals that are asymptotically standard normals. Thus, this transformation is asymptotically distribution free (see Delgado, Velasco, 2011). In theory it is not possible to estimate the sets of orthogonal autocorrelation vectors $\{\hat{\gamma}_{\theta_0}(i)\}_{i=1}^k$, $k = 2, \ldots, m$ without estimating the estimation error distribution. However, we show that our recursive projection operator estimated using the derivatives $\zeta_{\hat{\theta}}^{(m)}$ is going to transform the vector $\hat{\gamma}_{\hat{\theta}}^{(m)}$ into $\tilde{\gamma}_{\hat{\theta}}^{(m)}$ that is undistinguishable from $\tilde{\gamma}_{\theta_0}^{(m)}$ as far as the estimated operator is concerned. The idea of this sort of martingale transform was introduced by Brown, Durbin & Evans (1975) for CUSUM tests in the linear framework. The theoretical part reminds Khmaladze (1981) who obtained similar result for martingale part of a Gaussian process.

In order to provide the standarization of autocovariance vector $\hat{\gamma}_\theta^{(m)}$ for $\theta \in \Theta$ under H_0 and strong VAR(p) assumptions we may use may either use the estimate $\hat{\Sigma}_\varepsilon(\theta) = \hat{\Gamma}_{\hat{\theta}}(0)$ given by (62) or note that under **A1-A3** we have

$$\hat{\rho}_\theta^{(m)} = \hat{G}_\theta^{-\frac{1}{2}} \hat{\gamma}_\theta^{(m)} \tag{13}$$

with $\hat{G}_\theta = \text{diag}\left(\text{vec}(\hat{\Gamma}_\theta(1))\text{vec}(\hat{\Gamma}_\theta(1))', \ldots, \text{vec}(\hat{\Gamma}_\theta(m))\text{vec}(\hat{\Gamma}_\theta(m))'\right)$. Clearly under **A1** and H_0 we have $\hat{\Sigma}_\varepsilon = \Sigma_\varepsilon + O_P(N^{-\frac{1}{2}})$ and

$$\sqrt{N}\hat{\rho}^{(m)}_{\theta_0} \xrightarrow{d} \mathcal{N}(0, I_{md^2})$$

Now, the consistent estimation of derivatives of empirical autocovariances is needed to claim the following proposition

Proposition 1 *Under H_0 and A1-A3 we have*

$$\hat{\theta} = \theta_0 + O_P(N^{-\frac{1}{2}}), \tag{14}$$

$$\hat{G}_{\hat{\theta}} \xrightarrow{p} G_{\theta_0}, \tag{15}$$

$$G_{\theta_0} = \mathit{diag}\,(\mathit{vec}(\Gamma_{\theta_0}(1))\mathit{vec}(\Gamma_{\theta_0}(1))', \ldots, \mathit{vec}(\Gamma_{\theta_0}(m))\mathit{vec}(\Gamma_{\theta_0}(m))') \tag{16}$$

and the following holds

$$\hat{\gamma}^{(m)}_{\hat{\theta}} = \hat{\gamma}^{(m)}_{\theta_0} + \bar{\zeta}^{(m)}_{\theta_0}(\theta_0 - \hat{\theta}) + o_P\left(N^{-\frac{1}{2}}\right) \tag{17}$$

assuming that $\bar{\zeta}^{(m)}_{\theta_0}$ is a probability limit of $\zeta^{(m)}_{\theta}$ evaluated in a true parameter vector $\theta_0 \in \Theta$.

It should be noted that in *Proposition* 2 and 3 we show that $\bar{\zeta}^{(m)}_{\theta_0}$ may be estimated consistently with $\zeta^{(m)}_{\hat{\theta}}$. Referring to (17), before the procedure of eliminating the estimation error from the autocovariances vector $\hat{\gamma}_{\hat{\theta}}$ it is necessary to obtain the $md^2 \times md^2$ estimate $\hat{G}(\hat{\theta})$ that would allow to perform following standarization

$$\hat{G}^{-\frac{1}{2}}_{\theta_0}\hat{\gamma}^{(m)}_{\hat{\theta}} = \hat{G}^{-\frac{1}{2}}_{\theta_0}\hat{\gamma}^{(m)}_{\theta_0} + \hat{G}^{-\frac{1}{2}}_{\theta_0}\bar{\zeta}^{(m)}_{\theta_0}(\theta_0 - \widehat{\theta}) + o_P(N^{-\frac{1}{2}}) \tag{18}$$

with following holding by *Proposition* 2

$$\xi^{(m)}_{\hat{\theta}}(j) = G^{-\frac{1}{2}}_{\theta_0}\nabla\text{vec}\hat{\Gamma}_{\hat{\theta}}(j) \xrightarrow{p} \bar{\xi}^{(m)}_{\theta_0}(j) \text{ for } j = 1\ldots, m. \tag{19}$$

Thus Taylor expansion for autocovariances $\hat{\gamma}_{\hat{\theta}}$ implies the expansion for autocorrelations vector $\hat{\rho}^{(m)}_{\hat{\theta}}$

$$\hat{\rho}_{\hat{\theta}}^{(m)} = \hat{\rho}_{\theta_0}^{(m)} + \bar{\xi}_{\theta_0}^{(m)}(\theta_0 - \hat{\theta}) + o_P(N^{-\frac{1}{2}}) \tag{20}$$

Now, following (17) we are proposing the distribution free transformation of a vector of residual autocovariances $\hat{\gamma}_{\hat{\theta}}^{(m)}$ using the least squares projections removing the $\bar{\zeta}_{\theta_0}^{(m)}(\theta_0 - \hat{\theta})$ elements from the elements of each autocovariance $\text{vec}\hat{\Gamma}_{\hat{\theta}}(i), i = 1, \ldots, m$ in a recursive manner.

Let us take the sequence of vectors $\{\hat{\rho}_{\hat{\theta}}(i)\}_{i=1}^{m}$. According to (20) we have that the drift in equation for autocorrelations $\hat{\rho}_{\hat{\theta}}^{(m)}$ is asymptotically dependent only on estimation error. The term asymptotically refers to the fact that derivatives $\xi_{\theta_0}^{(m)}$ need to be evaluated in θ_0 and we are going to use the estimated $\hat{\theta}$ instead.

Our method is based on estimating the parameter vector β given the following set of equations using the recursive LS

$$\left\{ \begin{array}{l} \hat{\rho}_{\hat{\theta}}(j) = \xi_{\hat{\theta}}^{(m)}(j) \times \beta + \mu_{j,j} \\ \hat{\rho}_{\hat{\theta}}(j+1) = \xi_{\hat{\theta}}^{(m)}(j+1) \times \beta + \mu_{j,j+1} \\ \vdots \\ \hat{\rho}_{\hat{\theta}}(m) = \xi_{\hat{\theta}}^{(m)}(m) \times \beta + \mu_{j,m} \end{array} \right\} \tag{21}$$

for $j = 1, \ldots, m-1$.

Clearly $\hat{\beta}_j$ is going to be the estimate of $\text{vec}(\theta_0 - \hat{\theta})$ error using the information from the sequence $\{\hat{\rho}_{\hat{\theta}}(j), \ldots, \hat{\rho}_{\hat{\theta}}(m)\}$ for $j = 1, \ldots, m-1$ and $\mu_{j,i}$ are errors centered in zero for each j'th iteration of LS. The feasible recursive LS estimator of β_{j+1} is determined to be

$$\hat{\beta}_{j+1} = \left(\sum_{i=j+1}^{m} \xi_{\hat{\theta}}^{(m)}(i)' \xi_{\hat{\theta}}^{(m)}(i) \right)^{-1} \sum_{i=j+1}^{m} \xi_{\hat{\theta}}^{(m)}(i)' \hat{\rho}_{\hat{\theta}}(i)$$

Our method is based on doing projections block by block, which is more natural than estimating recursive LS on each element of $\hat{\rho}_{\hat{\theta}}(i), i = j, \ldots, m-1$.

Now, from (20) we would get that asymptotically

$$\tilde{\rho}_{\hat{\theta}}(j) = \Im_{\hat{\theta}}^{(m)}(\hat{\rho}_{\hat{\theta}}(j)) = \hat{\rho}_{\hat{\theta}}(j) - \xi_{\hat{\theta}}^{(m)}(j) \left(\sum_{i=j+1}^{m} \xi_{\hat{\theta}}^{(m)}(i)' \xi_{\hat{\theta}}^{(m)}(i) \right)^{-1} \sum_{i=j+1}^{m} \xi_{\hat{\theta}}^{(m)}(i)' \hat{\rho}_{\hat{\theta}}(i) \tag{22}$$

is going to approximate $\tilde{\rho}_{\theta_0}(j)$, the trtansformed vectorized autocorelation of empirical residuals at lag j, evaluated at θ_0.
The main motivation is based on following observation

$$\Im_{\theta_0}^{(m)}(\hat{\rho}_{\hat{\theta}}(j)) = \Im_{\theta_0}^{(m)}(\hat{\rho}_{\theta_0}(j)),\ j = 1, \ldots, m \tag{23}$$

and the fact that for N sufficiently large we are going to have

$$\Im_{\hat{\theta}}^{(m)}(\hat{\rho}_{\theta}(j)) \xrightarrow{p} \Im_{\theta_0}^{(m)}(\hat{\rho}_{\theta}(j)),\ j = 1, \ldots, m,$$

for $\theta \in \theta$. Thus having that $\Im_{\theta_0}^{(m)}(\cdot)$ may be consistently estimated it implies by (23) that $\Im_{\hat{\theta}}^{(m)}(\hat{\rho}_{\hat{\theta}}(j))$ is going to converge in probability to $\Im_{\hat{\theta}}^{(m)}(\hat{\rho}_{\theta_0}(j))$. In the light of above discussion, it is interesting to note that if we treat the derivatives $\xi_{\hat{\theta}}^{(m)}(i)$, $i = j+1, \ldots, m$ as coefficients then $\Im_{\hat{\theta}}^{(m)}(\hat{\rho}_{\hat{\theta}}(j))$ is a linear combination of vectors $\hat{\rho}_{\hat{\theta}}(j+1), \hat{\rho}_{\hat{\theta}}(j+2), \ldots, \hat{\rho}_{\hat{\theta}}(m)$.

The last step is concerned with covariance of $\Im^{(m)}(\hat{\rho}_{\hat{\theta}}(i))$ which should have been taken into account in calculating Box-Pierce statistic (9). In general we have that $\sqrt{N}\hat{\rho}_{\theta_0}(j)$ are distributed as iid standard normals for $j \geq 1$. This implies that $\Im^{(m)}(\hat{\rho}_{\hat{\theta}}(j))$ are asymptotically distributed as independent normals with variance

$$\widehat{Avar}\left(\Im^{(m)}(\hat{\rho}_{\hat{\theta}}(j))\right) = \Big[I_{d^2} + \xi_{\hat{\theta}}^{(m)}(j) \big(\sum_{i=j+1}^{m} \xi_{\hat{\theta}}^{(m)}(i)' \xi_{\hat{\theta}}^{(m)}(i) \big)^{-1} \xi_{\hat{\theta}}^{(m)}(j)' \Big] \tag{24}$$

for $j \geq 1$. Now the key observation is that in VAR(p) case, if we used explicitly only m autocovariances, then the condition that has to be met is

$$rank(\sum_{i=j+1}^{m} \zeta_{\hat{\theta}}^{(m)}(i)' \zeta_{\hat{\theta}}^{(m)}(i)) = pd^2 \tag{25}$$

with

$$rank(\sum_{i=j+1}^{m} \zeta_{\hat{\theta}}^{(m)}(i)' \zeta_{\hat{\theta}}^{(m)}(i)) = (p+q)d^2 \tag{26}$$

in VARMA(p, q) case. The implication of the above is that certain number of covariances of order less that m could not be corrected because of singularity problems.

The simplest solution to this problem is to use constant number of derivatives to transform a subset of m empirical autocorrelations. The number of autocorrelations used in correction algorithm is not bounded. Thus let us assume that we are going to use arbitrary number r of past autocorrelations in each of the equations. This modification of (22) would produce following projection operator

$$\Im_{\hat{\theta}}^{(m,r)}(\hat{\rho}_{\hat{\theta}}(j)) = \hat{\rho}_{\hat{\theta}}(j) - \xi_{\hat{\theta}}^{(m)}(j) \left(\sum_{i=j+1}^{j+r} \xi_{\hat{\theta}}^{(m)}(i)' \xi_{\hat{\theta}}^{(m)}(i) \right)^{-1} \sum_{i=j+1}^{j+r} \xi_{\hat{\theta}}^{(m)}(i)' \hat{\rho}_{\hat{\theta}}(i) \tag{27}$$

$j = 1, \ldots, m-r, \; H \leq m-r, \; r < m.$

The argument for fixing the number of derivatives used in adjusting the vectorized autocorrelations of order $j = 1, \ldots, m$ is that the excessive number of derivatives of autocorrelations used is going to decrease the goodness of fit. Note that given VARMA(p, q) specification, the autocorrelations and derivatives of autocorrelations of order $p+1, p+2, \ldots$ are going to converge to zero for increasing lags. So it is clear that the estimates of inverces of matrices of the form (25) and (26) are not asymptotically bounded. On the other hand this problem may be treated as a property of resursive projection technique. Note that the projection operator $\Im_{\hat{\theta}}^{(m,r)}$ is an estimate of $\Im_{\theta_0}^{(m,r)}$ and there is a tredeoff between the precision of the estimate and the number of lags considered r, given the number of observations N.

Now accounting for (24), $\hat{\gamma}_{\theta_0}(0)$ which is equal to $\mathrm{vec}(I_d)$ and lack of estimation effect in $\hat{\rho}_{\hat{\theta}}(0)$, we have according to (9)

$$\check{Q}_H(\hat{\theta}) = N \sum_{j=1}^{H} \left\{ \Im_{\hat{\theta}}^{(m,r)}(\hat{\rho}_{\hat{\theta}}(j))' \left[\widehat{Avar} \left(\Im_{\hat{\theta}}^{(m,r)}(\hat{\rho}_{\hat{\theta}}(j)) \right) \right]^{-1} \Im_{\hat{\theta}}^{(m,r)}(\hat{\rho}_{\hat{\theta}}(j)) \right\} \tag{28}$$

We are claiming the following

Theorem 1 *Under assumptions* ***A1-A3*** *and Proposition 2 and 3 we have the following*

$$\check{Q}_H(\hat{\theta}) = \check{Q}_H(\theta_0) + o_P(1), \tag{29}$$

$$\check{Q}_H(\hat{\theta}) \xrightarrow{D} \chi^2(Hd^2), \ \textit{with } N \to \infty \tag{30}$$

under H_0 *for* $H = 1, 2, \ldots, \ H \leq m - r, \ r < m$.

Above result is a main reason that justifies our approach and allows for eliminating the estimation error in asymptotic distribution of Box-Pierce(Ljung-Box) statistic. The crucial point is that we do not estimate the $\hat{\rho}_{\theta_0}^{m}$ vector, because it would be impossible without estimating the asymptotic distribution of estimation error. In the proof we are using the features of $\Im_{\theta_0}^{(m,r)}$ operator as it is presented in the appendix.

This way we have constructed the procedure of obtaining an analogue of BP statistic (9). It is interesting to note that approach of Francq & Raissi (2005) is based on the statistic (8) while our statistic (28) may not be trivially presented in this form. As opposed to the standard LB statistic the treshold distribution is not going to be $\chi^2((H-p-q)d^2)$ but $\chi^2(Hd^2)$. It should be noted that projection operator $\Im_{\hat{\theta}}^{(m,r)}(\cdot)$ does not asymptotically depend on asymptotical distribution of estimation error, so that we do not need to estimate it. It is the main difference between

our method and the alternative method proposed in Francq & Raïssi (2007) which is going to be compared in the next section.
Now in order to assure consistent estimation of $\Im_{\theta_0}^{(m,r)}(\cdot)$ we need to show that derivatives of vectorized residual autocovariances with respect to θ, evaluated in θ_0 may be consistently estimated. $\nabla \mathrm{vec}\hat{\Gamma}_\theta(j)$, $j = 1, \ldots, m$ for $\theta \in \Theta$ are obtained using the derivatives of $\varepsilon_n(\theta)$ for $n = 1, \ldots, N$ with respect to θ'. In the following proposition we are claiming that under H_0 in general strong VARMA framework we will obtain the $o_P(1)$ convergence of $\zeta_{\hat{\theta}}^{(m)}$ with the proof moved to the appendix.

Proposition 2 *Let us assume canonical parametrization of VARMA(p, q)*

$$\theta = vec[A_1, A_2, \ldots, A_p, B_1, B_2, \ldots, B_q] \tag{31}$$

Under H_0 and A1-A3 we have that derivatives matrix defined in (12) *will admit*

$$\begin{aligned} &\nabla vec\hat{\Gamma}_{\hat{\theta}}(i) \xrightarrow{p} \nabla vec\hat{\Gamma}_{\theta_0}(i), \quad \text{for } i = 1, \ldots, m \\ &\check{\zeta}_{\hat{\theta}}^{(m)} = \check{\zeta}_{\theta_0}^{(m)} + o_P(1) \end{aligned} \tag{32}$$

where matrix $\check{\zeta}_{\theta_0}^{(m)}$ for $\theta_0 \in \Theta$ satisfies the following equation

$$\check{\zeta}_{\theta_0}^{(m)} = \begin{bmatrix} \nabla vec\hat{\Gamma}_{\theta_0}(m) \\ \vdots \\ \nabla vec\hat{\Gamma}_{\theta_0}(1) \end{bmatrix} = -\frac{1}{N} \sum_{n=m}^{N} \left(\sum_{i=0}^{\infty} \mathbf{B}_{(m)}^i \mathbf{\Psi}_{n-i}^{(m)} \otimes \varepsilon_{n-m}(\theta_0) \right) \tag{33}$$

with

$$\mathbf{B}_{(m)}(\theta_0) = \overbrace{\begin{bmatrix} -B_1(\theta_0) & -B_2(\theta_0) & \dots & -B_q(\theta_0) & 0 & \dots & 0 \\ I_d & 0 & \dots & 0 & \dots & \dots & 0 \\ 0 & I_d & \dots & 0 & \dots & \dots & 0 \\ \vdots & & & & & & \\ 0 & 0 & \dots & I_d & \dots & \dots & 0 \\ \vdots & & & & & & \\ 0 & 0 & \dots & 0 & \dots & I_d & 0 \end{bmatrix}}^{m\times m \text{ blocks}},$$

$$\mathbf{\Psi}_k^{(m)}(\theta_0) = \overbrace{\begin{bmatrix} X'_{k-1}\otimes I_d & \dots & X'_{k-p}\otimes I_d & \varepsilon_{k-1}(\theta_0)'\otimes I_d & \dots & \varepsilon_{k-q}(\theta_0)'\otimes I_d \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \\ & & & & & \end{bmatrix}}^{m\times 1 \text{ blocks}}$$

with the constraint $m \geq q$.

It should be noted that condition $m \geq q$ is purely technical, the form of matrices $\mathbf{B}_{(m)}(\theta_0)$ and $\mathbf{\Psi}_k^{(m)}(\theta_0)$ in the unrealistic case when $m < q$ was presented in the appendix. The standard (canonical) parametrization of VARMA(p,q) assumed in *Proposition* 2 may be easily changed to general parametrization (2). It is enough to see that if we defined arbitary mapping ν in parameter space Θ

$$\nu : \theta \in \Theta \subset R^s \longrightarrow \mathrm{vec}[A_1, \dots A_p, B_1, \dots, B_q]$$

then we may write, using the chain rule

$$\nabla \mathrm{vec}\hat{\Gamma}_\theta(i) = \nabla \mathrm{vec}\hat{\Gamma}_{[A,B]}(i)\nabla\nu(\theta)$$

for any $i = 1, \dots, m$ where $\mathrm{vec}\hat{\Gamma}_{[A,B]}$ is the residual covariance with respect to canonical parametrization (31). Note that *Proposition* 2 would still hold because $\nabla_\theta \nu$ is merely a linear function of $\theta \in \Theta$.

Clearly from *Proposition* 2 and **A2** we have that derivatives of VARMA(p,q) residuals will be well defined for all $\theta \in \Theta$, not only for θ_0. It follows from stationarity of residuals $\varepsilon_n(\theta)$ and stationarity of the series X_n. Note that derivatives of residuals with respect to parameter vector $\theta \in \Theta$ are defined even if the residuals $\varepsilon_n(\theta)$ are not stationary. However without the stationarity the estimators $\check{\zeta}_\theta^{(m)}$ defined as averages would not make any sense and would not converge to any limit. In general *Proposition* 2 states that using equation (33) as the asymptotic limit, under H_0 empirical covariances derivatives $\check{\zeta}_{\hat{\theta}}^{(m)}$ are going to converge to $\check{\zeta}_{\theta_0}^{(m)}$ at least with $o_P(1)$ rate. In order to continue we need to state the following

Proposition 3 *Assuming A1-A3 and* (31) *it is true that*

$$\check{\zeta}_{\theta_0}^{(m)} \xrightarrow{p} -E\left(\sum_{i=0}^{n} \mathbf{B}_{(m)}^{i}(\theta_0)\mathbf{\Psi}_{n-i}^{(m)}(\theta_0) \otimes \varepsilon_{n-m}(\theta_0)\right)$$

We are going to show estimation of $\check{\zeta}_{\theta_0}^{(m)}$ in the simplest case of VAR(p) with $A = [A_1, A_2, \dots, A_p]$ matrix. According to *Proposition* 2 we may simplify the

$\mathbf{B}^i_{(m)}(\theta_0)$ matrix because $B_1, \ldots, B_q$ are equal to zero and

$$\mathbf{B}^0_{(m)} = \overbrace{\begin{bmatrix} I_d & 0 & \\ 0 & \ddots & 0 \\ & 0 & I_d \end{bmatrix}}^{m\times m \text{ blocks}}, \mathbf{B}^1_{(m)} = \overbrace{\begin{bmatrix} 0 & & & \\ I_d & 0 & & \\ & \ddots & \ddots & \\ & & I_d & 0 \end{bmatrix}}^{m\times m \text{ blocks}},$$

$$\mathbf{B}^2_{(m)} = \overbrace{\begin{bmatrix} 0 & & & & \\ 0 & 0 & & & \\ I_d & 0 & 0 & & \\ 0 & \ddots & \ddots & \ddots & \\ & & I_d & 0 & 0 \end{bmatrix}}^{m\times m \text{ blocks}}, \ldots, \mathbf{B}^p_{(m)} = 0_{md}$$

Thus we are going to obtain that

$$\check{\zeta}^{(m)}_{\hat{\theta}} = \frac{1}{N} \sum_{n=m}^{N} \left(\sum_{i=0}^{m-1} \mathbf{B}^i_{(m)} \mathbf{\Psi}^{(m)}_{n-i} \otimes \varepsilon_{n-m}(\hat{\theta}) \right) \tag{34}$$

with

$$\mathbf{\Psi}^{(m)}_n = \begin{bmatrix} X'_{n-1} \otimes I_d & \ldots & X'_{n-p} \otimes I_d \\ 0 & \ldots & 0 \\ 0 & \ldots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \ldots & 0 \end{bmatrix} = \begin{bmatrix} F_n \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

and $F_n = [X'_{n-1}, \ldots, X'_{n-p}] \otimes I_d$.

Clearly $\mathbf{B}^i_{(m)}$ $i = 0, 1, \ldots, m-1$ are simply row switching linear operators so we are going to get according to (34)

$$
\begin{aligned}
\breve{\zeta}^{(m)}_{\hat{\theta}} &= \frac{1}{N}\sum_{n=m}^{N}\begin{bmatrix} F_n \\ F_{n-1} \\ \vdots \\ F_{n-m+1}\end{bmatrix} \otimes \varepsilon_{n-m}(\hat{\theta}) = \\
&= \frac{1}{N}\sum_{n=m}^{N}\begin{bmatrix} X'_{n-1}\otimes I_d & \ldots & X'_{n-p}\otimes I_d \\ X'_{n-2}\otimes I_d & \ldots & X'_{n-p-1}\otimes I_d \\ \vdots & \vdots & \vdots \\ X'_{n-m}\otimes I_d & \ldots & X'_{n-m-p+1}\otimes I_d \end{bmatrix} \otimes \varepsilon_{n-m}(\hat{\theta}) = \qquad (35) \\
&= \frac{1}{N}\sum_{n=m}^{N}\begin{bmatrix} X'_{n-1}\otimes \varepsilon_{n-m}(\hat{\theta}) & \ldots & X'_{n-p}\otimes \varepsilon_{n-m}(\hat{\theta}) \\ X'_{n-2}\otimes \varepsilon_{n-m}(\hat{\theta}) & \ldots & X'_{n-p-1}\otimes \varepsilon_{n-m}(\hat{\theta}) \\ \vdots & \vdots & \vdots \\ X'_{n-m}\otimes \varepsilon_{n-m}(\hat{\theta}) & \ldots & X'_{n-m-p+1}\otimes \varepsilon_{n-m}(\hat{\theta}) \end{bmatrix} \otimes I_d.
\end{aligned}
$$

In general there is not one way to write the derivatives $\breve{\zeta}^{(m)}_{\theta}$ because Kronecker product $\otimes$ is not commutative. It is easy to show that (35) is going to be asymptotically equivalent to (43) for VAR(p), however (35) is consistent with VARMA(p, q) representation of $\breve{\zeta}^{(m)}_{\theta}$. As far as probability limit is concerned under the strong VAR(p) specification we are going to obtain under LLN

$$
\breve{\zeta}^{(m)}_{\hat{\theta}} \xrightarrow{p} \begin{bmatrix} \sigma^{\otimes}_{\epsilon}(m-1) & \ldots & \sigma^{\otimes}_{\epsilon}(m-p+1) & \sigma^{\otimes}_{\epsilon}(m-p) \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \sigma^{\otimes}_{\epsilon}(1) \\ \vdots & \vdots & \iddots & \sigma^{\otimes}_{\epsilon}(0) \\ \vdots & \sigma^{\otimes}_{\epsilon}(1) & \iddots & 0 \\ \sigma^{\otimes}_{\epsilon}(1) & \sigma^{\otimes}_{\epsilon}(0) & \ldots & 0 \\ \sigma^{\otimes}_{\epsilon}(0) & 0 & \ldots & 0 \end{bmatrix} \otimes I_d
$$

where $\sigma_\epsilon^{\otimes}(f)$ satisfies

$$\begin{aligned}
&\sigma_\epsilon^{\otimes}(f) = \sum_{s=1}^{f} \sum_{k_1+\cdots+k_s=f} \left[\varepsilon_n' \left(\prod_{i=1}^{s} A_{k_i} \right)' \otimes \varepsilon_n \right], \\
&f = 0, \ldots, m-1, \\
&0 < (k_1)_{i=1}^{s} \leq p : \sum_{i=1}^{s} k_i = f \text{ for any } s > 0,
\end{aligned}$$

Convergence rates of derivatives of general VARMA(p, q) are needed to write the empirical covariances as a sum of uncorrelated covariances $\hat{\gamma}_\theta$ and a stochastic drift dependent on the estimation error under H_0. *Proposition* 3 gives the probability limit of $\check{\zeta}_{\theta_0}^{(m)}$. From *Proposition* 2, however we know that $\check{\zeta}_{\hat{\theta}}^{(m)}$ will be converging in probanility to $\check{\zeta}_{\theta_0}^{(m)}$ hence the above result gives the probability limit of $\check{\zeta}_{\hat{\theta}}^{(m)}$ under assumptions of *Proposition* 2.

4.Comparison of our method of removing estimation error with the alternative solution in case of strong VAR(p) models

Now we are going to review our method on the example of strong VAR(p) model ie. taking assumptions **A1-A3**. It means that we are only going to deal with estimation error $(\hat{\theta} - \theta_0)$ but the true empirical residuals $\varepsilon_n(\theta_0)$ will be independent and uncorrelated. The point is to show the asymptotic equivalence of projection method and the alternative procedures based on the joint analysis of the vector

$$\omega_{\theta_0} = (\sqrt{N} \hat{\gamma}'^{(m)}_{\theta_0}, \sqrt{N}(\hat{\theta} - \theta_0)')'$$

The representative review of this method in case of weakly dependent VAR(p) model was introduced in Francq & Raïssi (2007), recently it was used in Escanciano, Lobato and Zhu (2010). As far as estimation is concerned, VAR(p) in R^d under **A1-A3** may be estimated using the standard OLS method assuming the canonical parametrisation of VAR(p), (31), following Francq & Raïssi (2007).
Let us define

$$\tilde{X}_n = [X_n', X_{n-1}', \ldots, X_{n-p+1}']'$$

and let us use the convention that $X_n = 0$, $\varepsilon_n(\theta) = 0$ for any $n < 1, n > N$. We have that

$$\widehat{\mathbf{Cov_X}}(p) = \frac{1}{N}\sum_{n=1}^{N} X_n \tilde{X}_n'$$

and under the assumption that

$$\widehat{\mathbf{Var_X}} = \frac{1}{N}\sum_{n=1}^{N} \tilde{X}_n \tilde{X}_n'$$

is invertible (see *e.g.* Lütkepohl, 1993) and **A1** we have the LS estimator of θ_0 is

$$\hat{\theta} = \mathrm{vec}\left(\widehat{\mathbf{Cov_X}}(p)(\widehat{\mathbf{Var_X}})^{-1}\right) \tag{36}$$

consistent and asymptotically normally distributed. Note that if we defined the residuals as in (5) then equation (36) gives the closed form solution for minimization of QML estimator (61).
The main contribution in solving the poblem of eliminating the asymptotic estimation error from Ljung-Box statistic has been proposed in Francq & Raïssi (2007). In order to relate it to our result we are going to give the brief decription of their solution. In Francq & Raïssi (2007) it is shown that for weakly dependent VAR(p)

model under H_0 vector $\omega_{\theta_0} = (\hat{\gamma'}^{(m)}_{\theta_0}, (\hat{\theta} - \theta_0)')'$ satisfies

$$\sqrt{N}\omega_{\theta_0} \xrightarrow{d} \mathcal{N}(0, \Xi_F).$$

Now, covariance matrix Ξ_F has the following structure

$$\Xi_F = \begin{pmatrix} \Sigma_{\hat{\gamma}(\theta_0)} & \Sigma_{(\hat{\gamma}(\theta_0),\hat{\theta})} \\ \Sigma'_{(\hat{\gamma}(\theta_0),\hat{\theta})} & \Sigma_{\hat{\theta}} \end{pmatrix} = \sum_{h=-\infty}^{\infty} E\Upsilon_n\Upsilon'_{n-h}, \tag{37}$$

with $\Upsilon_n = \begin{pmatrix} w_n \\ v_n \end{pmatrix}$ where

$$\begin{aligned} w_n &= \begin{pmatrix} \varepsilon_{n-1}(\theta_0) \\ \vdots \\ \varepsilon_{n-m}(\theta_0) \end{pmatrix} \otimes \varepsilon_n(\theta_0), \\ v_n &= \mathbf{Var_X}^{-1} \begin{pmatrix} X_{n-1} \\ X_{n-2} \\ \vdots \\ X_{n-p} \end{pmatrix} \otimes \varepsilon_n(\theta_0). \end{aligned} \tag{38}$$

where $\mathbf{Var_X} = \text{plim}\widehat{\mathbf{Var_X}}$. Now following *Theorem* 1 (Francq, Raïssi, 2007) it has been shown that

$$\sqrt{N}\hat{\gamma}^{(m)}_{\hat{\theta}} \xrightarrow{d} \mathcal{N}(0, \Sigma_{\gamma_{\hat{\theta}}(\hat{\theta})}) \tag{39}$$

with

$$\Sigma_{\gamma_{\hat{\theta}}(\hat{\theta})} = \Sigma_{\gamma_{\hat{\theta}}(\theta_0)} + \Phi_m\Sigma_{\hat{\theta}}\Phi'_m + \Sigma_{(\hat{\gamma}(\theta_0),\hat{\theta})}\Phi'_m + \Phi_m\Sigma'_{(\hat{\gamma}(\theta_0),\hat{\theta})} \tag{40}$$

where

$$\Phi_m = -E \begin{pmatrix} \varepsilon_{n-1}(\theta_0) \\ \vdots \\ \varepsilon_{n-m}(\theta_0) \end{pmatrix} \otimes \begin{pmatrix} X_{n-1} \\ \vdots \\ X_{n-p} \end{pmatrix}' \otimes I_d \tag{41}$$

In addition the following formula for correlation $\hat{\rho}_{\hat{\theta}}^{(m)}$ is shown to hold

$$\sqrt{N}\hat{\rho}_{\hat{\theta}}^{(m)} \xrightarrow{d} \mathcal{N}(0, \Sigma_{\hat{\rho}(\hat{\theta})})$$

where

$$\Sigma_{\hat{\rho}(\hat{\theta})} = (I_m \otimes (S_\varepsilon \otimes S_\varepsilon)^{-1})\Sigma_{\gamma_{\hat{\theta}}(\hat{\theta})}(I_m \otimes (S_\varepsilon \otimes S_\varepsilon)^{-1}) \tag{42}$$

where S_ε is the variance of d coordinates of ε_n, estimated by

$$\hat{\sigma}(i) = \sqrt{\frac{1}{N}\sum_{j=1}^{N} \varepsilon_{ji}^2(\hat{\theta})},\ \ i = 1, \dots, d$$
$$\hat{S}_{\hat{\varepsilon}} = \operatorname{diag}(\hat{\sigma}_\varepsilon(1), \hat{\sigma}_\varepsilon(2), \dots, \hat{\sigma}_\varepsilon(d)).$$

We have that $\hat{S}_\varepsilon$ does not have to be asymptotically equivalent to $\hat{\Sigma}_\varepsilon$ because it does allow only diagonal asymptotic covariance matrix of model errors.
It is clear that a main problem is concerned with estimation of $\Sigma_{\gamma_{\hat{\theta}}(\hat{\theta})}$. Equation (40) gives the relation between $\Sigma_{\hat{\rho}(\hat{\theta})}$ and submatrices of Ξ_F. Let us assume for the moment that variance matrix of ω_{θ_0}, Ξ_F may be consistently estimated. We are going to show how equation (40) is related to (17).

The first thing to be noted is that we have plim $(\zeta_{\hat{\theta}}^{(m)}) = E\zeta_{\theta_0}^{(m)} = \Phi_m$. Clearly it is true for $\theta \in \Theta$ that

$$\nabla \text{vec}\hat{\Gamma}_\theta(j) = -\frac{1}{N}\sum_{n=p}^{N} \varepsilon_{n-j}(\theta) \otimes \begin{pmatrix} X_{n-1} \\ X_{n-2} \\ \vdots \\ X_{n-p} \end{pmatrix}' \otimes I_d \quad \text{for } j = 1, \ldots, m.$$

It follows that for $\zeta_{\hat{\theta}}^{(m)}$ defined in (12) we have

$$\zeta_{\hat{\theta}}^{(m)} = -\frac{1}{N}\sum_{n=p}^{N} \begin{pmatrix} \varepsilon_{n-1}(\hat{\theta}) \\ \vdots \\ \varepsilon_{n-m}(\hat{\theta}) \end{pmatrix} \otimes \begin{pmatrix} X_{n-1} \\ \vdots \\ X_{n-p} \end{pmatrix}' \otimes I_d \tag{43}$$

with

$$\text{plim } \zeta_{\hat{\theta}}^{(m)} = -E \begin{pmatrix} \varepsilon_{n-1}(\theta_0) \\ \vdots \\ \varepsilon_{n-m}(\theta_0) \end{pmatrix} \otimes \begin{pmatrix} X_{n-1} \\ \vdots \\ X_{n-p} \end{pmatrix}' \otimes I_d = \Phi_m$$

with the first equality following from *Propositions* 2 and 3.
Now we have that under H_0, taking assumptions from *Proposition* 1, drift equation (17) holds which implies that

$$\begin{aligned} Avar(\hat{\gamma}_{\hat{\theta}}^{(m)}) =& Avar(\hat{\gamma}_{\theta_0}^{(m)}) + \\ &+ \text{plim}(\zeta_{\theta_0}^{(m)}) Avar(\theta_0 - \hat{\theta}) \text{plim} {\zeta'}_{\theta_0}^{(m)} + \\ &+ Acov(\hat{\gamma}_{\theta_0}^{(m)}, (\theta_0 - \hat{\theta})) \text{plim}({\zeta'}_{\theta_0}^{(m)}) + \\ &+ \text{plim}(\zeta_{\theta_0}^{(m)}) Acov(\theta_0 - \hat{\theta}, \hat{\gamma}_{\theta_0}^{(m)}), \end{aligned}$$

which leads in the notation used in (40) to

$$\Sigma_{\gamma_{\hat{\theta}}(\hat{\theta})} = \Sigma_{\gamma_{\hat{\theta}}(\theta_0)} + \text{plim}(\zeta_{\theta_0}^{(m)}) \Sigma_{\hat{\theta}} \text{plim}({\zeta'}_{\theta_0}^{(m)}) + \text{plim}(\zeta_{\theta_0}^{(m)}) \Sigma_{(\hat{\gamma}(\theta_0),\hat{\theta})} + \Sigma_{(\hat{\gamma}(\theta_0),\hat{\theta})} \text{plim}({\zeta'}_{\theta_0}^{(m)}) \tag{44}$$

Now, if $\text{plim}(\zeta_{\theta_0}^{(m)}) = \Phi_m$ then we have that (40) is equivalent to Francq & Raïssi equation (51) holding for components of matrix Ξ_F. In principle (40) is an implication of (17) which is also going to hold under weak VAR(p) assumptions. However, (17) does not impose $(\hat{\theta} - \theta_0)$ or $\hat{\Sigma}_{\hat{\theta}}$ while (40) is specific for particular estimator (see Den Haan, Levin, 1998).

We should note that in our procedure we are aiming at estimating the asymptotic distribution of a combination of past autocorrelations $\{\hat{\rho}_{\hat{\theta}}^{(m)}(h)\}_{h=i}^{i+r}$ for every $\hat{\rho}_{\hat{\theta}}^{(m)}(i)$, $i = 1, \ldots, m$ which is going to approximate the $\Im_{\hat{\theta}}^{(m)}(\hat{\rho}_{\theta_0}^{(m)}(i))$. In Francq & Raïssi (2007) the authors obtain asymptotic distribution of Box-Pierce statistic (8) evaluated for $\hat{\rho}_{\hat{\theta}}^{(m)}$ instead. Distribution of $\hat{Q}_m^{FR}(\hat{\theta})$ is obtained using the estimator $\hat{\Sigma}_{\hat{\rho}(\hat{\theta})}$ of $\Sigma_{\hat{\rho}(\hat{\theta})}$ defined in (42), using the feasible estimator of S_ε

$$\hat{\Sigma}_{\hat{\rho}(\hat{\theta})} = \left(I_m \otimes \hat{S}_\varepsilon^{-\frac{1}{2}} \otimes \hat{S}_\varepsilon^{-\frac{1}{2}}\right) \Sigma_{\gamma_{\hat{\theta}}(\hat{\theta})} \left(I_m \otimes \hat{S}_\varepsilon^{-\frac{1}{2}} \otimes \hat{S}_\varepsilon^{-\frac{1}{2}}\right) \tag{45}$$

following $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$. Now the distribution of $\hat{Q}_m^{FR}(\hat{\theta})$ is obtained as asymptotically equal to the distribution of following random variable

$$Z_m(\{\psi_i\}_{i=1}^{md^2}) = \sum_{i=1}^{md^2} \psi_i Z_i^2 \tag{46}$$

where $\{\psi_i\}_{i=1}^{md^2}$ is a set of md^2 eigenvalues of matrix $\hat{\Sigma}_{\hat{\rho}(\hat{\theta})}$ and the variables Z_i, $i = 1, \ldots, md^2$ are independent $\mathcal{N}(0,1)$. The p-values of a test under H_0 are thus obtained using (46). The origin of such a method goes back to 1951 and is called an Imhof algorithm.

This step is necessary because, due to approach used, the statistic of interest has an unknown distribution. Let us recall that (46) is a direct analogue of equation

(8) using the known fact that

$$\mathrm{tr}(A) = \mathrm{tr}(VJV^{-1}) = \sum_{i=1}^{md^2} \psi_i, \quad \text{for any } A$$

where J is a Jacobi representation matrix of A and ψ_i, $i = 1, \ldots, md^2$ are its eigenvalues. The main issue is consistent estimation of eigenvalues ψ_i. This problem is equivalent to consistent estimation of matrix Ξ_F which is going to be analized in the next section. As far as our method is concerned the most evident difference is that we do not need to estimate the covariance matrix $\hat{\Sigma}_{\hat{\theta}}$. Equivalently we do not need to include vectorized correlation matrix v_n in Υ_n.

5.Weak VAR(p) and VARMA(p, q) models

The strong formulation of VAR and VARMA models has been criticized as too restrictive as far as the assumption on strong white noise innovation process ε_n is concerned. The criticism comes from the fact that for a strong formulation of VARMA, 'under null' the true θ_0 exists, which means that the linear function of past observations and innovations is in fact the best predictor available. However, it is more realistic to assume that 'under null' the function of past observations and past innovations is the best predictor available, but this function does not neccesarily has to be linear. This observation leads to introducing a different assumption on innovations process $\{\varepsilon_n\}$, instead of strong white noise ie. independent, identically distributed vectors ε_n with homoskedastic covariance matrix Σ_ε it is assumed that $\{\varepsilon_n\}$ are non correlated but not necessarily independent vectors. This assumption together with formulations (4) and (1) define weak VAR(p) and weak VARMA(p, q) models. In this setup, true parameter vectors θ_0 do not exist per se, and are defined as parameters that define the best linear predictor using past observations available.

As noted before, there are two problems that might occur while performing Box-Pierce (Ljung-Box). The first one is concerned with estimation error. The second one is caused by allowing for lack of independence of innovations ε_n. The problem may be described as follows. Assuming ergodicity, stationarity and lack of correlation of $\{\varepsilon_n(\theta)\}$ we could write

$$\sqrt{N}\hat{\rho}_{\theta_0}^{(m)} = \sqrt{N}\begin{bmatrix}\mathrm{vec}(\hat{\rho}_{\theta_0}(1)) \\ \mathrm{vec}(\hat{\rho}_{\theta_0}(2)) \\ \vdots \\ \mathrm{vec}(\hat{\rho}_{\theta_0}(m))\end{bmatrix} \xrightarrow{d} N(0, \Sigma_{\hat{\rho}(m)}) \tag{47}$$

under some mixing conditions. Covariance matrix $\Sigma_{\hat{\rho}(m)}$, however is not going to be identity matrix under

H_0: There exists the VARMA model, given the dataset, with uncorrelated innovations for some $\theta_0 \in \Theta$.

with the alternative

H_1: There exists no VARMA model, given the dataset, with uncorrelated innovations for some $\theta_0 \in \Theta$.

as it would occur in iid $\{\varepsilon_n\}$ case (see *e.g.* Francq, Raïssi, 2007). It follows that statistics (8) will not have asymptotic χ^2 distribution. Before we move further in the discussion we have to analise the sufficient conditions for ergodicity in the weak VAR(p) and VARMA(p, q) models.

As far as weak VAR model is concerned we define it according to Francq & Raïssi (2007). Let us assume the model formulation (1) and (3) with the assumption that $\{\varepsilon_n\}$ is stationary and ergodic. In case of weak VAR(p) model this implies the ergodicity of $\{X_n\}$. Almost sure convergence of OLS estimation of weak VAR(p) is proven in *Proposition* 1 in Francq, Raïssi (2007). It has been proven that if we assume that Σ_ε is nonsingular then we have for LS estimator (36)

$$\hat{\theta} \xrightarrow{a.s} \theta_0,$$
$$\hat{\Sigma}_\varepsilon \xrightarrow{a.s} \Sigma_\varepsilon$$
$$\mathbf{Var_X} \text{ is invertible } a.s.$$

for $N \to \infty$. In order to obtain asymptotic normality of LS estimator of θ_0 we are using *Proposition* 2 from Francq & Raïssi (2007) which states that under specific mixing conditions $\hat{\theta}$ error is asymptotically normal. The necessary assumption needed for this result is

$$\sum_{h=0}^{\infty}\{\alpha_X(h)^{\frac{\nu}{2+\nu}}\} < \infty \quad (*)$$

for some $\nu > 0$, where $||\cdot||.$ is an euclidean norm and $\alpha_X(h)$ is a α-mixing coefficient defined as

$$\alpha_X(h) = \sup_{A\in\sigma(X_u,u\leq n),B\in\sigma(X_u,u\geq n+h)} |P(A\cap B) - P(A)P(B)|$$

It follows that a sufficient assumption replacing **A1** in case of weak VAR(p) is

A1': The process $\{\varepsilon_n(\theta)\}$ is stationary and ergodic for $\theta \in \Theta$. We are also assuming α-mixing condition on $\{X_n\}$, $(*)$. In addition we have to assume that $||X_n||_{4+2\nu} < \infty$ which together with $(*)$ is sufficient for consistent estimation of $\theta_0 \in \theta$. Thus **A1'** is clearly stronger than **A1**.

In case of weak VARMA model we are following Boubacar Mainassara & Francq (2009). The sufficient assumption for consistency of QML estimator is the following

A1": The process $\{\varepsilon_n(\theta)\}$ is stationary and ergodic for $\theta \in \Theta$. α-mixing condition $(*)$ holds. We have also $E||\varepsilon_n||^{4+2\nu} < \infty$ where $||\cdot||$ is euclidean norm.

Estimation of weak VARMA(p,q) model has been described in the Appendix. Asymptotic normality of $\hat{\theta}$ estimator comes from Boubacar Mainassara & Francq (2009, *Theorem* 3). Given that $\mathrm{Var}(\varepsilon_n) = \Sigma_\varepsilon$ is not dependent on structural parameters QML feasible estimator (63) is consistent and asymptotically normal. However, the canonical parametrization (31) is imposed. It is interesting to note that in Delgado & Velasco (2011) in order to obtain the asymototic result for weak nonlinear models assumes **A1"**.

Compared to the previous section, the central role is going to be played by covariance matrices of weakly dependent processes. Let us recall covariance matrix Ξ_F and equation for asymptotic covariance coefficients (37). This kind of estimator of covariance matrix of nonindependent processes has been analised in Hannan & Heyde (1972) and more recently by Romano & Thombs (1996). In Francq & Raïssi (2007) the authors are using Den Haan & Levin estimator we are going to present further. The idea of deriving the pivotal transform of Box-Pierce statistic in case of weak VAR(p) and VARMA(p,q) models is based in a simple observation.

If it would be possible to estimate the $md^2 \times md^2$ matrix G_{θ_0} that would satisfy under H_0 hypothesis

$$\sqrt{N}\hat{\gamma}_{\theta_0}^{(m)} \xrightarrow{d} \mathcal{N}(0, G_{\theta_0}), \tag{48}$$

then the general method of asymptotic Taylor approximation depicted by the equation (17) could hold also in a weak case. In fact under **A1'** or **A1"** with moment condition $||X_n||_{4+2\nu} < \infty$ for some $\nu > 0$ it is possible to estimate G_{θ_0} consistently by its empirical analog (see *eg*. Romano, Thombs, 1996).

We have that

$$
\begin{aligned}
&G_{\theta_0} = \big(g_{\alpha,\beta}\big),\\
&g_{\alpha,\beta} = E(\varepsilon_n^i(\theta_0)\varepsilon_{n+k}^j(\theta_0)\varepsilon_n^a(\theta_0)\varepsilon_{n+h}^b(\theta_0)),\\
&1 \le k, h \le m,\\
&1 \le i, j, a, b \le d,\\
&\alpha = (k-1)d^2 + (j-1)d + i,\\
&\beta = (h-1)d^2 + (b-1)d + a,
\end{aligned}
$$

where $\varepsilon_n^j(\theta_0)$ is the j'th element of $\varepsilon_n(\theta_0)$ vector. Now, in order to get more compact picture of estimator $\hat{G}_{\hat{\theta}}$, note that k, h denote the block position in $\hat{G}_{\hat{\theta}}$ matrix in a following way

$$
\hat{G}_{\hat{\theta}} = \begin{pmatrix}
\hat{C}_{11} & \hat{C}_{21} & \hat{C}_{31} & \dots & \hat{C}_{m1}\\
\hat{C}_{21} & \hat{C}_{22} & \dots & \dots & \hat{C}_{m2}\\
\hat{C}_{31} & \vdots & \ddots & \dots & \vdots\\
\vdots & \vdots & \vdots & \hat{C}_{kh} & \vdots\\
\hat{C}_{m1} & \hat{C}_{m2} & \dots & \dots & \hat{C}_{mm}
\end{pmatrix} \tag{49}
$$

$$
\hat{C}_{kh} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{vec}(\varepsilon_n(\hat{\theta})\varepsilon'_{n+k}(\hat{\theta}))\mathrm{vec}(\varepsilon_n(\hat{\theta})\varepsilon'_{n+h}(\hat{\theta}))' \tag{50}
$$

with $\hat{C}_{kh}$ being $d^2 \times d^2$ matrix for each $k, h = 1, 2, \dots, m$.

In general, above estimator of asymptotic covariance matrix G_{θ_0} for weak VARMA model allows for nonlinear effects in data generated innovations ε_n and it is consistent with estimator proposed in *Proposition* 2 for strong VARMA model. However, the problem is the number of coefficients that need to be estimated. For $G_{\theta_0} = \mathrm{diag}\,(\mathrm{vec}(\Gamma_{\theta_0}(1))\mathrm{vec}(\Gamma_{\theta_0}(1))', \dots, \mathrm{vec}(\Gamma_{\theta_0}(m))\mathrm{vec}(\Gamma_{\theta_0}(m))')$ the number of parameters is only md^2 while in the weak case we have a matrix to be estimated with m^2d^4 entries. This number may be reduced by imposing additional assump-

tions on the ε_n process. For example in GARCH setup only diagonal matrices $\hat{C}_{11}, \ldots, \hat{C}_{mm}$ would have to be estimated.
Our main result is based on following

Theorem 2 *Under VARMA(p,q) specification,* ***A2-A3*** *and weak condition for innovations,* ***A1"*** *we have that under* H_0

$$\hat{\theta} = \theta_0 + O_P(T^{-\frac{1}{2}})$$
$$\hat{G}_{\hat{\theta}} = \hat{G}_{\theta_0} + o_P(1)$$

and the following decomposition holds

$$\hat{\gamma}^{(m)}_{\hat{\theta}} = \hat{\gamma}^{(m)}_{\theta_0} + \bar{\zeta}^{(m)}_{\theta_0}(\theta_0 - \hat{\theta}) + o_P\left(N^{-1/2}\right) \tag{51}$$

with

$$\zeta^{(m)}_{\hat{\theta}}(j) = \nabla vec\hat{\Gamma}_{\hat{\theta}}(j) \xrightarrow{p} \bar{\zeta}^{(m)}_{\theta_0}(j) \tag{52}$$

(53)

Above theorem states that all the key equations that hold for strong VARMA(p,q) are going to hold also in weak case. The only difference is introducing asymptotic covariance matrix G_{θ_0} with available estimator revised in the beginning of the section. The detailed proof has been moved to the appendix.

Estimation of Ξ_F is based on autoregressive spectral estimator method derived in Den Haan & Levin (1998). Their main result is based on the observation that covariance of empirical resuduals, where ε_n are weakly dependent process, under assumption $(*)$ has an $AR(\infty)$ representation. They have proposed the estimator for general covariance matrices, specifically $\hat{\Gamma}_{\theta_0}(0)$, however Francq & Raïssi are using this result to construct the estimator of covariance matrix Ξ_F where

$$\Xi_F = Cov(\sqrt{N}\hat{\gamma}'^{(m)}_{\theta_0}, \sqrt{N}(\hat{\theta} - \theta_0)') = \sum_{h=-\infty}^{\infty} E\Upsilon_n\Upsilon'_{n-h}$$

with Υ_n defined in (38). According to Den Haan & Levin (1998) in above specification weakly dependent sequence $\{\tilde{\Upsilon}_n\}$ may be approximated as VAR(∞) process with $o_P(N^{-\frac{1}{2}})$ order of convergence.

$$\tilde{\Upsilon}_n = \sum_{i=1}^{\infty} \mathcal{A}_i \tilde{\Upsilon}_{n-i} + \mu_n + o_p(T^{-\frac{1}{2}}) \tag{54}$$

where μ_n is a residual. Now it has been proven that asymptotically we have

$$\Xi_F = \left(I_{d^2m} - \sum_{i=1}^{r} \mathcal{A}_i \right)^{-1} E(\mu_n \mu_n') \left(I_{d^2m} - \sum_{i=1}^{r} \mathcal{A}_i \right)^{\prime -1} \tag{55}$$

for r fixed and $N \to \infty$. It is important to note that our method does not require any specific estimator of parameters or autocovariance matrices and only requires any consistent, asymptotically normal estimator of parameters θ_0 and consistent estimator of asymptotic covariance matrix G_{θ_0}. In the previous section we have shown that the reason for adding the v_n component in Υ_n is allowing for concurrent estimation of matrices $\Sigma_{(\hat{\gamma}(\theta_0),\hat{\theta})}$ which will be essential for building up matrix $\Sigma_{\hat{\gamma}(\hat{\theta})}$ in weak VAR(p) specification. The key point is that Υ_n is a vector of dimension $d^2(m+p)$ that is already much higher then dimensions of derivatives that are essential in our method. This is going to affect the finite performance of VAR(r) spectral esimator of $\sum_{h=-\infty}^{\infty} E\Upsilon_n\Upsilon_{n-h}'$ given the observation number N. We are going to show that for finite sample Francq & Raïssi transformed LB test will suffer from worse size characteristics than our method.

5.Numerical ilustrations

In this section we are going to investigate the properties of our modified BP test and compare it with standard BP test and an alternative modification proposed by Francq & Raïssi, (2007) for VAR(p) models. Our main motivation is to show the advantage of our method (DV) over Francq & Raïssi (FR) at the lower range of lenght of the estimated series.

5.1. Empirical size.

Data generating process follows the empirical size experiment for weak VAR(1) model in Francq & Raïssi (2007). The true series $\{X_n\}_{n=1}^N$ is generated according to

$$\begin{aligned}
&X_{n+1} = A_1 X_n + \varepsilon_{n+1}, \\
&A_1 = 0.6 I_2, \\
&\varepsilon_n = \begin{pmatrix} \eta_n^1 \eta_{n-1}^1 \eta_{n-2}^1 \\ \eta_n^2 \eta_{n-1}^2 \eta_{n-2}^2 \end{pmatrix}, \\
&\begin{pmatrix} \eta_n^1 \\ \eta_n^2 \end{pmatrix} \sim \mathcal{N}(0, I_2), iid \quad n = 1, 2, \ldots, N
\end{aligned} \tag{56}$$

We simulated $n = 2000$ independent trajectories of lenght $N = 500$ of this weak VAR(1) model. For each replication we estimated matrix coefficients of VAR(1) model and then applied portmanteau tests to the residuals. We have chosen an asymptotic nominal level of the test to be $\alpha = 5\%$. The number of covariances used in computing $\check{Q}$ is $H = 1, \ldots, 12$. For DV and FR modifications of LB test the H_0 is rejected when $\check{Q}_H > \chi^2_{0.95}(4H)$. For the standard LB test the H_0 is rejected when $\check{Q}_H > \chi^2_{0.95}(4H - 4)$. As far as FR test is concerned the results have been presented for $r = 1, \ldots, 5$ which represent the order of VAR(r) spectral estimator in (55). In Francq & Raïssi (2007) the authors suggest using the Levinson estimation algorithm. We have presented also results using the standard Moore-Penrose inverse (pseudoinverse). For DV modification of BP test we have presented the results for $k = 1, \ldots, 4$ numbers of derivatives used in estimating the transform of each $\hat{\rho}_{\hat{\theta}}(j), j = 1, \ldots, H$, following operator (27).

Clearly FR modification of LB test looses its size properties dramatically with increasing order of $\check{Q}_H$ statistic. DV modification is significantly more stable than FR modification. In FR method there is no difference between Levinson algorithm and applying Moore Penrose pseudoinverse. Order of the autoregressive spectral estimator r does not play significant role.

Table 1: *Empirical size (in %) of the modified DV, modified FR and standard LB test of 5% nominal level in the case of the weak AR*(1) *model* (56)*, N=500*

H	1	2	3	4	5	6	7	8	9	10	11	12
Projection method, number of derivatives used, k												
k=1	4.25	2.60	2.45	2.85	2.25	2.00	1.70	2.45	1.60	1.75	2.35	1.60
k=2	3.85	3.40	3.60	2.70	1.85	2.25	2.10	2.55	1.80	2.55	1.95	1.55
k=3	3.35	2.25	3.05	2.45	2.30	2.55	2.30	2.85	1.60	2.40	1.60	1.50
k=4	2.50	3.45	2.90	2.40	2.40	1.95	2.40	2.15	1.95	1.85	1.95	2.10
Francq & Raïssi method, Moore Penrose pseudoinverse, VAR(r) routine												
r=1	4.20	2.95	1.95	1.45	1.30	1.15	0.65	0.60	0.30	0.65	0.50	0.50
r=2	3.55	3.20	2.00	1.35	1.05	1.10	0.90	0.80	0.55	0.35	0.55	0.15
r=3	3.70	3.10	1.70	1.20	1.40	0.90	0.45	0.70	0.70	0.80	0.30	0.20
r=4	3.70	2.80	1.85	1.30	0.95	1.45	0.70	0.60	0.50	0.30	0.25	0.65
r=5	4.40	2.70	1.40	1.70	1.45	0.65	0.85	0.60	0.65	0.45	0.55	0.35
Francq & Raïssi method, Levinson algorithm, VAR(r) routine												
r=1	4.05	2.75	2.45	1.35	1.30	0.90	0.50	0.30	0.75	0.55	0.55	0.35
r=2	3.35	1.90	2.30	1.60	0.75	1.10	1.00	0.80	0.45	0.30	0.40	0.35
r=3	3.60	3.05	1.70	1.65	1.10	0.75	0.85	0.50	0.45	0.45	0.30	0.40
r=4	4.50	2.95	2.05	1.35	1.05	0.90	0.90	0.70	0.35	0.75	0.55	0.30
r=5	3.85	2.65	1.90	2.50	1.15	0.95	0.45	0.65	0.80	0.45	0.45	0.45
Standart Box Pierce test												
LB	n.a	38.9	34.1	28.6	27.0	22.6	22.3	20.4	19.4	18.9	19.7	18.0

Next experiment is based on simulating $n = 1000$ VARMA$(1,1)$ model generated series of lenght $N = 1000$. In this case the model errors are generated as $GARCH(1)$.

$$\begin{aligned}
&X_{n+1} = A_1 X_n + B_1 \varepsilon_n + \varepsilon_{n+1},\\
&A_1 = 0.6 I_2, B_1 = 0.3 I_2,\\
&\varepsilon_n = \mathcal{N}(0, \Sigma_{\varepsilon,n}), \Sigma_{\varepsilon,n} = diag(\sigma_n, \sigma_n)\\
&\sigma_n = \sqrt{1 + 0.4\sigma_{n-1}^2}.
\end{aligned} \tag{57}$$

The series is estimated as VARMA$(1,1)$ and the H_0 hypothesis is being tested using DV test with $\check{Q}_H$ transformed Box-Pierce statistic. The results for FR method are not available because this method only applies for VAR(p) models. For DV modification of BP test we have presented the results for $k = 1, \dots, 5$ numbers of derivatives used in estimating the transform of each $\hat{\rho}_{\hat{\theta}}(j), j = 1, \dots, H$.

Table 2: *Empirical size (in %) of the modified DV of 5% nominal level in the case of the weak VARMA*$(1,1)$ *model* (57)*, N=1000*

H	1	2	3	4	5	6	7	8	9	10	11	12
Projection method, number of derivatives used, k												
k=1	74.5	88.4	80.4	78.6	76.8	71.7	70.5	69.1	67.5	68.1	69.4	70.6
k=2	5.1	5.1	3.6	6.1	6.1	6.2	7.2	7.5	8.7	10.3	8.4	11.7
k=3	5.4	4.4	6.3	6.1	7.2	5.2	9.7	8.7	10.3	8.6	11.9	11.8
k=4	5.2	5.8	5.9	6.7	8.3	8.0	7.7	9.2	11.0	11.3	13.2	13.3
k=5	6.3	5.6	6.7	9.5	10.0	7.5	8.2	9.0	13.0	12.3	14.5	13.1

Most notable characteristic of this experiment are very high size estimates for $k = 1$. However, this result agrees with asymptotic theory. In general in case of testing uncorrelation hypothesis using modified DV method the number of derivatives used has to be larger than 1.

5.2. Empirical power.

Data generating process follows the power size Monte Carlo experiment for weak VAR(2) model in Francq & Raïssi (2007). The true series $\{X_n\}_{n=1}^N$ is generated according to

$$\begin{aligned}
&X_{n+1} = A_1 X_n + A_2 X_{n-1} + \varepsilon_{n+1}, \\
&A_1 = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}, A_2 = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}, \\
&\varepsilon_n = \begin{pmatrix} \eta_n^1 \eta_{n-1}^1 \eta_{n-2}^1 \\ \eta_n^2 \eta_{n-1}^2 \eta_{n-2}^2 \end{pmatrix}, \\
&\begin{pmatrix} \eta_n^1 \\ \eta_n^2 \end{pmatrix} \sim \mathcal{N}(0, I_2), iid \quad n = 1, 2, \dots, N
\end{aligned} \tag{58}$$

We simulated $n = 2000$ independent trajectories of lenght $N = 1000$ of this weak VAR(2) model. For each replication we estimate matrix coefficient A_1 of

VAR(1) model and then apply portmanteau tests of H_0 on the residuals. The asymptotic level is $\alpha = 5\%$. For DV and FR modifications of LB test the H_0 is rejected when $\check{Q}_H > \chi^2_{0.95}(4H)$. For the standard LB test the H_0 is rejected when $\check{Q}_H > \chi^2_{0.95}(4H-4)$. The power is not corrected by size. (seems to follow Francq & Raïssi (2007)).

Table 3: *Empirical power (in %) of the modified DV, modified FR and standard LB test of 5% nominal level in the case of the weak AR(2) model* (58), *N=1000*

H	1	2	3	4	5	6	7	8	9	10	11	12
Projection method, number of derivatives used, k												
k=1	50.95	41.45	35.50	28.75	24.50	22.50	18.85	16.10	14.50	12.80	11.50	12.10
k=2	30.35	41.25	32.90	27.20	23.35	21.50	18.10	17.95	13.45	13.75	10.65	11.40
k=3	23.70	38.85	31.00	28.00	22.70	21.55	19.00	15.55	14.80	12.70	10.85	11.00
k=4	22.45	37.90	31.55	25.80	23.40	18.25	17.75	16.55	13.60	12.25	10.45	10.55
Francq & Raïssi method, Moore Penrose pseudoinverse, VAR(r) routine												
r=1	43.35	51.65	40.50	32.65	28.95	23.30	22.20	18.70	15.35	11.10	9.90	8.75
r=2	45.30	48.90	42.55	33.60	29.25	24.20	20.65	19.55	17.00	11.95	9.50	8.55
r=3	43.75	47.95	41.80	34.90	27.75	22.85	21.70	19.10	16.10	13.50	9.75	8.65
r=4	44.70	49.45	41.10	34.30	27.00	24.25	20.40	19.45	16.45	13.50	9.65	7.85
r=5	43.50	50.85	43.20	32.60	28.65	24.30	21.60	18.85	14.55	12.75	10.55	7.90
Francq & Raïssi method, Levinson algorithm, VAR(r) routine												
r=1	43.30	48.45	40.50	34.20	27.50	23.65	21.00	19.55	17.25	12.95	9.35	7.45
r=2	43.80	50.00	41.70	36.10	28.20	23.20	21.85	19.90	16.40	12.20	8.90	8.30
r=3	43.55	48.70	40.60	34.40	27.65	23.50	21.90	19.10	16.55	12.25	10.35	7.15
r=4	44.20	47.30	40.70	35.40	26.00	24.10	20.85	19.65	16.65	12.25	8.35	8.05
r=5	41.85	46.10	41.05	32.60	28.00	23.80	22.10	19.40	15.40	12.75	9.35	8.40
Standart Box Pierce test												
LB	n.a	84.15	79.90	74.05	68.70	66.15	61.70	59.50	58.50	54.45	55.25	51.90

Power of DV and FR modifications of LB test are comparable. Power of both tests is decreasing with increasing order of $\check{Q}_H$. The reason is the following. For VAR(p) model vectorized residual autocovariances of increasing lags are expotentially converging to zero. However, the error of estimation is not going to be affected by lag number. The same applies to estimating residuals in DV method. This may cause significant errors in the inverses that are necessary in both DV and FR modifications of LB statistic.

In (58) case finite sample performance of DV and FR modifications were comparable. However the lenght of the series $\{X_n\}$ was $N = 1000$, which may be treated as large in some applications. The following example is derived to show the difference in DV and FR methods for small sample $N = 100$ nontrivial case.

The series $\{X\}_{n=1}^{N}$ for $N = 100$ follows VARMA$(1,1)$ model

$$
\begin{aligned}
&X_{n+1} = A_1 X_n + B_1 \varepsilon_n + \varepsilon_{n+1}, \\
&A_1 = \frac{1}{3}\begin{pmatrix} 2 & 0.5 \\ 0.7 & 1 \end{pmatrix}, B_1 = \frac{1}{6}\begin{pmatrix} 2 & -1.5 \\ 0 & 3 \end{pmatrix}, \\
&\varepsilon_n = \begin{pmatrix} \eta_n^1 \eta_{n-1}^1 \eta_{n-2}^1 \\ \eta_n^2 \eta_{n-1}^2 \eta_{n-2}^2 \end{pmatrix}, \\
&\begin{pmatrix} \eta_n^1 \\ \eta_n^2 \end{pmatrix} \sim \mathcal{N}(0, I_2), iid \quad n = 1, 2, \ldots, N
\end{aligned}
\tag{59}
$$

We have simulated $n = 2000$ trajectiories and in each case estimated matrices A_1 of VAR(1) model. Then the DV and FR have tested the H_0 hypothesis. In general testing VARMA$(1,1)$ generating data with VAR(1) specification is one of the simplest tests that may be performed with modified Box statistics. The reason is straightforward, estimating VAR(1) model for VARMA$(1,1)$ has to produce autocorrelated residuals and at the same time the series ε_n is generated as a MDS. Nevertheless the FR modification of BP test is not going to detect the correlation of residuals. This does not apply to our modification of BP test.

Table 4: *Empirical power (in %) of the modified DV, modified FR test of 5% nominal level in the case of the weak VARMA*$(1,1)$ *model* (59)*, N=100*

H	1	2	3	4	5	6	7	8	9	10	11	12
	Projection method, number of derivatives used, k											
k=1	42.90	21.20	12.85	7.55	5.55	4.60	5.70	6.40	7.70	8.60	11.75	18.10
k=2	31.85	22.60	15.05	10.50	7.45	7.05	7.85	7.95	10.30	12.55	17.55	22.40
k=3	28.95	24.25	14.75	10.95	8.05	9.25	8.95	10.75	14.65	17.75	24.15	27.85
k=4	23.95	22.85	15.70	13.30	11.10	12.20	12.00	14.80	16.80	21.60	29.40	35.30
	Francq & Raïssi method, Moore Penrose pseudoinverse, VAR(r) routine											
r=1	18.50	20.80	11.80	6.05	3.95	2.45	1.75	1.50	1.40	1.10	0.75	0.50
r=2	19.25	22.10	12.40	6.30	4.20	2.05	2.05	1.60	1.30	0.90	0.85	1.00
r=3	20.55	22.05	13.20	6.85	4.20	3.05	2.25	1.35	1.25	0.80	1.25	0.70
r=4	20.50	21.40	13.65	7.30	3.90	2.45	2.20	1.15	1.25	0.80	0.45	0.80
r=5	19.00	21.70	13.95	6.30	4.25	2.95	2.50	1.25	1.70	1.65	1.10	0.75
	Francq & Raïssi method, Levinson algorithm, VAR(r) routine											
r=1	20.25	21.00	14.10	8.50	4.50	3.10	1.80	1.55	1.35	1.30	1.30	1.40
r=2	21.40	24.15	12.80	6.15	5.85	3.80	2.20	1.70	1.65	1.25	1.35	1.55
r=3	19.95	21.90	11.40	7.70	5.10	3.75	3.05	2.30	1.70	1.00	0.90	0.85
r=4	21.90	22.20	13.20	8.15	4.60	3.00	2.35	1.85	1.00	1.00	1.00	1.30
r=5	19.75	23.70	12.25	7.00	5.10	3.15	2.65	1.75	1.45	1.65	1.45	1.00

5.3. Economic example of application of our projection correction of BP statistic.

We have remade the example used in Lobato, Nankervis & Savin (2001). These authors tested null hypothesis that $\rho(1) = \cdots = \rho(5)$ of daily currency returns for the pound sterling for the data set $01.01.1993 - 31.12.1996$. The dataset taken into account is pressumed to be a random walk with weakly dependent errors. In order to check our method in multivariate setting we have taken the data set vector of Nominal Major Currencies Dollar Index and Nominal Other Important Trading Partners Dollar Index in the period $01.01.1999 - 31.12.2001$ with the lenght the series $N = 754$. Then we have estimated VAR(1) and VAR(2) models of the demeaned vectors

$$X_t = \begin{bmatrix} y_t \\ z_t \end{bmatrix},$$

$$y_t = NMDI_t - \frac{1}{754} \sum_{n=1}^{754} NMDI_n,$$

$$z_t = NODI_t - \frac{1}{754} \sum_{n=1}^{754} NODI_n,$$

in order to test the uncorrelation of residuals using standart BP test and DV modification of BP test. The results are presented in Table 5.

Table 5: *Values of projection method $\hat{Q}_H^{DV}$ statistic, standart BP statistic $\hat{Q}_H$ and their rejection tresholds for estimated VAR(1) and VAR(2) models on the dataset.*

H	1	2	3	4	5	6	7	8	9	10
Values of projection method $\hat{Q}_H^{DV}$, number of derivatives used, k, VAR(1)										
k=1	10.23	14.01	19.51	28.18	37.79	40.77	41.35	42.06	44.22	46.74
k=2	10.83	13.32	20.10	28.50	31.43	39.10	41.14	43.08	44.88	45.13
k=3	11.25	15.05	24.82	24.88	30.40	39.10	41.66	43.18	43.25	46.73
P-value treshold for $\hat{Q}_H^{DV}$										
	9.48	15.50	21.02	26.29	31.41	36.41	41.33	46.19	50.99	55.75
Standart Box Pierce statistic $\hat{Q}_H$ value										
BP	17.69	20.04	22.92	27.17	34.93	40.38	42.88	44.35	44.70	47.06
P-value treshold for $\hat{Q}_H$										
	n.a.	9.48	15.50	21.02	26.29	31.41	36.41	41.33	46.19	50.99
Values of projection method $\hat{Q}_H^{DV}$, number of derivatives used, k, VAR(2)										
k=1	n.a.	4.47	9.45	19.05	22.36	25.90	27.87	28.11	30.86	33.79
k=2	4.04	9.95	19.05	22.36	25.90	27.87	28.11	30.86	33.79	38.88
k=3	3.92	13.13	19.24	23.44	21.81	27.80	30.01	32.86	34.50	43.13
P-value treshold for $\hat{Q}_H^{DV}$										
	9.48	15.50	21.02	26.29	31.41	36.41	41.33	46.19	50.99	55.75
Standart Box Pierce statistic $\hat{Q}_H$ value										
BP	0.08	4.48	7.27	11.85	19.96	25.08	28.03	29.60	29.87	32.28
P-value treshold for $\hat{Q}_H$										
	n.a.	9.48	15.50	21.02	26.29	31.41	36.41	41.33	46.19	50.99

Clearly as in in Lobato, Nankervis & Savin (2001) projection method statistic $\hat{Q}_H^{DV}$ in VAR(1) case suggests the lack of autocorrelation of residuals in VAR(1) case. We have that transition matrix of estimated VAR(1) is statistically significantly equal to zero. However, standart BP statistic $\hat{Q}_H$ is rejecting null hypothesis suggesting wrong model specification. In this sense our modification of BP test serves its purpose of accepting the hypothesis of two dimensional vector X_t following a random walk. Using the same rule we have estimated a VAR(2) model and tested lack of autocorrelation of residuals. It may be treated as an extreme case because if null hypothesis and random walk hypothesis are both true than we are going to have overfitting of the model. Similarily as in VAR(1) the transition matrices of VAR(2) are significantly equal to zero, but it is worth noting that for $H = 1$ first derivative needed to obtain $\hat{Q}_1^{DV}$ is equal to zero - we may not obtain the inverts. Also note that this exactly is responsible for the same values of $\hat{Q}_H^{DV}$ statistic diagonally between $k = 1$ and $k = 2$. However, our method still outperforms the standart BP statistic for $H \geq 2$.

6. Summary

TO BE WRITTEN

Appendix

- **Estimation of weak VARMA model**

In case of VARMA(p, q) the Least Squares procedure used in VAR(p) estimation can not be used due to recursive nature of residuals $\varepsilon_n(\theta)$. Considering this, it is possible to approximate, given observations $\{X_n\}_{n=1}^N$, the θ_0 as a minimum of the gaussian quasi-likelihood

$$L_N(\theta_0, \Sigma_\varepsilon) = \prod_{n=1}^{N} \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma_\varepsilon}} \exp\left\{-\frac{1}{2}\varepsilon_n(\theta)'\Sigma_\varepsilon^{-1}\varepsilon_n(\theta)\right\}$$

Now QMLE of $(\theta_0, \Sigma_\varepsilon)$ is a measurable solution $(\hat{\theta}, \hat{\Sigma}_\varepsilon)$

$$(\hat{\theta}, \hat{\Sigma}_\varepsilon) = \underset{(\theta, \Sigma_\varepsilon)}{\arg\min} \left\{ \log(\det\Sigma_\varepsilon) + \frac{1}{N} \sum_{n=1}^{N} \varepsilon_n'(\theta) \Sigma_\varepsilon^{-1} \varepsilon_n(\theta) \right\}. \tag{60}$$

Note that if we are looking only for estimator of θ_0 parameters than we may write the following

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} \left\{ \log(\det\hat{\Sigma}_\varepsilon(\theta)) + \frac{1}{N} \sum_{n=1}^{N} \varepsilon_n'(\theta) \hat{\Sigma}_\varepsilon^{-1}(\theta) \varepsilon_n(\theta) \right\} \tag{61}$$

where

$$\hat{\Sigma}_\varepsilon(\theta) = \frac{1}{N} \sum_{n=1}^{N} \varepsilon_n(\theta) \varepsilon_n(\theta)' \tag{62}$$

Procedures (60) and (61) constitute the QML estimator of $(\theta, \Sigma_\varepsilon)$ for strong VARMA model.
Asymptotic normality of $\hat{\theta}$ estimator comes from Boubacar Mainassara & Francq (2009, *Theorem* 3) which states that under **A1"-A3** the LSE estimator of θ_0

$$\hat{\theta}_{LS} = \underset{\theta \in \Theta}{\arg\min} \log \det \left(\sum_{n=1}^{N} \varepsilon_n(\theta) \varepsilon_n(\theta)' \right) \tag{63}$$

coincides with QML estimators (60) and (61). This implies that under assumptions **A1"-A3** we have that $\hat{\theta} = \theta_0 + O_p(N^{-\frac{1}{2}})$ for QML estimators (60) and (61).

- **Proof of Proposition 1**

In order to prove the claim it is sufficient to show that for each $j = 1, \ldots, m$

$$\text{vec}\left(\hat{\Gamma}_{\hat{\theta}}(j)\right) - \text{vec}\left(\hat{\Gamma}_{\theta_0}(j)\right) = \text{plim} \nabla \hat{\Gamma}_{\theta_0}(j)(\hat{\theta} - \theta_0) + D_N(j)$$

where $\text{plim}\nabla\hat{\Gamma}_{\theta_0}(j) = \nabla\Gamma_{\theta_0}(j)$ by assumption and and $D_N(j) = o_P(N^{-\frac{1}{2}})$.
Now we can not use in this case the standard approximation

$$D_N(j) = (\hat{\theta} - \theta_0)' \left\{ \frac{\partial^2 \text{vec}(\Gamma_\theta(j))}{\partial\theta\partial\theta'}(\theta^*) \right\} (\hat{\theta} - \theta_0)$$

with θ^* satisfying $||\theta^* - \theta_0||_d \leq ||\hat{\theta} - \theta_0||_d$ because $\frac{\partial^2 \text{vec}(\Gamma(j))}{\partial\theta\partial\theta'}(\theta^*)$ is second order tensor. However, we may show the following instead

$$\left[\text{vec}\left(\hat{\Gamma}_{\hat{\theta}}(j)\right)\right]_{(i)} - \left[\text{vec}\left(\hat{\Gamma}_{\theta_0}(j)\right)\right]_{(i)} = [\nabla\Gamma_{\theta_0}(j)]_{i'\text{th row}} (\hat{\theta} - \theta_0) + D_N(j)_{(i)}$$

where (i) denotes position in vector and $i = 1, 2, \ldots, d^2$, with $D_N(j)_{(i)} = o_P(N^{-\frac{1}{2}})$.
We are going to show that

$$D_N(j)_{(i)} = (\hat{\theta} - \theta_0)' \frac{\partial^2 \text{vec}(\Gamma(j))_{(i)}}{\partial\theta\partial\theta'}(\theta^*)(\hat{\theta} - \theta_0) \tag{64}$$

Now in order to prove that (64) holds it is enough to show that $\frac{\partial^2 \text{vec}(\Gamma(j))_{(i)}}{\partial\theta\partial\theta'}(\theta^*) = O_P(1)$. The easiest way to show this condition is use the definition of derivative of $\frac{\partial \text{vec}(\Gamma(j))}{\partial\theta'}_{(i)}$ with respect to θ evaluated in θ^*. First we need to show that in the neighbourhood of θ_0 we have

$$\frac{\partial \text{vec}(\Gamma(j))}{\partial\theta'}_{(i)}(\theta_*) = O_P(1), \text{ for } j = 1, \ldots, m, i = 1, \ldots, d^2 \tag{65}$$

The above condition follows from *Proposition* 2 and **A3** because θ_* is close to θ_0 and *Proposition* 3 because $\text{plim}\nabla\hat{\Gamma}_{\theta_0}(j) = E\nabla\hat{\Gamma}_{\theta_0}(j)$.
To show that a second derivative of i'th row of $\nabla\Gamma_{\theta_*}(j)$ is bounded we may use **A3** ie. smoothness of θ function in Θ for the derivative of $\text{vec}\nabla\Gamma_{\theta_*}(j)$. Note that

$$\nabla\Gamma_{\theta_*}(j) \xrightarrow{p} E\frac{\partial\varepsilon_{n+j}(\theta_0)}{\partial\theta'} \otimes \varepsilon_n(\theta_0) = E\nabla_{\theta_0}\varepsilon_{n+j} \otimes \varepsilon_n(\theta_0)$$

Now we may write

$$\frac{\partial \mathrm{vec}(\nabla \Gamma_{\theta_*}(j))}{\partial \theta'} \xrightarrow{p} E \frac{\partial \mathrm{vec}(\nabla_{\theta_0} \varepsilon_{n+j} \otimes \varepsilon_n(\theta_0))}{\partial \varepsilon_n'} \times \frac{\partial \varepsilon_n(\theta_0)}{\partial \theta'}$$

where $\frac{\partial \varepsilon_n(\theta_0)}{\partial \theta'} = O_P(1)$ in general by **A3** and it may be shown that

$$\frac{\partial \mathrm{vec}(\nabla_{\theta_0} \varepsilon_{n+j} \otimes \varepsilon_n(\theta_0))}{\partial \varepsilon_n'} = \mathrm{vec}(\nabla_{\theta_0} \varepsilon_{n+j}) \otimes I_d = O_P(1)$$

QED

- **Proof of Proposition 2**

The proof of statement

$$G_{\theta_0} = \mathrm{diag}\left(\mathrm{vec}(\Gamma_{\theta_0}(1))\mathrm{vec}(\Gamma_{\theta_0}(1))', \ldots, \mathrm{vec}(\Gamma_{\theta_0}(m))\mathrm{vec}(\Gamma_{\theta_0}(m))'\right)$$

is straightforward. Note that following formulation of G_{θ_0} in (49) under iid ε_n we would obtain that only submatrices $C_{11}, \ldots, C_{mm}$ are nonzero and we have

$$\mathrm{vec}(\varepsilon_n \varepsilon_{n+i}')\mathrm{vec}(\varepsilon_n \varepsilon_{n+i}')' = \mathrm{vec}(\Gamma_{\theta_0}(i))\mathrm{vec}(\Gamma_{\theta_0}(i))'.$$

Consistency follows from consistency of $\hat{\Gamma}_{\hat{\theta}}(i), i = 1, \ldots, m$.

Now in order to show the main claim we have to present the form of derivatives in multivariate setting. From (1) we may write the system of equations

$$\frac{\partial \varepsilon_n(\theta)}{\partial \mathrm{vec}(A_i)'}(\theta) = -X_{n-i}' \otimes I_d - \sum_{j=1}^{q} B_j \frac{\partial \varepsilon(\theta)_{n-j}}{\partial \mathrm{vec}(A_i)'}(\theta) \tag{66}$$

and

$$\frac{\partial \varepsilon(\theta)_n}{\partial \mathrm{vec}(B_j)'}(\theta) = -\frac{\partial B_j \varepsilon(\theta)_{n-j}}{\partial \mathrm{vec}(B_j)'}(\theta) - \sum_{i \neq j}^{q} B_i \frac{\partial \varepsilon(\theta)_{n-i}}{\partial \mathrm{vec}(B_j)'}(\theta) \tag{67}$$

for $i = 1, \ldots, p$ and $j = 1, \ldots, q$. Now, we may write the first expression on the right side in (67) as

$$\frac{\partial B_j \varepsilon_{n-j}(\theta)}{\partial \mathrm{vec}(B_j)'}(\theta) = -\varepsilon_{n-j}(\hat{\theta})' \otimes I_d - B_j \frac{\partial \varepsilon_{n-j}(\theta)}{\partial \mathrm{vec}(B_j)'}(\theta)$$

to get the final expression

$$\frac{\partial \varepsilon_n(\theta)}{\partial \mathrm{vec}(B_j)'}(\theta) = -\varepsilon_{n-j}(\theta)' \otimes I_d - \sum_{i=1}^{q} B_i \frac{\partial \varepsilon_{n-i}(\theta)}{\partial \mathrm{vec}(B_j)'}(\theta) \tag{68}$$

In order to write it compactly let us introduce the following notation for derivatives evaluated in $\hat{\theta}$

$$\nabla_{A_i} \varepsilon_{n-i}(\hat{\theta}) = X'_{n-i} \otimes I_d - \sum_{j=1}^{q} B_j(\hat{\theta}) \nabla_{A_i} \varepsilon_{n-j}(\hat{\theta}) \text{ for } i = 1, \ldots, p$$
$$\nabla_{B_i} \varepsilon_{n-i}(\hat{\theta}) = \varepsilon_{n-i}(\hat{\theta})' \otimes I_d - \sum_{j=1}^{q} B_j(\hat{\theta}) \nabla_{B_i} \varepsilon_{n-j}(\hat{\theta}) \text{ for } i = 1, \ldots, q$$

where the entire jacobian matrix may be written as

$$\nabla_\theta \varepsilon_n(\hat{\theta}) = \nabla_{[A,B]} \varepsilon_n(\hat{\theta}) =$$
$$\left[\nabla_{A_1} \varepsilon_n(\hat{\theta}), \nabla_{A_2} \varepsilon_n(\hat{\theta}), \ldots, \nabla_{A_p} \varepsilon_n(\hat{\theta}), \ldots, \nabla_{B_1} \varepsilon_n(\hat{\theta}), \nabla_{B_2} \varepsilon_n(\hat{\theta}), \ldots, \nabla_{B_q} \varepsilon_n(\hat{\theta})\right]$$

with

$$\nabla_A \varepsilon_n(\hat{\theta}) = \left[\nabla_{A_1} \varepsilon_n(\hat{\theta}), \nabla_{A_2} \varepsilon_n(\hat{\theta}), \ldots, \nabla_{A_p} \varepsilon_n(\hat{\theta})\right]$$
$$\nabla_B \varepsilon_n(\hat{\theta}) = \left[\nabla_{B_1} \varepsilon_n(\hat{\theta}), \nabla_{B_2} \varepsilon_n(\hat{\theta}), \ldots, \nabla_{B_q} \varepsilon_n(\hat{\theta})\right]$$

Now, it leads to the following

$$\nabla_A \varepsilon(\theta)_n = -\begin{bmatrix} X_{n-1} \\ X_{n-2} \\ \vdots \\ X_{n-p} \end{bmatrix}' \otimes I_d - [B_1, B_2, \ldots, B_q]_{(\hat{\theta})} \begin{bmatrix} \nabla_{A_1}\varepsilon_{n-1}(\hat{\theta}) & \nabla_{A_2}\varepsilon_{n-1}(\hat{\theta}) & \ldots & \nabla_{A_p}\varepsilon_{n-1}(\hat{\theta}) \\ \nabla_{A_1}\varepsilon_{n-2}(\hat{\theta}) & \nabla_{A_2}\varepsilon_{n-2}(\hat{\theta}) & \ldots & \nabla_{A_p}\varepsilon_{n-2}(\hat{\theta}) \\ \vdots & \vdots & \vdots & \vdots \\ \nabla_{A_1}\varepsilon_{n-q}(\hat{\theta}) & \nabla_{A_2}\varepsilon_{n-q}(\hat{\theta}) & \ldots & \nabla_{A_p}\varepsilon_{n-q}(\hat{\theta}) \end{bmatrix}$$

which is equivalent to

$$\nabla_A \varepsilon_n(\hat{\theta}) = -\begin{bmatrix} X_{n-1} \\ X_{n-2} \\ \vdots \\ X_{n-p} \end{bmatrix}' \otimes I_d - [B_1, B_2, \ldots, B_q]_{(\hat{\theta})} \begin{bmatrix} \nabla_A\varepsilon_{n-1}(\hat{\theta}) \\ \nabla_A\varepsilon_{n-2}(\hat{\theta}) \\ \vdots \\ \nabla_A\varepsilon_{n-q}(\hat{\theta}) \end{bmatrix} \tag{69}$$

and in a similar manner

$$\nabla_B \varepsilon_n(\hat{\theta}) = -\begin{bmatrix} \varepsilon_{n-1}(\hat{\theta}) \\ \varepsilon_{n-2}(\hat{\theta}) \\ \vdots \\ \varepsilon_{n-q}(\hat{\theta}) \end{bmatrix}' \otimes I_d - [B_1, B_2, \ldots, B_q]_{(\hat{\theta})} \begin{bmatrix} \nabla_B\varepsilon_{n-1}(\hat{\theta}) \\ \nabla_B\varepsilon_{n-2}(\hat{\theta}) \\ \vdots \\ \nabla_B\varepsilon_{n-q}(\hat{\theta}) \end{bmatrix} \tag{70}$$

Finally according to (69) and (70) it is possible to write full derivative $\nabla_\theta \varepsilon_n(\hat{\theta})$ as

$$\nabla_\theta \varepsilon_n(\hat{\theta}) = -\left[\begin{bmatrix} X_{n-1} \\ X_{n-2} \\ \vdots \\ X_{n-p} \end{bmatrix}', \begin{bmatrix} \varepsilon_{n-1}(\hat{\theta}) \\ \varepsilon_{n-2}(\hat{\theta}) \\ \vdots \\ \varepsilon_{n-q}(\hat{\theta}) \end{bmatrix}'\right] \otimes I_d - [B_1, B_2, \ldots, B_q]_{(\hat{\theta})} \begin{bmatrix} \nabla_\theta\varepsilon_{n-1}(\hat{\theta}) \\ \nabla_\theta\varepsilon_{n-2}(\hat{\theta}) \\ \vdots \\ \nabla_\theta\varepsilon_{n-q}(\hat{\theta}) \end{bmatrix} \tag{71}$$

Now, going back to summation notation equation (71) may be written as

$$\nabla_\theta \varepsilon_n(\hat{\theta}) = -F_n - \sum\nolimits_{i=1}^{q} B_i \nabla_\theta \varepsilon_{n-i}(\hat{\theta}) \tag{72}$$

where

$$F_n = \overbrace{\left[\begin{bmatrix} X_{n-1} \otimes I'_d \\ X_{n-2} \otimes I'_d \\ \vdots \\ X_{n-p} \otimes I'_d \end{bmatrix}' , \begin{bmatrix} \varepsilon_{n-1}(\hat{\theta}) \otimes I'_d \\ \varepsilon_{n-2}(\hat{\theta}) \otimes I'_d \\ \vdots \\ \varepsilon_{n-q}(\hat{\theta}) \otimes I'_d \end{bmatrix}' \right]}^{1\times 1 \text{ block}}$$

In order to give the explicit solution to $\nabla_\theta \varepsilon_n(\theta)$ for $n = 1, \ldots, N$, $\theta \in \Theta$ let us define the following matrices

$$D_n = \begin{bmatrix} \nabla_\theta \varepsilon_n(\theta) \\ \nabla_\theta \varepsilon_{n-1}(\theta) \\ \vdots \\ \nabla_\theta \varepsilon_{n-q+1}(\theta) \end{bmatrix}, \tag{73}$$

$$\mathbf{\Psi}_n^{(q)} = \overbrace{\begin{bmatrix} F_n \\ 0 \\ \vdots \\ 0 \end{bmatrix}}^{q\times 1 \text{ blocks}}, \tag{74}$$

$$\mathbf{B}_{(q)} = \overbrace{\begin{bmatrix} -B_1 & \ldots & -B_{q-1} & -B_q \\ I_d & & & \\ & \ddots & & \\ & & I_d & 0 \end{bmatrix}_\theta}^{q\times q \text{ blocks}} \tag{75}$$

Using the above we may write equation (72) as

$$D_n = \mathbf{B}_{(q)} D_{n-1} - \mathbf{\Psi}_n^{(q)}, \tag{76}$$

which gives unique solution

$$D_n = -\sum_{i=0}^{\infty} \mathbf{B}_{(q)}^i \mathbf{\Psi}_{n-i}^{(q)} \tag{77}$$

It should be noted that for any $1 \leq n \leq N$ D_n is converging in probability to the vector $D_n(\theta_0)$ depending on values $\{X_1, \dots, X_{n-1}\}$ under weak and strong VARMA specification. It follows immediately from the expressions (72) and (77). On the other hand it is clear that the sequence $\{\nabla_\theta \varepsilon_n(\theta_0)\}$ is autocorrelated because it admits the multivariate expotential expansion. It is easy to see that in a VAR(p,q) formulation we have following (72), $\nabla_\theta \varepsilon_n(\hat{\theta}) = -F_n$. Note that in VARMA$(p,q)$ case the sequence $\{D_n - D_n^{var}\}_{n=1}^{\infty}$, where D_n^{var} represents the covariance derivatives matrix corresponding to VAR(p) part of a VARMA(p,q) model, is an MDS. This gives an interesting result, while VARMA(p,q) and VAR(p) are structurally entirely different, because VARMA(p,q) is nonlinear, the derivatives of residuals with respect to parameters are different merely by the martingale difference process.
We are mainly interested in obtaining the derivatives of residual autocovariances $\nabla_{\hat{\theta}} \hat{\Gamma}(i)$. In general VARMA$(p,q)$ the object of the form $\nabla_{\hat{\theta}} \varepsilon$ does not exist because $\nabla_{\hat{\theta}} \varepsilon_n$ is dependent on $\{\nabla_{\hat{\theta}} \varepsilon_i, X_i\}_{i=1}^{n-1}$. However we are going to show that it is possible to obtain $\nabla_\theta \hat{\Gamma}(i)$ for arbitrary $j = 1, \dots, m$ and ay $\theta \in \Theta$. We may write the following

$$\begin{aligned}
\nabla_\theta \hat{\Gamma}_n(j) =& \frac{\partial \mathrm{vec}(\varepsilon_{n+j}(\theta)\varepsilon_n'(\theta))}{\partial \theta'} = \\
=& \frac{\partial \varepsilon_{n+j}(\theta)}{\partial \theta'} \otimes \varepsilon_n(\theta) + \varepsilon_{n+j}(\theta) \otimes \frac{\partial \varepsilon_n(\theta)}{\partial \theta'} = \\
=& \nabla_\theta \varepsilon_{n+j} \otimes \varepsilon_n(\theta) + \varepsilon_{n+j}(\theta) \otimes \nabla_\theta \varepsilon_n
\end{aligned} \tag{78}$$

Now, from equations (77) and (78) we may obtain

$$D_n \otimes \varepsilon_{n-m}(\theta) = -\sum_{i=0}^{\infty} \mathbf{B}_{(q)}^{i} \mathbf{\Psi}_{n-i}^{(q)} \otimes \varepsilon_{n-m}(\theta) = \begin{bmatrix} \nabla_\theta \mathrm{vec}\hat{\Gamma}_n(m) \\ \nabla_\theta \mathrm{vec}\hat{\Gamma}_n(m-1) \\ \vdots \\ \nabla_\theta \mathrm{vec}\hat{\Gamma}_n(m-q) \end{bmatrix}_{(\theta)} \tag{79}$$

Note that in *Proposition* 3 implies that we would have

$$-\frac{1}{N}\sum_{n=p+q}^{N}\sum_{i=0}^{\infty} \mathbf{B}_{(q)}^{i}(\theta_0) \mathbf{\Psi}_{n-i}^{(q)}(\theta_0) \otimes \varepsilon_{n-m}(\theta_0) \xrightarrow{p} E \begin{bmatrix} \nabla_\theta \mathrm{vec}\hat{\Gamma}(m) \\ \nabla_\theta \mathrm{vec}\hat{\Gamma}(m-1) \\ \vdots \\ \nabla_\theta \mathrm{vec}\hat{\Gamma}(m-q) \end{bmatrix}_{(\theta_0)} \tag{80}$$

which is also going to be satisfied for $\theta \in \Theta$ by stationarity of $\varepsilon_n(\theta)$.
The way to obtain the form of matrices $\mathbf{\Psi}_k^{(m)}$ and $\mathbf{B}_{(m)}^{i}$ from *Proposition* 2 is a straightforward application of (79) and (77) for matrix

$$D_n^{(m)}(\theta_0) = \begin{bmatrix} \nabla_\theta \varepsilon_n(\theta) \\ \nabla_\theta \varepsilon_{n-1}(\theta) \\ \vdots \\ \nabla_\theta \varepsilon_{n-q}(\theta) \\ \vdots \\ \nabla_\theta \varepsilon_{n-m+1}(\theta) \end{bmatrix}_{(\theta_0)}$$

assuming that $m \geq q$ and proposing a natural estimator

$$\check{\zeta}_{\theta_0}^{(m)} = \begin{bmatrix} \nabla_\theta \mathrm{vec}\hat{\Gamma}(m) \\ \vdots \\ \nabla_\theta \mathrm{vec}\hat{\Gamma}(1) \end{bmatrix}_{(\theta_0)} = \frac{1}{N}\sum_{n=p+q}^{N} (\check{\zeta}_{\theta_0}^{(m)})_n$$

with

$$(\check{\zeta}^{(m)}_{\theta_0})_n = \begin{bmatrix} \nabla_\theta \mathrm{vec}\hat{\Gamma}_n(m) \\ \vdots \\ \nabla_\theta \mathrm{vec}\hat{\Gamma}_n(1) \end{bmatrix}_{(\theta_0)}$$

Now repeating the same reasoning for matrix $D^{(m)}_n$ as in (79) and (80) will yield equation (33), expanded matrix $\mathbf{B}_{(m)}(\theta_0)$ and matrix $\mathbf{\Psi}^{(m)}_k(\theta_0)$ that does not change compared to (74).

In order to prove convergence in probability (32) we are going to use equations (66) and (68). Let us use the convention that for all $\theta \in \Theta$ we have $\varepsilon_n(\theta) = 0$ given that $n \leq 0$ or $n \geq N$.

(This proof is not correct, the reasoning will break in the limit)

Now the sufficient condition that must hold for derivative of order $h = 1, \ldots, m$ is

$$\begin{aligned}
&\frac{1}{N}\sum_{n=i+h}^{N} (X'_{n-i} \otimes I_d) \otimes (\varepsilon_{n-h}(\hat{\theta}) - \varepsilon_{n-h}(\theta_0)) + \\
&+ \frac{1}{N}\sum_{n=i+h}^{N} \left\{ \sum_{j=1}^{q} B_j(\hat{\theta}) \frac{\partial \varepsilon_{n-j}(\hat{\theta})}{\partial \mathrm{vec}(A_i)'} \right\} \otimes (\varepsilon_{n-h}(\hat{\theta}) - \varepsilon_{n-h}(\theta_0)) + \\
&+ \frac{1}{N}\sum_{n=i+h}^{N} \left\{ \sum_{j=1}^{q} B_j(\hat{\theta}) \frac{\partial \varepsilon_{n-j}(\hat{\theta})}{\partial \mathrm{vec}(A_i)'} - B_j(\theta_0) \frac{\partial \varepsilon_{n-j}(\theta_0)}{\partial \mathrm{vec}(A_i)'} \right\} \otimes \varepsilon_{n-h}(\theta_0) \\
&= o_P(1)
\end{aligned} \tag{81}$$

for any $i = 1, \ldots, p$ after using $A_1 \otimes B_1 - A_2 \otimes B_2 = A_1 \otimes (B_1 - B_2) + (A_1 - A_2) \otimes B_2$. Note that the bound

$$\sup_{0 \leq n \leq N} ||\varepsilon_{n-h}(\hat{\theta}) - \varepsilon_{n-h}(\theta_0)|| = O_P(N^{-\frac{1}{2}}) \tag{82}$$

comes from the observation that estimation error $(\hat{\theta}-\theta_0)$ is asymptotically normal and at the same time by **A3** we have that the function

$$\theta \to [\mathrm{vec}A_1', \ldots, \mathrm{vec}A_p', \mathrm{vec}B_1', \ldots, \mathrm{vec}B_q']'$$

is at least three times differentiable. Thus by smoothness the distance between $\varepsilon_n(\hat{\theta})$ and $\varepsilon_n(\theta_0)$ is going to be bounded uniformly by the distance between $\hat{\theta}$ and θ_0 multiplied by some constant (intermediate value theorem). Now by (82) we have

$$\begin{aligned}
&||\frac{1}{N}\sum_{n=i+h}^{N}({X'}_{n-i}\otimes I_d)\otimes(\varepsilon_{n-h}(\hat{\theta})-\varepsilon_{n-h}(\theta_0))|| \leq \\
&\leq \sup_{0\leq n\leq N}||{X'}_{n-i}\otimes I_d|| \sup_{0\leq n\leq N}||\varepsilon_n(\hat{\theta})-\varepsilon_n(\theta_0)|| = o_P(1)
\end{aligned}$$

because $\sup_{0\leq n\leq N}||{X'}_{n-i}\otimes I_d|| = O_P(1)$. In the second expression we may use similar argument but we need the following bound

$$\sup_{0\leq n\leq N}\sum_{j=1}^{q}B_j(\hat{\theta})\frac{\partial\varepsilon_{n-j}(\hat{\theta})}{\partial\mathrm{vec}(A_i)'} = O_P(1)$$

for any $h = 1\ldots, m$ which comes again from assumption **A3**. In the third expression due to convention and (68) we may write

$$\sup_{0\leq n\leq N}||\frac{\partial\varepsilon_n(\hat{\theta})}{\partial\mathrm{vec}(A_i)'}-\frac{\partial\varepsilon_n(\theta_0)}{\partial\mathrm{vec}(A_i)'}|| = f(\{X_j\}_{k=1}^{n-1}, \{B_i(\theta_0)\}_{i=1}^{q}, \{B_i(\hat{\theta})\}_{i=1}^{q})$$

where f is a bounded linear, measurable and smooth function. Now we have obviously that $B_i(\hat{\theta}) = B_i(\theta_0)+O_P(N^{-\frac{1}{2}})$ so by this, **A3** and continuous mapping theorem we will get $f(\cdot) = O_P(N^{-\frac{1}{2}})$ and

$$\sup_{0\leq n\leq N}||B_j(\hat{\theta})\frac{\partial\varepsilon_n(\hat{\theta})}{\partial\mathrm{vec}(A_i)'}-B_j(\theta_0)\frac{\partial\varepsilon_n(\theta_0)}{\partial\mathrm{vec}(A_i)'}|| = O_P(N^{-\frac{1}{2}})$$

for each $n = 1, \ldots, N$. So by rough approximation we will get $o_P(1)$ convergence rate for the second expression in (81). Using the same arguments for equation (68) we get

$$\begin{aligned}
&\frac{1}{N}\sum_{n=i+h}^{N} {\varepsilon'}_{n-i}(\hat{\theta}) \otimes I_d \otimes (\varepsilon_{n-h}(\hat{\theta}) - \varepsilon_{n-h}(\theta_0)) + \\
&+ \frac{1}{N}\sum_{n=i+h}^{N} ({\varepsilon'}_{n-i}(\hat{\theta}) - {\varepsilon'}_{n-i}(\theta_0)) \otimes I_d \otimes \varepsilon_{n-h}(\theta_0) + \\
&+ \frac{1}{N}\sum_{n=i+h}^{N} \left\{ \sum_{j=1}^{q} B_j(\hat{\theta}) \frac{\partial \varepsilon_{n-j}(\hat{\theta})}{\partial \mathrm{vec}(B_i)'} \right\} \otimes (\varepsilon_{n-h}(\hat{\theta}) - \varepsilon_{n-h}(\theta_0)) + \\
&+ \frac{1}{N}\sum_{n=i+h}^{N} \left\{ \sum_{j=1}^{q} B_j(\hat{\theta}) \frac{\partial \varepsilon_{n-j}(\hat{\theta})}{\partial \mathrm{vec}(B_i)'} - B_j(\theta_0) \frac{\partial \varepsilon_{n-j}(\theta_0)}{\partial \mathrm{vec}(B_i)'} \right\} \otimes \varepsilon_{n-h}(\theta_0) \\
&= o_P(1)
\end{aligned} \tag{83}$$

For the first, second and third expressions in (83) the reasoning is the same as for previous case because we may use the uniform bound (82). Forth expression is an analog of fourth expression in (81).
QED

- **Proof of Proposition 3**

From *Proposition 2* for any $1 \leq k, l \leq N$ we have that under **A1-A3** it is possible to write following (33)

$$\check{\zeta}_{\theta_0}^{(m)} = -\frac{1}{N}\sum_{n=p+q}^{N} \left(\sum_{i=0}^{\infty} \mathbf{B}_{(m)}^{i} \mathbf{\Psi}_{n-i}^{(m)} \otimes \varepsilon_{n-m}(\theta_0) \right)$$

where in general we have

$$\mathbf{\Psi}_n^{(m)} \otimes \varepsilon_{n'}(\theta_0) =$$

$$\begin{bmatrix} X'_{n-1} \otimes I_d \otimes \varepsilon_{n'} & \dots & X'_{n-p} \otimes I_d \otimes \varepsilon_{n'} & \varepsilon'_{n-1} \otimes I_d \otimes \varepsilon_{n'} & \dots & \varepsilon'_{n-q} \otimes I_d \otimes \varepsilon_{n'} \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \end{bmatrix}_{(\theta_0)}$$

It is easy to check that in this case kronecker product is partially commutative so we may write

$$\varepsilon(\theta)'_k \otimes I_d \otimes \varepsilon(\theta)_l = \varepsilon(\theta)'_k \otimes \varepsilon(\theta)_l \otimes I_d$$

for $1 \leq k, l \leq N$ so we have

$$\mathbf{\Psi}_n^{(m)} \otimes \varepsilon_{n'}(\theta_0) = \begin{bmatrix} X'_{n-1} \otimes \varepsilon_{n'} & \dots & X'_{n-p} \otimes \varepsilon_{n'} & \varepsilon'_{n-1} \otimes \varepsilon_{n'} & \dots & \varepsilon'_{n-q} \otimes \varepsilon_{n'} \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \end{bmatrix}_{(\theta_0)} \otimes I_d \tag{84}$$

Now let us divide each matrix $[\mathbf{\Psi}_n^{(m)} \otimes \varepsilon_{n'}(\theta_0)]$ into blocks

$$[\mathbf{\Psi}_n^{(q)} \otimes \varepsilon_{n'}(\theta_0)] = \big([\mathbf{\Psi}_n^{(m)} \otimes \varepsilon_{n'}(\theta_0)]_1, [\mathbf{\Psi}_n^{(m)} \otimes \varepsilon_{n'}(\theta_0)]_2\big)$$

where

$$[\mathbf{\Psi}_n^{(m)} \otimes \varepsilon_{n'}(\theta_0)]_1 = \begin{bmatrix} X'_{n-1} \otimes \varepsilon_{n'} & \dots & X'_{n-p} \otimes \varepsilon_{n'} \\ 0 & \dots & 0 \\ 0 & \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \dots & 0 \\ & & \end{bmatrix}_{(\theta_0)} \otimes I_d \tag{85}$$

and

$$[\mathbf{\Psi}_n^{(m)} \otimes \varepsilon_{n'}(\theta_0)]_2 = \begin{bmatrix} \varepsilon'_{n-1} \otimes \varepsilon_{n'} & \dots & \varepsilon'_{n-q} \otimes \varepsilon_{n'} \\ 0 & \dots & 0 \\ 0 & \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \dots & 0 \\ & & \end{bmatrix}_{(\theta_0)} \otimes I_d \tag{86}$$

Now considering (33) we have that matrices of the form (85) and (86) are responsible respectively for derivatives of residual covariances with respect to vec$[A_1, \dots, A_p]$ and vec$[B_1, \dots, B_q]$. Taking the assumption that $m \geq q$ we have for (85) by LLN

$$
\begin{aligned}
&-\frac{1}{N}\sum_{n=m}^{N}\left(\sum_{i=0}^{\infty}\mathbf{B}_{(m)}^{i}[\mathbf{\Psi}_{n-i}^{(m)}\otimes\varepsilon_{n-m}(\theta_0)]_1\right)\xrightarrow{p}\\
&-\mathbf{B}_{(m)}^{m-p+1}\begin{bmatrix} 0 & \dots & 0 & \Sigma_{\varepsilon}^{\otimes} \\ 0_{(m-1)\times 1} & \dots & 0_{(m-1)\times 1} & 0_{(m-1)\times 1}\end{bmatrix}+\\
&-\mathbf{B}_{(m)}^{m-p+2}\begin{bmatrix} 0 & \dots & \Sigma_{\varepsilon}^{\otimes} & 0 \\ 0_{(m-1)\times 1} & \dots & 0_{(m-1)\times 1} & 0_{(m-1)\times 1}\end{bmatrix}+\dots\\
&-\mathbf{B}_{(m)}^{m}\begin{bmatrix} \Sigma_{\varepsilon}^{\otimes} & \dots & 0 & 0 \\ 0_{(m-1)\times 1} & \dots & 0_{(m-1)\times 1} & 0_{(m-1)\times 1}\end{bmatrix}
\end{aligned}
$$

where $\Sigma_{\varepsilon}^{\otimes}=E[\varepsilon_n(\theta_0)'\otimes\varepsilon_n(\theta_0)]\otimes I_d$. In a similar manner we are going to obtain for (86)

$$
\begin{aligned}
&-\frac{1}{N}\sum_{n=m}^{N}\left(\sum_{i=0}^{\infty}\mathbf{B}_{(m)}^{i}[\mathbf{\Psi}_{n-i}^{(m)}\otimes\varepsilon_{n-m}(\theta_0)]_2\right)\xrightarrow{p}\\
&-\mathbf{B}_{(m)}^{m-q+1}\begin{bmatrix} 0 & \dots & 0 & \Sigma_{\varepsilon}^{\otimes} \\ 0_{(m-1)\times 1} & \dots & 0_{(m-1)\times 1} & 0_{(m-1)\times 1}\end{bmatrix}+\\
&-\mathbf{B}_{(m)}^{m-q+2}\begin{bmatrix} 0 & \dots & \Sigma_{\varepsilon}^{\otimes} & 0 \\ 0_{(m-1)\times 1} & \dots & 0_{(m-1)\times 1} & 0_{(m-1)\times 1}\end{bmatrix}+\dots\\
&-\mathbf{B}_{(m)}^{m}\begin{bmatrix} \Sigma_{\varepsilon}^{\otimes} & \dots & 0 & 0 \\ 0_{(m-1)\times 1} & \dots & 0_{(m-1)\times 1} & 0_{(m-1)\times 1}\end{bmatrix}
\end{aligned}
$$

QED

- **Proof of Theorem 1**

Note that (30) is implied immediately by (30). Let us derive the asymptotic distribution of $\check{Q}(\theta_0)$, using the asymptotic limit of projection operator, $\Im_{\theta_0}^{(m,r)}$. Thus from (27) we have

$$\tilde{\rho}_{\theta_0}(j) = \Im_{\theta_0}^{(m,r)}(\rho_{\theta_0}(j)) = \rho_{\theta_0}(j) - \bar{\xi}_{\theta_0}^{(m)}(j) \left(\sum_{i=j+1}^{j+r} \bar{\xi}_{\theta_0}^{(m)}(i)' \bar{\xi}_{\theta_0}^{(m)}(i) \right)^{-1} \sum_{i=j+1}^{j+r} \bar{\xi}_{\theta_0}^{(m)}(i)' \rho_{\theta_0}(i)$$

$j = 1, \ldots, m - r,\ H \leq m - r,\ r < m.$

where $\rho_{\theta_0}(i) \sim \mathcal{N}(0, I_{d^2})$ iid for $i = 1, \ldots, m$. However, $\tilde{\rho}_{\theta_0}(j)$ is not going have an asymptotically normal distribution because of weighting by coefficients of the form

$$\xi_{\theta_0}^{(m)}(j) \left(\sum_{i=j+1}^{j+r} \xi_{\theta_0}^{(m)}(i)' \xi_{\theta_0}^{(m)}(i) \right)^{-1} \sum_{i=j+1}^{j+r} \xi_{\theta_0}^{(m)}(i)',$$

which are not identity matrices. Now in order to standarize the vector $\tilde{\rho}_{\theta_0}^{(m)}$ we may write

$$\begin{aligned} &\bar{\rho}_{\theta_0}(j) = \tilde{\rho}_{\theta_0}(j) \times Avar\left(\tilde{\rho}_{\theta_0}(j)\right)^{-1/2}, \\ &\bar{\rho}_{\theta_0}(j) \sim \mathcal{N}(0, I_{d^2}), iid \text{ for } j = 1, \ldots, m. \end{aligned}$$

Thus we get that $\check{Q}(\theta_0) \to \chi^2(Hd^2)$.

In order to show (29) it is sufficient to note that $\Im_{\hat{\theta}}^{(m,r)}(\cdot) \to \Im_{\theta_0}^{(m,r)}(\cdot)$ because from *Proposition* 2 and 3 we have that

$$\xi_{\theta_0}(j) = \bar{\xi}_{\theta_0}(j) + O_p(N^{-\frac{1}{2}}), \quad j = 1, \ldots, m$$

and $\Im_{\theta_0}^{(m,r)}(\cdot)$ is a finite sum of derivatives and identity operators. Now $\hat{\rho}_{\hat{\theta}}^{(m)} \xrightarrow{p} \bar{\rho}_{\theta_0}^{(m)}$ so also $Avar\left(\tilde{\rho}_{\theta_0}(j)\right)$ is estimated consistently.
QED

References

1. Billingsley, P. (1999) *Convergence of probability measures* (2nd ed.), New York: Wiley.

2. Box, G.E.P., and Cox, D. (1964), "An Analysis of Transformations", *Journal of the Royal Statistical Society*, Ser.B, 26, 211-252.

3. Box, G.E.P., and Pierce, D. A. (1970), "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models", *Journal of the American Statistical Association*, 65, 1509-1526.

4. Box, G.E.P., and Jenkins, M. J. (1976), *Time Series Analysis: Forecasting and Control*, San Francisco, CA: Holden-Day.

5. Brockwell, P. J., and Davies, R. A. (1991), *Time Series: Theory and Methods.* New York: Springer Verlag.

6. Chittori, R. V. (1974), "Distribution of Residual Autocorrelations in Multiple Autoregressive Schemes", *Journal of the American Statistical Association*, 69, 928-34.

7. Davidson, J. E. H. (1994) *Stochastic Limit Theory*, New York: Oxford University Press.

8. Delgado, and Velasco, C. (2012), "An Asymptotically Pivotal Transform of the Residuals Sample Autocorrelations With Application To Model Checking", *Journal of the American Statistical Association*, 106:495, 946-958.

9. Delgado, M. A., and Velasco, C. (2010), "Distribution-Free Test for Time Series Model Specification", *Journal of Econometrics*, 155, 128-137.

10. Delgado, M. A., Hidalgo, J., and Velasco, C. (2005), "Distribution Free Goodness-of-Fit Tests for Linear Processes", *Annals of Statistics*, 33, 2568-2609.

11. Duchesne P., and Francq C., (2008), "On Diagnostic Checking Time Series Models With Portmanteau Test Statistics Based on Generalized Inverses and {2}-Inverses", in COMSTAT 2008, *Proceedings in Computational Statistics*, ed.P.Brito, Heidelberg: Physica-Verlag, pp. 143-154.

12. Dufour, J.-M., and Pelletier, D. (2005), "Practical Methods for Modelling weak VARMA Processes: Identification, Estimation, and Specification with Macroeconomic Application". Discussion Paper, CIRANO and CIREQ, Universitè de Montrèal.

13. Durbin, J. (1970), "Testing for Serial Correlation in Least-Squares Regression When Some of the Regressors Are Lagged Dependent Variables", *Econometrica*, 38, 410-421.

14. Francq, C., and Raïssi, H. (2007), "Multivariate Portmanteau Test For Autoregressive Models With Uncorrelated but Nonindependent Errors", *Journal of Time Series Analysis*, 28:3, 454-470.

15. Francq, C., Roy, R., and Zakoïan, J.-M. (2005), "Diagnostic Checking in ARMA Models with Uncorrelated Errors", *Journal of the American Statistical Association*, 100, 532-544.

16. Francq, C., and Zakoïan, J.-M. (1998), "Estimating Linear Representations of Nonlinear Processes", *Journal of Statistical Planning and Inference*, 68, 145-65.

17. Francq, C., and Zakoïan, J.-M. (2005), "Recent Results for Linear Time Series Models With Non Independent Innovations". In *Statistical Modelling and Analysis for Complex Data Problems*, Chap. 12 (eds. Duchesne, P., and Rèmillard). New York: Springer Verlag, 137 - 161.

18. Hosking, J. R. M. (1980). "The Multivariate Portmanteau Statistic", *Journal of American Statistical Association*, 75, 343-86.

19. Hosking, J. R. M. (1981). "Equivalent Forms of the Multivariate Portmanteau Statistic", *Journal of the Royal Statistical Society*, B43, 219-30.

20. Ljung, G. M. (1986), "Diagnostic Testing of Univariate Time Series Models", *Biometrica*, 73, 725-730.

21. Ljung, G. M., and Box, G. E. P. (1978), "On Measure of Lack Of Fit in Time Series Models", *Biometrica*, 65, 297-303.

22. Lütkepohl, H. (1993), *Introduction to Multiple Time Series Analysis*, Berlin: Springer Verlag.

23. Boubacar Mainassara, Y., Francq, C.(2009), "Estimating Structural VARMA Models With Uncorrelated but Non-Independent error Terms", MPRA paper No. 15141.