UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

Departamento de Fundamentos de Análisis Económico II

HERRAMIENTAS DE MODELIZACIÓN PARA SERIES TEMPORALES MULTIVARIANTES

MEMORIA PARA OPTAR AL GRADO DE DOCTOR
PRESENTADA POR

Ignacio Arbués Lombardía

Bajo la dirección de la doctora

María Dolores Robles Fernández

Madrid, 2013

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES
Departamento de Fundamentos del Análisis Económico II

HERRAMIENTAS DE MODELIZACIÓN PARA SERIES TEMPORALES MULTIVARIANTES.

MEMORIA PARA OPTAR AL GRADO DE DOCTOR
PRESENTADA POR

Ignacio Arbués Lombardía

Bajo la dirección de la doctora

María Dolores Robles Fernández

Madrid, 2012

TESIS DOCTORAL

HERRAMIENTAS DE MODELIZACIÓN PARA SERIES TEMPORALES MULTIVARIANTES

Ignacio Arbués Lombardía

Directora: María Dolores Robles Fernández

Universidad Complutense de Madrid
Facultad de Ciencias Económicas y Empresariales
Departamento de Fundamentos del Análisis Económico II

Abril 2012

AGRADECIMIENTOS

Hay dos grupos de personas a las que tengo mucho que agradecer. El primero está formado por aquellos sin los que nunca habría terminado esta tesis. El segundo, por aquellos sin los que el tiempo transcurrido desde que la empecé habría sido bastante peor. En el primer grupo están en un lugar destacado mi directora, Lola Robles, José Luis Fernández y Pedro Revilla. El segundo es demasiado numeroso para citarlos a todos. Hay sin embargo dos personas que están en el primer grupo y en el segundo: Nélida y Menchu.

Índice general

Capítulo 1

Introducción

La mayor parte de las contribuciones que se hacen en esta tesis se pueden describir como inferencia sobre modelos lineales de series temporales multivariantes. También incluye una novedad teórica que, aunque se sale de este marco pues es un resultado de procesos estocásticos, se utiliza para demostrar dos proposiciones que sí encajan en la descripción anterior.

Para poner estos resultados en contexto debemos hacer una introducción a los principales campos de conocimiento con los que están conectados. Las series temporales tanto univariantes como multivariantes están relacionadas, por un lado con multitud de disciplinas aplicadas en las que están el estudio de las manchas solares, la hidrología o el béisbol (ver [1], [2] y [3]). Por otro lado, están conectadas con las disciplinas teóricas que proporcionan las herramientas para tratarlas. En la sección siguiente haremos un pequeño recorrido por algunos de de esos campos.

1.1. Contextualización histórica

De todos los campos relevantes, nos vamos a concentrar en tres: i) la macroeconomía y en particular los modelos autorregresivos vectoriales

(VAR, *vector autoregression*) y los modelos factoriales dinámicos (DFM, *Dynamic Factor Models*); ii) la teoría estadística de sistemas lineales y iii) la predicción.

1.1.1. Los modelos VAR y DFM en la macroeconomía

Hasta los años setenta del siglo pasado, para estudiar el comportamiento conjunto de las variable macroeconómicas se empleaban fundamentalmente los Modelos de Ecuaciones Simultáneas. Estos modelos se habían generalizado a partir de los trabajos de la Comisión Cowles en los años cuarenta y consisten en sistemas de ecuaciones lineales en los que aparecen las variables cuyo comportamiento se quiere estudiar, posiblemente retardos de estas variables y perturbaciones aleatorias. Un modelo de este tipo se puede escribir de la forma

$$A(L)y_t = u_t, \tag{1.1}$$

donde y_t es un vector que contiene las variables del modelo, $A(z) = A_0 + A_1 + \ldots + A_p z^p$ es un polinomio cuyos coeficientes son matrices cuadradas (que podemos considerar también como una matriz cuyos elementos son polinomios), L es el operador de retardos (es decir, $L^j y_t = y_{t-j}$) y u_t es un vector de perturbaciones aleatorias. Por desgracia, un modelo de la forma (1.1) tiene un inconveniente: no está identificado. Esto significa que puede haber dos conjuntos diferentes de parámetros que dan lugar a modelos que no pueden ser distinguidos a partir de los datos porque sus distribuciones de probabilidad son idénticas. Los problemas de identificación se pueden entender como la consecuencia de buscar el modelo en un espacio demasiado grande[1]. Por ello, una manera de proceder es reducir este espacio imponien-

[1] Aquí seguimos la práctica habitual de emplear la palabra modelo tanto para lo general que representamos en (1.1) como para lo particular que se obtiene dando unos valores concretos a los parámetros. El contexto permite distinguir ambos significados.

do restricciones a los parámetros. Un ejemplo de restricción es obligar a que la matriz de covarianza de las perturbaciones sea diagonal (es decir, que los *shocks* estén incorrelados o incluso que sean independientes en el caso gaussiano). Otro tipo de restricción consiste en introducir la distinción entre variable endógena y exógena. Si ahora denotamos por y_t al vector que contiene las llamadas variable endógenas y por x_t al que contiene las exógenas, entonces planteamos el modelo de la forma

$$B(L)y_t = C(L)x_t + u_t,$$

lo que que es equivalente a suponer que en la matriz A eran nulos los coeficientes de x en las ecuaciones de y.

Introducir restricciones en los parámetros de un modelo aumenta la precisión de las estimaciones. Eso hace que exista la tentación de incluir en los modelos más restricciones de las estrictamente necesarias para obtener identificabilidad. Pero si las restricciones son incorrectas, entonces el modelo está mal especificado. Esto sugiere que las restricciones deberían incluirse con algún tipo de respaldo empírico o teórico. Sin embargo, como dice Sims en el influyente artículo "Macroeconomics and reality"([4]):

> *"To the extent that models end up with very different sets of variables on the right-hand-sides of these equations, they do so not by invoking economic theory, but (in the case of demand equations) by invoking an intuitive, econometrician's version of psychological and sociological theory ..."*[2]

En la referencia que acabamos de citar, Sims aboga por un enfoque distinto, según el cuál todas las variables se consideran endógenas y en el que

[2]En tanto que los modelos se quedan con muy diferentes conjuntos de variable en el lado derecho de las ecuaciones, no lo hacen según la teoría económica, sino (en el caso de las ecuaciones de demanda) según la versión intuitiva de un económetra de la teoría psicológica y sociológica ...

el problema de la identificación se trata pasando el modelo a la *forma reducida*. Para obtener la forma reducida multiplicamos a izquierda el modelo por A_0^{-1} y lo reescribimos como

$$\Phi(L)y_t = \varepsilon_t,$$

donde $\Phi(z) = I + \Phi_1 z + \ldots \Phi_p z^p$, $\Phi_j = A_0^{-1}A_j$ y $\varepsilon_t = A_0^{-1}u_t$.

Este tipo de modelo es conocido como VAR y se convirtió en una de las herramientas más habituales para analizar el comportamiento dinámico de las variables macroeconómicas.

Desde el punto de vista matemático, podemos considerar el modelo VAR como un caso particular del modelo VARMA (*Vector AutoRegression and Moving Average*). Este modelo es la versión multivariante del modelo ARMA, ampliamente usado para modelizar series univariantes, especialmente desde el trabajo de Box y Jenkins ([5]). Los modelos VARMA tienen la forma

$$\Phi(L)y_t = \Theta(L)\varepsilon_t, \tag{1.2}$$

donde Θ es otra matriz de polinomios. La parte derecha de (1.2) es lo que se llama una media móvil (*moving average* o MA). Los modelos VARMA nunca han gozado de la misma popularidad que los VAR o sus hermanos pequeños, los ARMA univariantes. Entre otras razones, porque la dificultad computacional de estimarlos es muy superior. Sin embargo debemos mencionarlos por razones que serán evidentes más adelante.

Aproximadamente en la misma época en la que Sims publicaba [4]–en realidad, un poco antes–, él mismo con Thomas Sargent ([6]) y John Geweke, en su tesis doctoral dirigida por el primero ([7]), introducían otro tipo de modelo que intenta responder a algunos de los problemas de los modelos clásicos y a otros más. En [6], además de la crítica de las restricciones a priori[3], se plantea uno de los inconvenientes tanto de los Modelos de Ecua-

[3]Los autores dicen *"... very little of the a priori theory embodied in macroeconometric*

ciones Simultáneas como de los VAR, la proliferación de parámetros. Esto decían Sargent y Sims:

> *"Cyclical interactions among macroeconomic variables, probably commonly involve lags of eight or more quarters. A ten-equation, tenth-order autoregression of general form (ten lags of ten variables in each equation) leaves zero degrees of freedom, approximately, in U.S. postwar data. Rather than reduce the dimensionality of our models by restricting particular equations a priori, as in the standard methodology, we proceed by imposing simplifying conditions which are symmetric in the variables"*[4].

El problema reside en que el número de parámetros, por ejemplo de un modelo VAR $n-$variante de orden p, es $pn^2 + (n+1)n/2$. La presencia de términos cuadráticos en n hace que el número de variables que se pueden incluir quede muy limitado. Como decíamos anteriormente, las restricciones pueden mitigar este problema, pero en lugar de introducir restricciones intuitivas y específicas para cada ecuación o variable, los autores plantean una restricción general (simétrica en el sentido de que trata por igual a todas las variables, o en términos matemáticos, que es invariante respecto a

models is based explicitly on models of the behavior of individuals" (muy poca de la teoría a priori incluida en los modelos macroeconométricos se basa explícitamente en modelos del comportamiento de los individuos). Con la introducción de los Modelos Estocásticos Dinámicos de Equilibrio General, se corregiría esto.

[4]Las interacciones cíclicas entre variables macroeconómicas de manera probablemente habitual implican retardos de ocho o más trimestres. Un modelo autorregresivo con diez ecuaciones y de orden diez en forma general (diez retardos de diez variables en cada ecuación deja aproximadamente cero grados de libertad con los datos de EE.UU. posteriores a la II Guerra Mundial. En lugar de reducir la dimensión de nuestros modelos imponiendo restricciones a priori a ciertas ecuaciones en particular, como se hace en la metodología estándar, nosotros imponemos condiciones simplificadoras que son simétricas en las variables.

permutaciones de las variables). Esta restricción la podemos expresar informalmente de la forma siguiente: la dinámica de las n variables se puede explicar mediante un número reducido (fijo) de factores comunes.

Formalmente, podemos escribir el DFM como

$$y_t = A(L)f_t + v_t,$$

donde A es nuevamente una matriz de polinomios, f_t es un vector de factores comunes de dimensión k típicamente pequeña y v_t es un vector de factores específicos. Al igual que en el análisis factorial tradicional, los factores específicos se consideran incorrelados, de tal forma que las relaciones entre las componentes de y_t se establecen solo a través de los factores comunes. Esto es lo que corta drásticamente la proliferación de parámetros. Los factores tienen sus propios modelos dinámicos, por ejemplo, VAR para f_t y AR unidimensionales para las componentes de v_t

$$\begin{aligned} \Phi(B)f_t &= \xi_t, \\ \varphi^i(B)v_t^i &= \eta_t^i. \end{aligned}$$

Supongamos que los órdenes de los modelos dinámicos de los factores y el grado de A son a lo sumo p. Entonces el número de parámetros es $nkp + k^2p + k(k+1)/2 + n(p+1)$. Ahora no hay términos cuadráticos en n, sino solo de grado uno en n con coeficiente kp, luego si se mantiene controlado kp se puede –en principio– hacer n grande.

También se han desarrollado en los últimos años otros tipos de modelos factoriales en los que se relaja la hipótesis de que los factores específicos estén incorrelados. En estos *modelos factoriales aproximados* se permite un cierto grado de correlación y a cambio se exige que cuando $n \to \infty$, esta correlación esté acotada en cierta forma. Sin entrar en detalles innecesarios, podemos decir que la correlación entre las variables que aportan los factores específicos se vaya haciendo despreciable comparada con la que aportan los

factores comunes. Este tipo de modelos fue introducido por Chamberlain y Rothschild ([8]) en un marco estático, mientras que en un marco dinámico distintas variantes han sido estudiadas por un lado, por Bai y Ng (en [9] y [10]) y Onatski ([11]) y por otro lado por Forni, Hallin, Lippi y Reichlin ([12], [13] y [14]).

1.1.2. Inferencia en sistemas lineales

Mientras se producían los cambios en la forma de modelizar las variables macroeconómicas que describíamos antes, de forma paralela tenían lugar ciertos avances en la teoría estadística de sistemas lineales que son relevantes para nuestra exposición. Esta teoría se fundamenta en un resultado teórico conocido como descomposición de Wold que podemos encontrar, por ejemplo, en [15]. Éste afirma que cualquier proceso linealmente regular[5] se puede representar como

$$x_t = \Psi(L)\varepsilon_t$$

donde $\sum_j \|\Psi_j\|^2 < +\infty$ y ε_t es ruido blanco, es decir, que es débilmente estacionario y ε_t está incorrelado con ε_s para $s \neq t$.

Este resultado permite buscar modelos lineales para explicar la dinámica de cualquier proceso estacionario[6], aunque se puede emplear al menos de dos maneras:

(a) La más habitual es en combinación con una hipótesis adicional: que la función de transferencia Ψ es racional, es decir, que se puede escribir como $\Psi = \Phi^{-1}\Theta$ donde Φ y Θ son dos polinomios. Esto es equivalente a decir que el proceso satisface un modelo ARMA.

[5] un proceso estacionario x_t es linealmente regular si $\mathbb{E}[x(t)] = 0$ y $\lim_{\tau\to\infty} x(t+\tau|t) = 0$ donde $x(t+\tau|t)$ es el predictor lineal mínimo cuadrático de $x(t+\tau)$ que usa $\{x(s) : s \leq t\}$

[6] Esto no implica que no pueda haber modelos no lineales con mejores propiedades.

(b) Sin esa hipótesis adicional, la construcción de un modelo lineal, en este caso se plantea como un problema de aproximación, en el que la relación que antes se considera exacta, ahora se considera aproximada. Por ejemplo, se puede aproximar Ψ mediante una fracción recional $\Phi^{-1}\Theta$, aunque es más habitual transformar la representación de Wold en una representación autorregresiva infinita, por ejemplo $\Pi(B)x_t$ y aproximar la serie de potencias Π por un polinomio Φ.

Mientras que con el primer enfoque, se estima un modelo lineal $\hat{\Phi}^{-1}\hat{\Theta}$ de tal forma que la diferencia entre la verdadera función de transferencia del proceso y la estimada es debida solamente al uso de los parámetros muestrales en lugar de los poblacionales, en el segundo enfoque es también el orden finito de los polinomios el que separa al modelo del comportamiento verdadero del proceso. Esto tiene implicaciones importantes en la forma de trabajar con estos modelos: (i) mientras que bajo la hipótesis de racionalidad tiene sentido hablar de identificación del modelo, en el caso aproximado sólo podemos hablar de selección del modelo, ya que ninguno es correcto; (ii) en el primer caso, podemos usar el BIC para seleccionar el modelo, en el segundo es más conveniente el AIC; (iii) en el caso aproximado no tiene sentido hacer contrastes de bondad de ajuste. En este trabajo se hacen aportaciones en ambos marcos, como veremos más tarde.

En 1970 se publicó el libro "Time series analysis: Forecasting and control"de Box y Jenkins ([5]), que resultó tremendamente influyente. Este libro proporcionaba a todo el que trabajaba con series temporales univariantes un programa en varias fases para construir modelos de un tipo conocido como ARIMA. Los modelos ARIMA son modelo ARMA a los que se les añade en el lado izquierdo factores de diferencias, o bien *regulares* $\nabla = (1 - L)$ o bien *estacionales* $\nabla_s = (1 - L^s)$, donde s es el número de observaciones, por cada ciclo de estacionalidad (para datos económicos, normalmente, un año).

Esta metodología está respaldada por una teoría matemática muy completa que proporciona condiciones para la consistencia de las estimaciones de los parámetros y de los órdenes de los modelos, distribuciones asintóticas, predicción, control, etc. En particular, nos interesa un tipo de contrastes de hipótesis que constituyen una herramienta teórica muy ampliamente empleada para validar los modelos. Nos referimos al contraste de autocorrelación residual de Box y Pierce ([16]) y al de Ljung y Box ([17]), que es una corrección para muestras pequeñas del primero.

Estos contrastes se basan en lo siguiente: tras estimar el modelo, se obtienen los residuos $\hat{\varepsilon}_t$, se calcula su función de autocovarianza, $\hat{\gamma}_j$ y se obtiene el estadístico del contraste, $Q_k = \sum_{j=1}^{k} T^2(T-j)^{-1}\hat{\gamma}_j^2$ en el caso de Ljung-Box. Bajo la hipótesis de que el modelo está correctamente identificado, Q_k se distribuye asintóticamente, *para k grande*, como una chi-cuadrado con $k-r$ grados de libertad, donde r es el número de parámetros estimados. La condición de que k sea grande se expresa de esta manera informal tanto en [16] como en [17], de manera que en realidad no hay ningún enunciado preciso de las propiedades asintóticas del test. Tampoco se enuncian todas las hipótesis requeridas. De esta forma llegamos a la primera de las aportaciones de este trabajo:

Aportación 1: Enunciamos con precisión propiedades asintóticas y las hipótesis de los contrastes de autocorrelación residual[7].

Expresar con más precisión la convergencia tiene consecuencias prácticas sobre la forma de elegir el máximo retardo que se incluye en el contraste.

A lo largo de los años setenta y principios de los ochenta, se fueron adaptando al caso multivariante muchos de los resultados teóricos que había para los modelos ARMA univariantes. Así, Dunsmuir y Hannan ([18]) y Deistler, Dunsmuir y Hannan ([19]) demostraron propiedades asintóticas

[7] Capítulo 2, sección 3 y capítulo 3, apartado 4.1. de la Tesis

de los estimadores máximo-verosímiles de modelos lineales parametrizados con un número finito de parámetros y en particular, de los VARMA. Estos resultados requerían un trabajo muy grande de descripción de la topología de los modelos lineales (ver [20]).

También se adaptaron al caso multivariante los contrastes de autocorrelación residual. La primera versión multivariante fue propuesta por Hosking ([21]) y sus propiedades asintóticas fueron estudiadas por Poskitt y Tremayne ([22]). En estos artículos se mantiene la imprecisión a la que nos referíamos antes en el caso univariante. Generalmente a este tipo de contrastes en el caso multivariante se les llama de Portmanteau (en ocasiones, también se llama así al contraste univariante). Unos años después, en 1988, Ahn demostró en [23] que el contraste de Portmanteau podía aplicarse a modelos VAR con restricciones, siempre que éstas afectasen solo a los coeficientes de la parte autorregresiva y no a los de la matriz de covarianzas de las perturbaciones. En este caso, para calcular el número de grados de libertad hay que descontar solo los parámetros libres, es decir, que al número de autocorrelaciones que se incluyen se le resta el número de parámetros y se le suma el de restricciones. En la literatura se da por sentado que este resultado es válido para modelos VARMA con restricciones (así se hace en una referencia tan relevante como el libro de Lütkepohl, [24], que da como referencia a [23]) aunque no nos consta que eso hubiera sido demostrado.

Como mostraremos, el contraste de Portmanteau no puede aplicarse en general al caso de modelos VARMA con restricciones que afectan a la vez a los coeficientes de las partes AR o MA y a la matriz de covarianzas (restricciones así se denominan *mixing*). Por el contrario, la teoría de estimación máximo verosímil para modelos lineales sí que es perfectamente aplicable a ese caso (por ejemplo, [15]). Podría parecer que esto es un detalle de importancia solo teórica, pero hay una razón por la que es relevante: los Modelos

Factoriales Dinámicos equivalen a modelos VARMA con ciertas restricciones y éstas son precisamente del tipo *mixing*. Por tanto, una de las herramientas de diagnóstico más importantes para modelos lineales no puede aplicarse a los DFM. Aquí llegamos a la segunda aportación de esta tesis:

Aportación 2: Damos condiciones para que el contraste de Portmanteau pueda aplicarse con restricciones *mixing* y proponemos un contraste alternativo (contraste de Portmanteau extendido) para cuando esto no es posible. También mostramos como esto se aplica al caso de los DFM[8].

Hay otros tipos de modelos a los que se puede aplicar nuestro contraste extendido y no el contraste tradicional, como el modelo vectorial autorregresivo estructural factorial de Stock y Watson (FSVAR, [25]) y el modelo de Peña y Box ([26]).

1.1.3. Aproximación por funciones racionales

Como decíamos en la sección anterior, se puede trabajar sin la hipótesis de que la función de transferencia es racional. En este caso, un modelo ARMA solo puede ser una aproximación a la dinámica verdadera del proceso.

Supongamos que un proceso x_t se puede representar como VAR infinito, de la forma $x_t = \sum_{k=1}^{\infty} \Pi_k x_{t-k} + \varepsilon_t$ y nosotros estimamos un modelo VAR finito de la forma $x_t = \sum_{k=1}^{p} \hat{\Phi}_k x_{t-k} + \varepsilon_t$, entonces el modelo estimado se aleja del verdadero en dos sentidos, que están relacionados respectivamente con los conceptos de:

> **Aproximación**: buscamos el modelo que más se aproxima al verdadero, entre los modelos de la familia VAR(p). Si conociéramos la dinámica exacta del modelo verdadero, podríamos obtener el modelo aproximado de orden p minimizando cierta medida de distancia que no

[8] Capítulo 2.

especificaremos. Pero en lugar de la dinámica verdadera tenemos una muestra del proceso, con lo que entra en juego el concepto siguiente.

Estimación: elegimos los coeficientes del modelo aproximado como aquellos en los que se alcance el óptimo de una determinada función de los valores de la muestra, como la función de verosimilitud o el error cuadrático agregado.

Este marco es muy distinto del de la estimación bajo la hipótesis de que el proceso realmente satisface un modelo de la clase donde estimamos. En particular, la parte de aproximación y la de estimación tienen consecuencias opuestas respecto a la elección del orden del modelo p. Un p grande es bueno desde el punto de vista de la aproximación, puesto que en una clase de modelos más grandes es posible acercarse más al modelo verdadero. Sin embargo, desde el punto de vista de la estimación, p grande es malo, porque deteriora la relación entre el número de parámetros y el de observaciones, haciendo que las estimaciones sean más ruidosas. De este modo, la elección del orden del modelo consiste en poner en la balanza ambos efectos.

La teoría sobre esta cuestión generalmente propone condiciones al crecimiento de p en función de T. La condición $p/(T \log T)^{1/2} \to 0$ garantiza consistencia ([27]) y $p^3/T \to 0$ garantiza normalidad para combinaciones lineales de los coeficientes ([28] y [29]). Si queremos ir más allá y probar otras propiedades asintóticas de los estimadores, como la normalidad conjunta del vector de parámetros, nos tropezamos con una dificultad. En la mayoría de situaciones, los parámetros del modelo se pueden representar conjuntamente con un vector en un determinado espacio vectorial de dimensión d, $\mathbb{R}^d$. Entonces, podemos aplicar resultados teóricos como distintos Teoremas Centrales del Límite. Sin embargo, si el orden del modelo p diverge, no podemos encerrar los parámetros en ningún espacio vectorial de dimensión finita. Este problema de la *dimensión creciente* aparece en otras

situaciones (por ejemplo, en [30] y en los modelos factoriales aproximados anteriormente mencionados) y obliga a emplear técnicas alternativas. Los resultados anteriores sobre normalidad se limitan a demostrar que ciertas combinaciones lineales de los coeficientes son asintóticamente normales.

Con objeto de probar la normalidad asintótica conjunta del vector de parámetros, desarrollamos un resultado previo que da lugar a la tercera de las aportaciones de esta Tesis:

Aportación 3: Acotamos superiormente la rapidez con la que una sucesión de martingalas de dimensión creciente se acerca a la normalidad. Para ello generalizamos un resultado anterior para martingalas en espacios de Banach[9].

Aplicamos este resultado a la normalidad de los parámetros estimados de un modelo AR de orden p que tiende a infinito con T, lo que da lugar a la cuarta aportación.

Aportación 4: Damos condiciones sobre el crecimiento de p para que las estimaciones sean asintóticamente conjuntamente normales[10].

Esta técnica se puede aplicar también a otros problemas. Por ejemplo, volviendo a los contrastes de autocorrelación residual que introdujimos en la sección 1.1.2, la forma de demostrar que el estadístico Q_k se distribuye como una χ^2 es probar que el vector de autocovarianzas es asintóticamente normal, pero como para que esto se cumpla k debe ser grande (es decir, que tienda a infinito con T), entonces no se pueden aplicar los resultados de normalidad asintótica tradicionales para martingalas, lo que probablemente explique que no se hubieran enunciado con precisión las propiedades del contraste. En esta tesis, empleando nuestra aportación 3 conseguimos lo quinta aportación:

Aportación 5: Damos condiciones sobre el crecimiento de k para

[9] Capítulo 3, secciones 2 y 3.

[10] Capítulo 3, apartado 4.2.

que el estadístico Q_k se distribuya asintóticamente como una χ^2 cuando k y T divergen simultáneamente[11].

Sin la herramienta teórica que introdujimos en la aportación 3, solo podemos probar el resultado secuencial, es decir, cuando primero $T \to \infty$ y después $k \to \infty$. Esto es menos relevante para la práctica, ya que en realidad, uno nunca puede esperar a que T haya llegado a infinito para elegir k. Al establecer condiciones conjuntas para el crecimiento k y T, se da una indicación sobre la elección de k dado T, que habitualmente viene prefijado por las circunstancias.

1.1.4. Selección de modelos para predicción

Hasta ahora hemos descrito modelos y herramientas de diagnóstico sin hacer distinción entre los usos que se le pretenden dar al modelo. Sin embargo, la teoría que vamos a tratar en esta sección está orientada específicamente a los modelos lineales como instrumentos para predecir ciertos procesos. Como decíamos al principio de esta introducción, la predicción es una de las funciones de los modelos de series temporales y sin duda una de las más relevantes.

De todos modos, antes de entrar de lleno en la cuestión de la predicción, vamos a detenernos en la cuestión de qué hacer para elegir un modelo lineal para una serie temporal multivariante, digamos $\mathbf{x}_t = (x_t^0, \ldots, x_t^n)'$. Una de las formas de hacerlo es emplear un criterio de selección de modelos. Los más conocidos tienen la forma siguiente:

$$-2 \log L + kg(T) \tag{1.3}$$

donde L es la verosimilitud del modelo, k es el número de parámetros libres, T es el número de observaciones y $g(\cdot)$ es una función no decreciente. Si eligiéramos el modelo que hace menor el primer término de (1.3), estaríamos

[11] Capítulo 3, apartado 4.1.

buscando el mejor ajuste, pero el segundo término penaliza la complejidad del modelo. La función g nos dice cómo varía esta penalización al aumentar el número de observaciones. Cuando usamos la verosimilitud gaussiana, podemos reescribir (1.3) como

$$\log \hat{\sigma}^2 + k\frac{g(T)}{T}, \tag{1.4}$$

donde $\hat{\sigma}^2$ es la varianza estimada de las perturbaciones. Las elecciones más habituales son $g(T) = \log T$, con lo que tenemos el BIC (*Bayesian Information Criterion*, [32]) y $g(T) = 2\log\log T$, que corresponde al criterio HQ (de Hannan y Quinn, [31]). Se puede probar que estos dos criterios proporcionan una identificación consistente del modelo bajo la hipótesis de función de transferencia racional. También se emplea el AIC (*Akaike Information Criterion*, [33], [34]) con $g(T) = 2$, que es inconsistente para el caso racional, pero que proporciona predicciones asintóticamente óptimas para el caso no racional (ver [35] y [36] para una propiedad de optimalidad espectral).

Hay otras herramientas que, en ciertos contextos, sirven para seleccionar modelos, como los contrastes de bondad de ajuste, o los de razón de verosimilitudes, pero estos otros enfoques presentan importantes limitaciones. La principal es que emplearlos para seleccionar un modelo de entre varios posibles requeriría hacer contrastes secuencialmente y habría que controlar la probabilidad de error en la selección del modelo, que depende de las probabilidades de error en cada contraste. Esto sería tan complicado que es irrealizable en la práctica. Por tanto, los contrastes son más útiles como herramienta de diagnóstico una vez hecha la selección o al menos cuando se ha reducido el conjunto de modelos posibles a un número muy pequño de ellos. Esta es una razón por la que los métodos de modelización automática como el del programa TRAMO emplean criterios de selección (en ese caso, el BIC; ver [37]).

Ahora supongamos que nuestro interés está en encontrar un modelo que

funcione bien para predecir, x_t^0. Entonces, no tenemos por qué limitarnos a probar modelos para una serie multivariante $\mathbf{x}_t = (x_t^0, \ldots, x_t^n)'$ en particular. Podemos construir series multivariantes con distintos subconjuntos de variables o series univariantes de entre todas aquellas que tenemos a nuestra disposición.

Por supuesto, si pudiéramos conseguir el modelo verdadero y los valores exactos de los parámetros, entonces la decisión sobre qué series incluir sería irrelevante porque las predicciones obtenidas con más variables serían al menos tan buenas como las que conseguiríamos con menos. Sin embargo, en el mundo real los modelos son estimados y por tanto no son exactos. Emplear un modelo de más complejidad, como un VAR, aumenta la indeterminación debida a la estimación y hacerlo innecesariamente cuando el modelo más complejo –incluso con valores poblacionales– no mejora las predicciones del más sencillo podría perjudicar gravemente los resultados. Por el contrario, también puede ocurrir que incluir más series mejore la capacidad predictiva del modelo. Por tanto, por razones semejantes a las que hacen conveniente estimar bien los órdenes de un modelo ARMA en lugar de elegir unos suficientemente grandes como para contener el modelo verdadero, aquí interesa elegir el subconjunto mínimo de variables que sean relevantes para la predicción.

Así el proceso de selección del modelo se podría descomponer en dos partes:

(A) elegir el subconjunto de variables, y

(B) elegir el modelo adecuado para ellas.

Hasta ahora, la manera más habitual de tratar esta cuestión es considerar los modelos independientemente del subconjunto de variables para el que están definidos y emplear contrastes de capacidad predictiva. La literatura sobre contrastes de capacidad predictiva despega con el artículo de

Diebold y Mariano ([38]). Aquí aparecen algunas de las ideas fundamentales de este tipo de contrastes, en especial el uso de predicción postmuestral. El contraste que proponían los autores en [38] es, en concreto, un contraste de igual capacidad predictiva. Es decir, la hipótesis nula es que entre dos modelos, no hay diferencia en cuanto a la calidad de sus predicciones, medidas según cierta función de pérdida. El contraste tiene forma de t de Student, donde el numerador mide la diferencia entre las funciones de pérdida y el denominador es la raíz cuadrada de una estimación de la media cuadrática de esa diferencia.

Desgraciadamente, se comprobó que el estadístico de Diebold y Mariano solo tenía como distribución límite una normal estándar cuando los modelos que se comparan son no anidados, mientras que para modelos anidados, converge a una distribución que es una función integral de un proceso browniano (Clark y McCracken, [40]). La teoría sobre la forma de tratar modelos anidados ha experimentado un desarrollo considerable (principalmente [39], [40], [41], [42] y [43]).

Como decíamos antes, al seleccionar un modelo entre muchas posibilidades, el empleo de contrastes puede ser problemático. De hecho, en el capítulo 4, se muestra cómo la aplicación sucesiva de contrastes hace que la elección del subconjunto óptimo sea inconsistente con probabilidad uno. Para resolver esta dificultad, nosotros proponemos una familia de criterios de selección, semejante a la dada por (1.4), pero con la diferencia de que sirve para decidir no el modelo (parte B de la decisión según la descomposición que presentamos más arriba), sino el subconjunto de series para el que se construye el modelo (parte A de la decisión). Esto constituye la última aportación:

Aportación 6: Proponemos una familia de criterios para seleccionar el subconjunto óptimo de series y demostramos su con-

sistencia. También comparamos su rendimiento con el de varios contrastes de capacidad predictiva[12].

1.2. Estructura de la tesis

En esta sección, describimos el contenido de los tres trabajos que la forman y la manera en la que está estructurados. En cada uno de los apartados siguientes nos centramos en uno de los capítulos del 2 al 4.

1.2.1. Capítulo 2: Contraste de *portmanteau* extendido

El objeto principal de este capítulo es proponer y evaluar un contraste de bondad de ajuste basado en la autocorrelación residual para modelos VARMA con restricciones no lineales. Este contraste es una modificación del contraste de *portmanteau* multivariante, introducido por Hosking ([21]) cuyo estadístico es

$$Q_k = T \sum_{j=1}^{k} \mathrm{tr}\left[\hat{C}_j' \hat{C}_0^{-1} \hat{C}_j \hat{C}_0^{-1}\right],$$

donde $\hat{C}_j$ es la j-ésima matriz de autocovarianza residual.

Se sabe que distribución asintótica del estadístico de *portmanteau* clásico (que llamamos así para distinguirlo del nuestro) es una chi-cuadrado cuando se aplica a los residuos de modelos VARMA sin restricciones o VAR con restricciones ([21] y [23]). Sin embargo, la convergencia de la distribución del estadístico falla cuando se estiman modelos con restricciones que afectan simultáneamente a los coeficientes de la matriz de covarianza de las innovaciones y a los de las partes autorregresiva y de medias móviles (restricciones *mixing*). Este tipo de restricciones, que pueden parecer artificiosas,

[12] Capítulo 4.

aparecen de manera natural si consideramos, por ejemplo, Modelos Factoriales Dinámicos, ya que éstos se pueden considerar como modelos VARMA con restricciones *mixing*. También encontramos este tipo de restricciones en modelos de espacio de estados. Para entender la relación entre los modelos VARMA con restricciones y los modelos factoriales o de espacio de estados, tenemos que tener en cuenta que un modelo VARMA cuyos coeficientes son funciones de un cierto conjunto de parámetros (parametrización estructurada en la terminología de [23]) se puede entender como un modelo VARMA con una restricción que consiste en la pertenencia a la imagen de la parametrización. Por otro lado, mediante el teorema de la función implícita, también un modelo con restricciones se puede entender como un modelo con parametrización estructurada. En la sección 3 se detalla la equivalencia entre la parametrización estructurada y las restricciones.

Por otra parte, cuando se introducen restricciones en la matriz de covarianzas de las innovaciones, parece conveniente medir la bondad de ajuste del modelo comprobando si la matriz de covarianzas residuales es compatible con esas restricciones. El contraste que proponemos, Q_k^*, tiene dos ventajas; (a) comprueba esta incompatibilidad y (b) a diferencia del *portmanteau* clásico, converge a una chi-cuadrado en el caso de restricciones *mixing*. La espresión del estadístico es,

$$Q_k^* = \frac{T}{2}\mathrm{tr}\left[(\hat{C}_0\hat{\Sigma}^{-1} - I_n)(\hat{C}_0\hat{\Sigma}^{-1} - I_n)'\right] + T\sum_{j=1}^{k}\mathrm{tr}\left[\hat{C}_j'\hat{C}_0^{-1}\hat{C}_j\hat{C}_0^{-1}\right], \quad (1.5)$$

donde $\hat{\Sigma}$ es la matriz de covarianzas estimada bajo las restricciones. También consideramos una variante de (1.5) en la que $\hat{C}_0$ es sustituida por $\hat{\Sigma}$ en el segundo término.

Un segundo objetivo del capítulo es hacer más precisa la teoría sobre la convergencia del estadístico clásico, ya que los resultados teóricos anteriores en la literatura son vagos tanto en la especificación de las hipótesis,

como en el sentido de la convergencia asintótica, ya que solo se considera la convergencia en T, mientras que es necesario tener también en cuenta k.

Las aportaciones teóricas del artículo son, pues, dos: (1) la demostración de la convergencia de la distribución de nuestro estadístico en el caso general (teorema 2.1) y (2) la especificación de unas condiciones bajo las cuales se puede emplear el estadístico clásico (teorema 2.2).

En las secciones subsiguientes se analiza la aplicación de los contrastes a varias clases de modelos: los DFM, el conocido como modelo de Peña y Box ([26]), y el modelo de autorregresión vectorial con estructura factorial de Stock y Watson ([44]). Para demostrar que todos ellos entran en el marco de nuestros resultados introducimos una clase más general de modelos que llamamos Modelo de Impulsos Factoriales (MIF), que se diferencian de los VARMA en que la parte de medias móviles tiene una estructura factorial. A continuación comprobamos que: (a) los MIF cumplen las condiciones requeridas y (b) todos los modelos anteriores se pueden expresar como MIF con restricciones.

Aun queda una dificultad por resolver en cuanto a la teoría. Los modelos factoriales están identificados salvo transformaciones lineales de los factores comunes, es decir, que un proceso que se pueda representar mediante un determinado DFM también admite otras representaciones cuyos factores comunes son combinaciones lineales de los del primer DFM. Demostramos que esta indeterminación no impide que la distribución asintótica del estadístico sea una chi-cuadrado, pero obliga a introducir una corrección en el número de grados de libertad.

Aplicamos el contraste a datos simulados para comprobar que la frecuencia de rechazo está suficientemente cerca de la teórica cuando se cumple la hipótesis nula y para comparar la potencia del contraste con la del estadístico clásico bajo varias hipótesis alternativas. Finalmente, presentamos los

resultados de una aplicación a los datos del Índice de Producción Industrial de España.

1.2.2. Capítulo 3: Alejamiento de la normalidad de martingalas de dimensión creciente

Este capítulo se puede dividir en dos partes: primero (secciones 2 y 3) se introduce un resultado sobre velocidad de convergencia a la normalidad de un tipo de martingalas y a continuación (sección 4) se aplica esta herramienta a dos problemas de inferencia sobre modelos de series temporales multivariantes.

Las martingalas que se estudiamos están numeradas de la forma X_{ni}, donde $i = 1, \ldots, n$. Tenemos, pues, de una sucesión triangular, con la peculiaridad de que cada X_{ni} es un vector cuya dimensión depende de n. En particular, nos interesa el caso en el que su dimensión tiende a infinito, pues en caso contrario se puede tratar con las herramientas tradicionales. Concretamente, queremos encontrar cotas para la distancia entre la distribución de la media de las diferencias y la distribución normal de la dimensión correspondiente.

Hay muchas maneras de medir la distancia entre dos distribuciones, pero para nuestros propósitos, la más conveniente es la métrica de Projórov (ver [45], la referencia más exhaustiva sobre métricas de distribuciones). Por otro lado, la métrica de Kantoróvich proporciona una cota para la de Projórov y tiene buenas propiedades. Debido a esto, tomamos como punto de partida para nuestro trabajo los resultados de Rachev y Rüschendorf ([46]) para martingalas en espacios de Banach.

El primer resultado del nuestro trabajo es una versión del teorema principal de [46] con algunas modificaciones técnicas. A continuación empleamos este resultado para acotar la distancia (medida con la métrica de Kan-

toróvich) desde la distribución de la media de las diferencias de las martingalas a la distribución normal multivariante de la dimensión correspondiente. Más concretamente, probamos que cuando la dimensión de las martingalas no crece demasiado rápido, entonces la distancia decrece proporcionalmente a una potencia del número de términos.

La primera aplicación de estos resultados se centra en los contrastes de autocorrelación residual. Como decíamos en el apartado anterior, una característica general de la teoría conocida sobre estos contrastes es que sus propiedades están expresadas de manera bastante vaga. Generalmente, se afirma que el estadístico, Q_k, donde k corresponde a la autocorrelación de mayor orden, se aproxima a una chi-cuadrado con $d(k)$ grados de libertad, donde d es una función que se especifica, para k grande y longitud de la serie T también grande. Para argumentarlo se prueba que Q_k es una suma de funciones cuadráticas de una media de diferencias de martingalas más términos que tienden a cero cuando $T \to \infty$. Entonces, se aplica el Teorema Central del Límite, pero con la dificultad de que la matriz de covarianzas asintótica solo es aproximadamente idempotente cuando k es grande. Por tanto, un resultado de convergencia debería tener en cuenta el límite tanto en k como en T.

El primer resultado de este tipo parece ser el teorema 2.1, pero ahí el limite es sucesivo, primero en T y después en k. Este tipo de propiedad asintótica no es la más conveniente, sino que es más adecuado un resultado que asegure la convergencia cuando $k, T \to \infty$ y se cumpla alguna relación entre k y T, ya que esto da una indicación sobre cómo elegir k cuando T está determinado de antemano por la disponibilidad de los datos. La relevancia de las propiedades asintóticas del estadístico se debe a que el nivel de significación empírico del contraste se aproxima mejor o peor al teórico según la distribución del estadístico esté más cerca de su límite o

menos. Por tanto, una buena propiedad sería que el error debido a emplear la región de rechazo teórica en lugar de la verdadera tienda a cero. Si $F_T(x)$ es la función de distribución del estadístico y $G_T(x)$ es la de una chi-cuadrado con el número correspondiente de grados de libertad, entonces queremos que para un cierto p,

$$\lim_{T\to\infty} F_T(G_T^{-1}(p)) = p.$$

Este es justamente el resultado que demostramos.

La segunda aplicación está relacionada con la inferencia de modelos autorregresivos cuando el verdadero modelo es un AR(∞). Si estimamos un modelo autorregresivo para una serie temporal de longitud T generada por un proceso AR(∞), entonces el orden del modelo no debe crecer demasiado con relación a T para que las estimaciones tengan buenas propiedades. Se ha probado que si el orden k del modelo cumple que $k^3/T \to 0$ y la serie de potencias es absolutamente sumable, entonces se cumple una propiedad de normalidad asintótica de los estimadores. Desgraciadamente, esta propiedad no se ha establecido para el vector de parámetros, sino para una combinación lineal suya. Si $\hat{\phi}(k)$ es un vector con los coeficientes estimados y $l(k)$ es una sucesión de vectores que satisface ciertas condiciones, entonces $l(k)'\hat{\phi}(k)$ es asintóticamente normal. Nosotros, por el contrario, probamos que la distribución de $\hat{\phi}$ está cerca de una normal multivariante, lo que permite, por ejemplo, construir regiones de confianza para los parámetros.

1.2.3. Capítulo 4: Determinación de la sección cruzada óptima en el sentido del error medio cuadrático de predicción

En este capítulo proponemos una familia de criterios para extraer de entre un conjunto de series temporales, el subconjunto óptimo para predecir otra serie dada. El sentido en el que consideramos un conjunto como "ópti-

mo" es que haga mínimo el error cuadrático medio (ECM) de predicción a h periodos de distancia y a igualdad de ECM, que sea el subconjunto de menor tamaño. En la parte teórica presentamos la familia de criterios y demostramos su consistencia bajo ciertas hipótesis, después comparamos mediante simulaciones su eficacia con la de otros métodos conocidos y finalmente mostramos su aplicación a un caso real extraído de la literatura sobre predicción.

Para construir nuestros criterios seguimos un principio semejante a los criterios de selección de modelos, como el de Akaike, el de Schwarz o el de Hannan y Quinn, pero no partiendo de la verosimilitud, sino del error cuadrático medio de predicción $\hat{\sigma}_h^2(I)$. Por I, denotamos un cierto subconjunto de entre todas las variables que tenemos a nuestra disposición. La forma de estos criterios es:

$$\mathrm{FC}(I) = \log \hat{\sigma}_h^2(I) + \delta(I)\frac{S_T}{T},$$

donde $\delta(I)$ es una medida del tamaño de I, T es el número de predicciones y S_T es una función creciente que determina si la selección es más o menos parsimoniosa.

A continuación demostramos que el subconjunto seleccionado mediante cualquiera de estos criterios converge hacia el óptimo bajo ciertas hipótesis. Estas hipótesis son poco restrictivas por cuanto permiten emplear una clase bastante amplia de modelos, siempre que estén estimados consistentemente y pueden ser tanto modelos anidados como no anidados. También permite calcular los criterios con predicciones tanto postmuestrales como intramuestrales. Por el contrario, se requiere que los modelos estén bien especificados. A modo de ejemplo, mostramos que estas hipótesis se cumplen para modelos VARMA.

Seguidamente, presentamos dos generalizaciones. La primera consiste en demostrar que el método funciona cuando se seleccionan los subconjuntos

de entre una clase aleatoria. Este caso se presenta cuando de entre todos los posibles subconjuntos se hace una preselección a partir de los datos. La segunda generalización consiste en una versión de los criterios adaptada al caso en que nos interese predecir más de una serie. En este caso, hay que sustituir el ECM de predicción por una cierta función de la matriz de covarianzas de las predicciones. Demostramos que si esta función cumple ciertas condiciones, entonces se mantiene la consistencia del caso univariante.

En la sección de simulaciones comparamos la probabilidad de seleccionar el subconjunto óptimo mediante nuestros criterios con la que se obtiene mediante varios contrastes de hipótesis: el de Diebold y Mariano; los contrastes ENC-T y ENC-NEW de Clark y McCracken, el contraste condicional de Giacomini y White y el de causalidad de Granger (ver [38], [40], [47] y [48]). Para ello generamos 5.000 realizaciones de varios procesos bivariantes de forma que en algunos casos el subconjunto óptimo solo incluye la propia serie a predecir y en otros incluye la segunda serie. También comprobamos la eficacia del método en condiciones de mala especificación. Los resultados indican que estos criterios mejoran algunos de los contrastes y no son mejorados por ninguno.

Para terminar, aplicamos los mismos métodos a datos reales. Siguiendo [40] intentamos determinar si la inflación subyacente de EE.UU. puede predecirse mejor empleando la tasa de paro. Nuestros criterios, al igual que los de Clark y McCracken y a diferencia del de Diebold y Mariano y del de Giacomini y White indican que sí es útil el dato del paro para predecir la inflación a un trimestre de distancia.

1.3. Principales conclusiones

La idea que unifica los trabajos que componen esta tesis es la de proporcionar herramientas para afrontar las dificultades que aparecen según

aumenta la dimensión de las series temporales multivariantes. Ante este problema se pueden adoptar al menos dos estrategias distintas: bien elegir modelos cuya complejidad crezca moderadamente en función de la dimensión o bien mantener esta última en valores pequeños.

Con relación a la primera estrategia, hemos presentado un contraste de bondad de ajuste para un tipo de modelos, los DFM, que pueden ser aplicados a series de dimensión relativamente grande (por ejemplo, comparando con los VAR). En cuanto a la segunda, proponemos una familia de criterios de selección que permiten extraer de entre un conjunto de variables, un subconjunto óptimo en el sentido de que sirve para predecir otra variable de forma que el error cuadrático medio se haga mínimo.

Por otra parte, las técnicas que hemos desarrollado para obtener esos dos resultados nos han permitido también mejorar la teoría de los contrastes de autocorrelación para modelos VARMA y la de estimación de modelos autorregresivos aproximados.

Bibliografía

[1] Yule, G. U., (1927), On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers, *Philosophical Transactions of the Royal Society of London. Series A*, 226, 267–298.

[2] Hosking, J. R. M. (1984), Modeling persistence in hydrological time series using fractional differencing, *Water resources research*, 20, 1898–1908.

[3] Kaplan, D. (2008), Univariate and Multivariate Autoregressive Time Series Models of Offensive Baseball Performance: 1901-2005, *Journal of Quantitative Analysis in Sports*, 4.

[4] Sims, C. A. (1980), Macroeconomics and reality, *Econometrica*, 48, 1–48.

[5] Box, G. E. P. y Jenkins, G. M. (1976), *Time Series Analysis:Forecasting and Control*, Holden-Day, San Francisco.

[6] Sargent, T. J., y Sims, C. A. (1977), Business cycle modeling without pretending to have too much a priori economic theory in C. Sims et al., *New Methods in Business Cycle Research*. Minneapolis: Federal Reserve Bank of Minneapolis.

[7] Geweke, J. (1977), The Dynamic Factor Analysis of Economic Time Series, en D. J. Aigner y A. S. Goldberger (eds.) *Latent Variables in Socio-Economic Models*. Amsterdam: North Holland. Cap. 19.

[8] Chamberlain, G. y Rothschild, M. (1983), Arbitrage, Factor Structure and Mean-Variance Analysis in Large Assets Markets. *Econometrica* 51, 1305–1324.

[9] Bai, J y Ng, S. (2002), Determining the Number of Factors in Approximate Factor Models. *Econometrica* 70, 191–221.

[10] Bai, J. (2003), Inferential Theory for Factor Models of Large Dimensions. *Econometrica* 71, 135–171.

[11] Onatski, A. (2009), Testing hypotheses about the number of factors in large factor models. *Econometrica* 77, 1447–1479.

[12] Forni, M., Hallin, M., Lippi, F. y Reichlin, L. (2000), The Generalized Dynamic Factor Model: Identification and Estimation. *Review of Economics and Statistics* 82, 540–554.

[13] Forni, M., Hallin, M., Lippi, F. y Reichlin, L. (2004), The Generalized Dynamic Factor Model: Consistency and Rates. *Journal of Econometrics* 119, 231–255.

[14] Forni, M., Hallin, M., Lippi, F. y Reichlin, L. (2005), The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting. *Journal of the American Statistical Association* 100, 830–840.

[15] Hannan, E. J. y Deistler, M. (1988), *The Statistical Theory of Linear Systems*, New York: John Wiley and Sons.

[16] Box, G. E. P. y Pierce, D. (1970), Distribution of Autocorrelations in Autoregressive Moving Average Time Series Models. *Journal of the American Statistical Association* 65, 1509–1526.

[17] Ljung, G. M. y Box, G. E. P. (1978), On a Measure of Lack of Fit in Time Series Models. *Biometrika* 65, 297–303.

[18] Dunsmuir, W. y Hannan, E. J. (1976), Vector linear time series models. *Advances in Applied Probability* 8, 339–364.

[19] Deistler, M. y Dunsmuir, W. y Hannan, E. J.(1978), Vector linear time series models: corrections and extensions. *Advances in Applied Probability* 10, 360–372.

[20] Deistler, M. y Pötscher, B. M. (1984), The behaviour of the likelihood function for ARMA models. *Advances in Applied Probability* 16, 843–865.

[21] Hosking, J. R. M. (1980), The Multivariate Portmanteau Statistic. *Journal of the American Statistical Association* 371, 602–608.

[22] Poskitt, D. S. y Tremayne, A. R. (1982), Diagnostic Tests for Multiple Time Series Models. *The Annals of Statistics* 10, 114–120.

[23] Ahn, S. K. (1988), Distribution for Residual Autocovariances in Multivariate Autoregressive Models With Structured Parameterization. *Biometrika* 75, 590–593.

[24] Lütkepohl, H. (1991), *Introduction to Multiple Time Series*, Berlin: Springer-Verlag.

[25] Stock, J. H. y Watson, M. W. (2005), Implcations of Dynamic Factor Models for VAR Analysis, NBER WP 11467.

[26] Peña, D. y Box, G. E. P. (1987), Identifying a Symplifying Structure in Time Series *Journal of the American Statistical Association* 82, 836–843.

[27] Hannan, E. J. y Kavalieris, L. (1986), Regression, autoregression models. *Journal of Time Series Analysis* 7, 27–50.

[28] Berk, K. N. (1974), Consistent autoregressive spectral estimates. *The Annals of Statistics* 2, 489–502.

[29] Lewis, R. y Reinsel, G. C. (1985), Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis* 16, 393–411.

[30] Mammen, E. (1989) Asymptotics with increasing dimension for the robust regression with applications to the bootstrap. *The Annals of Statistics* 17, 382–400.

[31] Hannan E. J. y Quinn B. G. (1979), The determination of the order of an autoregression. *Journal of the Royal Statistical Society B* 41, 190–195.

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

[33] Akaike H. (1973), Information Theory and an extension of the Maximum Likelihood Principle, in: B. N. Petrov and F. Csaki, (eds.), *Second international symposium on information theory.* Academiai Kiado: Budapest, pp. 267–281.

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

[35] Shibata, R. (1980), Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics* 8, 147–164.

[36] Shibata, R. (1981), An optimal autoregressive spectral estimate. *The Annals of Statistics* 9, 300–306.

[37] Maravall, A. (2008) Notes on programs TRAMO and SEATS© www.bde.es/webbde/es/secciones/servicio/software/tramo/Part_II_Tramo.pdf

[38] Diebold F. y Mariano R. (1995), Comparing Predictive Accuracy. *Journal of Business and Economics Statistics* 13, 252–263.

[39] Harvey, D. I., Leybourne, S. J. y Newbold, P. (1998), Tests for Forecast Encompassing. *Journal of Business and Economic Statistics* 16, 254–259.

[40] Clark T. E. y McCracken M. W. (2001), Tests of Equal Forecast Accuracy and Encompassing for Nested Models, *Journal of Econometrics* 105, 85–110.

[41] Clark T. E. y McCracken M. W. (2005), Evaluating Direct Multistep Forecasts, *Econometric Reviews* 24, 369–404 .

[42] Clark T. E. y West K. D. (2007), Approximately Normal Tests for Equal Predictive Accuracy in Nested Models, *Journal of Econometrics* 138, 291–311.

[43] Clark T. E. y McCracken M. W. (2011), Reality Checks and Comparisons of Nested Predictive Models, *Journal of Business and Economics Statistics* doi:10.1198/jbes.2011.10278 .

[44] Stock, J. H. y Watson, M. W. (2003), Understanding Changes in International Business Cycle Dynamics. *NBER Working Paper* No. W9859.

[45] Rachev, S. T. (1991), *Probability metrics and the stability of stochastic models.* Wiley, Nueva York.

[46] Rachev, S. T. y Rüschendorf, L. (1994). On the rate of convergence in the CLT with respect to the Kantorovich metric. In *Probability in Banach spaces.* (Hoffmann- Jorgensen, J., Kuelbs, J. y Marcus, B., eds.) 9, 193–207. Birkhäuser, Londres.

[47] Giacomini R. y White H. (2006), Tests of conditional predictive ability. Econometrica 74(6), 1545–1578.

[48] Granger C. W. J. (1969), Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424–438.

Capítulo 2

An extended portmanteau test for VARMA models with mixing nonlinear constraints*

2.1. Introduction

Diagnostic tools based on residual autocorrelations are among the most frequently used to analyze time series models. In the univariate case, the residuals of estimated ARMA models can be used to measure the goodness of fit by using the tests described by Box and Pierce (1970) and Ljung and Box (1978). In this paper, we focus on the multivariate ARMA model,

$$y_t = \Phi_1 y_{t-1} + \ldots + \Phi_p y_{t-p} + \varepsilon_t + \Theta_1 \varepsilon_{t-1} + \ldots + \Theta_q \varepsilon_{t-q} \qquad (2.1)$$

where $y_t = (y_t^1, \ldots, y_t^n)$, $E[\varepsilon_t] = 0$, $E[\varepsilon_t \varepsilon_t'] = \Sigma$. The model can be written more succinctly as $\Phi(B)y_t = \Theta(B)\varepsilon_t$, where $\Phi(z) = I - \Phi_1 z - \ldots - \Phi_p z^p$, $\Theta(z) = I + \Theta_1 z + \ldots + \Theta_p z^p$ and B is the backshift operator.

* *Journal of Time Series Analysis*, 29, (2008) 741–761

If T residuals $\hat{\varepsilon}_t$ are available, we can compute the residual covariances $\hat{C}_j = \sum \hat{\varepsilon}_t \hat{\varepsilon}_{t+j}/T$, and then, the classical multivariate portmanteau statistic is obtained as,

$$Q_k = T \sum_{j=1}^{k} \mathrm{tr}\left[\hat{C}_j' \hat{C}_0^{-1} \hat{C}_j \hat{C}_0^{-1}\right] \tag{2.2}$$

When the coefficients in (2.1) are freely estimated, it was established by Hosking (1980) and Li and McLeod (1981) that Q_k is asymptotically distributed as a chi-square with $(k-p-q)n^2$ degrees of freedom. This result was generalized by Ahn (1988) to the autoregressive case when the matrices are parameterized by a certain vector, that is, when $\Phi(z)$ is constrained to the image of the parameterization. In this case, the asymptotic distribution is a chi-square with $kn^2 - b$, where b is the number of free parameters.

In this paper, all matrices in (2.1) including Σ depend on a $b \times 1$ vector $\dot{\beta}$. When it is necessary to highlight this dependency, we write $\Sigma(\dot{\beta})$, $\Phi_j(\dot{\beta})$, $\Theta_j(\dot{\beta})$, $\Phi(\dot{\beta})$ and $\Theta(\dot{\beta})$ (that indicate not the polynomial evaluated at $\dot{\beta}$ but the dependency of the coefficients). Thus, the whole system including the covariance matrix is constrained to a certain class of models. The generalization of the asymptotic distribution results of Q_k to this case is not possible in general.

The fact that Σ is not freely estimated suggests that the goodness of fit should also be tested using the zero lag covariances. If $\hat{\beta}$ is an estimate of the true β, it seems convenient to measure how much $\hat{\Sigma} = \Sigma(\hat{\beta})$ differs from $\hat{C}_0$, so instead of (2.2) we propose,

$$Q_k^* = \frac{T}{2}\mathrm{tr}\left[(\hat{C}_0\hat{\Sigma}^{-1} - I_n)(\hat{C}_0\hat{\Sigma}^{-1} - I_n)'\right] + T\sum_{j=1}^{k} \mathrm{tr}\left[\hat{C}_j'\hat{C}_0^{-1}\hat{C}_j\hat{C}_0^{-1}\right] \tag{2.3}$$

Under the null hypothesis, $\hat{\Sigma}$ and $\hat{C}_0$ are both consistent estimates of Σ, and thus any of them can be used in the second term of (2.3). Then, we can

also define,

$$R_k^* = \frac{T}{2}\mathrm{tr}\left[(\hat{C}_0\hat{\Sigma}^{-1} - I_n)(\hat{C}_0\hat{\Sigma}^{-1} - I_n)'\right] + T\sum_{j=1}^{k}\mathrm{tr}\left[\hat{C}_j'\hat{\Sigma}^{-1}\hat{C}_j\hat{\Sigma}^{-1}\right] \quad (2.4)$$

For small samples we can apply the usual correction factor $T/(T-j)$ to the j-th term. We denote the corrected statistics by $\tilde{Q}_k$, $\tilde{Q}_k^*$ and $\tilde{R}_k^*$.

In section 2.2 we prove the asymptotic distribution of (2.3) and (2.4). In section 2.3 sufficient conditions for the classical portmanteau (2.2) are presented. Some cases in which mixing constraints arise are the Dynamic Factor Model (DFM), the Factor-Structural Vector Autoregression (FSVAR) and the Peña-Box model. In section 2.4, we show how to apply the test to these models. In sections 2.5 and 2.6, results with simulated and real data are presented. Section 2.7 includes some final remarks.

The relation of Q_k^* and R_k^* to the Lagrange Multiplier tests, analyzed by Poskitt and Tremayne (1982) for Q_k remains for future work.

Let us introduce some notation. We write ρ for different constants in $(0,1)$ and M for positive constants. We have already used the dot notation $\dot{\beta}$ which means any possible value of the parameters, which as a particular case can be equal to the true values β or the estimates $\hat{\beta}$. This notation is also used for objects depending on $\dot{\beta}$, such as the residuals $\dot{\varepsilon}_t$, while $\partial/\partial\beta$ stands for $\partial/\partial\dot{\beta}$ evaluated at β. The euclidean norm and the matrix norm induced by it are denoted by $|\ldots|$ and $\|\ldots\|$ respectively. We use $o_p(.)$ and $O_p(.)$ to indicate order in probability and $o(.)$ and $O(.)$ for order in $E|.|^2$, when applied to random variables.

2.2. Extended Portmanteau Statistic

Let us assume the following,

A1 $\Phi_j(\dot{\beta})$, $\Theta_j(\dot{\beta})$ and $\Sigma(\dot{\beta}) > 0$ are $n \times n$ twice continuously differentiable functions of a $b \times 1$ vector $\dot{\beta}$ in a set Ω.

A2 For $\dot{\beta} = \beta$ in the interior of Ω, the stationary process y_t is generated by (2.1) where ε_t is a iid process with fourth-order moments that depend on the first and second-order ones in the same way as a multivariate gaussian.

A3 $|\Phi(z)|$, $|\Theta(z)| \neq 0$, for any z, $|z| \leq 1$.

Under A3, we can also define $\Psi(z) = \Phi(z)^{-1}\Theta(z)$ and $\Pi(z) = \Theta(z)^{-1}\Phi(z)$. Let us call $\phi = \text{vec}(\Phi')$, $\theta = \text{vec}(\Theta')$, $\pi = \text{vec}(\Pi')$, $\psi = \text{vec}(\Psi')$, $\sigma = \text{vec}(\Sigma)$, $\pi^* = [\sigma', \pi']'$ and $\psi^* = [\sigma', \psi']'$.

A4 $\partial\pi^*/\partial\beta'$ is of full rank b.

Under A3, assumption A4 is equivalent to $\partial\psi^*/\partial\beta'$ being of rank b. A4 ensures uniqueness of β in a neighborhood of the true (Φ, Θ, Σ).

Theorem 2.1. *Let $F^*_{k,T}(x)$ be the distribution function of R^*_k or Q^*_k and $\chi^2_{\nu,\alpha}$ the $1-\alpha$ quantile of a chi-square distribution with ν degrees of freedom. Then, under A1-A4, there exists some $\rho \in (0,1)$ such as,*

$$\limsup_T |F^*_{k,T}(\chi^2_{n(n+1)/2+kn^2-b,\alpha}) - (1-\alpha)| = O(\rho^k) \tag{2.5}$$

NOTE: (2.5) does not imply $\lim_{k,T} |\dots| = 0$. On the other hand, for practical applications, T is usually given and a certain $k(T)$ has to be chosen. We would like this choice to be such as when $T \to \infty$ and $k = k(T)$, $|\dots|$ converges to zero as fast as possible. The difficulty of obtaining convergence rates for the Central Limit Theorem prevents us to give a precise proof of the optimal $k(T)$. Nevertheless, $O_p(T^{-1})$ terms appear in the proof of theorem 2.1. Therefore, no choice of $k(T)$ can make the convergence faster than that. The minimum k for which $O(\rho^{k(T)})$ has order T^{-1} is

$k(T) = -\log(T)/\log(\rho)$. The coefficient ρ depends on the decay rate of $\Psi(z)$. Therefore, if $|\Phi(z)|$ has roots near the unit circle, greater values of k are required.

2.3. The Classical Portmanteau

Further assumptions are needed for using the portmanteau (2.2). A sufficient assumption is that it is possible to decompose $\dot{\beta}$ as $(\dot{\beta}^0, \dot{\beta}^1)$ so that $\dot{\Sigma}$ depends on $\dot{\beta}^0$ and $\{\dot{\Phi}_j, \dot{\Theta}_j\}$ on $\dot{\beta}^1$ (non-mixing parameterization), but this assumption can be relaxed. We can allow mixing dependence if the derivatives of $\dot{\Sigma}$ on $\dot{\beta}^1$ and $\{\dot{\Phi}_j, \dot{\Theta}_j\}$ on $\dot{\beta}^0$ are null at the true values. Then, the parameterization is asymptotically non-mixing as $T \to \infty$ and $\hat{\beta}$ converges to β.

On the other hand, let us consider an equivalent parameterization, in the sense that $\tilde{\Sigma}(\dot{\lambda})$,$\tilde{\Phi}(\dot{\lambda})$,$\tilde{\Theta}(\dot{\lambda})$ is equivalent to $\Sigma(\dot{\beta})$,$\Phi(\dot{\beta})$,$\Theta(\dot{\beta})$ when there exists some diffeomorphism h from a set Λ onto Ω such that $\tilde{\Sigma}(\dot{\lambda}) = \Sigma(h(\dot{\lambda}))$,$\tilde{\Phi}(\dot{\lambda}) = \Phi(h(\dot{\lambda}))$,$\tilde{\Theta}(\dot{\lambda}) = \Theta(h(\dot{\lambda}))$, $h(\dot{\lambda}) = \dot{\beta}$. Since the asymptotic distribution of (2.2) is the same with any two equivalent parameterizations, it follows that the existence of one such parameterization fulfilling the above condition is sufficient.

Another sufficient condition would be that the constraint is equivalent to two independent constraints, one on the ARMA coefficients and another one on the covariance matrix. Again, this condition can be relaxed to one on the first order derivatives.

Let us then enunciate more accurately the additional assumptions,

A5 There is a local equivalent parameterization, $\tilde{\Sigma}(\dot{\lambda})$, $\tilde{\Phi}(\dot{\lambda})$, $\tilde{\Theta}(\dot{\lambda})$ with $\dot{\lambda} = (\dot{\lambda}^0, \dot{\lambda}^1) \in \mathbb{R}^{b_0} \times \mathbb{R}^{b_1}$ such as at λ, $\partial\sigma/\partial\dot{\lambda}^1 = \partial\phi_j/\partial\dot{\lambda}^0 = \partial\theta_j/\partial\dot{\lambda}^0 = 0$ for any $j > 0$.

A5' The tangent space to the set $\{(\sigma(\dot{\beta})', \phi(\dot{\beta})', \theta(\dot{\beta})') | \dot{\beta} \in \Omega\}$ in β is equal to $E_\sigma \times E_{\phi,\theta}$ where $E_\sigma \subset \mathbb{R}^{n^2}$ and $E_{\phi,\theta} \subset \mathbb{R}^{(p+q)n^2}$ have dimensions b_0 and b_1 respectively.

Proposition 2.1. *Assumptions A5 and A5' are equivalent.*

Any of these assumptions allows to use the classical portmanteau test.

Theorem 2.2. *Let $F_{k,T}(x)$ be the distribution function of Q_k and $\chi^2_{\nu,\alpha}$ as in theorem 2.1. Then, under A1-A5, there exists some $\rho \in (0,1)$ such as,*

$$\limsup_T |F_{k,T}(\chi^2_{n(n+1)/2+kn^2-b_1,\alpha}) - (1-\alpha)| = O(\rho^k) \tag{2.6}$$

NOTE: it is also true that the first term in (2.4), is asymptotically distributed as a chi-square with $n(n+1)/2 - b_0$. Since the gaussian assumption is required only to establish the asymptotic distribution of the zero lag autocorrelations, it is not necessary for theorem 2.2, so it is a generalization of the results by Hosking (1980) and Ahn (1988).

For the case of non-mixing parameterization, the substantial part of the rank condition A4 is to check that the derivative of ψ with respect to β^1 is of full rank b_1. This is true, for example, in the case of affine restrictions defining (with the left-coprimeness condition) an identifiable class as in theorem 2.7.3 of Hannan and Deistler (1988). Let us assume that $\Phi(z), \Theta(z)$ are left coprime. If the rank is less than b_1 then it can be proved that there exist $[U(z), V(z)] \neq 0$ in the tangent space to the constraint manifold such as $U(z)\Phi(z)^{-1}\Theta(z) = V(z)$ and then for any real γ, $\tilde{\Phi}(z)\Phi(z)^{-1}\Theta(z) = \tilde{\Theta}(z)$ holds with $\tilde{\Phi}(z) = \Phi(z) + \gamma U(z)$ and $\tilde{\Theta}(z) = \Theta(z) + \gamma V(z)$. For γ small enough, $\tilde{\Phi}(z)$ is invertible and $[\tilde{\Phi}(z), \tilde{\Theta}(z)]$ is also left coprime and satisfies the restrictions, but $\tilde{\Phi}(z)^{-1}\tilde{\Theta}(z) = \ Phi(z)^{-1}\Theta(z)$, which contradicts the identifiability.

2.4. Testing Dynamic Factor Models

Since they were introduced by Geweke (1977), Sargent and Sims (1977) and Engle and Watson (1981), Dynamic Factor Models have become very popular for analyzing multivariate series. The idea underlying these models is that the cross-correlation between several variables y_t^i, $i = 1, \ldots, n$ can be explained by assuming the existence of some unobserved or latent variables (common factors). Then, y_t^i is a linear combination of the common factors plus a specific or idiosyncratic factor. The model is dynamic because autoregressive (eventually ARMA) structures for the common and idiosyncratic factors are assumed. We write the DFM as,

$$y_t^i = L^i f_t + v_t^i \tag{2.7}$$

$$\Phi(B) f_t = \xi_t \tag{2.8}$$

$$\varphi^i(B) v_t^i = \eta_t^i \tag{2.9}$$

where L^i is the loading vector of the series y_t^i, f_t is the vector of common factors and η_t^i, ξ_t are uncorrelated white noise processes with variances σ_i^2 and covariance matrix Ξ respectively.

As a result of the specification of the model, the covariance matrix of the forecasting error Σ cannot be any symmetric positive definite matrix, but it is constrained on a manifold. Thus, unlike in the ordinary VARMA case, as important as to check the absence of autocorrelation of the residuals it is to check whether their covariance matrix is consistent with the model.

These models are usually estimated by maximum likelihood using a state-space representation. This method is computationally inadequate for very large number of series. Quah and Sargent (1993) estimated a model for a quite large set of series by using the Expectation Maximization Algorithm.

In a static framework, Chamberlain and Rothschild (1983) introduced the approximate factor models, allowing the idiosyncratic factors to be cor-

related (then it is necessary to assume $n \to \infty$ and certain conditions on the covariance matrices of the factors and on the loadings). These approximate models have been translated to the dynamic case and have recently received wide attention, remarkably in Bai and Ng (2002), Bai (2003), Stock and Watson (1998, 2002) and Forni, Hallin, Lippi and Reichlin (2000, 2004 and 2005). The relaxed assumptions of these models make them unsuited for the portmanteau test. The generalized factor model of Forni et. al. is strongly nonparametric, and thus, even less suited to our test. Consequently, in this paper we focus on the exact or strict factor models.

A related model is the one by Peña and Box (Peña and Box, 1987; Hu and Chou, 2004), which allows a more general structure for the dynamic factors (e. g., VARMA) and assumes v_t^i as white noise process.

We can also consider the Factor-Structural Vector Autoregression (FSVAR) (Stock and Watson, 2003). This model, consists of a VAR model for y_t with factor structure for the innovations,

$$\Phi(B)y_t = L\xi_t + \eta_t \tag{2.10}$$

Thus, the correlations between variables are explained partially by the common factors, and partially by the off-diagonal elements of $\Phi(z)$.

In the next subsection, we analyze a Factor Shock Model (FSM) that can be represented as a constrained ARMA. In subsection 2.4.2, we see that the DFM, Peña-Box and FSVAR can be regarded as constrained FSM. In 2.4.3 we show how to deal with the identification issues of the factor models.

2.4.1. The Factor Shock Model

Let us consider the model with r common shocks,

$$\Phi(B)y_t = \Theta_c(B)\xi_t + \Theta_s(B)\eta_t \tag{2.11}$$

where $\Phi(z)$ and $\Theta_s(z)$ are $n \times n$ polynomial matrices of degrees p and q, $\Theta_c(z)$ is $n \times r$ of degree q and $\Theta_s(z)$ is diagonal; ξ_t and η_t are uncorrelated gaussian processes of dimensions $r \times 1$ and $n \times 1$; ξ_t contains the common shocks and η_t the idiosyncratic ones. We can assume that the common and idiosyncratic factor processes have unit variances by allowing the components of Θ_c and Θ_s to have free zero degree terms, so that $\Theta_j(z) = \Theta_{j,0} + \Theta_{j,1}z + \ldots + \Theta_{j,q}z^q$, for $j = c, s$.

In order to apply our tests to the FSM, we need to check that it has a VARMA representation and that the parameterization fulfils assumptions A1-A4. The existence of such a representation is guaranteed by Lütkepohl (1984), but since regularity properties are required, we need to make our own derivation.

We can write the model (2.11) as,

$$\Phi(B)y_t = \bar{\Theta}(B)\chi_t \tag{2.12}$$

with $\bar{\Theta}(B) = [\Theta_c(B), \Theta_s(B)]$ and $\chi_t = (\xi_t', \eta_t')'$.

Then, y_t has the following state-space representation,

$$\begin{aligned} y_t &= HY_t \\ Y_t &= FY_{t-1} + U\chi_t \end{aligned} \tag{2.13}$$

with $H = (I_n 0 \ldots 0)'$, $U = (\bar{\Psi}_0', \ldots, \bar{\Psi}_s')'$, $s = max\{p, q+1\}$, $\bar{\Psi}(z) = \Phi(z)^{-1}\bar{\Theta}(z)$ and

$$F = \begin{pmatrix} 0 & I_q \otimes I_n \\ \Phi_s & \Phi_{s-1} \ldots \Phi_1 \end{pmatrix} \tag{2.14}$$

The Kalman Filter theory (for example, Hannan and Deistler, 1988) implies that if (2.13) holds, the process y_t can be represented as a $MA(\infty)$, $y_t = [I + H(I - FB)^{-1}FKB]\varepsilon_t$, where ε_t are innovations with covariance matrix V and K and V are the asymptotic Kalman gain and covariance matrix of the filter.

We can use this to analyze the VARMA representation of y_t. It suffices to check that the right side of (2.11) is a MA.

If we set $a_t = \bar{\Theta}(B)\chi_t$, then (2.13) holds for a_t instead of y_t with $\Phi_j = 0$ for $j > 0$. In this case the matrix F is nilpotent and this has two consequences: i) $(I - FB)^{-1}$ has only a finite number of terms and thus, the $MA(\infty)$ is in fact a finite MA and ii) the filter converges in a finite number of steps and then, K_t equals the asymptotic K at a certain t. Since K_t is result of algebraic operations, it is infinitely differentiable as a function of the coefficients in $\bar{\Theta}$. Then, also and the matrices of the MA representation are smooth. The autoregressive part is parameterized by itself, so A1 and A2 hold.

We can see also a sufficient condition for A3. The usual stability condition for $\Phi(z)$ is required. Since $\Theta(z)$ is obtained through the Kalman Filter, $|\Theta(z)| \neq 0$ when $|z| < 1$, so it only remains the case $|z| = 1$. It suffices to check that the spectral density matrix of the process is of full rank for all frequencies. Let us consider the spectral density matrix of y_t, $f(\omega) = (2\pi)^{-1} \sum_{-\infty}^{\infty} \Gamma(k) \exp(-ik\omega)$, being $\Gamma(k)$ the autocovariance function of y_t. We can write it also as $f(\omega) = (2\pi)^{-1}\Phi(z)^{-1}\bar{\Theta}(z)[\Phi(z)^{-1}\bar{\Theta}(z)]^*$, with $z = \exp(i\omega)$. If the rank of $\bar{\Theta}(z)$ is full when $|z| = 1$, then $|\Theta(z)| \neq 0$ for $|z| = 1$. We only need that none of the polynomials in the diagonal matrix $\Theta_s(z)$ has unit modulus roots.

Summarizing the discussion above, there exists a function fulfilling A1-A3 that maps the parameters in the FSM to the corresponding VARMA models.

2.4.2. DFM and FSVAR as constrained FSM

The FSVAR model is clearly a FSM with $q = 1$. It is easy to see how the DFM is related to the FSM. Since $\Phi(z)^{-1} = |\Phi(z)|^{-1}\text{adj}(\Phi(z))$ we can write

(2.8) as $|\Phi(B)|f_t = \text{adj}(\Phi(B))\xi_t$ and (2.9) as $v_t = \varphi^i(B)^{-1}\eta_t^i$. Then, substituting in (2.7), we obtain $|\Phi(B)|\varphi^i(B)y_t^i = \varphi^i(B)L^i\text{adj}(\Phi(B))\xi_t + |\Phi(B)|\eta_t^i$, which is included in the model (2.11). Note that allowing (2.8) and (2.9) to be ARMA structures rather than AR does not prevent the DFM to be included in this scheme. This argument can also be easily applied to the Peña-Box model.

The function that maps the coefficients of the DFM, FSVAR or Peña-Box into the corresponding VARMA models via FSM can be considered as a VARMA parameterization which inherits A1-A3 from the transformation analyzed in the previous subsection. Unfortunately, A4 is not fulfilled in general, but in 2.4.3 we will show that the portmanteau statistic can be used nevertheless.

2.4.3. Deficiency of rank in the DFM

The model (2.7)-(2.9) is not identified because of the invariance under nonsingular linear transformations. Without identification constraints, we cannot expect the rank of the expression in A4 to be greater than $d = b - r^2$. We will present now sufficient conditions under which the rank is exactly d.

Proposition 2.2. *Let us assume that* $\Xi > 0$, $\sigma_i^2 > 0$, $|\Phi(z)|$ *and* $\varphi^i(z)$ *have their roots outside the unit circle for any* i *and*

$$\text{rank}[\ (I_{n^2} + K_{nn})(I_n \otimes L) \quad L \otimes L \quad J_n\ ] = n(r+1) \tag{2.15}$$

where $J_n = [A_1 \dots A_n]'$, $A_j = \text{diag}(\delta_{j1}, \dots, \delta_{jn})$. *Then,* $\text{rank}(\partial\pi^*/\partial\beta') = b - r^2$.

K_{nn} is the commutation matrix (Magnus and Neudecker, 1988, pag. 46).

NOTE: Proposition 2.2 can be easily adapted for the Peña-Box and FSVAR models.

Provided that the assumptions in proposition 2.2 hold, since the rank cannot be greater than d, there exists a neighborhood of β where the rank is exactly d. Then, we will show that there exists a parameterization with d parameters whose image contains the same VARMA models and fulfils A1-A4. This will allow us to prove that the distribution of the extended portmanteau statistic in this case is the one indicated in theorem 2.1 but with d instead of b.

We can identify the transfer functions with the elements in l^∞, which is a Banach space. Then, by the implicit function theorem, the mapping $\dot{\beta} \mapsto \pi^*(\dot{\beta})$ is smooth. We can adapt the proof of the Rank Theorem (Bröcker and Jänich, 1982), to see that there are C^2 diffeomorphisms ν and κ such as $\bar{\pi}^* = \nu \circ \pi^* \circ \kappa^{-1}$ has the form $\bar{\pi}^*(\dot{\lambda}_1, \ldots, \dot{\lambda}_b) = (\dot{\lambda}_1, \ldots, \dot{\lambda}_d, 0, \ldots)$.

If j_d is the injection $j_d(\dot{\lambda}_1, \ldots, \dot{\lambda}_d) = (\dot{\lambda}_1, \ldots, \dot{\lambda}_d, 0, \ldots, 0) \in \mathbb{R}^b$, the parameterization $\tilde{\Sigma} = \Sigma \circ \kappa^{-1} \circ j_d, \tilde{\Phi} = \Phi \circ \kappa^{-1} \circ j_d, \tilde{\Theta} = \Theta \circ \kappa^{-1} \circ j_d$ has the following properties,

(i) $\operatorname{rank}(\partial \tilde{\pi}^* / \partial \lambda') = d$.

(ii) There exists a neighborhood W of $\pi^*(\beta)$ such as $\tilde{\pi}^*(\tilde{\Omega}) \cap W = \pi^*(\Omega) \cap W$.

Consequently, the maximum likelihood estimate of π^* and thus, the values of Q_k^* and R_k^* do not depend on the parameterization in a neighborhood of the true values. The $\dot{\lambda}$ parameterization fulfils A1-A4, so Q_k^* and R_k^* are asymptotically chi-square distributed with d degrees of freedom.

2.5. Simulation Results

A detailed analysis of the power of the test under a great variety of cases is out of the scope of this paper. Nevertheless, we have done a little

experimentation for checking the asymptotic distribution of the test under the null hypothesis and for assessing its power under some alternative ones.

2.5.1. Detecting additional common factors

For the null hypothesis, we have simulated with MATLAB for different values of T, $N = 2500$ realizations of a process with one common factor,

$$[A_0] : y_t^i = \theta_{c,0,i1}\xi_t^1 + \theta_{c,1,i1}\xi_{t-1}^1 + \theta_{s,0,ii}\eta_t^i + \theta_{s,1,ii}\eta_{t-1}^i \tag{2.16}$$

where $\theta_{c,0,i1} = \theta_{s,0,ii} = 1, \theta_{c,1,i1} = i/(n+2), \theta_{s,1,ii} = (n+1-i)/(n+2) i = 1, \ldots, n,\ n = 5$. ξ_t is the common factor, η_t^i are the idiosyncratic factors. This model is estimated by maximum likelihood (the calculation of the likelihood function was programmed in C for greater speed). In table 2.1, we compare the rejection frequencies of $\tilde{Q}_k$, $\tilde{R}_k^*$ and $\tilde{Q}_k^*$ for different significance levels $\alpha = 0.1, 0.05$ and 0.01. Even if there is not an exact criterion to assign a number of degrees of freedom for $\tilde{Q}_k$, we can compute the critical value assuming a $\chi^2_{kn^2-2n}$ because in this case the conditions of theorem 2.2 are approximately met, since Σ depend mainly on $\theta_{c,0,i1}, \theta_{s,0,i1}$ and the MA coefficients on $\theta_{c,1,i1}/\theta_{c,0,i1}, \theta_{s,1,i1}/\theta_{s,0,i1}$.

Now, we consider a process generated with an additional common factor,

$$[A_2] : y_t^i = \theta_{c,0,i1}\xi_t^1 + \theta_{c,1,i1}\xi_{t-1}^1 + \theta_{c,0,i2}\xi_{t-1}^2 + \theta_{s,0,ii}\eta_t^i + \theta_{s,1,ii}\eta_{t-1}^i \tag{2.17}$$

Where $\theta_{c,0,i2} = 1$.

For A_2 we use the same equation as A_1 but with the coefficients $\theta_{c,0,i2} = i/5$.

Finally, we generate data with two additional common factors,

$$[A_3] : y_t^i = \theta_{c,0,i1}\xi_t^1 + \theta_{c,1,i1}\xi_{t-1}^1 + \theta_{c,0,i2}\xi_{t-1}^2 + \theta_{c,0,i3}\xi_{t-1}^3 + \theta_{s,0,ii}\eta_t^i + \theta_{s,1,ii}\eta_{t-1}^i \tag{2.18}$$

Where $\theta_{c,0,i2} = i/5$ and $\theta_{c,0,i3} = (6-i)/5$.

We estimate the model A_0 using data generated with $A_0 - A_3$. In table 2.1, we present the rejection frequencies under $A_0 - A_3$. The frequencies under the null hypothesis A_0 are the expected ones.

In A_1, the additional factor is difficult to detect, since its loadings are the same as the zero lag of ξ_t. Nevertheless, for very large series (T=256 or greater), $\tilde{R}_k^*$ outperforms both $\tilde{Q}_k^*$ and $\tilde{Q}_k$.

For A_2, it is easier to detect the lack of fit, so the three tests have greater rejection probabilities even for series of moderate size (T=128). Again, $\tilde{R}_k^*$ seems to be the most powerful test, $\tilde{Q}_k^*$ being in the middle. For large series, the performance of $\tilde{R}_k^*$ and $\tilde{Q}_k^*$ is similar, and $\tilde{Q}_k$ is clearly worse.

For A_3, the increase of rejection frequencies of $\tilde{R}_k^*$ and $\tilde{Q}_k^*$ is significant even for T=64. The performance of $\tilde{Q}_k$ is far below.

2.5.2. Detecting a lag of the common factor

In this case, null hypothesis will be that the series is generated by a model with a common factor without lags,

$$[B_0] : y_t^i = \theta_{c,0,i1}\xi_t^1 + \theta_{s,0,ii}\eta_t^i + \theta_{s,1,ii}\eta_{t-1}^i \tag{2.19}$$

And the alternative hypothesis is the model with lagged factor,

$$[B_1] : y_t^i = \theta_{c,0,i1}\xi_t^1 + \theta_{c,1,i1}\xi_{t-1}^1 + \theta_{s,0,ii}\eta_t^i + \theta_{s,1,ii}\eta_{t-1}^i \tag{2.20}$$

with the coefficients from the model A_0 in 2.5.1. Finally, we consider the case B_2 with the same model as B_0 but for ξ_t^1 an AR(1) instead of iid. We generate it as $\xi_t^1 = \phi\xi_{t-1}^1 + \varepsilon_t$ with ε_t iid of variance $1 - \phi^2$, $\phi = 0.3$.

The rejection frequencies (table 2.2) show that the power of $\tilde{R}_k^*$ is greater that $\tilde{Q}_k^*$ and $\tilde{Q}_k$, although the differences are not so large as in 2.5.1.

2.6. Real Data Example

We will show an application of the extended portmanteau test for the modelization of the industrial production of Spain 1992:1-2005:9 using a factor model. The following series are considered: Consumer durables, Consumer non-durables, Capital, Intermediate and Energy. We denote them by $x_t^i, i = 1, \ldots, 5$. All series are working-day corrected and are available from the web site of the Instituto Nacional de Estadística (www.ine.es). The univariate analysis of the five variables suggests the adequacy of an airline model for the logarithm-transformed series. Thus, we try a Factor Model with one common factor and seasonal MA idiosyncratic factors. If we put $y_t^i = (1-B)(1-B^{12}) \log x_t^i$, then we can write the model as,

$$[M_0] : y_t^i = \theta_{c,0,i1}\xi_t + (1 - \vartheta_{s,1,i}B)(1 - \vartheta_{s,12,i}B^{12})\eta_t^i \qquad (2.21)$$

where the η_t^i's are independent white noise processes with zero mean and variance σ_i^2. No dynamic model is proposed for the common factor ξ_t, which is instead assumed as another white noise process independent from the η_t^i's and of unit variance.

The maximum likelihood estimation yields the values in table 2.3. For $k = 24$, if we have used the classical portmanteau test, we would accept the null hypothesis of good fit even at 90% (see table 2.4). On the contrary, with the extended one, we reject even at 99%. Then, we modify the model according to the correlation of the residuals with the common factor. Concretely, we add some lags of the common factor, arriving to a new model M_1 (table 2.3). We can see that the p-value for $\tilde{R}_k^*$ allows accepting the null hypothesis at 99%.

The previous analysis, including estimation and tests, has been done with the data from 1992:1 to 2004:9. Thus, the last year of data can be used for measuring the out-of-sample forecasting performance of both models,

resulting a mean squared error of 9.30 for M_0 and 8.11 for M_1.

2.7. Conclusions

We have proved that the statistics R_k^* and Q_k^* can be used for testing the goodness of fit in constrained VARMA models when the constraints affect simultaneously to all coefficients, including those of the covariance matrix of the innovations. Under the null hypothesis, they are distributed as a chi-square. The classical portmanteau test can be used when the constraints do not *locally* mix the ARMA parameters and the coefficients of the innovation covariance matrix.

Simulated and real data suggest that the power of the test is greater with $\tilde{R}_k^*$ than with $\tilde{Q}_k^*$. If we apply the corrected classical portmanteau test $\tilde{Q}_k$ despite of the lack of theoretical basis, our results show that its power is less that with the extended ones.

2.A. Annex: Proofs

Lemma 2.1. *If $\Phi(z)$ and $\Theta(z)$ are matrix polynomials such as $|\Theta(z)| \neq 0$ for $|z| \leq 1$, then $\Pi = \Theta^{-1}\Phi$ satisfies for any k,*

$$\left\| \frac{\partial \Pi_j}{\partial \beta_k} \right\| < M\rho^j \tag{2.22}$$

with $M > 0$, $0 < \rho < 1$.

Proof. If we differentiate $\Pi(z)'\Theta(z)' = \Phi(z)'$ and multiply by $(\Theta^{-1} \otimes I_n)$ we get,

$$\frac{\partial \mathrm{vec}(\Pi'(z))}{\partial \beta'} = (\Theta(z)^{-1} \otimes I_n)\left[-(I_n \otimes \Pi(z)')\frac{\partial \mathrm{vec}(\Theta(z)')}{\partial \beta'} + \frac{\partial \mathrm{vec}(\Phi(z)')}{\partial \beta'}\right] \tag{2.23}$$

that has exponential decay.

Lemma 2.2. *The asymptotic covariance matrix of* $T^{1/2}(\hat{\beta}-\beta)$ *is,*

$$\left[\frac{1}{2}\frac{\partial\sigma'}{\partial\beta}\Sigma^{-1}\otimes\Sigma^{-1}\frac{\partial\sigma}{\partial\beta'}+\sum_{i,j=1}^{\infty}\frac{\partial\pi_i'}{\partial\beta}\Sigma^{-1}\otimes\Gamma(i-j)\frac{\partial\pi_j}{\partial\beta'}\right]^{-1} \tag{2.24}$$

Proof. If we denote by τ the varying elements on Φ and Θ and by λ the Lagrange multipliers of the constrained maximization of the log-likelihood, from theorem 4.3.1 in Hannan and Deistler (1988), we can obtain the asymptotic covariance of $T^{1/2}(\hat{\tau}'-\tau',\hat{\sigma}'-\sigma',\hat{\lambda}')'$. By a first-order Taylor expansion, we can obtain the asymptotic covariance of $T^{1/2}(\hat{\beta}-\beta)$ as,

$$I(\beta)^{-1}\left\{\frac{\partial\sigma'}{\partial\beta}s_0\frac{\partial\sigma}{\partial\beta'}+\frac{\partial\tau'}{\partial\beta}\frac{\partial^2 l}{\partial\tau\partial\tau'}\frac{\partial\tau}{\partial\beta'}\right\}I(\beta)^{-1} \tag{2.25}$$

where $s_0=4^{-1}(\Sigma^{-1}\otimes\Sigma^{-1})M_4(\Sigma^{-1}\otimes\Sigma^{-1})$, $M_4=var(\varepsilon_t\otimes\varepsilon_t)$ and,

$$I(\beta)=\frac{\partial\sigma'}{\partial\beta}\frac{\partial^2 l}{\partial\sigma\partial\sigma'}\frac{\partial\sigma}{\partial\beta'}+\frac{\partial\tau'}{\partial\beta}\frac{\partial^2 l}{\partial\tau\partial\tau'}\frac{\partial\tau}{\partial\beta'} \tag{2.26}$$

Under the gaussian assumption, $M_4=(I+K_{nn})(\Sigma\otimes\Sigma)$, and then,

$$\frac{\partial\sigma'}{\partial\beta}\frac{\partial^2 l}{\partial\sigma\partial\sigma'}\frac{\partial\sigma}{\partial\beta'}=\frac{\partial\sigma'}{\partial\beta}s_0\frac{\partial\sigma}{\partial\beta'} \tag{2.27}$$

Thus, the asymptotic covariance matrix becomes $I(\beta)^{-1}$ and to conclude we only need to derive the likelihood function and to use,

$$\frac{\partial^2 l}{\partial\tau\partial\tau'}=\lim_T TE\frac{\partial l_T}{\partial\tau}\frac{\partial l_T}{\partial\tau'} \tag{2.28}$$

Lemma 2.3. *Let* X_i, $i=0,1$ *be* $b\times n_i$ *matrices of rank* b_i, *such as* $b=b_0+b_1$ *and* $X_0'X_1=0$. *Then,* $X_0'(X_0X_0'+X_1X_1')^{-1}X_1=0$.

Proof. Let $X_i=U_iD_iV_i'$ be the Singular Value Decomposition (SVD) of X_i, with D_i diagonal and U_i, V_i orthogonal. By reordering, we can assume that D_0 is a diagonal matrix with nonzero values only in the first b_0 places of the main diagonal and D_1 in the b_1 last ones. Since the last b_1 columns of U_0 and the first b_0 ones of U_1 are multiplied by zero in the

SVD, we can substitute them for arbitrary values and the identities hold. Then, we can build a new orthogonal matrix U with the first b_0 columns of U_0 and the last b_1 ones of U_1, and we have $X_i = UD_iV_i'$. Now, it holds $X_0'(X_0X_0'+X_1X_1')^{-1}X_1 = X_0'U(D_0D_0'+D_1D_1')^{-1}U'X_1$. Then, $D_0D_0'+D_1D_1'$ is a diagonal matrix, while the last b_1 columns of $U'X_0 = D_0V_0'$ and the first b_0 ones of $U'X_1 = D_1V_1'$ are null, so the lemma follows.

Proof of Theorem 2.1. If A1-A4 hold, the assumptions on theorem 4.2.1 of Hannan and Deistler (1988) also hold. Consequently, $\Sigma(\hat{\beta})$, $\Phi_i(\hat{\beta})$ and $\Theta_j(\hat{\beta})$ are consistent and then, for T large enough they enter in any neighborhood of the true values. The rank condition implies that there exists a neighborhood U where the derivative,

$$\frac{\partial(\sigma', \phi', \theta')}{\partial\beta} \tag{2.29}$$

has full rank b. The Rank Theorem (Bröcker and Jänich, 1982) guarantees the existence of twice differentiable constraints such as the theorem 4.3.1 can be applied and the estimates satisfy the Central Limit Theorem. On the other hand, in this neighborhood the relations $\dot{\Sigma} = \Sigma(\dot{\beta})$, $\dot{\Phi}_i = \Phi_i(\dot{\beta})$,$\dot{\Theta}_i = \Theta(\dot{\beta})$ can be solved for $\dot{\beta}$ with regularity, so the Central Limit Theorem holds for $\hat{\beta}$. In lemma 2.2 we obtain the asymptotic covariance matrix of $T^{1/2}(\hat{\beta} - \beta)$.

We can assume that y_t is defined for any t from $-\infty$ to T. In order to simplify subsequent calculations, it is useful to see that the theoretical residuals $\dot{\varepsilon}_t$, i. e. those which could be computed by using the whole process y_t from $-\infty$ are, up to $O_p(T^{-1})$ terms, equal to the actual ones $\dot{e}_t$ computed using T terms. Both $\dot{\varepsilon}_t$ and $\dot{e}_t$, satisfy the relation $\dot{\Theta}(B)\varepsilon_t = \dot{\Phi}(B)y_t$. This implies that the under the stability hypothesis, $E|\dot{e}_t - \dot{\varepsilon}_t|^2 = O(\rho^t)$. Now, put $\dot{C}_j = T^{-1}\sum \dot{\varepsilon}_t\dot{\varepsilon}_{t+j}'$ and $\dot{D}_j = T^{-1}\sum \dot{e}_t\dot{e}_{t+j}'$. Then, the difference between $\dot{C}_j$ and $\dot{D}_j$ is an $O_p(T^{-1})$.

We need to calculate the asymptotic distribution of C_j. Let us stack

the autocovariances up to lag k into $c = [\text{vec}(C_1)', \ldots, \text{vec}(C_k)']'$ and $c^* = [\text{vec}(C_0)', c']'$. It is not difficult to see that $T^{1/2}(c^* - \sigma^*)$, where $\sigma^* = [\sigma', 0, \ldots, 0]'$, is asymptotically distributed as a normal with zero mean and covariance matrix,

$$M_k^* = \begin{pmatrix} (\Sigma \otimes \Sigma)(I_{n^2} + K_{nn}) & 0 \\ 0 & I_k \otimes \Sigma \otimes \Sigma \end{pmatrix} \tag{2.30}$$

In order to calculate the asymptotic distribution of the estimate $\hat{c}^*$ we will use its Taylor expansion around c^*. First of all, let us compute the derivatives of the residuals with respect to $\dot{\beta}$. We can write the residuals as $\dot{\varepsilon}_t = \sum_{i=1}^{\infty}(I_n \otimes y'_{t-i})\dot{\pi}_i$, where $\dot{\pi}_i = \text{vec}(\dot{\Pi}'_i)$ and $\dot{\Pi}(z) = \dot{\Theta}(z)^{-1}\dot{\Phi}(z)$. By lemma 2.1, the series can be differentiated term by term in the invertibility region and we obtain

$$\frac{\partial \dot{\varepsilon}'_t}{\partial \dot{\beta}} = \sum_{i=1}^{\infty} \frac{\partial \dot{\pi}'_i}{\partial \dot{\beta}}(I_n \otimes y_{t-i}) \tag{2.31}$$

We write the vectorized covariance matrix $\text{vec}(C_j)$ as $\sum(\varepsilon_{t+j} \otimes I_n)\varepsilon_t$. Taking derivatives with respect to $\dot{\beta}$, we obtain,

$$\frac{\partial \text{vec}(C_j)'}{\partial \beta} = \frac{1}{T}\sum_{t=1}^{T-j}\sum_{i=1}^{\infty} \frac{\partial \pi'_i}{\partial \beta}\Big((I_n \otimes y_{t+j-i}\varepsilon'_t) + (\varepsilon'_{t+j} \otimes y_{t-i})\Big) \tag{2.32}$$

Due to the ergodicity of y_t, the sum above converges to $E_j = \sum_{i=1}^{j} \partial \pi'_i / \partial \beta (I_n \otimes \Psi_{j-i}\Sigma)$ for $j > 0$ and to zero for $j = 0$. Thus, since $\hat{c}^* = c^* + \partial c/\partial \beta(\hat{\beta} - \beta) + O_p(T^{-1})$ and defining $\tilde{X} = [0, E_1, \ldots, E_k]' = [0, X']'$, then $\partial c/\partial \beta = X + o_p(1)$, we find that

$$\hat{c}^* = c^* + \tilde{X}(\hat{\beta} - \beta) + o_p(T^{-1/2}) \tag{2.33}$$

Before using the expression above to obtain the asymptotic distribution of $\hat{c}^*$, we need the asymptotic covariance of c^* and $\hat{\beta}$. The log-likelihood of $\dot{\beta}$ is up to constants, normalizing by T, $L_T(\dot{\beta}) = -(2T)^{-1}\sum_{t=1}^{T} \log|\dot{\Sigma}_t| -$

$(2T)^{-1}\sum_{t=1}^{T}\dot{e}_t'\dot{\Sigma}_t^{-1}\dot{e}_t$, where $\dot{\Sigma}_t$ is the covariance matrix of $\dot{e}_t$. Under A3, $\Sigma_t = \Sigma + O(\rho^t)$ and thus, only $O_p(T^{-1})$ terms are neglected in the following calculations if instead of L_T we use,

$$l_T(\dot{\beta}) = -\frac{1}{2}\log|\dot{\Sigma}| - \frac{1}{2T}\sum_{t=1}^{T}\dot{\varepsilon}_t'\dot{\Sigma}^{-1}\dot{\varepsilon}_t \tag{2.34}$$

By a first order Taylor expansion of $\partial l_T/\partial\dot{\beta}$,

$$\hat{\beta} - \beta = I(\beta)^{-1}\frac{\partial l_T}{\partial\beta} + o_p(T^{-1/2}) \tag{2.35}$$

If we denote $\mathrm{vec}(\Sigma^{-1})$ by σ^{-1}, the derivative $\partial l_T/\partial\beta$ equals,

$$-\frac{1}{2}\frac{\partial\sigma'}{\partial\beta}\sigma^{-1} - \frac{1}{2T}\sum_{t=1}^{T}\left\{2\frac{\partial\varepsilon_t'}{\partial\beta}\Sigma^{-1}\varepsilon_t - \frac{\partial\sigma'}{\partial\beta}(\Sigma^{-1}\varepsilon_t\otimes\Sigma^{-1}\varepsilon_t)\right\} \tag{2.36}$$

The first term in (2.36) is deterministic and then, its covariance with $\mathrm{vec}(C_j)$ is zero. It also holds that $\mathrm{cov}(\Sigma^{-1}\varepsilon_t\otimes\Sigma^{-1}\varepsilon_t, \mathrm{vec}(C_j)) = 0$ for $j > 0$, due to the independence of the ε's. Then, using (2.31) we obtain as in Ahn (1988) for $j > 0$

$$\mathrm{cov}(T\frac{\partial l_T}{\partial\beta}, \mathrm{vec}(C_j)) = \frac{1}{T}\sum_{s=1}^{T-j}\sum_{i=1}^{\infty}E\frac{\partial\pi_i'}{\partial\beta}(I_n\otimes y_{s+j-i}\varepsilon_s') \tag{2.37}$$

which converges to E_j when $T\to\infty$. For the case $j = 0$, the first term between braces $\{\ldots\}$ in (2.36) is uncorrelated with $\mathrm{vec}(C_0)$, so $\mathrm{cov}(T\partial l_T/\partial\beta, \mathrm{vec}(C_0))$ equals,

$$\frac{1}{2T}\sum_{s=1}^{T-j}\sum_{t=1}^{T}\mathrm{cov}\left(\frac{\partial\sigma'}{\partial\beta}(\Sigma^{-1}\varepsilon_t\otimes\Sigma^{-1}\varepsilon_t), (I_n\otimes\varepsilon_s)\varepsilon_s\right) \tag{2.38}$$

It can be proved that $E(\Sigma^{-1}\varepsilon_t\otimes\Sigma^{-1}\varepsilon_t)\varepsilon_t'(I_n\otimes\varepsilon_t') = \sigma^{-1}\sigma' + I_{n^2} + K_{nn}$, so (2.38) yields,

$$\frac{1}{2T}\sum_{t=1}^{T}\frac{\partial\sigma'}{\partial\beta}\left(\sigma^{-1}\sigma' + I_{n^2} + K_{nn} - \sigma^{-1}\sigma\right) \tag{2.39}$$

Which by the symmetry of Σ, equals $\partial\sigma'/\partial\beta$. Let us call $E_0 = \partial\sigma'/\partial\beta$, $X^* = [E_0, X']'$, and then, $\lim_{T\to\infty} \mathrm{cov}(T\partial l_T/\partial\beta, c^*) = X^*$. Now, from the equation above and (2.35), $\mathrm{cov}(\hat{\beta}, c^*) = T^{-1}I(\beta)^{-1}X^*$. Since Σ is smooth enough, we can write,

$$\hat{\sigma}^* = \sigma^* + \begin{pmatrix} \frac{\partial\sigma}{\partial\beta'} \\ 0 \end{pmatrix}(\hat{\beta} - \beta) + o_p(T^{-1/2}) \tag{2.40}$$

From (2.33) and (2.40), $T^{1/2}(\hat{c}^* - \hat{\sigma}^*) = T^{1/2}(c^* - \sigma^*) - T^{1/2}X^*(\hat{\beta} - \beta) + o_p(1)$ and then, the asymptotic covariance matrix is,

$$M_k^* - X^* I(\beta)^{-1} X^{*\prime} \tag{2.41}$$

Since Σ is positive definite, there exists some S such as $S'S = \Sigma^{-1}$ and $S\Sigma S = I_n$. We define,

$$G_k = \begin{pmatrix} \frac{1}{\sqrt{2}} S \otimes S & 0 \\ 0 & I_k \otimes S \otimes S \end{pmatrix} \tag{2.42}$$

Let us call $\hat{G}_k$ a consistent estimate of G_k. If we define $\tilde{c}^* = \hat{G}_k(\hat{c}^* - \hat{\sigma}^*)$ then $T^{1/2}\tilde{c}^*$ is distributed asymptotically with covariance matrix, $V_k = Z_k - G_k X^* I(\beta)^{-1} X^{*\prime} G_k'$, where,

$$Z_k = \begin{pmatrix} \frac{1}{2}(I_{n^2} + K_{nn}) & 0 \\ 0 & I_{kn^2} \end{pmatrix} \tag{2.43}$$

The expression $T(\tilde{c}^*)'\tilde{c}^*$ equals Q_k^* or R_k^* depending on which $\hat{G}_k$ we choose among the following,

$$\begin{pmatrix} \frac{1}{\sqrt{2}} \hat{\Sigma}^{-1/2} \otimes \hat{\Sigma}^{-1/2} & 0 \\ 0 & I_k \otimes C_0^{-1/2} \otimes C_0^{-1/2} \end{pmatrix} \tag{2.44}$$

$$\begin{pmatrix} \frac{1}{\sqrt{2}} \hat{\Sigma}^{-1/2} \otimes \hat{\Sigma}^{-1/2} & 0 \\ 0 & I_k \otimes \Sigma^{-1/2} \otimes \Sigma^{-1/2} \end{pmatrix} \tag{2.45}$$

Let us check that for large k, V_k is nearly an idempotent matrix of the required rank, but we still need some calculations. First, we can see that $J(\beta) = X^{*\prime} G_k' G_k X^*$ equals,

$$\frac{1}{2}\frac{\partial \sigma'}{\partial \beta}\Sigma^{-1} \otimes \Sigma^{-1}\frac{\partial \sigma}{\partial \beta'} + \sum_{l=1}^{k}\sum_{i,j=1}^{\infty}\frac{\partial \pi_i'}{\partial \beta}\Sigma^{-1} \otimes (\Psi_{l-i}\Sigma\Psi_{l-j})\frac{\partial \pi_j}{\partial \beta'} \tag{2.46}$$

which is the same as $I(\beta)$ save for $\sum_{l=1}^{k}\Psi_{l-i}\Sigma\Psi_{l-j}$ instead of $\Gamma(i-j) = \sum_{l=1}^{\infty}\Psi_{l-i}\Sigma\Psi_{l-j}$. Thus, the difference $J(\beta) - I(\beta)$ is a $O(\rho^k)$. Then we can write V_k as,

$$\begin{aligned} V_k &= W_k + U_k \\ W_k &= Z_k - G_k X^* J(\beta)^{-1} X^{*\prime} G_k' \\ U_k &= G_k X^* [J(\beta)^{-1} - I(\beta)^{-1}] X^{*\prime} G_k' \end{aligned} \tag{2.47}$$

Where $\|U_k\| \leq O(\rho^k)\|X^{*\prime} G_k'\|^2$. The matrix $(I_{n^2} + K_{nn})/2$ is idempotent and has rank $n(n+1)/2$ (see Magnus and Neudecker, 1982, pag. 48). On the other hand, $Z_k G_k X^* = G_k X^*$ due to $K_{nn}(S \otimes S) = (S \otimes S)K_{nn}$ and $K_{nn}\partial\sigma/\partial\beta' = \partial\sigma/\partial\beta'$. Then, W_k is the difference of commuting idempotent matrices and then, it is idempotent itself and its rank equals rank(Z_k) − rank(X^*) (using that for idempotent matrices, the rank equals the trace). The rank of X^* is b for large k due to A4 together with (2.23).

Since W_k is idempotent of rank $d = n(n+1)/2 + kn^2 - b$, there exists some $(k+1)n^2 \times (k+1)n^2$ matrix H such as $H'H = I_{(k+1)n^2}$ and,

$$HW_kH' = \begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix} \tag{2.48}$$

If $H = (H_0', H_1')'$ with H_0 sized $d \times (k+1)n^2$, then $u = H_0\tilde{c}^*$ satisfies, $T(\tilde{c}^*)'\tilde{c}^* = Tu'u + \Delta_1$, where $E|\Delta_1|^2 \leq M\|U_k\|$. Now, we define $v = (I_d + H_0 U_k H_0')^{-1/2}u$ and then, $TEvv' = I_d$ and, $Tv'v = Tu'u + \Delta_2$ with,

$$\Delta_2 = T\sum_{j=1}^{\infty}(-1)^j u'(H_0 U_k H_0')^j u \tag{2.49}$$

and then for large k,

$$E|\Delta_2| \leq \sum_{j=1}^{\infty} Mk\rho^{kj}\|X^{*\prime}G_k'\|^j \leq M\frac{k\rho^k\|X^{*\prime}G_k'\|}{1-\rho^k\|X^{*\prime}G_k'\|} \tag{2.50}$$

Let us see that $\|X^{*\prime}G_k'\|$ is bounded uniformly in k. It is easy to see that the norm of G_k is bounded. For X^*, we can use that,

$$\|E_j\| \leq \sum_{i=1}^{j} \|\frac{\partial \pi_i'}{\partial \beta}\| \times \|I_n \otimes \Psi_{j-i}\Sigma\| \tag{2.51}$$

Since Ψ and Π are both $O(\rho^k)$, then $\|E_j\| \leq M\rho^i\rho^{j-i}$, and thus, the norm of X^* is uniformly bounded.

From (2.33), (2.35) and (2.36) we find that, up to $o_p(T^{-1/2})$ terms, any linear combination of $(v', \hat{\beta}' - \beta')'$ is a martingale difference. The Central Limit Theorem for martingales (Billingsley, 1995, p. 476) applies, since the martingale version of the Lindeberg condition holds under bounded fourth-order moments. Thus, $T^{1/2}v$ converges in distribution to a multivariate normal with zero mean and unit covariance matrix. Therefore, $Tv'v$ is asymptotically distributed as a chi-square with d degrees of freedom. Summarizing, we have that the extended portmanteau statistic equals the sum of $Tv'v$, plus $O(\rho^k)$ terms. It holds that for any $\alpha \in (0,1)$, $\delta > 0$,

$$\begin{aligned} P[Q_k^* < x_{k,\alpha}] &\leq P[Tv'v \leq \chi^2_{d,\alpha} + \delta] + P[|O(\rho^k)| > \delta] \\ P[Tv'v < x_{k,\alpha} - \delta] &\leq P[Q_k^* \leq \chi^2_{d,\alpha}] + P[|O(\rho^k)| > \delta] \end{aligned} \tag{2.52}$$

Letting $T \to \infty$ and using that the maximum of the density function of a χ^2_l is $O(l^{-1})$,

$$\limsup_T |F_{k,T}(\chi^2_{d,\alpha}) - (1-\alpha)| \leq O(k^{-1})\delta + O(\rho^k)/\delta^2 \tag{2.53}$$

If we choose $\delta = \zeta^k$ with $\rho < \zeta^2 < 1$, the theorem yields.

□

Proof of Proposition 2.1. It is easy to see that A5 implies A5'. Since the jacobian matrix of the parameterization $\partial(\sigma, \phi, \theta)/\partial\lambda$ is block diagonal, the image of the linear transformation is the product of the images of $\partial\sigma/\partial\lambda_0$ and $\partial(\phi, \theta)/\partial\lambda_1$.

Conversely, if A5' holds and $\Delta = (\sigma(\dot{\beta}), \phi(\dot{\beta}), \theta(\dot{\beta}))$ is a local parameterization with $\dot{\beta} \in \Omega$, we can say that the image of $\partial\Delta/\partial\lambda$ is equal to $E_\sigma \times E_{\phi,\theta}$, so it has a base $\{u_1, \ldots, u_{b_0}\} \cup \{u_{b_0+1}, \ldots, u_b\}$, such as $u_j \in E_\sigma \times \{0_{(p+q)n^2}\}$ for $j = 1, \ldots, b_0$ and $u_j \in \{0_{n^2}\} \times E_\sigma$ for $j = b_0 + 1, \ldots, b$. We can complete the base with vectors in the orthogonal complement of the image, so if we build U with the $u's$ arranged in columns and V with the $v's$, being $v_j = (\partial\Delta/\partial\beta)^{-1}u_j$, we can say that $\partial\Delta/\partial\beta = UV^{-1}$. Then, the parameterization $\Delta(V\dot{\lambda})$ for $\dot{\lambda} \in \Lambda = V^{-1}\Omega$ fulfils A5.

□

Proof of Theorem 2.2. If we prove that for any $j > 0$, $E_0' I(\beta)^{-1} E_j = 0$, then (2.41) is block diagonal and the asymptotic covariance matrix of $T^{1/2}\hat{c}$ is $I - XI(\beta)^{-1}X'$, which is an approximately idempotent matrix of rank $n^2k - rank(X)$. The proof of theorem 2.1 can be easily adapted from (2.41) onwards.

Now, A5 implies,

$$\frac{\partial\sigma}{\partial\lambda'}\frac{\partial\phi_j'}{\partial\lambda} = \frac{\partial\sigma}{\partial\lambda'}\frac{\partial\theta_j'}{\partial\lambda} = 0 \tag{2.54}$$

so we can use (2.23) to see that,

$$\frac{\partial\sigma}{\partial\lambda'}\frac{\partial\pi_j'}{\partial\lambda} = 0 \tag{2.55}$$

For $l > 0$, let us define the matrix W with $n^2 \times n^2$ block i, j as $\Sigma^{-1} \otimes \Gamma(i-j)$. The block Toeplitz matrix with (i, j) block $\Gamma(i - j)$ and Σ are positive semidefinite, so we know that there exists $W^{1/2}$, so we can apply lemma 2.3 to $X_0 = 2^{1/2}(\partial\sigma'/\partial\lambda)(S \otimes S)$ and $X_1 = [\partial\pi_1'/\partial\lambda, \ldots, \partial\pi_l'/\partial\lambda]W^{1/2}$ to

conclude that for fixed $j \leq l$,

$$\frac{\partial \sigma}{\partial \lambda'}(X_0 X_0' + X_1 X_1')^{-1} \frac{\partial \pi_j'}{\partial \lambda} = 0 \tag{2.56}$$

If we let $l \to \infty$ we find,

$$\frac{\partial \sigma}{\partial \lambda'} I(\lambda)^{-1} \frac{\partial \pi_j'}{\partial \lambda} = 0 \tag{2.57}$$

Now, by using that

$$I(\lambda)^{-1} = \frac{\partial \lambda}{\partial \beta'} I(\beta)^{-1} \frac{\partial \lambda'}{\partial \beta} \tag{2.58}$$

then,

$$\frac{\partial \sigma}{\partial \beta'} I(\beta)^{-1} \frac{\partial \pi_j'}{\partial \dot{\beta}} = \frac{\partial \sigma}{\partial \lambda'} \frac{\partial \lambda}{\partial \beta'} I(\beta)^{-1} \frac{\partial \lambda'}{\partial \beta} \frac{\partial \pi_j'}{\partial \lambda} = \frac{\partial \sigma}{\partial \lambda'} I(\lambda)^{-1} \frac{\partial \pi_j'}{\partial \lambda} = 0 \tag{2.59}$$

From the equation above, we find that $E_0' I(\beta)^{-1} E_j = 0$, so we conclude the theorem.

□

Proof of Proposition 2.2. We will prove that $\text{rank}(\partial \psi^* / \partial \beta') = d$. Let $\dot{\psi}^* = (\dot{\sigma}', \dot{\psi}')'$ be a vector containing a generic covariance matrix and a generic transfer function in vector form. We can define $\mathcal{S}_A(\dot{\psi}^*) = f$, a spectral density function defined by $f(\omega) = \dot{\Psi}(z) \dot{\Sigma} \dot{\Psi}(z^{-1})$, with $z = \exp(i\omega)$. If ψ^* corresponds to y_t in (2.7)-(2.9), then $\mathcal{S}_A(\psi^*) = f_y(\omega)$ can be also computed from (2.7) as $f_y(\omega) = L f_c(\omega) L' + \text{diag}(f_s(\omega))$, where f_c is the spectral density matrix of the common factors and f_s is a vector containing the spectral densities of the idiosyncratic factors. Let us define the mappings $\mathcal{S}_F(\dot{\beta}) = (\dot{L}, \dot{f}_c, \dot{f}_s)$ and $\mathcal{F}(a, b, c) = aba' + \text{diag}(c)$. Since $\mathcal{S}_A \circ \psi^* = \mathcal{F} \circ \mathcal{S}_F$, we get by the cha in ru le,

$$d\mathcal{S}_A(\psi^*(\beta)) \frac{\partial \psi^*}{\partial \beta'} = d\mathcal{F}(\mathcal{S}_F(\beta)) \frac{\partial \mathcal{S}_F}{\partial \beta'} \tag{2.60}$$

If we prove that the right side has rank d, then necessarily $\partial\psi^*/\partial\beta'$ has rank at least d. Let us see that indeed the product above has rank d. First of all, the derivative of $\mathcal{F}$ can be expressed as,

$$\frac{\partial \text{vec}(\mathcal{F})}{\partial(a', b', c')}(L, f_c, f_s) = [(I_{n^2} + K_{nn})(Lf_c \otimes I_n), L \otimes L, J_n] \tag{2.61}$$

Since $f_c(\omega)$ is nonsingular for any ω, then the rank of the matrix above, say M, is $n(r+1)$. We can see that the kernel of M is spanned by the columns of the matrix,

$$N = \begin{bmatrix} -I_r \otimes L \\ (I_{r^2} + K_{rr}) f_c \otimes L \\ 0_{n \times r^2} \end{bmatrix} \tag{2.62}$$

The space spanned by the columns of N, $im(N)$ is included in $\ker(M)$ because $MN = 0$. The dimension of the kernel is the number of columns $n(r+1) + r^2$, minus $\text{rank}(M) = n(r+1)$, that is, r^2. If L is of full rank, then the dimension of $im(N)$ is also r^2 and then $im(N) = \ker(M)$. The case $\text{rank}(L) < r$ can be ruled out because it contradicts $\text{rank}(M) = n(r+1)$.

We denote by $(\beta_1', \beta_2', \beta_3')'$ a partition of β such as $\beta_1 = \text{vec}(L)$ and β_2 and β_3 contain the coefficients of the common and idiosyncratic factors respectively. In order to see that the right hand side of (2.60) has rank d, let $u = (u_1', u_2', u_3')$ be a vector in the kernel. Then, it holds,

$$[(I_{n^2} + K_{nn})(Lf_c \otimes I_n), L \otimes L, J_n] \begin{bmatrix} u_1 \\ \frac{\partial \text{vec}(f_c)}{\partial \beta_2'} u_2 \\ \frac{\partial \text{vec}(f_s)}{\partial \beta_3'} u_3 \end{bmatrix} = 0 \tag{2.63}$$

The relation above depends on ω through the spectral densities. Since it holds for any ω, there exists a certain $v(\omega)$ such as,

$$\begin{bmatrix} u_1 \\ \frac{\partial \text{vec}(f_c)}{\partial \beta_2'} u_2 \\ \frac{\partial \text{vec}(f_s)}{\partial \beta_3'} u_3 \end{bmatrix} = N v(\omega) \tag{2.64}$$

Since L is of full rank, we can put $v(\omega) = -(I_k \otimes (L'L)^{-1}L')u_1$, so $v(\omega)$ is in fact constant. If $\partial \text{vec}(f_c)/\partial\beta_2'$ and $\partial \text{vec}(f_s)/\partial\beta_3'$ are of full rank we can also solve (2.64) for u_2 and u_3, obtaining u as a linear transformation of v and thus, the dimension of the kernel of (2.60) cannot be greater than r^2.

The condition on the rank of $\partial \text{vec}(f_c)/\partial\beta_2'$ is easily checked when $\Xi > 0$ and $\Phi(z)$ is stable because the differential of the mapping $\mathcal{S}(\Sigma, A(z)) = (I + zA(z))\Sigma(I + zA(z))^*$ is injective when $\Sigma > 0$ and $I + zA(z)$ stable. If $d\mathcal{S}(\Sigma, A(z))(\Omega, B(z)) = \mathcal{S}(\Omega, A(z)) + (I + zA(z))\Sigma(zB(z))^* + zB(z)\Sigma(I + zA(z))^* = 0$, then, $\Omega + \Sigma((I + zA(z))^{-1}zB(z))^* + (I + zA(z))^{-1}zB(z)\Sigma = 0$. When $|z| = 1$, the second term has only negative powers of z, and the third term only positive ones. Then, we conclude that $\Omega = 0$ and $B(z) = 0$. For $A(z) = \Phi_1 z + \Phi_2 z^2 + \dots$ and $\Sigma = \Xi$, this means that the differential of the spectral mapping $\mathcal{S}$ at the true values is injective, while the differential of the mapping from the parameters to the transfer functions is trivially injective. The argument also hol ds fo r the elements of f_s.

□

Bibliography

[1] Ahn, S. K. (1988) Distribution for Residual Autocovariances in Multivariate Autoregressive Models With Structured Parameterization. *Biometrika* 75, 590-593.

[2] Bai, J. (2003) Inferential Theory for Factor Models of Large Dimensions. *Econometrica* 71, 135-171.

[3] Bai, J and Ng, S. (2002) Determining the Number of Factors in Approximate Factor Models. *Econometrica* 70, 191-221.

[4] Billingsley, P. (1995) *Probability and Measure* (3rd ed.). New York: John Wiley and Sons.

[5] Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994). *Time Series Analysis. Forecasting and Control* (3rd ed.). Englewood Cliffs, New Jersey: Prentice Hall.

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

[7] Bröcker, T. and Jänich, K. (1982) Introduction to Differential Topology. New York, Cambridge University Press.

[8] Chamberlain, G. and Rothschild, M. (1983) Arbitrage, Factor Structure and Mean-Variance Analysis in Large Assets Markets. *Econometrica* 51, 1305-1324.

[9] Engle, R. And Watson, M. (1981) A One-Factor Multivariate Time Series Model of Metropolitan Wage Rates. *Journal of the American Statistical Association* 376, 774-781.

[10] Forni, M., Hallin, M., Lippi, F., Reichlin, L. (2000) The Generalized Dynamic Factor Model: Identification and Estimation. *Review of Economics and Statistics* 82, 540-554.

[11] Forni, M., Hallin, M., Lippi, F., Reichlin, L. (2004) The Generalized Dynamic Factor Model: Consistency and Rates. *Journal of Econometrics* 119, 231-255.

[12] Forni, M., Hallin, M., Lippi, F., Reichlin, L. (2005) The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting. *Journal of the American Statistical Association* 100, 830-840.

[13] Geweke, J. (1977) The Dynamic Factor Analysis of Economic Time-Series Model. *Latent Variables in Socio-Economic Models*, eds. D. J. Aigner and A. S. Goldberger, Amsterdam: North Holland.

[14] Hannan, E. J. and Deistler, M. (1988) *The Statistical Theory of Linear Systems*, New York: John Wiley and Sons.

[15] Hosking, J. R. M. (1980) The Multivariate Portmanteau Statistic. *Journal of the American Statistical Association* 371, 602-608.

[16] Hu, Y.-P. and Chou, R.-J. (2004) On the Peña-Box Model. *Journal of Time Series Analysis* 25, 811-830.

[17] Li, W. K. and McLeod, A. I. (1981) Distribution of the Residual Autocorrelations in Multivariate ARMA Time Series Models. *Journal or the Royal Statistical Society*, Ser. B, 43, 231-239.

[18] Ljung, G. M. and Box, G. E. P. (1978) On a Measure of Lack of Fit in Time Series Models. *Biometrika* 65, 297-303.

[19] Lütkepohl, H. (1984) Linear Transformations of vector ARMA processes. *Journal of Econometrics* 26, 283-293.

[20] Lütkepohl, H. (1991) *Introduction to Multiple Time Series*, Berlin: Springer-Verlag.

[21] Magnus, J. R. and Neudecker, H. (1988) Matrix Differential Calculus. New York: John Wiley and Sons.

[22] McLeod, A. I. (1978) On the Distribution of Residual Autocorrelations in Box-Jenkins Models. *Journal of the Royal Statistical Society* Ser. B, 40, 296-302.

[23] Peña, D. and Box, G. E. P. (1987) Identifying a Symplifying Structure in Time Series *Journal of the American Statistical Association* 82, 836-843.

[24] Poskitt, D. S. and Tremayne, A. R. (1982) Diagnostic Tests for Multiple Time Series Models. *The Annals of Statistics* 10, 114-120.

[25] Quah, D. and Sargent, T. J. (1993) A Dynamic Index Model for Large Cross-Sections. *Business Cycles, Indicators and Forecasting*, eds. J. H. Stock and M. H. Watson, Chicago: University of Chicago Press.

[26] Sargent, T. J. and Sims, C. A. (1977) Business Cycle Modelling Without Pretending to Have Too Much A Priori Economic Theory. *New Methods*

in Business Cycle Research: Proceedings from a Conference, ed, C. A. Sims, Minneapolis: Federal Reserve Bank of Minneapolis.

[27] Stock, J. H. and Watson, M. W. (1998) Diffusion Indexes. *NBER Working Paper* No. W6702.

[28] Stock, J. H. and Watson, M. W. (2002) Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association* 97, 1167-1179.

[29] Stock, J. H. and Watson, M. W. (2003) Understanding Changes in International Business Cycle Dynamics. *NBER Working Paper* No. W9859.

Table 2.1: Rejection frequencies of $\tilde{R}_k^*$, $\tilde{Q}_k^*$ and $\tilde{Q}_k$ under A_0, A_1, A_2 and A_3.

		$\tilde{R}_k^*$			$\tilde{Q}_k^*$			$\tilde{Q}_k$		
T	model	90%	95%	99%	90%	95%	99%	90%	95%	99%
32	A_0	0.190	0.106	0.028	0.052	0.020	0.002	0.116	0.048	0.006
64	A_0	0.122	0.064	0.016	0.065	0.028	0.004	0.095	0.048	0.008
128	A_0	0.105	0.046	0.008	0.070	0.029	0.005	0.091	0.040	0.006
256	A_0	0.109	0.054	0.013	0.093	0.043	0.010	0.105	0.053	0.011
512	A_0	0.098	0.054	0.008	0.091	0.050	0.007	0.094	0.052	0.008
1024	A_0	0.092	0.055	0.009	0.089	0.048	0.009	0.092	0.047	0.009
32	A_1	0.167	0.092	0.019	0.044	0.020	0.002	0.104	0.044	0.004
64	A_1	0.124	0.068	0.017	0.068	0.032	0.003	0.103	0.050	0.009
128	A_1	0.115	0.059	0.013	0.083	0.037	0.008	0.102	0.047	0.008
256	A_1	0.128	0.064	0.015	0.104	0.052	0.012	0.104	0.050	0.012
512	A_1	0.141	0.075	0.022	0.127	0.066	0.018	0.114	0.059	0.014
1024	A_1	0.180	0.106	0.028	0.169	0.100	0.024	0.139	0.077	0.016
32	A_2	0.092	0.052	0.010	0.054	0.020	0.003	0.113	0.050	0.008
64	A_2	0.121	0.058	0.010	0.102	0.044	0.008	0.130	0.059	0.012
128	A_2	0.185	0.102	0.030	0.161	0.089	0.021	0.144	0.083	0.015
256	A_2	0.352	0.240	0.090	0.328	0.216	0.072	0.220	0.125	0.035
512	A_2	0.669	0.534	0.304	0.628	0.492	0.254	0.363	0.239	0.081
1024	A_2	0.956	0.920	0.797	0.938	0.889	0.745	0.676	0.539	0.285
32	A_3	0.115	0.062	0.012	0.080	0.033	0.004	0.118	0.054	0.007
64	A_3	0.192	0.110	0.036	0.167	0.098	0.024	0.122	0.065	0.017
128	A_3	0.352	0.240	0.086	0.344	0.226	0.071	0.147	0.076	0.018
256	A_3	0.697	0.586	0.346	0.691	0.560	0.306	0.210	0.112	0.030
512	A_3	0.976	0.953	0.863	0.970	0.943	0.839	0.377	0.250	0.083
1024	A_3	1.000	1.000	1.000	1.000	1.000	1.000	0.701	0.565	0.311

Table 2.2: Rejection frequencies of $\tilde{R}_k^*$, $\tilde{Q}_k^*$ and $\tilde{Q}_k$ under B_0, B_1 and B_2.

		$\tilde{R}_k^*$			$\tilde{Q}_k^*$			$\tilde{Q}_k$		
T	model	90%	95%	99%	90%	95%	99%	90%	95%	99%
32	B_0	0.074	0.038	0.009	0.048	0.018	0.002	0.089	0.045	0.007
64	B_0	0.082	0.035	0.006	0.067	0.029	0.005	0.092	0.045	0.009
128	B_0	0.086	0.039	0.009	0.078	0.033	0.006	0.085	0.039	0.009
256	B_0	0.085	0.040	0.007	0.077	0.038	0.007	0.085	0.040	0.006
512	B_0	0.086	0.038	0.009	0.085	0.039	0.009	0.090	0.039	0.008
1024	B_0	0.084	0.038	0.010	0.082	0.038	0.009	0.080	0.037	0.008
32	B_1	0.126	0.069	0.020	0.084	0.042	0.006	0.140	0.078	0.016
64	B_1	0.260	0.158	0.056	0.204	0.112	0.028	0.226	0.130	0.031
128	B_1	0.540	0.413	0.192	0.497	0.350	0.133	0.438	0.301	0.106
256	B_1	0.928	0.862	0.664	0.891	0.808	0.571	0.821	0.698	0.425
512	B_1	1.000	0.998	0.994	0.999	0.998	0.986	0.996	0.990	0.948
1024	B_1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
32	B_2	0.099	0.054	0.017	0.059	0.023	0.007	0.121	0.052	0.011
64	B_2	0.135	0.086	0.032	0.116	0.065	0.016	0.152	0.088	0.024
128	B_2	0.262	0.179	0.084	0.227	0.152	0.070	0.249	0.169	0.076
256	B_2	0.430	0.324	0.179	0.392	0.286	0.148	0.400	0.291	0.151
512	B_2	0.721	0.594	0.388	0.666	0.545	0.329	0.667	0.544	0.325
1024	B_2	0.978	0.956	0.871	0.968	0.938	0.828	0.967	0.934	0.820

Table 2.3: Maximum Likelihood estimates of M_0 and M_1

model	i				σ_i
M_0	1	x_t^1	=	$(1-0.391B)(1-0.502B^{12})\eta_t^1 + 0.042\xi_t$	0.034
M_0	2	x_t^2	=	$(1-0.682B)(1-0.676B^{12})\eta_t^2 + 0.011\xi_t$	0.029
M_0	3	x_t^3	=	$(1-0.563B)(1-0.297B^{12})\eta_t^3 + 0.034\xi_t$	0.037
M_0	4	x_t^4	=	$(1-0.995B)(1-0.575B^{12})\eta_t^4 + 0.025\xi_t$	0.009
M_0	5	x_t^5	=	$(1-0.474B)(1-0.792B^{12})\eta_t^5 + 0.006\xi_t$	0.029
M_1	1	x_t^1	=	$(1-0.309B)(1-0.400B^{12})\eta_t^1 +$	0.042
				$+0.042(1-0.464B+0.196B^2)(1-0.045B^{12})\xi_t$	
M_1	2	x_t^2	=	$(1-0.689B)(1-0.637B^{12})\eta_t^2 + 0.009\xi_t$	0.029
M_1	3	x_t^3	=	$(1-0.564B)(1-0.278B^{12})\eta_t^3 + 0.034(1-0.202B)\xi_t$	0.037
M_1	4	x_t^4	=	$(1-0.999B)(1-0.563B^{12})\eta_t^4 + 0.025(1-0.167B)\xi_t$	0.009
M_1	5	x_t^5	=	$(1-0.477B)(1-0.774B^{12})\eta_t^5 + 0.006\xi_t$	0.029

Table 2.4: p-values of $\tilde{R}_k^*$ and $\tilde{Q}_k$ for M_0 and M_1.

model	k (lag)	$\tilde{R}_k^*$	$\tilde{Q}_k$	k (lag)	$\tilde{R}_k^*$	$\tilde{Q}_k$
M_0	3	0.9977	0.9098	15	0.9909	0.7791
M_0	6	0.9453	0.6128	18	0.9969	0.8892
M_0	9	0.9091	0.4084	21	0.9969	0.8998
M_0	12	0.9886	0.7445	24	0.9951	0.8635
M_1	3	0.9821	0.5945	15	0.9799	0.5866
M_1	6	0.8422	0.2898	18	0.9886	0.7187
M_1	9	0.7800	0.1543	21	0.9855	0.7212
M_1	12	0.9584	0.4623	24	0.9832	0.7001

Capítulo 3

Departure from normality of increasing-dimension martingales*

3.1. Introduction

Many versions of the Central Limit Theorems and convergence rate estimates have been proved under different conditions. The case here considered is that of a triangular array X_{ni} of vector martingale differences when the dimension k depends on the length n of the martingale. We are interested, in particular, in the case that $k(n) \to \infty$ because otherwise, all X_{ni} could be considered as vectors of dimension $\max\{k(n) : n \in \mathbb{N}\}$. Even in the case that k diverges, the behaviour of X_{ni} can in principle be analyzed by considering them as infinite sequences completed with zeros and using Banach Space techniques. Nevertheless, for some applications, to establish the convergence to an infinite random sequence is not so useful as to measure

* *Journal of Multivariate Analysis*, 100 (2009), 1304–1315.

how much the $k(n)-$variate distribution differs from a $k(n)-$variate gaussian. Therefore, we focus on calculating bounds to the distance from the distribution of $Z_n = n^{-1/2} \sum_i X_{ni}$ to the $k(n)-$variate normal distribution measured with a certain metric.

With respect to which metric to use, the most commonly used one in results of this kind, beginning with the classical Berry-Esséen theorem, is probably the uniform metric (for example, [9] and [4]). Unfortunately, this metric is not so convenient for inference. As we will show, bounds on the Prokhorov allow us to prove some results we discuss in the applications (section 3.4). On the other hand, the Kantorovich metric provides an upper bound on the Prokhorov metric and behaves well with respect to Lipschitz transformations of the variables. For these reasons, we take as the starting point of our work the results stated by Rachev and Rüschendorf in [18] for martingales in Banach Spaces. However, we cannot use directly their theorem, but a generalized version we prove in section 3.2.

Our results are not just another theoretical turn of screw. In fact, they have been worked out to fill a theoretical gap in the diagnostic of time series models. There is a number of tests for residual autocorrelation, beginning with Box and Pierce ([6], [13]) and Ljung and Box ([12]) for the univariate case. The multivariate case was analyzed by Hosking in [10]. Ahn generalized the multivariate test for the constrained autoregressive case in [1]. More recently, it has been proved in [7] that the test can be applied to Vector Error Correction models. A variation of the test is proposed in [15]. A common feature to all these papers is the vagueness with which the asymptotic distribution property is stated (with the exception of the $\hat{D}_m$ statistic of [15], which has a different form and distribution). Generally, it is claimed that the distribution of the statistic, say Q_k, where k is the greatest autocorrelation order, can be approximated by a chi-square of $d(k)$ degrees of

freedom for large k and T, where T is the number of observations. This is argued by proving that Q_k is equal to the sum of squares of an average of martingale differences plus terms that vanish when $T \to \infty$. Then, the CLT is applied to that average, say E_T, but there is the difficulty that the covariance matrix of E_T is only approximately idempotent for k large. Hence, any convergence result has to consider both the limits $T \to \infty$ and $k \to \infty$.

The first result of this kind that provides a precise convergence result seems to be [2], but it pays the price of taking a sequential limit, first in T, and then in k. This kind of asymptotic property is not the most adequate for applications because it is not realistic. In real life, T is usually given and k chosen by the analyst, so the desired result is one that provides convergence when $k, T \to \infty$ and some joint condition is satisfied by k and T. In which sense should this convergence be established? Clearly, a good convergence should ensure that the error due to the use of the theoretical rejection region instead of the true one, converges to zero. If we choose $k = k(T)$ satisfying the joint convergence condition, $F_T(x)$ is the distribution function of the true statistic Q_k and $G_T(x)$ the one of a chi-square with the theoretical degrees of freedom corresponding to $k(T)$, then we want that for any $p \in (0, 1)$,

$$\lim_{T \to \infty} F_T(G_T^{-1}(p)) = p.$$

In subsection 3.4.1 we prove that the relation above holds under some assumptions.

The second application is presented in subsection 3.4.2 and it is related to the inference of autoregressive models when the true model is an $AR(\infty)$. If we fit an autoregressive model to a time series of length T that is generated by an $AR(\infty)$ process, then the order k of the model and T have to satisfy some joint conditions for the estimates to have good properties. These conditions were analyzed in [5] for the univariate case and in [11] for the multivariate case. In these articles, the asymptotic normality of the

estimates was established when $k^3/T \to 0$ and $T^{1/2} \sum_{j=k}^{\infty} \|\Phi_j\| \to 0$, where Φ_j are the autoregressive coefficients of the true model. Unfortunately, the asymptotic normality is not established for the vector of estiamtes but for a linear combination of its components. If $\hat{\phi}(k)$ is a vector containing the estimate coefficients and $l(k)$ is a sequence of constant vectors satisfying certain conditions, then $l(k)'\hat{\phi}(k)$ is asymptotically normal. Instead of this, we will establish the asymptotic normality of $\hat{\phi}$ by proving that the distance from the distribution of $\hat{\phi}$ to a certain $k-$variate gaussian converges to zero under some assumptions.

3.2. CLT rates for martingales in Banach spaces

First we need a generalization of Theorem 3.6 in [18]. This generalization consists of relaxing an assumption that can usually be checked only when the conditional second order moments of the ith martingale difference with respect to the $(i-1)$th field are almost surely constant. This condition is too strong for our applications. Hence, we will only impose that the moments with respect to the $(i-\nu)$th field are constant for a certain $\nu \geq 0$. Besides this generalization, we need to state the main proposition in such a form that all the constants appearing in the bound are absolute. This will allow us to apply in section 3.3 the result to different spaces for each n. No substantial modifications of the proof in [18] are needed for this.

In order to make our results easier to relate to the ones in [18], we adhere as much as possible to their notation, both in this section and in the next one. In these two sections, we will use the first upper case roman letters with or without numeral subscripts $(A, B, C, A_1, \ldots)$ to denote absolute constants. We use also capital roman letters with subscripts that indicate dependence with respect to variables or parameters, such as C_θ or L_r. The last capital roman letters $U, V, \ldots, Z$ are reserved for random variables.

For two random variables X and Y defined on a probability space $(\Omega, \mathcal{F}, P)$ taking values in a separable Banach Space $(\mathcal{X}, \|\cdot\|)$, let us denote by $\ell_1(X, Y)$ the Kantorovich metric,

$$\ell_1(X, Y) = \sup\{|E(f(X) - f(Y))| : f \in \mathcal{L}_1^B\}$$

where $\mathcal{L}_1^B$ is the set of the bounded real functions f defined on $\mathcal{X}$ such that $|f(x) - f(y)| \leq \|x - y\|$. For $r > 0$, we define the smoothing distance,

$$\ell_r(X, Y) = \sup_{h>0} h^{r-1} \ell_1(X + h\theta, Y + h\theta)$$

where θ is a symmetric $\alpha-$stable random variable independent from X and Y. We also need the total variation metric $\sigma(X, Y) = \sup\{|E(f(X) - f(Y))| : f \in C^0(\mathcal{X}; [0, 1])\}$, where $C^0(\mathcal{X}; [0, 1])$ denotes the set of the continuous functions defined from $\mathcal{X}$ to $[0, 1]$. In a similar way to ℓ_r the following smoothing metric is defined,

$$\sigma_r(X, Y) = \sup_{h>0} h^r \sigma(X + h\theta, Y + h\theta).$$

Let X_i be a sequence of martingale differences and θ_i a sequence of independent variables distributed as θ, which has a symmetric $\alpha-$stable distribution. Let us define for $\nu \geq 0$,

$$\begin{array}{ll} X_{i,\nu} = \sum_{j=i-\nu}^{i} X_j & X_{i,-\nu} = \sum_{j=i}^{i+\nu} X_j \\ \tilde{X}_{i,\nu} = \sum_{j=i-\nu}^{i-1} X_j + W_i & \tilde{X}_{i,-\nu} = W_i + \sum_{j=i+1}^{i+\nu} X_j \\ \hat{X}_{i,\nu} = \sum_{j=i-\nu}^{i-1} X_j + \theta_i & \hat{X}_{i,-\nu} = \theta_i + \sum_{j=i+1}^{i+\nu} X_j \end{array}$$

where W_i is a random variable with the same distribution as X_i and independent from $\{X_j : j \neq i\}$. In order to establish our results, we will need

the following constants,

$$\begin{aligned}
\ell_r &= \sup_i \ell_r(X_i, \theta_i) \\
\tau_{r,\nu} &= \sup_i E\ell_r(P_{X_{i,\nu}|\mathcal{F}_{i-\nu}}, P_{\hat{X}_{i,\nu}|\mathcal{F}_{i-\nu}}) \\
\tilde{\tau}_{r,\nu} &= \sup_i E\ell_r(P_{X_{i,\nu}|\mathcal{F}_{i-\nu-1}}, P_{\tilde{X}_{i,\nu}|\mathcal{F}_{i-\nu-1}}) \\
\hat{\tau}_{r,\nu} &= \sup_i E\ell_r(P_{X_{i,-\nu}|\hat{\mathcal{G}}_{i+\nu+1}}, P_{\hat{X}_{i,-\nu}|\hat{\mathcal{G}}_{i+\nu+1}}) \\
\sigma_r &= \sup_i \sigma_r(X_i, \theta_i) \\
t_{r,\nu} &= \max\{\ell_1, \sigma_1, \sigma_r^{(1/(r-2))}, \hat{\tau}_{r,\nu}^{(1/(r-2))}, \tilde{\tau}_{1,\nu}\} \\
\tilde{\ell}_{r,\nu} &= \max\{\ell_r, \tau_{r,\nu}\}
\end{aligned}$$

where $P_{X|\mathcal{F}}$ is the conditional distribution of X with respect to the $\sigma-$field $\mathcal{F}$ and $\mathcal{F}_i = \sigma(X_j : j \leq i), \hat{\mathcal{G}}_i = \sigma(X_j : j \geq i)$. For $Z_n = n^{-1/\alpha}\sum_{i=1}^n X_i$, we can state,

Proposition 3.1. *If $E\|\theta\| < +\infty$, then there exists a constant C_θ such that,*

$$\ell_1(Z_n, \theta) \leq C_\theta(n^{1-r/\alpha}\tilde{\ell}_{r,\nu} + n^{-1/\alpha}t_{r,\nu}). \tag{3.1}$$

Moreover, there exist M, N such that C_θ can be chosen satisfying $C_\theta \leq M + NE\|\theta\|$.

If $\nu = 0$, we obtain theorem 3.6 from [18] as a particular case. Before going on to the proof of proposition 3.1, we present a modified version of lemma 3.3 in [18],

Lemma 3.1. *Let $(X_i, \mathcal{F}_i)$ be a stochastic sequence and $(\mathcal{G}_i)$ a decreasing sequence of sub $\sigma-$fields such that Y_j are $\mathcal{G}_i$ measurable for $j \leq i$. Then, for $\nu \geq 0$,*

$$\ell_r(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i) \leq \sum_{i=1}^n E\ell_r(P_{U|\mathcal{F}_{i-\nu-1}\vee\mathcal{G}_{i+\nu+1}}, P_{V|\mathcal{F}_{i-\nu-1}\vee\mathcal{G}_{i+\nu+1}})$$

where $U = \sum_{j=i-\nu}^{i} X_i + \sum_{j=i+1}^{i+\nu} Y_i$ and $V = \sum_{j=i-\nu}^{i-1} X_i + \sum_{j=i}^{i+\nu} Y_i$.

The proof of our lemma 3.1 is essentially the same as lemma 3.3 in [18], whereas the proof of theorem 3.6 in [18] has to be modified only in three points. We summarize the modifications in lemma 3.2. Let us define,

$$\gamma_1 = \ell_1\Big(n^{-1/\alpha}\sum_{i=1}^{n} X_i + \varepsilon\theta, n^{-1/\alpha}\Big[\sum_{i=1}^{n-1} X_i + W_n\Big] + \varepsilon\theta\Big) \tag{3.2}$$

$$\gamma_2 = \sum_{j=1}^{m} \ell_1\Big(n^{-1/\alpha}\Big[\sum_{l=1}^{j}\theta_l + \sum_{l=j+1}^{n} X_l\Big] + \varepsilon\theta, n^{-1/\alpha}\Big[\sum_{l=1}^{j}\theta_l + W_{j+1} + \sum_{l=j+2}^{n} X_l\Big] + \varepsilon\theta\Big) \tag{3.3}$$

$$\gamma_3 = \ell_1\Big(n^{-1/\alpha}\Big[\sum_{i=1}^{m+1}\theta_i + \sum_{i=m+2}^{n} X_i\Big] + \varepsilon\theta, n^{-1/\alpha}\sum_{i=1}^{n}\theta_i + \varepsilon\theta\Big) \tag{3.4}$$

Lemma 3.2. *There exists a constant A such that,*

$$\gamma_1 \leq n^{-1/\alpha}\tilde{\tau}_{1,\nu} \tag{3.5}$$

$$\gamma_2 \leq Aa^{-r+1+\alpha}n^{-1/\alpha}\hat{\tau}_{r,\nu}^{1/(r-\alpha)} \tag{3.6}$$

$$\gamma_3 \leq An^{1-r/\alpha}\tau_{r,\nu} \tag{3.7}$$

where $a = \varepsilon n^{1/\alpha}/\max\{\sigma_1, \sigma_r^{1/(r-\alpha)}, \ell_r^{1/(r-\alpha)}, \hat{\tau}_{r,\nu}^{1/(r-\alpha)}\}$.

Proof. Using the dependence metric defined for a metric μ as,

$$\mu(X, Y|\mathcal{F}) = \sup_{V\in\mathcal{F}} \mu(X+V, Y+V)$$

where $V \in \mathcal{F}$ means that V is $\mathcal{F}-$measurable, we get,

$$\begin{aligned}
\gamma_1 &\leq \ell_1(n^{-1/\alpha}X_{n,\nu} + \varepsilon\theta, n^{-1/\alpha}\tilde{X}_{n,\nu} + \varepsilon\theta|\mathcal{F}_{n-\nu-1}) \leq \\
&\leq \ell_1(n^{-1/\alpha}X_{n,\nu}, n^{-1/\alpha}\tilde{X}_{n,\nu}|\mathcal{F}_{n-\nu-1}) \leq \\
&\leq n^{-1/\alpha}E\ell_1(P_{X_{n,\nu}|\mathcal{F}_{n-\nu-1}}, P_{\tilde{X}_{n,\nu}|\mathcal{F}_{n-\nu-1}}) \leq n^{-1/\alpha}\tilde{\tau}_{1,\nu}
\end{aligned}$$

where the first inequality is due to the definition of the dependence metric, the second is due to the regularity of ℓ_1 and the third to its homogeneity and to the property $\ell_r(X, Y|\mathcal{F}) \leq E\ell_r(P_{X|\mathcal{F}}, P_{Y|\mathcal{F}})$ (lemma 3.2 in [18]).

Using the $\alpha-$stability of θ and the inequality $\ell_r(X,Y) \geq h^{r-1}\ell_1(X + h\theta, Y + h\theta)$,

$$\begin{aligned}
\gamma_2 &\leq \sum_{j=1}^{m} \Big(\frac{j}{n} + \varepsilon^\alpha\Big)^{(1-r)/\alpha} \ell_r\big(n^{-1/\alpha} \sum_{l=j+1}^{n} X_l, n^{-1/\alpha}[\tilde{X}_{j+1} + \sum_{l=j+2}^{n} X_l]\big) \leq \\
&\leq \sum_{j=1}^{m} \frac{(j + n\varepsilon^\alpha)^{(1-r)/\alpha}}{n^{(1-r)/\alpha}} n^{-r/\alpha} \ell_r(X_{j+1,-\nu}, \tilde{X}_{j+1,-\nu} | \hat{\mathcal{G}}_{j+2+\nu}) \leq \\
&\leq \sum_{j=1}^{m} (j + na^\alpha \hat{\tau}_{r,\nu}^{\alpha/(r-\alpha)} n^{-1})^{(1-r)/\alpha} n^{-1/\alpha} \hat{\tau}_{r,\nu} \leq \\
&\leq An^{-1/\alpha} \frac{1}{a^{r-1-\alpha}} \tilde{\tau}_{r,\nu}^{1/(r-\alpha)}
\end{aligned}$$

where the last inequality is due to the following approximate identity (see [3], p. 379),

$$\sum_{k=m+1}^{\infty} \frac{1}{k^s} \approx \frac{1}{(s-1)m^{s-1}}$$

Finally,

$$\begin{aligned}
\gamma_3 &\leq \ell_1\Big(\Big(\frac{m+1}{n}\Big)^{1/\alpha}\theta + n^{-1/\alpha} \sum_{j=m+1}^{n} X_j, \Big(\frac{m+1}{n}\Big)^{1/\alpha}\theta + n^{-1/\alpha} \sum_{j=m+1}^{n} \theta_j\Big) \leq \\
&\leq \Big(\frac{m+1}{n}\Big)^{(1-r)/\alpha} n^{-r/\alpha} \sum_{j=m+1}^{n} E\ell_r(P_{X_{i,\nu}|\mathcal{F}_{i-\nu-1}\vee\mathcal{G}_{i+\nu+1}}, P_{\hat{X}_{i,\nu}|\mathcal{F}_{i-\nu-1}\vee\mathcal{G}_{i+\nu+1}})
\end{aligned}$$

where we have used lemma 3.1 with $\mathcal{G}_i = \sigma(\theta_j : j \geq i)$. Since $\{\theta_i\}$ are independent among them and from $\{X_i\}$, then $P_{X_{i,\nu}|\mathcal{F}_{i-\nu-1}\vee\mathcal{G}_{i+\nu+1}} = P_{X_{i,\nu}|\mathcal{F}_{i-\nu-1}}$ and $P_{\tilde{X}_{i,\nu}|\mathcal{F}_{i-\nu-1}\vee\mathcal{G}_{i+\nu+1}} = P_{\tilde{X}_{i,\nu}|\mathcal{F}_{i-\nu-1}}$. Hence,

$$\gamma_3 \leq An^{1-r/\alpha}\tau_{r,\nu}.$$

□

The proof of proposition 3.1 is the same as that of theorem 3.6 in [18] except that we use the inequalities of lemma 3.2 at the appropriate points, with $\gamma_1 = \Delta_3$, $\gamma_2 = \Delta_6$ and $\gamma_3 = \Delta_7$. The fact that the constant C_θ satisfies $C_\theta \leq M + NE\|\theta\|$ is due to how the constant C is chosen in [18].

3.3. Increasing-dimension martingales

We will use the result of the previous section to prove that the Kantorovich distance from $n^{-1/2}$ times a sum of n vector martingale differences to a gaussian distribution converges to zero when the dimension $k(n)$ of the vectors grows with n in a certain way. This means that we will focus on the gaussian case $\alpha = 2$, $r = 3$. The fact that the constants N and M in proposition 3.1 are absolute is critical because we need to apply the inequality (3.1) for spaces that have different dimension for each n. This is highlighted by an additional subscript n in some places.

Let us consider the real random variables $\{Y_i^j\}_{i,j\in\mathbb{N}}$ defined in a probability space $(\Omega, \mathcal{F}, P)$ and $n \mapsto k(n) \in \mathbb{N}$ a nondecreasing function. We define $X_{ni} = (Y_i^1, \ldots, Y_i^{k(n)})$ and $Z_n = n^{-1/2} \sum_{i=1}^n X_{ni}$. We denote by θ_n and θ_{ni}, random variables defined in the same probability space and distributed as a $k(n)-$dimensional gaussians with zero mean and unit covariance matrix. In order to measure the difference between the distributions of Z_n and θ_n, we consider them as elements of the Banach space $\mathbb{R}^{k(n)}$ endowed with the the norm $\|(x_1, \ldots, x_{k(n)})\|_p = (\sum_{j=1}^{k(n)} |x_i|^p)^{1/p}$.

Assumption 3.1. *For any n, the sequence $\{X_{ni}\}_i \in \mathbb{N}$ is a martingale difference sequence w.r.t the sequence of $\sigma-$fields $\mathcal{F}_i$.*

Assumption 3.2. *The following holds,*

(i) $EX_{ni}X'_{ni} = I_{k(n)}$.

(ii) There exists a function $\nu(n)$ such that,

(a) $\mathrm{Cov}[X_{ni}|\mathcal{F}_{i-\nu(n)-1}] = I_{k(n)}$, *a.s.*

(b) $\mathrm{Cov}[X_{ni}|X_{nj} : j \geq i + \nu(n) + 1] = I_{k(n)}$, *a.s.*

Where $I_{k(n)}$ is the unit matrix of dimension $k(n)$. We need the following lemma before before stating the main result of this section,

Lemma 3.3. *There exist constants B_j, for $j = 1, 3$ such that if $\varphi(x)$ is the density function of a $m-$variate gaussian distribution with zero mean and unit covariance matrix, then,*

$$\sup_{x\in\mathbb{R}^m}\sup_{\|z\|\leq 1}|\varphi^{(j)}(x)(z)| \leq B_j(2\pi)^{-m/2} \tag{3.8}$$

where $\varphi^{(j)}(x)(z)$ is the j-th derivative of φ as a j-linear form, evaluated at $(z,\ldots,z)$. Besides that, the following inequality holds,

$$\int_{\mathbb{R}^m}\sup_{\|z\|\leq 1}|\varphi^{(3)}(x)(z)|dx \leq m^{3/2}+3m^{1/2}. \tag{3.9}$$

Proof. The partial derivatives of φ are,

$$\begin{aligned}
\frac{\partial\varphi}{\partial x_i} &= (2\pi)^{-m/2}\exp\Big(-\frac{\|x\|^2}{2}\Big)x_i \\
\frac{\partial^3\varphi}{\partial x_i\partial x_j\partial x_k} &= (2\pi)^{-m/2}\exp\Big(-\frac{\|x\|^2}{2}\Big)\{-x_ix_jx_k+x_j\delta_{ik}+x_k\delta_{ij}+x_i\delta_{jk}\}
\end{aligned}$$

where δ_{uv} is the Kronecker delta.

Then, if $\|z\|\leq 1$,

$$\begin{aligned}
|\varphi'(x)(z)| &= (2\pi)^{-m/2}\exp\Big(-\frac{\|x\|^2}{2}\Big)|(x,z)| \leq (2\pi)^{-m/2}\exp\Big(-\frac{\|x\|^2}{2}\Big)\|x\| \\
|\varphi^{(3)}(x)(z)| &= (2\pi)^{-m/2}\exp\Big(-\frac{\|x\|^2}{2}\Big)\big|-(x,z)^3+3(x,z)\big| \leq \\
&\leq (2\pi)^{-m/2}\exp\Big(-\frac{\|x\|^2}{2}\Big)\big\{\|x\|^3+3\|x\|\big\} \qquad (3.10)
\end{aligned}$$

Consequently,

$$\begin{aligned}
\sup_{x\in\mathbb{R}^m}\sup_{\|z\|\leq 1}|\varphi'(x)(z)| &\leq (2\pi)^{-m/2}S_1 \\
\sup_{x\in\mathbb{R}^m}\sup_{\|z\|\leq 1}|\varphi^{(3)}(x)(z)| &\leq (2\pi)^{-m/2}\left(S_3+3S_1\right)
\end{aligned}$$

where $S_j = \sup_{r>0} r^j\exp\{-r^2/2\}$.

On the other hand, (3.10) implies,

$$\int_{\mathbb{R}^m}\sup_{\|z\|\leq 1}|\varphi^{(3)}(x)(z)|dx \leq E\|\theta\|^3+3E\|\theta\|$$

We get (3.9) using that $E\|\theta\|^3 \leq \{E(\sum_i \theta_i^2)^3\}^{1/2}$ and $E\|\theta\| \leq \{E\sum_i \theta_i^2\}^{1/2}$ and this completes the proof of the lemma. □

Let us write $\mu_q = \sup_{i,j} E|Y_i^j|^q$. Then,

Proposition 3.2. *If assumptions 3.1 and 3.2 hold and $\mu_{3p} < +\infty$, then there exist constants $D_{1,Y}$ and $D_{2,Y}$ such that,*

$$\ell_1(Z_n, \theta_n) \leq \frac{D_{1,Y} k(n)^{3/2+4/p} + D_{2,Y} k(n)^{3/2+2/p} \nu(n)}{n^{1/2}} \tag{3.11}$$

Proof. Using that the constants M and N in proposition 3.1 are absolute, we can use inequality (3.1) for the case that the Banach space $\mathcal{X}$ varies with n. Therefore, we assume that $\mathcal{X}_n = (\mathbb{R}^{k(n)}, \|\cdot\|_p)$.

We will need bounds for some moments of X_{ni},

$$E\|X_{ni}\|_p \leq \left(k(n)\mu_p\right)^{1/p} \quad E\|X_{ni}\|_p^3 \leq \left(k(n)^3\mu_{3p}\right)^{1/p}.$$

If the conditions of proposition 3.1 hold with $\alpha = 2$ and $r = 3$, we get,

$$\ell_1(Z_n, \theta_n) \leq (M + NE\|\theta_n\|_p)\frac{\tilde{\ell}_{3,\nu} + t_{3,\nu}}{n^{1/2}}$$

In order to estimate these constants, we will use that for any integer r the metrics ℓ_r and σ_r satisfy the bounds,

$$\ell_r(X, Y) \leq G_{r,\theta}\zeta_r(X, Y) \tag{3.12}$$

$$\sigma_r(X, Y) \leq L_{r,\theta}\zeta_r(X, Y) \tag{3.13}$$

where ζ_r is the Zolotarev metric (defined in [17]) and $G_{r,\theta} = \int \sup_{\|z\|_p \leq 1} |\varphi^{(r)}(x)(z)|dx$, $L_{r,\theta}$ is a Lipschitz constant of $\varphi^{(r-1)}$ and φ is the density function of θ_n. For the inequality on σ_r, see proposition 4.4 in [14]; for ℓ_r, see [19].

From (3.9), the constant $G_{3,\theta}$ is bounded by $k(n)^{3/2} + 3k(n)^{1/2}$ and from (3.8), the constants $L_{1,\theta}$ and $L_{1,\theta}$ are bounded respectively by $B_1(2\pi)^{-k(n)/2}$ and $B_3(2\pi)^{-k(n)/2}$.

In turn, $\zeta_r(X,Y)$ is bounded by $2s+r$ times the pseudomoment metric $\kappa_r(X,Y)$ when $s=\max\{j\in\mathbb{N}: j<r\}$ and the moments of order j of X and Y are equal for any integer $j\leq s$ (see [17] or [14]). The pseudomoment metric is defined as,

$$\kappa_r(X,Y)=\sup\{|E(f(X)-f(Y))|: f\in\mathcal{M}_r\}$$

where $\mathcal{M}_r=\{f\in\mathbb{R}^{\mathcal{X}}: |f(x)-f(y)|\leq \|\|x\|^{r-1}x-\|y\|^{r-1}y\|\}$.

We will use relations (3.12) and (3.13) together with the obvious $\kappa_r(X,Y)\leq E\|X\|^r+E\|Y\|^r$ to find bounds for all constants involved in proposition 3.1.

$$\begin{array}{cccccccc} \text{(a)} & \ell_1 & \text{(b)} & \ell_3 & \text{(c)} & \sigma_1 & \text{(d)} & \sigma_3 \\ \text{(e)} & \hat{\tau}_{3,\nu} & \text{(f)} & \tilde{\tau}_{1,\nu} & \text{(g)} & \tau_{3,\nu} & & \end{array}$$

(a) Let us begin with $\ell_1=\sup_{ni}\ell_1(X_{ni},\theta_{ni})$. Since $\ell_1(X,Y)\leq E\|X-Y\|$,

$$\ell_1(X_{ni},\theta_{ni})\leq E\|X_{ni}-\theta_{ni}\|_p\leq E\|X_{ni}\|_p+E\|\theta_{ni}\|_p\leq (k(n)\mu_p)^{1/p}+(k(n)\mu_p(\theta))^{1/p}$$

where $\mu_p(\theta)=E\|\theta\|_p$.

(b) We can use the inequalities involving ℓ_r, ζ_r and κ_r to get,

$$\begin{aligned} \ell_3(X_{ni},\theta_{ni})\leq G_{3,\theta}\zeta_3(X_{ni},\theta_{ni})\leq G_{3,\theta}7\kappa_3(X_{ni},\theta_{ni})\leq \\ \leq G_{3,\theta}7\left(E\|X_{ni}\|_p^3+E\|\theta_{ni}\|_p^3\right)\leq G_{3,\theta}7\left(\left(k(n)^3\mu_{3p}\right)^{1/p}+\left(k(n)^3\mu_{3p}(\theta)\right)^{1/p}\right)\leq \\ \leq 7\left(k(n)^{3/2}+3k(n)^{1/2}\right)k(n)^{3/p}\left(\mu_{3p}^{1/p}+\mu_{3p}(\theta)^{1/p}\right) \end{aligned} \tag{3.14}$$

The identity of the first and second-order moments of X_{ni} and θ_{ni} is a consequence of assumptions 3.1 and 3.2 (i).

(c) The metric σ_1 is bounded in the following way,

$$\begin{aligned} \sigma_1(X_{ni},\theta_{ni})\leq L_{1,\theta}\zeta_1(X_{ni},\theta_{ni})\leq B_1(2\pi)^{-k(n)/2}\zeta_1(X_{ni},\theta_{ni})\leq \\ \leq B_1(2\pi)^{-k(n)/2}\kappa_1(X_{ni},\theta_{ni})\leq B_1(2\pi)^{-k(n)/2}\{(k(n)\mu_p)^{1/p}+(k(n)\mu_p(\theta))^{1/p}\} \end{aligned}$$

(d) The case of σ_3 is similar to ℓ_3.

$$\sigma_3(X_{ni},\theta_{ni}) \le L_{3,\theta}7\big(k(n)^3\mu_{3p}\big)^{1/p} \le 7B_3(2\pi)^{-k(n)/2}\big(k(n)^3\mu_{3p}\big)^{1/p}$$

(e) For the bound on $\hat{\tau}_{3,\nu}$, we have to use again the bound of ℓ_3 in terms of ζ_3 and from assumption 3.2 (ii), also the bound of ζ_3 in terms of κ_3. Thus,

$$\ell_3(P_{X_{i,-\nu}|\hat{\mathcal{G}}_{i+\nu+1}}, P_{\hat{X}_{i,-\nu}|\hat{\mathcal{G}}_{i+\nu+1}}) \le$$

$$\le G_{3,\theta}7\left(2E\Big[\|\sum_{j=i+1}^{i+\nu} X_j\|_p^3|\mathcal{G}_{i+\nu(n)+1}\Big] + E\Big[\|X_i\|_p^3|\mathcal{G}_{i+\nu(n)+1}\Big] + E\Big[\|\theta_{ni}\|_p^3|\mathcal{G}_{i+\nu(n)+1}\Big]\right)$$

Taking expectation in both sides we obtain,

$$E\ell_r(P_{X_{i,-\nu}|\hat{\mathcal{G}}_{i+\nu+1}}, P_{\hat{X}_{i,-\nu}|\hat{\mathcal{G}}_{i+\nu+1}}) \le 14\big(k(n)^{3/2}+3k(n)^{1/2}\big)\nu(n)k(n)^{1/p}\big(\mu_{3p}^{1/p}+\mu_{3p}(\theta)^{1/p}\big)$$

(f) Let us now estimate $\tilde{\tau}_{1,\nu}$.

$$\tilde{\tau}_{1,\nu} = \sup_i E\ell_1(P_{X_{i,\nu}|\mathcal{F}_{i-\nu}}, P_{\tilde{X}_{i,\nu}|\mathcal{F}_{i-\nu}}) \le E\|X_{i,\nu}\|_p + E\|\tilde{X}_{i,\nu}\|_p \le$$

$$\le \sum_{j=i-\nu}^{i-1} E\|X_j\|_p + E\|X_i\|_p + E\|W_i\|_p \le 2\nu(n)\big(k(n)\mu_p\big)^{1/p}$$

(g) The case of $\tau_{3,\nu}$ is similar to $\hat{\tau}_{3,\nu}$.

Gathering all these bounds, we get,

$$\ell_1(Z_n,\theta_n) \le \Big(M + N(k(n)\mu_p)^{1/p}\Big)\frac{H_{1,Y}k(n)^{3/2+3/p} + H_{2,Y}\nu(n)k(n)^{3/2+1/p}}{n^{1/2}}$$

so (3.11) yields. □

3.4. Applications

In this section, we present two applications of proposition 3.2. The first is a proof of the asymptotic distribution of the statistic of Box and Pierce ([6]) Q_k when it is computed on the residuals of an estimated autoregressive and the second is an application to approximate confidence regions for the coefficients of an autoregressive model when the process is an $AR(\infty)$.

3.4.1. Residual autocorrelation tests

There are many results related to the asymptotic distribution of the autocorrelation tests. To avoid inessential complications, we will consider that the $r-$variate process x_t satisfies, rather than a general ARMA, an autoregressive model,

$$x_t = \sum_{j=1}^{p} \Phi_j x_{t-j} + \varepsilon_t$$

where $\varepsilon_t = (\varepsilon_{t1}, \ldots, \varepsilon_{tr})'$ has zero mean and covariance matrix Σ. We denote by $\hat{\varepsilon}_t$ the residuals of a model with coefficients $\hat{\Phi}_1, \ldots, \hat{\Phi}_p$ estimated by gaussian maximum likelihood using a series of length T. The residual autocovariances are $\hat{C}_j = T^{-1} \sum_t \hat{\varepsilon}_t \hat{\varepsilon}'_{t-j}$ (we also use the notation $\hat{\Sigma}$ for $\hat{C}_0$) and the statistic of the test is defined as,

$$Q_k = T \sum_{j=1}^{k} \mathrm{tr}(\hat{C}_0^{-1} \hat{C}_j \hat{C}_0^{-1} \hat{C}'_j).$$

Assumption 3.3. *The following holds,*

(i) The polynomial $|I_r - \Phi_1 z - \ldots - \Phi_p z^p|$ has its roots $z_1, \ldots, z_{rp}$ outside the unit circle.

(ii) ε_t is i.i.d. and $E|\varepsilon_{ti}|^{6\beta} < +\infty$, $\beta \geq 2$.

Let us call $\rho_* = (\min_j |z_j|)^{-1}$.

Proposition 3.3. *If assumption 3.3 holds and $T\rho^k, k^\alpha T^{-1} \to 0$ then, for any $\rho > \rho_*$ and $u \in (0,1)$,*

$$F_T(G_T^{-1}(u)) = u + O\Big(\Big[T\rho^k + \frac{k^{\alpha/2}}{T^{1/2}} + k^{-k/\tau}\Big]^{1/2}\Big) \tag{3.15}$$

where $\alpha = (\beta-2)/\beta + \max\{5+4/\beta, 3+8/\beta\}$ and $\tau > 0$, F_T is the distribution function of Q_k and G_T is the distribution function of a chi-square with $r^2(k-p)$ degrees of freedom.

Remark 3.1. *The fastest convergence rate is achieved in (3.15) when* $k = O(\log T)$*. Hence, the growth of* k *is not effectively limited by the condition* $k^{\alpha}T^{-1} \to 0$ *with respect to the convergence under the null. Notwithstanding this, a faster growth of* k *cound be convenient for increasing the power of the test under an alternative hypothesis.*

Before going on to the proof of proposition 3.3, we need an auxiliary lemma.

Lemma 3.4. *Let* $\{F_T\}_{T\in\mathbb{N}}$ *and* $\{G_T\}_{T\in\mathbb{N}}$ *be two classes of distribution functions in* $\mathbb{R}$ *and the elements of* $\{G_T\}_{T\in\mathbb{N}}$ *satisfy a Lipschitz condition with the common constant* L*. Then for any* $u \in (0,1)$*,*

$$|F_T(G_T^{-1}(u)) - u| \leq (1+L)\pi(F_T, G_T)$$

where $\pi(\cdot,\cdot)$ *denotes the Prohorov metric.*

Proof of lemma 3.4. Let δ be such that $\pi(F_T, G_T) < \delta$. Then,

$$\mathbb{P}_T(A) \leq \mathbb{Q}_T(A^{\delta}) + \delta \qquad \mathbb{Q}_T(A) \leq \mathbb{P}_T(A^{\delta}) + \delta \tag{3.16}$$

where $\mathbb{P}_T$ and $\mathbb{Q}_T$ are the probability measures of F_T and G_T respectively, A is any borel set of $\mathbb{R}$ and $A^{\gamma} := \{x : d(x,A) \leq \gamma\}$. If we put for $u \in (0,1)$, $A = (-\infty, {G_T}^{-1}(u)]$ in the first inequality of (3.16) and $A = (-\infty, {G_T}^{-1}(u) - \delta]$ in the second, we get,

$${F_T}({G_T}^{-1}(u)) \leq u + L\delta + \delta \qquad u - L\delta \leq {F_T}({G_T}^{-1}(u)) + \delta$$

Then, $|{F_T}({G_T}^{-1}(u)) - u| \leq \delta + L\delta$, so we conclude. □

Proof of proposition 3.3. In what follows, we denote by $O_m(\cdot)$ order in absolute mean –in general, with respect to T and k–, that is, for any random variable $X_{T,k}$, the identity $X_{T,k} = O_m(g(T,k))$ means that $g(T,k)^{-1}E\|X_{T,k}\|$ is bounded independently from k and T. In this section, the notation $\|\cdot\|$

means the euclidean norm, while the $\beta-$norm is denoted by $\|\cdot\|_\beta$. When the ℓ_1 metric is referred to a $\beta-$norm with $\beta \neq 2$, we write ℓ_1^β.

Assumption 3.3 implies that the process x_t has a Wold representation,

$$x_t = \sum_{l=0}^{\infty} \Psi_l \varepsilon_{t-l}.$$

such that the coefficients decay exponentially. Moreover, for any $\rho > \rho_*$, there exists a constant M_ρ such that $\|\psi_l\| \leq M_\rho \rho^l$.

Another consequence of assumption 3.3 is that the maximum likelihood estimates are consistent (see, for example, [8]) and satisfy a CLT. Thus, if we stack all the coefficients in $\phi = (\mathrm{vec}(\Phi_1)', \ldots, \mathrm{vec}(\Phi_p)')'$ and $\hat{\phi} = (\mathrm{vec}(\hat{\Phi}_1)', \ldots, \mathrm{vec}(\hat{\Phi}_p)')'$, then $\hat{\phi} - \phi = O_m(T^{-1/2})$.

By a Taylor expansion, we get,

$$\hat{c}_j = c_j + \frac{\partial c_j}{\partial \phi'}(\hat{\phi} - \phi) + \frac{1}{2} D^2 c_j(\xi_{\hat{\phi}})(\hat{\phi} - \phi) \otimes (\hat{\phi} - \phi) \tag{3.17}$$

where $D^2 c_j(\xi_{\hat{\phi}})$ is the matrix of the second derivatives of the elements of c_j with respect to ϕ arranged in the appropriate way, $\xi_{\hat{\phi}}$ is an element in the segment from ϕ to $\hat{\phi}$ and $c_j = \mathrm{vec}(C_j)$, $\hat{c}_j = \mathrm{vec}(\hat{C}_j)$. It can be proved that,

$$\frac{\partial c_j}{\partial \phi'} = -\hat{W}_j$$

where,

$$\hat{W}_j = (\hat{W}_{j1}, \ldots, \hat{W}_{jp})$$
$$\hat{W}_{ji} = (1/T) \textstyle\sum_t \{(I_r \otimes \varepsilon_t)(x'_{t-i-j} \otimes I_r) + (\varepsilon_{t-j} \otimes I_r)(x'_{t-i} \otimes I_r)\}$$

The quadratic part of the Taylor expansion (3.17) satisfies,

$$\sum_{j=1}^{k} \|D^2 c_j(\xi_{\hat{\phi}})(\hat{\phi} - \phi) \otimes (\hat{\phi} - \phi)\|^2 = O\Big(\frac{k}{T^2}\Big)$$

On the other hand, if we write $c = (c_1, \ldots, c_k)'$, $\hat{W} = [\hat{W}'_1, \ldots, \hat{W}'_k]'$ and $W = [W'_1, \ldots, W'_k]'$, with $W_j = [\Sigma\Psi_j \otimes I_r, \ldots, \Sigma\Psi_{j-p} \otimes I_r]$, then it can be

proved that $\|W\|$ is bounded uniformly in k (due to the exponential decay of Ψ_l) and $\hat{W} = W + O([k/T]^{1/2})$. Consequently, we can write,

$$\hat{c} = c - W(\hat{\phi} - \phi) + O_m(k^{1/2}T^{-1})$$

The statistic Q_k can be written as,

$$Q_k = T\hat{c}'(I_k \otimes \hat{\Sigma}^{-1} \otimes \hat{\Sigma}^{-1})\hat{c}$$

If we put $\Omega = (I_k \otimes \Sigma^{-1} \otimes \Sigma^{-1})$ and $\hat{\Omega} = (I_k \otimes \hat{\Sigma}^{-1} \otimes \hat{\Sigma}^{-1})$, using that $\hat{\Sigma} = \Sigma + O_m(T^{-1/2})$ we get,

$$Q_k = T(c - W(\hat{\phi} - \phi))'\Omega^{-1}(c - W(\hat{\phi} - \phi)) + O_m(k^{1/2}T^{-1/2}) \tag{3.18}$$

By a Taylor expansion of the log-likelihood l, it can be proved that,

$$\hat{\phi} - \phi = \frac{1}{T} I(\phi)^{-1} \frac{\partial l}{\partial \phi} + O_m(T^{-1}) \tag{3.19}$$

where $I(\phi) = (\Sigma \otimes \gamma_{u-v})_{u,v}$ is the information matrix, with $\gamma_l = Ex_t x'_{t-l}$. The derivative of the log-likelihood is given by,

$$\frac{\partial l}{\partial \phi_j} = -\sum_t (x_{t-j} \otimes I_r)\Sigma^{-1}\varepsilon_t = -\sum_t \sum_{u=0}^{\infty} (\Psi_u \varepsilon_{t-j-u} \otimes I_r)\Sigma^{-1}\varepsilon_t =$$

$$= -\sum_{u=0}^{k-j} (\Psi_u \Sigma \otimes I_r)(\Sigma^{-1} \otimes \Sigma^{-1}) \sum_t \mathrm{vec}(\varepsilon_{t-j-u}\varepsilon'_t) + O(\rho^{k+1-j})O_m(T^{1/2})$$

Then,

$$\frac{\partial l}{\partial \phi} = -TW'\Omega^{-1}c + O(\rho^k)O_m(T^{1/2}) \tag{3.20}$$

Let us put $E_T := T^{-1/2}\sum_{t=1}^{T} e_t$, where e_t is the increasing-dimension martingale difference $e_t = \Omega^{-1/2}(\mathrm{vec}(\varepsilon_t\varepsilon'_{t-1})', \ldots, \mathrm{vec}(\varepsilon_t\varepsilon'_{t-k})')'$. If we substitute $\partial l/\partial\phi$ in (3.19) by the right and side of (3.20) we get,

$$\hat{\phi} - \phi = -T^{-1/2} I(\phi)^{-1} W \Omega^{-1/2} E_T + O_m(T^{-1}) + O(\rho^k)O_m(T^{1/2}) \tag{3.21}$$

Now, using (3.18) and (3.21) we obtain,

$$Q_k = E_T'(I_{kr^2} - \Omega^{-1/2}WI(\phi)^{-1}W'\Omega^{-1/2})'(I_{kr^2} - \Omega^{-1/2}WI(\phi)^{-1}W'\Omega^{-1/2})E_T + \\ + O_m(T^{-1/2} + k^{1/2}T^{-1} + \rho^k T) \quad (3.22)$$

We can write more succinctly,

$$Q_k = \xi_T^2 + \omega_T \quad (3.23)$$

where $\xi_T^2 = E_T'A_T'A_TE_T$, $A_T = I_{kr^2} - \Omega^{-1/2}WI(\phi)^{-1}W'\Omega^{-1/2}$ and $\omega_T = O_m(T^{-1/2}+k^{1/2}T^{-1}+\rho^k T)$. The matrix $U_T = I_k - \Omega^{-1/2}W(W'\Omega^{-1}W)^{-1}W'\Omega^{-1/2}$ is idempotent and,

$$\|A_T - U_T\| \leq \|\Omega^{-1}\| \cdot \|W\|^2 \cdot \|I(\phi)^{-1} - (W'W)^{-1}\| \leq \\ \leq \|\Omega^{-1}\| \cdot \|W\|^2 \cdot \|I(\phi)^{-1}\| \cdot \|(W'W)^{-1}\| \cdot \|I(\phi) - W'W\| = O(\rho^k) \quad (3.24)$$

Since ε_t are i.i.d., assumption 3.2(ii) holds for E_T with, $n = T$ and $\nu = k$ and thus, by proposition 3.2, $\ell_1(E_T, \theta_T) = O\big([k^\varpi/T]^{1/2}\big)$, with $\varpi = \max\{5 + 4/\beta, 3 + 8/\beta\}$. Here we consider E_T, θ_T as elements of the Banach space $\mathbb{R}^k$ with the norm $\|\cdot\|_\beta$. From now onwards, k is a function of T, but in order to simplify the notation, we omit this dependency. By the triangular inequality,

$$\ell_1(A_TE_T, U_T\theta_T) \leq \ell_1(A_TE_T, U_TE_T) + \ell_1(U_TE_T, U_T\theta_T)$$

We can estimate both terms as,

$$\begin{aligned} \ell_1(A_TE_T, U_TE_T) &\leq \big\|A_T - U_T\big\| \cdot E\|E_T\| = O(\rho^k) \\ \ell_1(U_TE_T, U_T\theta_T) &\leq \big\|U_T\big\|\ell_1(E_T, \theta_T) \leq \\ &\leq \big\|U_T\big\|k^{(\beta-2)/(2\beta)}\ell_1^\beta(E_T, \theta_T) = k^{(\beta-2)/(2\beta)}O\Big(\Big[\frac{k^\varpi}{T}\Big]^{1/2}\Big) \end{aligned}$$

Where we have used that $\|z\|_2 \leq k^{(\beta-2)/(2\beta)}\|z\|_\beta$ and then, for $\eta_T = (\theta_T'U_T\theta_T)^{1/2}$,

$$\ell_1(\xi_T, \eta_T) = O\Big(\rho^k + \Big[\frac{k^\alpha}{T}\Big]^{1/2}\Big)$$

with $\alpha = (\beta - 2)/\beta + \varpi$. The variable η_T is the square root of a chi-square with $k - p$ degrees of freedom, that is, a chi distribution. Using that $\pi(X, Y)^2 \leq \ell_1(X, Y)$ (see [18]), we get,

$$\pi(\xi_T, \eta_T) = O\Big(\Big[\rho^k + \frac{k^{\alpha/2}}{T^{1/2}}\Big]^{1/2}\Big) \tag{3.25}$$

Now put $g_1(k, T) = [\rho^k + T^{-1/2}k^{\alpha/2}]^{1/2}$ and $g_2(T, k) = [T^{-1/2} + k^{1/2}T^{-1} + \rho^k T]^{1/2}$. From (3.23), we have $\ell_1(Q_k, \xi_T^2) = O(g_2(T, k)^2)$ and then, again from $\pi(X, Y)^2 \leq \ell_1(X, Y)$, we get,

$$\pi(Q_k, \xi_T^2) = O(g_2(T, k)). \tag{3.26}$$

On the other hand, since $\ell_1(\xi_T, \eta_T) = O(g_1(T, k)^2)$, for f defined as,

$$f(x) = \begin{cases} 1 & x \in (-\infty, 1] \\ 2 - x & x \in (1, 2] \\ 0 & x \in (2, +\infty) \end{cases}$$

we have that $|Ef(\xi_T) - Ef(\eta_T)| = O(g_1(T, k)^2)$. The properties of the distribution of η_T imply that $Ef(\eta_T) = O(k^{-k/\tau})$, and then $Ef(\xi_T) = O(g_1(T, k)^2 + k^{-k/\tau})$. Since $\xi_T \geq 0$, then $P[\xi \leq 1] \leq Ef(\xi_T)$. From, (3.26), we know that there exists a sequence $\epsilon_{T,k}$ such that $\lim_{k,T} \epsilon_{T,k} g(T, k)^{-1/2} = 0$ and for any Borel set $A \subset \mathbb{R}_+$,

$$P[Q_k \in A] \leq P[\xi_T^2 \in A^{\epsilon_{T,k}}] + \epsilon_{T,k}$$

which implies,

$$\begin{aligned} P[Q_k^{1/2} \in A] = P[Q_k \in A^2] &\leq P[\xi_T^2 \in (A^2)^{\epsilon_{T,k}}] + \epsilon_{T,k} \leq \\ &\leq P[\xi_T \in A^{\epsilon_{T,k}}] + \epsilon_{T,k} + P[\xi_T \leq 1] \leq \\ &\leq P[\xi_T \in A^{\epsilon_{T,k} + P[\xi_T \leq 1]}] + \epsilon_{T,k} + P[\xi_T \leq 1] \end{aligned} \tag{3.27}$$

where the second inequality is due to the relation $|\xi_T - \zeta| \leq |\xi_T^2 - \zeta^2|$, when $\xi_T \geq 1$ and $\zeta > 0$. Consequently,

$$\pi(Q_k^{1/2}, \xi_T) = O(g_2(T, k) + g_1(T, k)^2 + k^{-k/\tau}). \tag{3.28}$$

From (3.25) and (3.28),

$$\pi(Q_k^{1/2}, \eta_T) \leq \pi(Q_k^{1/2}, \xi_T) + \pi(\xi_T, \eta_T) = O(g_1 + g_2 + k^{-k/\tau})$$

A chi distribution with d degrees of freedom has as density function,

$$g_d(x) = \frac{2^{1-d/2}}{\Gamma(d/2)} x^{d-1} e^{-x^2/2}$$

It is easy to see that the class $\{g_d : d > 1\}$ is uniformly bounded by, say, M. This implies that the distribution functions satisfy a common Lipschitz condition with constant M. If F_T^0 is the distribution function of $Q_k^{1/2}$ and G_T^0 is the distribution function of a chi with $k - p$ degrees of freedom, that is, the distribution function of η_T, then by lemma 3.4,

$$|F_T^0({G_T^0}^{-1}(u)) - u| = O(g_1 + g_2 + k^{-k/\tau})$$

Since $F_T^0({G_T^0}^{-1}(u)) = F_T(G_T^{-1}(u))$, (3.15) is established. □

3.4.2. Confidence regions for approximate autoregressive models

We will consider stationary processes x_t such that,

Assumption 3.4. *x_t satisfies,*

$$x_t = \sum_{j=0}^{\infty} \Psi_j \varepsilon_{t-j}.$$

where ε_t is a sequence of i.i.d. random variables of zero mean and covariance matrix Σ. We also assume that $\lim_n n \sum_{j=n+1}^{\infty} \|\Psi_j\|^2 < +\infty$ and $|\sum_{j=0}^{\infty} \Psi_j z^j|$ has no roots in the closed unit circle $\{z : |z| \leq 1\}$.

Then, relationship (3.4) can be inverted as,

$$x_t = \sum_{j=1}^{\infty} \Phi_j x_{t-j} + \varepsilon_t$$

where $\sum_{j=0}^{\infty} \|\Phi_j\| < +\infty$ and $|\sum_{j=0}^{\infty} \Phi_j z^j|$ has also no roots in the closed unit circle. For a certain k, we compute the estimates $\hat{\Phi}_1, \ldots, \hat{\Phi}_k$ using the Yule-Walker equations. Let us stack them into a vector $\hat{\phi}(k) = (\mathrm{vec}(\hat{\Phi}_1)', \ldots, \mathrm{vec}(\hat{\Phi}_k)')'$ and the true values into $\phi(k) = (\mathrm{vec}(\Phi_1)', \ldots, \mathrm{vec}(\Phi_k)')'$. Let us also define the matrix $\Gamma_k = (\gamma_{u-v})_{u,v=0}^{k-1}$, with $\gamma_j = Ex_t x'_{t+j}$.

Proposition 3.4. *If 3.4 holds, ε_t satisfies assumption 3.3 (ii) and k is such that $T^{1/2} \sum_{j=k}^{\infty} \|\Phi_j\| \to 0$ and $k^{\alpha}/T \to 0$ where α is defined as in proposition 3.3, then,*

$$\ell_1((T-k)^{1/2}(\Gamma_k^{1/2} \otimes \Sigma^{-1/2})(\hat{\phi}(k) - \phi(k)), Z_k) \to 0 \tag{3.29}$$

where Z_k is a kr^2-dimensional normal-distributed r. v. with zero mean and unit covariance matrix.

Proof. In this proof we use the notation $\xi_T = o_m(u_T)$ to indicate that $\lim_T u_T^{-1} E\|\xi_T\| = 0$. It is very easy to modify the proof of theorem 2 in [11] to establish,

$$(T-k)^{1/2}(\hat{\Phi}(k) - \Phi(k)) = \frac{1}{(T-k)^{1/2}} \mathrm{vec}\Big[\sum_{t=k}^{T-1} \varepsilon_{t+1} X'_{t,k} \tilde{\Gamma}_k^{-1}\Big] + o_m(1) \tag{3.30}$$

where $X_{t,k} = (x'_t, x'_{t-1}, \ldots, x'_{t-k+1})'$. Let us denote by s_T the first term of the right hand side in (3.30). We can use the relationship,

$$X_{t,k} = \sum_{u=0}^{\infty} (I_k \otimes \Psi_u) \epsilon_{t-u,k}$$

with $\epsilon_{t,k} = (\varepsilon_t', \varepsilon_{t-1}', \ldots, \varepsilon_{t-k+1}')'$ to write,

$$s_T = \frac{1}{(T-k)^{1/2}}\mathrm{vec}\Big[\sum_{t=k}^{T-1}\sum_{u=0}^{\infty}\varepsilon_{t+1}\epsilon_{t-u,k}'(I_k\otimes\Psi_u')\Gamma_k^{-1}\Big] =$$

$$= \frac{1}{(T-k)^{1/2}}\mathrm{vec}\Big[\sum_{t=k}^{T-1}\sum_{u=0}^{j}\varepsilon_{t+1}\epsilon_{t-u,k}'(I_k\otimes\Psi_u')\Gamma_k^{-1}\Big] +$$

$$+\frac{1}{(T-k)^{1/2}}\mathrm{vec}\Big[\sum_{t=k}^{T-1}\sum_{u=j+1}^{\infty}\varepsilon_{t+1}\epsilon_{t-u,k}'(I_k\otimes\Psi_u')\Gamma_k^{-1}\Big] =$$

$$= \frac{1}{(T-k)^{1/2}}(\Gamma_k^{-1}\otimes I_r)\sum_{t=k}^{T-1}Ge_t + \eta_T \tag{3.31}$$

where $e_t = (\mathrm{vec}(\varepsilon_{t+1}\varepsilon_t'), \mathrm{vec}(\varepsilon_{t+1}\varepsilon_{t-1})', \ldots, \mathrm{vec}(\varepsilon_{t+1}\varepsilon_{t-k-j+1})')'$ and G is defined as the block matrix $(G_{uv})_{u=0,v=0}^{k-1,k+j-1}$ with $G_{uv} = \Psi_{v-u}\otimes I_r$ if $0 \le v-u \le j$ and $G_{uv} = 0$ otherwise. Let us see that $\eta_T = o_m(1)$,

$$E\big\|\mathrm{vec}\big[\sum_{t=k}^{T-1}\sum_{u=j+1}^{\infty}\varepsilon_{t+1}\epsilon_{t-u,k}'(I_k\otimes\Psi_u')\big]\big\|^2 \le$$

$$\le E\sum_{l=0}^{k}\sum_{t=k}^{T-1}\sum_{u=j+1}^{\infty}\big\|\mathrm{vec}\big[\varepsilon_{t+1}\varepsilon_{t-l-u}'\Psi_u'\big]\big\|^2 \le$$

$$\le \sum_{l=0}^{k}\sum_{t,s=k}^{T-1}\sum_{u,v=j+1}^{\infty}\mathrm{vec}(\Psi_u')'E\big[(I_r\otimes\varepsilon_{t+1}\varepsilon_{t-l-u}')'(I_r\otimes\varepsilon_{s+1}\varepsilon_{s-l-v}')\big]\mathrm{vec}(\Psi_v') =$$

$$= O(Tk\sum_{u=j+1}^{\infty}\|\psi_u\|^2)$$

If we set $j \propto k$, given that the norm of Γ_k^{-1} is uniformly bounded ([5], p. 491), we find that $E\|\eta_T\| \to 0$ and thus,

$$(T-k)^{1/2}(\hat{\Phi}(k) - \Phi(k)) = \frac{1}{(T-k)^{1/2}}(\Gamma_k^{-1}\otimes I_r)\sum_{t=k}^{T-1}Ge_t + o_m(1) \tag{3.32}$$

Consequently, we can write,

$$(T-k)^{1/2}(\Gamma_k^{1/2}\otimes\Sigma^{-1/2})(\hat{\Phi}(k)-\Phi(k)) = \frac{1}{(T-k)^{1/2}}(\Gamma_k^{-1/2}\otimes\Sigma^{-1/2})\sum_{t=k}^{T-1}Ge_t+o_m(1) \tag{3.33}$$

The process $(G\Omega G')^{-1/2}Ge_t$, where Ω is defined as in the proof of proposition 3.3 has unit covariance matrix and then proposition 3.2 can be applied to it. Therefore, we have to prove that $\|(\Gamma_k \otimes \Sigma)^{-1/2} - (G\Omega G')^{-1/2}\| \to 0$ and then, since $\|G\|$ and $E\|(T-k)^{-1/2}\sum_t e_t\|$ are bounded, we can substitute $(\Gamma_k^{-1/2} \otimes \Sigma^{-1/2})Ge_t$ by $u_t = (G\Omega G')^{-1/2}Ge_t$ in (3.33) getting,

$$(T-k)^{1/2}(\Gamma_k^{1/2} \otimes \Sigma^{-1/2})(\hat{\Phi}(k) - \Phi(k)) = \frac{1}{(T-k)^{1/2}} \sum_{t=k}^{T-1} u_t + o_m(1) \quad (3.34)$$

Let us first see that $\|(\Gamma_k \otimes \Sigma) - (G\Omega G')\| \to 0$. It can be proved that $\|(\Gamma_k \otimes \Sigma) - (G\Omega G')\| \leq \|\Sigma\| \cdot \|\Gamma_k - \hat{\Gamma}_k\|$ where $\hat{\Gamma}_k = (\hat{\gamma}_{u-v})_{u,v}$ and,

$$\hat{\gamma}_{u-v} = \sum_{w=\max(u,v)}^{\min(u,v)+j} \Psi_{w-u}\Sigma\Psi'_{w-v}$$

On the other hand, γ_{u-v} can be written as the sum above but with ∞ as the upper limit. If we put $j = 2k$, then at least the first k terms of the sum are common for γ_{u-v} and $\hat{\gamma}_{u-v}$. For a symmetric matrix A, $\|A\| \leq \|A\|_\infty = \max_u \sum_v |a_{uv}|$. Then,

$$\|\Gamma_k - \hat{\Gamma}_k\| \leq \max_{u=1,\ldots,k} \sum_{v=1}^{k} \|\gamma_{u-v} - \hat{\gamma}_{u-v}\| \leq 2\sum_{v=1}^{k} \|\gamma_v - \hat{\gamma}_v\| \leq$$
$$\leq 2k\Big\| \sum_{\mu=k+1}^{\infty} \Psi_\mu \Sigma \Psi'_{\mu+v}\Big\| \leq 2k\|\Sigma\| \sum_{\mu=k+1}^{\infty} \|\Psi_\mu\|^2$$

that by the assumptions converges to zero. Now, let us consider all matrices Γ_k and $\hat{\Gamma}_k$ as $\infty \times \infty$ matrices completed with zeros or equivalently as elements of the Banach space of the self-adjoint bounded linear operators of ℓ^2 (the Banach Space of the sequences of real numbers $x = (x_n)_n$ with the norm $\|x\| = \{\sum_n |x_n|^2\}^{1/2}$) into itself. Both sequences $\{\Gamma_k\}_k$ and $\{\hat{\Gamma}_k\}_k$ converge to the limit Γ_∞. Let us now consider the mapping $s : A \mapsto s(A) = (AA)^{-1}$. The differential of s at $\Gamma_\infty^{1/2}$ is the linear operator $ds(\Gamma_\infty^{1/2})(A) = \Gamma_\infty^{-1}(\Gamma_\infty^{1/2}A + A\Gamma_\infty^{1/2})\Gamma_\infty^{-1}$. Under the assumptions, Γ_∞ is nonsingular (see

[16]). Thus, ff $ds(\Gamma_\infty^{1/2})(A) = 0$, then $\Gamma_\infty^{1/2}A + A\Gamma_\infty^{1/2} = 0$. Consequently for any $x \in \ell^2$, $x'(\Gamma_\infty^{1/2}A + A\Gamma_\infty^{1/2})x = 2x'A\Gamma_\infty^{1/2}x = 0$, and then, $A\Gamma_\infty^{1/2} = 0$, but then $A = 0$. Consequently, we can apply the Inverse Function Theorem to s, so there is a differentiable inverse s^{-1} in a neighborhood of Γ_∞ getting $\|s^{-1}(\Gamma_k) - s^{-1}(\hat{\Gamma}_k)\| \leq \|ds^{-1}(\Gamma_\infty)\| \cdot \|\Gamma_k - \hat{\Gamma}_k\| + o(\|\hat{\Gamma}_k - \Gamma_\infty\|) + o(\|\Gamma_k - \Gamma_\infty\|) \to 0$.

On the other hand,

$$\|T^{-1/2}\sum_{t=1}^{T} u_t - (T-k)^{-1/2}\sum_{t=k}^{T-1} u_t\| \leq$$

$$\leq T^{-1/2}\{E\|u_T\| + \sum_{t=1}^{k-1} E\|u_t\|\} + \Big((T-k)^{-1/2} - T^{-1/2}\Big)\sum_{t=k}^{T-1} E\|u_t\| \leq$$

$$\leq C\frac{k(k+j)^{1/2}}{T^{1/2}} + D\frac{k(T-k-1)(k+j)^{1/2}}{T^{1/2}(T-k)^{1/2}(T^{1/2}+(T-k)^{1/2})} \tag{3.35}$$

Again, with $j = 2k$, the expression above converges to zero.

Then, we get from (3.34) and (3.35) that,

$$(T-k)^{1/2}(\hat{\Phi}(k) - \Phi(k))\sigma^{-2}\tilde{\Gamma}_k^{1/2} = \frac{1}{T^{1/2}}\sum_{t=1}^{T} u_t + o_m(1) \tag{3.36}$$

We can now apply proposition 3.2 to u_t and we get, as in the proof to proposition 3.3,

$$\ell_1(T^{-1/2}\sum_{t=1}^{T} u_t, Z_k) \to 0. \tag{3.37}$$

Finally (3.29) is a consequence of (3.36) and (3.37). □

We can apply proposition 3.4 to compute confidence regions for the parameter vectors $\Phi(k)$. Let $\varphi_{k,u}$ be such that if $\chi^2_{kr^2}$ is a chi-square of kr^2 degrees of freedom, then $P\{\chi^2_{kr^2} \leq \varphi_{k,u}\} = u$. We can build a confidence region $C_k = \{\Phi : (T-k)(\Phi - \hat{\Phi})'(\Gamma_k \otimes \Sigma^{-1})(\Phi - \hat{\Phi})' \leq \varphi_{k,u}\}$. If we proceed as in the proof of proposition 3.3, we can check that $\lim_T P\{\Phi \in C_k\} = u$.

Bibliography

[1] Ahn, S. K. (1988). Distribution for residual autocovariances in multivariate autoregressive models with structured parameterization. *Biometrika* **75,** 590–593.

[2] Arbués, I. (2008). An extended portmanteau test for VARMA models with mixing nonlinear constraints. *Journal of Time Series Analysis*

[3] Bender, C. M. and Orszag, S. A. (1999). *Advanced mathematical methods for scientists and engineers.* Springer, New York.

[4] Bentkus, V. M. and Götze, F. (1997). Uniform rates of convergence in the CLT for quadratic forms in multidimensional spaces. *Probability Theory and Related Fields* **109,** 367–416.

[5] Berk, K. N. (1974). Consistent autoregressive spectral estimates. *The Annals of Statistics* **2,** 489–502.

[6] Box, G. E. P. and Pierce, D. (1970). Distribution of autocorrelations in autoregressive moving average time series models. *Journal of the American Statistical Association* **65,** 1509–1536.

[7] Brügemann, R., Lütkepohl, H. and Saikkonen, P. (2006). Residual autocorrelation testing for vector error correction models. *Journal of Econometrics* **134,** 579–604.

[8] Hannan, E. J. and Deistler, M. (1988). *The statistical theory of linear systems.* Wiley, New York.

[9] Heyde, C, C.. and Brown, G. C. (1970). On the departure of normality of a certain class of martingales. *Annals of Mathematical Statistics* **41,** 2161–2165.

[10] Hosking, J. R. M. (1980). The multivariate portmanteau statistic. *Journal of the American Statistical Association* **75,** 602–608.

[11] Lewis, R. and Reinsel, G. C. (1985). Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis* **16,** 393–411.

[12] Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika* **65,** 297–303.

[13] McLeod, A. I. (1978). On the distribution of residual autocorrelations in Box-Jenkins models. *Journal of the Royal Statistical Society B* **40,** 296–302.

[14] Neininger, R. and Rüschendorf, L. (2004). A general limit for recursive algorithms and combinatorial structures. *Annals of Applied Probability* **14,** 378–418.

[15] Peña, D. and Rodríguez, J. (2002). A powerful portmanteau test of lack of fit for time series. *Journal of the American Statistical Association* **97,** 601–619.

[16] Hannan E. J. and Poskitt D. S. (1987). Unit canonical correlation between future and past. *The Annals of Statistics* **16,** 784-790.

[17] Rachev, S. T. (1991). *Probability metrics and the stability of stochastic models.* Wiley, New York.

[18] Rachev, S. T. and Rüschendorf, L. (1994). On the rate of convergence in the CLT with respect to the Kantorovich metric. In *Probability in Banach spaces.* (Hoffmann- Jorgensen, J., Kuelbs, J. and Marcus, B., eds.) **9,** 193–207. Birkhäuser, London.

[19] Zolotarev, V. M. (1983). Probability metrics. *Theory of Probability and Applications* **28,** 278–302.

Capítulo 4

Determining the MSE-optimal cross section to forecast*

4.1. Introduction

If we want to forecast a time series x_t^0 using time series models, we have to decide whether to use a univariate or a multivariate model, and in this latter case, which variables to include in the model. Once the composition of the vector time series has been decided, there are many tools to identify and estimate the model. Consequently, in this paper, we focus on the first two decisions. More precisely, if we have a certain set of time series, which subset (hereinafter, 'subset' will mean a certain subset of the whole set of time series available) is the most convenient to forecast x_t^0? An easy answer to this question is to use all the time series, but if the number of series is large, the number of parameters of the models is also large (usually growing faster

* *Journal of Econometrics* (aceptado).

than linearly). In that case, the parameters are computationally difficult to estimate and even if the computational difficulties are overcome, the estimates may have large variances that render the models useless. Methods based in factor models try to limit the proliferation of parameters while using all the information of the panel. See, for example, Stock and Watson (2002) or Forni, Hallin, Lippi and Reichlin (2005).

If we want to use a model that does not allow large cross sections, a usual approach to the problem is to use hypothesis tests. For example, a two-sided test is described in Diebold and Mariano (1995), to compare the predictive efficiency of two models. In Clark and McCracken (2001, 2007), one-side tests are presented, that allow us to decide between nested models, rejecting the null when the most parsimonious one does not encompass the other. Granger-Causality tests (see Granger, 1969) are designed to determine if some series included in a certain subset are indeed useful to produce forecasts. Giacomini and White (2006) proposed a test of conditional predictive ability. In Peña and Sánchez (2007), a method to compare univariate and multivariate forecasts was presented.

In this paper, we present a different approach. We select the subset, or cross section, using a selection criterion rather than a hypothesis test. A great variety of model selection criteria have been proposed, for example, the AIC by Akaike (1973 and 1974), Schwarz's (1978) SBC or the HQ criterion by Hannan and Quinn (1979). The case of misspecification has been analyzed, among others, by Nishii (1988) and Sin and White (1996).

Instead of the penalized log-likelihood, our criteria consist of the logarithm of the mean squared $h-$step prediction error of x_t^0 plus penalty terms that take into account, not the number of parameters, but the size of the cross section. Thus, it is not covered by most of the references cited above. The exception is section 6 of Sin and White (1996), but we impose less

stringent conditions on the penalty terms.

In section 4.2, we describe in detail the problem of the optimal cross section selection. We present in section 4.3 the class of criteria. Strong and weak consistency results are proved for a relatively general class of models in section 4.4 and in section 4.4.3 we show that the assumptions are satisfied in the case of VARMA models. In section 4.5 we consider the cases that the subset is selected among a random class and when we want to forecast more than one series.

In order to assess the performance of the method, we have used Monte Carlo simulations to compare the criteria to some hypothesis tests. Specifically, we have considered the test by Diebold and Mariano and the ENC-T and ENC-NEW test of Clark and McCracken and the conditional predictive ability test by Giacomini and White. The results of this experiment are discussed in section 4.6. Finally, section 4.7 reports the results of an empirical application.

4.2. The optimal cross section

Let $(\Omega, \mathcal{F}, p)$ be a probability space and $\{x_t^0\}_{t\in\mathbb{Z}}$ a discrete-time stochastic process. Suppose we want to forecast x_t^0 at horizon h. For this purpose, besides x_t^0 itself, we have at our disposal a set of processes $\{x_t^i\}$ with $i = 1, \ldots, N$.

We also assume that for any subset $I \subset \mathcal{S} = \{0, \ldots, N\}$, such that $0 \in I$, there is a forecast $x_{t+h|t}^{0,I}$ of x_{t+h}^0 computed with the information contained in the variables indexed by the elements of I up to t. In other words, $x_{t+h|t}^{0,I}$ is $\mathcal{F}_t(I)-$measurable, where $\mathcal{F}_t(I)$ is the $\sigma-$field generated by $\{x_s^I : s \leq t\}$, $x_s^I = (x_s^{i_1}, \ldots, x_s^{i_n})'$ and $I = \{i_1, \ldots, i_n\}$. In particular, $x_{t+h|t}^{0,I}$ is chosen as

the optimal predictor in the sense that it minimizes

$$\sigma_h^2(I) = \mathbf{E}[(x_{t+h}^0 - x_{t+h|t}^{0,I})^2] \tag{4.1}$$

among a certain class of predictors (later, it will be the class of the linear predictors). The generalization to other loss functions remains for future investigation. The expression (4.1) is finite if x_{t+h}^0 and $x_{t+h|t}^{0,I}$ have bounded second-order moments and it is independent from t if x_t^I is strictly stationary. In the case of linear predictors, the condition can be relaxed to weak stationarity.

It is possible that some choices of I are ruled out in advance. Consequently, the selection is restricted to a certain class of subsets $\mathcal{I} \subset P(\{0, \ldots, N\})$. Now, we can state our problem, that is, to minimize σ_h^2 in $\mathcal{I}$. Let us denote by $\mathcal{I}_0$ the class of minimizers. In general, for any $I \in \mathcal{I}_0$, $I \subset J$ implies $J \in \mathcal{I}_0$, so the solution will not be unique. Therefore, it is natural to choose, among the multiple solutions, those most parsimonious in some sense. Let $\delta(I)$ be an integer function of I, such that if $I \subset J$, then $\delta(I) \leq \delta(J)$ and if $I \subsetneq J$, then $\delta(I) < \delta(J)$. For the sake of generality, we allow different possibilities for δ, but in our experiments we use the cardinality of I.

Consequently, our aim is to consistently estimate a subset I_0 that minimizes δ in $\mathcal{I}_0$. We call $\mathcal{I}_{00}$ the set of such minimizers. In general, even $\mathcal{I}_{00}$ may have more that one element, but in some cases uniqueness can be proved.

4.3. Criteria

In real life, rather than the optimal predictor of x_{t+h}^0, we will have an approximation, $\hat{x}_{t+h|t}^{0,I}$, typically computed with an estimated model, say for $t = 1, \ldots, T$.

We can define,

$$\hat{\varepsilon}_{t,h}^{0,I} = x_{t+h}^{0} - \hat{x}_{t+h|t}^{0,I}, \tag{4.2}$$

$$\hat{\sigma}_h^2(I) = \frac{1}{T}\sum_{t=1}^{T-h}(\hat{\varepsilon}_{t,h}^{0,I})^2 \tag{4.3}$$

and the family of criteria

$$\mathrm{FC}(I) = \log \hat{\sigma}_h^2(I) + \delta(I)\frac{S_T}{T}, \tag{4.4}$$

where S_T is a nondecreasing function of T whose properties will be prescribed in the following sections.

With this criteria, we choose the set $\hat{I}_T$ as

$$\hat{I}_T = \arg\min_{I\in\mathcal{I}} \mathrm{FC}(I). \tag{4.5}$$

The necessity of restraining the choice of $\hat{I}_T$ in (4.5) to a certain class $\mathcal{I}$ is due to the fact that the growth of $\#P(\{1,\ldots,N\}) = 2^N$ makes, even for moderate values of N, unfeasible to try all subsets. On the other hand, the assumption that $\mathcal{I}$ is always fixed in advance is not realistic. In some cases, $\mathcal{I}$ will be determined using the data of the series and thus it will be random. Nevertheless, in order to introduce the main ideas of the consistency results, we will present in section 4.4 the case of deterministic $\mathcal{I}$ and in 4.5 we describe the changes necessary to deal with the random case.

We have excluded the possibility of using more than one model for each subset I. In that case, a natural extension would be to replace $\hat{\sigma}_h^2(I)$ by the minimum MSE across models. This variation remains for future research.

4.4. Consistency

In this section we will establish some conditions under which the estimate $\hat{I}_T$ described in the previous section is consistent. Given that the set of

optimal values, $\mathcal{I}_{00}$, may contain more than one element, we say that $\hat{I}_T$ is almost sure or strongly consistent if there exists with probability 1 some T_0 such that for any $T > T_0$, $\hat{I}_T \in \mathcal{I}_{00}$. Then, we write $\hat{I}_T \xrightarrow{a.s.} \mathcal{I}_{00}$. We say that $\hat{I}_T$ is consistent in probability if $P[\hat{I}_T \in \mathcal{I}_{00}] \to 1$ and we write $\hat{I}_T \xrightarrow{p} \mathcal{I}_{00}$.

4.4.1. Assumptions

First of all, we establish an assumption on the structure of $\mathcal{I}$.

Assumption 4.1. *The class $\mathcal{I}$ is closed with respect to union.*

This assumption is not unreasonable. If two sets, I and J contain relevant information to predict x^0_{t+h}, it is natural to try $I \cup J$, so that the predictions use both the information from I and J.

Assumption 4.2. *All x^I_t are weakly stationary and linearly regular*[1].

If assumption 4.2 holds, the Wold decomposition of x^I_t can be written as

$$x^I_t = \Psi^I(L)\varepsilon^I_{t-k} = \sum_{k=0}^{\infty} \Psi^I_k \varepsilon^I_{t-k}, \tag{4.6}$$

where ε^I_t are the linear innovations of x^I_t and L is the lag operator.

Assumption 4.3. *The following holds.*

(a) ε^I_t is ergodic white noise, with bounded fourth-order moments, $\mathbf{E}[\varepsilon^I_t | \mathcal{F}_{t-1}(I)] = 0$, $\mathbf{E}[\varepsilon^I_t \varepsilon^{I\prime}_t | \mathcal{F}_{t-1}(I)] = \Sigma^I$, with $\Sigma^I > 0$.

(b) The continuation of $\Psi^I(z)$ to the unit circle has no unit modulus roots and $\sum_k k^{1/2} \|\Psi_k\| < +\infty$.

[1] As defined in Hannan and Deistler (1988). This property is also known as "linearly, purely nondeterministic".

Under these assumptions, the Best Linear Predictor (BLP) of x^0_{t+h} using $\{x^I_s : 1 \leq s \leq t\}$ is the best predictor, in the sense of mean squared error. Also, the process x^I_t satisfies an infinite autoregressive representation $\Pi^I(L)x^I_t = \varepsilon^I_t$ with $\Pi^I(z) = \Psi^I(z)^{-1}$ (see chapter 4 in Hannan and Deistler, 1988; hereafter referred as HD). In practice, we have an estimated model determined by, say, $\hat{\Pi}^I(z)$ and $\hat{\Sigma}^I$. The prediction $\hat{x}^{0,I}_{t+h|t}$ is computed using the BLP corresponding to the estimated model. We need some assumptions about the asymptotic behavior of the estimated model.

Assumption 4.4. *Either of the two following conditions hold:*

(a) $\sum_k k \sum_{j=k}^{\infty} \|\hat{\Pi}^I_j\|, Q_T^{-1} k^{\alpha} \sum_{j=k}^{\infty} \|\hat{\Pi}^I_j - \Pi^I_j\| = O(1)$, *with* $\alpha > 0$ *and* $Q_T = [\log\log T/T]^{1/2}$. *For large* T, *with probability 1,* $\hat{\Sigma}^I \geq r\mathbb{I}$, *with* $r > 0$ *and* $\mathbb{I}$ *is the identity matrix.*

(b) $\sum_k k \sum_{j=k}^{\infty} \|\hat{\Pi}^I_j\| = O_p(1)$ *and* $\forall \epsilon > 0, \exists \Omega_\epsilon$ *such that* $P(\Omega_\epsilon) \geq 1 - \epsilon$ *and* $T^{1/2} \sum_{j=k}^{\infty} \mathbf{E}\big[\|\hat{\Pi}^I_j - \Pi^I_j\|^2 1_{\Omega_\epsilon}\big]^{1/2} = O(1)$, *where* 1_{Ω_ϵ} *is the indicator function of* Ω_ϵ. *Also* $P[\hat{\Sigma}^I \geq r\mathbb{I}] \to 1$.

We use $O(\cdot)$ and $o(\cdot)$ for almost sure order when applied to random variables and $O_p(\cdot)$ for order in probability.

NOTE: we do not impose any relationship between the data used to compute the estimates and the data used to forecast, besides that the convergence of the estimated predictors depends on T. Thus, if we have a time series at our disposal, we can use the whole series to estimate the model and to obtain the forecasting residuals of (4.3) or we can split the series into a length-T_e part to estimate the model and another one of length T to obtain the forecasting residuals. The assumptions hold as long as T_e and T are in an adequate relationship, e.g. $\lim_T T_e/T \in (0, +\infty)$. On the other hand, the models can be nested or nonnested and they can be identified by whatever method is preferred, as long as consistency is ensured.

4.4.2. Consistency properties

We establish first the following rate of convergence.

Proposition 4.1. *If assumptions 4.1, 4.2, 4.3 and 4.4(a) hold, then for any* $I, J \in \mathcal{I}_0$, $\hat{\sigma}_h^2(I) - \hat{\sigma}_h^2(J) = O(Q_T^2)$. *If assumptions 4.1, 4.2, 4.3 and 4.4(b)* [2], *then* $\hat{\sigma}_h^2(I) - \hat{\sigma}_h^2(J) = O_p(T^{-1})$.

We can state now the main result in this section,

Proposition 4.2. *The following holds.*

(i) If assumptions 4.1, 4.2, 4.3 and 4.4(a) hold and $S_T/T \to 0$, $S_T/\log\log T \to +\infty$, *then* $\hat{I}_T \xrightarrow{a.s.} \mathcal{I}_{00}$.

(ii) If assumptions 4.1, 4.2, 4.3 and 4.4(b) and $S_T/T \to 0$, $S_T \to +\infty$, *then* $\hat{I}_T \xrightarrow{p} \mathcal{I}_{00}$.

NOTE: if we extended our framework to allow for an infinite time series class $\mathcal{I}$, we could analyze the case that there is not a finite optimal cross section. In that case, the asymptotic optimality of the predictors could have more practical relevance than consistency. Something similar happens in the context of order determination of autoregressive models when the process is an AR(∞) (see Shibata, 1980).

With an additional assumption we can also prove uniqueness.

Proposition 4.3. *If assumptions 4.1–4.3 hold and in addition,* $\mathcal{I}$ *is closed with respect to intersection, then* $\mathcal{I}_{00}$ *has only one element. Furthermore, if* $\mathcal{I}_{00} = \{I\}$, *then* $\forall J \in \mathcal{I}_0, I \subseteq J$.

[2] If instead of 4.4(b), it holds $\sum_k k \sum_{j=k}^{\infty} \|\hat{\Pi}_j^I\|, Q_T^{-1} k^{\alpha} \sum_{j=k}^{\infty} \|\hat{\Pi}_j^I - \Pi_j^I\| = O_p(1)$, we get $O_p(Q_T^2)$.

4.4.3. The VARMA case

We can see that the consistency of $\hat{I}_T$ is easy to prove in the case of VARMA models under some usual conditions that guarantee the consistency of the estimates of the model coefficients.

Assumption 4.5. *The following holds.*

(a) Let Φ^I and Θ^I be matrix polynomials of degrees p^I and q^I with their roots outside the unit circle and such that x_t^I satisfies the VARMA model $\Phi^I(L)x_t^I = \Theta^I(L)\varepsilon_t^I$, where the covariance matrix of ε_t^I satisfies $\Sigma^I > r\mathbb{I}$ and $r > 0$.

(b) For $\mu, \nu, \tau \in I$, $\lim_{k\to\infty} \mathbf{E}[\varepsilon_{\mu,t}^I \varepsilon_{\nu,t}^I \varepsilon_{\tau,t}^I | \mathcal{F}_{t-k}(I)] = \mathbf{E}[\varepsilon_{\mu,t}^I \varepsilon_{\nu,t}^I \varepsilon_{\tau,t}^I]$, where $\varepsilon_{\mu,t}^I$ is the μth component of ε_t^I.

Then, $x_t^I = \Psi^I(L)\varepsilon_t^I$ with $\Psi^I = (\Phi^I)^{-1}\Theta^I$ and $\Pi^I(L)x_t^I = \varepsilon_t^I$, with $\Pi^I = (\Theta^I)^{-1}\Phi^I$. Let $\hat{\Phi}^I$ and $\hat{\Theta}^I$ be maximum likelihood estimates and $\hat{\Psi}^I = (\hat{\Phi}^I)^{-1}\hat{\Theta}^I$, $\hat{\Pi}^I = (\hat{\Theta}^I)^{-1}\hat{\Phi}^I$.

With assumption 4.5, we can apply theorems 4.2.1, 4.3.1 and 4.3.2 from [10] and we get that the maximum likelihood estimates $\hat{\Phi}^I$ and $\hat{\Theta}^I$ satisfy

$$T^{1/2}\mathrm{vec}(\hat{\Phi}^I - \Phi^I) \xrightarrow{d} N(0, \Xi_\Phi), \quad \hat{\Phi}^I = \Phi^I + O(Q_T), \tag{4.7}$$

$$T^{1/2}\mathrm{vec}(\Theta^I - \Theta^I) \xrightarrow{d} N(0, \Xi_\Theta), \quad \hat{\Theta}^I = \Theta^I + O(Q_T), \tag{4.8}$$

for certain matrices Ξ_Φ and Ξ_Θ. Then, for large T, $\hat{\Theta}^I$ has its roots bounded away from the unit circle and then, the identity $\Pi^I - \hat{\Pi}^I = (\Theta^I)^{-1}(\Phi^I - \hat{\Phi}^I) + (\Theta^I)^{-1}(\hat{\Theta}^I - \Theta^I)(\hat{\Theta}^I)^{-1}\hat{\Phi}^I$ entails that there are constants $\rho < 1, c > 0$ such that

$$\|\Pi_i\|, \|\Psi_i\| \leq c\rho^i \tag{4.9}$$

$$\|\Pi_i - \hat{\Pi}_i\| \leq c\Big(\sum_{j=1}^{p^I} \|\Phi_j^I - \hat{\Phi}_j^I\|\rho^{i-j} + \sum_{j=1}^{q^I} \|\Theta_j^I - \hat{\Theta}_j^I\|\rho^{i-j}\Big). \tag{4.10}$$

Since both parts of assumption 4.4 are straightforward consequences from (4.7)–(4.10), we can state the following proposition.

Proposition 4.4. *If assumptions 4.3a and 4.5 hold, then assumptions 4.2, 4.3b, 4.4a and 4.4b also hold.*

4.5. Generalizations

In this section, we consider some variants of the framework described in the previous sections.

In 4.5.1, we consider the case that instead of choosing the subset among a fixed $\mathcal{I}$, we have a random $\hat{\mathcal{I}}$. This generalization is necessary if there are so many series that a preliminary work is done in order to discard some choices before using FC. If the pre-selection is done using the data of the series, then the subset is in fact selected among a random class. We provide some natural assumptions under which the FC-selected $\hat{I}$ is still consistent.

In 4.5.2, we analyze the case that we are interested in forecasting several of the time series available with the same multivariate model.

4.5.1. Random $\mathcal{I}$

We will now choose $\hat{I}$ as $\arg\min_{I \in \mathcal{I}_T} \mathrm{FC}(I)$, where $\mathcal{I}_T$ is random and possibly depending on the time series, whence the subscript. In order to achieve consistency, it is necessary to impose some constraints in the behavior of $\mathcal{I}_T$. In particular, we have to avoid the case that there are optimal subsets, say I, such that $I \in \mathcal{I}_T$ and $I \notin \mathcal{I}_T$ infinitely many times. For the sake of brevity, we restrict the analysis to the strong convergence results, but it can be easily adapted to the weak case.

Let us define,

$$\overline{\mathcal{I}}_{\infty}^{p} = \{I \in P(\{0,\ldots,N\}) : P[I \in \limsup_T \mathcal{I}_T] = p\}, \tag{4.11}$$

$$\underline{\mathcal{I}}_{\infty}^{p} = \{I \in P(\{0,\ldots,N\}) : P[I \in \liminf_T \mathcal{I}_T] = p\}, \tag{4.12}$$

where

$$\limsup_T A_T = \bigcap_{T=1}^{\infty} \bigcup_{S=T}^{\infty} A_S, \tag{4.13}$$

$$\liminf_T A_T = \bigcup_{T=1}^{\infty} \bigcap_{S=T}^{\infty} A_S. \tag{4.14}$$

We can now express more precisely the condition on $\mathcal{I}_T$.

Assumption 4.6. *$\underline{\mathcal{I}}_{\infty}^{1} \neq \emptyset$ and for any $J \in \bigcup_{p>0} \overline{\mathcal{I}}_{\infty}^{p}$, $\sigma_h^2(I) \geq \sigma_{h,*}^2$, where*

$$\sigma_{h,*}^2 = \min_{I \in \underline{\mathcal{I}}_{\infty}^{1}} \sigma_h^2(I). \tag{4.15}$$

If the inequality holds as equality, then $\delta(J) > \min\{\delta(I) : I \in \underline{\mathcal{I}}_{\infty}^{1}\}$.

With this assumption, we can focus on the set $\underline{\mathcal{I}}_{\infty}^{1}$. Let us denote by $\mathcal{I}_{\infty,0}$ the minimizers of σ_h^2 in $\underline{\mathcal{I}}_{\infty}^{1}$ and by $\mathcal{I}_{\infty,00}$ the minimizers of δ in $\mathcal{I}_{\infty,0}$. Then, we can state the following proposition,

Proposition 4.5. *If 4.1, 4.2, 4.3, 4.4 and 4.6 hold, then $\hat{I}_T \xrightarrow{a.s.} \mathcal{I}_{\infty,00}$.*

With the generalization to the random $\mathcal{I}_T$, we can analyze some examples. We present one in which a scheme to build $\mathcal{I}_T$ from the data is combined with FC to provide a consistent estimate of the —in this case unique— element of $\mathcal{I}_{\infty,00}$. Then, we analyze other method that produces inconsistent estimates.

Example 4.1. *A consistent method to select $\hat{I}_T$.*

Let us consider the following scheme to build $\mathcal{I}_T$.

(i) $\tilde{I}^0 = \{0\}$.

(ii) For $k > 0$, $\tilde{I}^k = \tilde{I}^{k-1} \cup \{j\}$, where $j = \arg\min_{j \notin \tilde{I}^{k-1}} \hat{\sigma}_h^2(\tilde{I}^{k-1} \cup \{j\})$.

The process ends with $\tilde{I}^N = \{0, \ldots, N\}$ and then, $\mathcal{I}_T = \{\tilde{I}^0, \ldots, \tilde{I}^N\}$. In order to study the asymptotic behavior of $\mathcal{I}_T$, we define $I^0, \ldots, I^N$ according to (i) and (ii) but with σ_h^2 instead of $\hat{\sigma}_h^2$. Let us write $\sigma_h^2(I^0) \geq \sigma_h^2(I^1) \geq \ldots \geq \sigma_h^2(I^{k_0}) = \ldots = \sigma_h^2(I^N)$, where for the sake of simplicity we also assume that the inequalities are strict.

It is easy to see that under the assumptions of section 4.4, w.p. 1, for any $j \leq k_0$, $\tilde{I}^j \to I^j$. Thus, $\{I^0, \ldots, I^{k_0}\} \subset \underline{\mathcal{I}}_\infty^1$ and for $\delta(I) = \#I$, assumption 4.6 holds. Then, $\hat{I}_T$ chosen among the elements of $\mathcal{I}_T$ according to FC is a consistent estimate of I^{k_0}.

Example 4.2. *An almost surely inconsistent method to select* $\hat{I}_T$.

We want to forecast x_{t+1}^0 using information up to t with VAR models. We can use the scheme (i)-(ii) from the previous example, but now, we make a hypothesis test on $H_0 : \Phi_{0,k,l} = 0$, for $l = 1, \ldots, p$. That is, we test that the coefficients of x_t^l in the equation of x_t^0 are null. Then if we reject H_0, we go on to the next iteration of (ii), but if we accept, then the process terminates and $\hat{I}_T = \tilde{I}^k$.

We can see that with probability 1, $\hat{I}_T \nrightarrow I^{k_0}$, where k_0 is as in example 4.1. Under the assumptions of proposition 4.2, the estimates $\hat{\phi}_{0,k,l}$ satisfy a Law of the Iterated Logarithm and then, w.p. 1,

$$\limsup_T \frac{\hat{\Phi}_{0,k,l} - \Phi_{0,k,l}}{Q_T} = a^{1/2}, \tag{4.16}$$

where a is the variance of the asymptotic (gaussian) distribution of $T^{1/2}(\hat{\Phi}_{0,k,l} - \Phi_{0,k,l})$. On the other hand, we reject at $100 \times (1-\alpha)\%$, when

$$\left| \frac{\hat{\Phi}_{0,k,l}}{T^{-1/2} a^{1/2}} \right| > \xi_\alpha, \tag{4.17}$$

where ξ_α is the value that leaves a $\alpha/2$ right tail of a zero-mean, unit-variance gaussian. From (4.16), we know that with probability 1, even if the null assumption holds, there exists a subsequence such that the left side of (4.17) diverges to infinity as $\log\log T$. Thus, with probability 1, $\hat{I}_T \neq I^{k_0}$ infinitely many times as $T \to \infty$.

4.5.2. Forecasting Multiple Series

Let us now introduce the case that we want to forecast several time series. In order to maintain as much as possible the notation of the previous sections, in this section the symbol x_t^0 mean $(z_t^1, \ldots, z_t^m)'$ and thus, $x_t^I = (z_t^1, \ldots, z_t^m, x_t^{i_1}, \ldots, x_t^{i_n})'$.

If we intend to apply the techniques of the previous sections to this case, it is necessary to adapt the criterion of optimality $\sigma_h^2(I) = \min_{J\in\mathcal{I}} \sigma_h^2(J)$. We can use the order relationship defined in the set S^+ of the symmetric positive semidefinite matrices by

$$A \prec B \Longleftrightarrow \exists C \in S^+, B = A + C. \tag{4.18}$$

Now, we can adapt the results of the $m = 1$ case. If $\Sigma_h(I)$ is now the matrix

$$\Sigma_h(I) = \mathbf{E}[\varepsilon_{t+h|t}^0 {\varepsilon_{t+h|t}^0}'], \tag{4.19}$$

where $\varepsilon_{t+h|t}^0$ is as in section 4.2, but now a vector, the symbol $\mathcal{I}_0$ denotes the set all $I \in \mathcal{I}$ such that $\Sigma_h(I) \prec \Sigma_h(J)$ for any J. If $\{0, \ldots, N\} \in \mathcal{I}$, then $\mathcal{I}_0$ is nonempty. Again, $\mathcal{I}_{00}$ is defined as in section 4.2.

The relationship $\prec$ in S^+ does not directly provide a criteria to select $\hat{I}$ because it is not a total order relationship, so we have to summarize the information of $\hat{\Sigma}_h(I)$ into some scalar. Let us consider a function $\nu : S^+ \mapsto \mathbb{R}$ to do that.

We will study the asymptotic behavior of $\hat{I}$ with a quite general family of such functions. We only restrict ν in the following way,

Assumption 4.7. *ν satisfies the following properties,*

(a) It is strictly increasing, that is, for any $A, B \in S^+$ such that $A \prec B$, $\nu(A) \leq \nu(B)$ and if $A \prec B$ and $\nu(A) = \nu(B)$, then $A = B$.

(b) In any region $\mathcal{A}$ such that $\forall A \in \mathcal{A}, \det A \geq \mu > 0$, ν is Lipschitz with respect to some matrix norm $\|\dots\|$, that is, there exists $L(\mathcal{A}) > 0$ such that $|\nu(A) - \nu(B)| \leq L(\mathcal{A})\|A - B\|$.

It is obvious that the elements of $\mathcal{I}_0$ are minimizers of ν, but also if $\mathcal{I}_0 \neq \emptyset$, then every minimizer of ν, belongs to $\mathcal{I}_0$.

Hence, if we define the criteria

$$\mathrm{FC}(I) = \nu(\hat{\Sigma}_h(I)) + \delta(I)\frac{S_T}{T}, \tag{4.20}$$

then we can consistently estimate $\mathcal{I}_{00}$ with $\hat{I}$ as in section 4.3.

Proposition 4.6. *The following holds,*

(i) If assumptions 4.1, 4.2, 4.3, 4.4 and 4.7 hold and $S_T/T \to 0$, $S_T/\log\log T \to +\infty$, then $\hat{I} \xrightarrow{a.s.} \mathcal{I}_{00}$.

(ii) If assumption 4.4bis holds instead of 4.4 and $S_T/T \to 0$, $S_T \to +\infty$, then $\hat{I} \xrightarrow{p} \mathcal{I}_{00}$.

Some examples of ν are,

$\nu_1(\Sigma) = \log\det\Sigma$.

$\nu_2(\Sigma) = \log \mathrm{tr}\Sigma$.

$\nu_3(\Sigma) = \log \Pi_{j=1}^m \Sigma_{jj}$.

Proposition 4.7. *Functions ν_1, ν_2 and ν_3 satisfy assumption 4.7.*

4.6. Monte Carlo experimentation

We have made a simulation experiment to assess the performance of the criteria in selecting the optimal cross section. Let us consider the following bivariate Data Generating Process,

$$\text{DGP}_1 \begin{cases} x_t^0 &= .5x_{t-1}^0 + & bx_{t-1}^1 & +\varepsilon_t^0 \\ x_t^1 &= & .5x_{t-1}^1 & +\varepsilon_t^1 \end{cases}, \tag{4.21}$$

where $(\varepsilon_t^0, \varepsilon_t^1)'$ is a zero-mean Gaussian white noise with unit covariance matrix. When $b > 0$, the optimal cross section to forecast x_t^0 at horizons $h = 1, 2, 3$ is $I = \{0, 1\}$ and when $b = 0$, it is $I = \{0\}$. We want to assess the performance of the criteria for the selection of the optimal cross section by comparing them to several tests, namely the S_1 test of Diebold and Mariano (1995); the ENC-T and ENC-NEW tests described by Clark and McCracken (2001); the conditional predictive ability test by Giacomini and White (2006, GW) and a Granger-Causality test (GC) by Granger (1969). Note that the ENC-T and ENC-NEW cannot be applied to forecasting horizons greater than 1, whereas the Granger-Causality test does not make an explicit distinction between forecasting horizons.

The performance of the tests is usually measured in terms of power and empirical size, but we are interested in measuring the frequency of correct selection of the optimal cross section. The acceptance/rejection is related to the choice of the cross section in different ways depending on the specific test. In the case of the encompassing and Granger-Causality tests, we select $\hat{I}_T = \{0, 1\}$ if the test rejects the null and $\hat{I}_T = \{0\}$ if it does not. In the case of the equal forecast accuracy test S_1, we select $\hat{I}_T = \{0, 1\}$ if the test rejects and the error with $I = \{0\}$ is larger and we select $\hat{I}_T = \{0\}$ otherwise. For the GW tests of conditional predictive ability, we use the decision rule proposed in Giacomini and White (2006) with $c = 0$. We have to specify

in advance the significance levels of the tests. For each of them, we set significance levels at 90% and at 95%.

Regarding the criteria, we have chosen $\delta(I)$ as the cardinality of I. Two penalty functions are considered, the BIC-like, $S_T = \log T$ and the HQ-like, $S_T = 2 \log \log T$. We denote the first as FC_1 and the second as FC_2. For the latter, we have not established strong consistency but only consistency in probability, but this is not relevant to the experiment.

As we explained in the note before proposition 4.1, the criteria can be computed either using the whole series to estimate and forecast or splitting the series into in-sample and out-of-sample parts. In this section and in the following one, we denote by FC_1^* and FC_2^* the criteria with out-of-sample forecasts. The parameter estimates for FC_1^* and FC_2^* and for tests S_1, ENC-T, ENC-NEW are obtained with the first 5/7ths of the observations and the out-of-sample forecasts with the last 2/7ths. The GW test is designed for a fixed or at least bounded estimation window. In our simulations, the window comprises the first 40 observations for $T \leq 100$ and the first 100 for $T > 100$.

In order to obtain the forecasts, we have to determine models for $I = \{0\}$ and for $I = \{0, 1\}$. We have run the simulations in two different ways, (a) fitting AR(1) and VAR(1) models for $I = \{0\}$ and $I = \{0, 1\}$ respectively and (b) fitting AR(p) and VAR(p), with p selected by the BIC. When the order is selected by BIC, the order of the univariate model may be greater than the order of the multivariate one. Then, the models are nonnested and some of the tests cannot be applied. Furthermore, the results do not show a significantly different behavior with respect to the ones with fixed p. Thus, we have not included their results but they can be obtained from the author.

We generated $M = 5,000$ realizations of DGP_1 for $b = 0, 0.05, 0.1, 0.2$ and for different series lengths (50, 100, 200, 400, 800). In table 4.2, we

represent the frequencies of selecting $\{0,1\}$ for each of the combinations of b and length and for each of the tests or criteria.

We have designed additional scenarios to check the performance of the criteria under different conditions. In order to check the effect of heavy-tailed noise, we have run simulations of a process with the same autoregressive structure as DGP_1, but ε_t^0 and ε_t^1 are $t-$distributed with 4 degrees of freedom. We call this, DGP_2.

Part (a) of assumption 4.4 involves consistency of the estimated predictors to the optimal ones. As we saw in section 4.4.3, this holds for well-specified ARMA models. However, we want to assess the performance of the criteria when the models are misspecified and thus, consistency is not guaranteed by our theoretical results.

In the first misspecifiacion scenario, the DGP is a VAR(2). In this case, it only makes sense the fixed order ($p = 1$) estimation in order to preserve the misspecification condition.

$$\text{DGP}_3 \begin{cases} x_t^0 = 1/3x_{t-1}^0 + bx_{t-1}^1 + 2/9x_{t-2}^0 + \varepsilon_t^0 \\ x_t^1 = 1/3x_{t-1}^1 + 2/9x_{t-2}^1 + \varepsilon_t^1. \end{cases} \tag{4.22}$$

The next case, DGP_4 is as DGP_1, but ε_t^0 and ε_t^1 are GARCH,

$$\varepsilon_t^j = \sqrt{h_t}\xi_t \qquad h_t = \mu_h + \varphi h_{t-1} + \alpha(\varepsilon_{t-1}^j)^2, \tag{4.23}$$

with $\mu_h = .05$, $\varphi = .8$, $\alpha = .15$ and $\xi_t \sim WN(0,1)$.

Finally, we generate a MA(1),

$$\text{DGP}_5 \begin{cases} x_t^0 = \varepsilon_t^0 + .5\varepsilon_{t-1}^0 + b\varepsilon_{t-1}^1 \\ x_t^1 = \varepsilon_t^1 + .5\varepsilon_{t-1}^1. \end{cases} \tag{4.24}$$

In tables 4.2-4.6 we present the frequency of selecting $\hat{I}_T = \{0,1\}$.

The comparison between different selection methods, either tests or criteria involves both the probability of correctly selecting $\{0,1\}$ when $b > 0$

and the probability of wrongly selecting $\{0,1\}$ when $b=0$. If both probabilities are greater for method A than for method B, we say that it is less conservative. If the probabilities of correct selection are greater for method A in both cases, we can say that A outperforms B.

By looking at the tables 4.2 and 4.3 we see that FC_1 is more conservative than FC_2 (but for $b=0.2$), ENC-T at 90%, ENC-NEW at 90% (but for b=0.2), and GC and outperforms ENC-NEW and ENC-T at 95%, GW and DM except for $h=3$ and b, T small.

On the other hand, FC_2 is less conservative than DM at 90% for $h=1,2$, DM at 95% and ENC-NEW, more conservative than GC and outperforms ENC-T and GW. The criteria FC_1^* and FC_2^*, not surprisingly, are more conservative versions of FC_1 and FC_2.

When the model is not correctly specified (tables 4.4-4.6), we see that, the relations are generally not very different, but if we are interested in forecasting at horizon $h>0$, the GC can produce extremely bad results when the model is misspecified as in DGP_5 (in fact, this example was intentionally included to illustrate the risks of using the GC-test for this purpose).

4.7. Empirical example

In this section, following Clark and McCracken (2001), we will try to determine whether the unemployment rate is useful to improve the 1-step forecast of inflation. More specifically, the variable to forecast x_t^0 will now be the second difference of the logarithm of the USA CPI index without food and energy and x_t^1 will be the first difference of the unemployment rate among males between 25 and 54. The quarterly series run from 1958:Q3 to 1998:Q1.

Using the notation of the previous sections, we want to choose among $I=\{0\}$ and $I=\{0,1\}$ as the optimal cross section. We will make our choice

using the same criteria and tests as in section 4.6.

As in the previous section, we have obtained the criteria both using the whole series to estimate and forecast and splitting it into in-sample and out-of-sample parts.

All the tests but the Granger-Causality test require out-of-sample forecasts. We divide the series into periods 1958:Q3 to 1987:Q1 and 1987:Q2 to 1998:Q1, so that their lengths are in relationship of 1 to 0.4. We use a fixed scheme, that is, the models are identified and the parameters estimated with the data of the in-sample period and they remain fixed for all the out-of-sample period.

Instead of using only autoregressive models, we use ARMA and VARMA with autoregressive and moving average orders up to 3. According to the BIC criterion, with the data of the in-sample period we choose a MA(1) model for $I = \{0\}$ and a VMA(1) for $I = \{0, 1\}$. This is very convenient because then, the models are nested and we can use the encompassing tests ENC-T and ENC-NEW. The models identified and estimated with the in-sample period data are,

$$x_t^0 = \varepsilon_t^0 - 0.3298(0.0959)\varepsilon_{t-1}^0,$$

where the variance of ε_t^0 is 2.7882 and

$$\begin{cases} x_t^0 = \varepsilon_t^0 - 0.4415(0.0943)\varepsilon_{t-1}^0 - 1.3051(0.3798)\varepsilon_{t-1}^1 \\ x_t^1 = \varepsilon_t^1 + 0.0446(0.0187)\varepsilon_{t-1}^0 + 0.6864(0.0710)\varepsilon_{t-1}^1 \end{cases}$$

where the covariance matrix of $(\varepsilon_t^0, \varepsilon_t^1)'$ is,

$$\Sigma_\varepsilon = \begin{pmatrix} 2.5170 & -0.0844 \\ -0.0844 & 0.1178 \end{pmatrix}.$$

The numbers in parentheses are the standard errors of the estimates. The results of the tests and criteria that are computed with out-of-sample forecasts are in the upper part of table 4.1.

On the other hand, for the GC test and FC_1 and FC_2 criteria, we have identified and estimated the models with the whole series from 1958:Q3 to 1998:Q1. According to the BIC criterion, we choose now a MA(1) model for the univariate and a VAR(1) for the multivariate. The estimated univariate model is

$$x_t^0 = \varepsilon_t^0 - 0.3470(0.0546)\varepsilon_{t-1}^0,$$

where the variance of ε_t^0 is 1.4281 and the multivariate one is

$$\begin{cases} x_t^0 &= -0.3450(0.0749)x_{t-1}^0 \quad -1.1229(0.2888)x_{t-1}^1 + \varepsilon_t^0 \\ x_t^1 &= \quad 0.0550(0.0164)x_{t-1}^0 \quad +0.6548(0.0629)x_{t-1}^1 + \varepsilon_t^1, \end{cases}$$

with covariance matrix

$$\Sigma_\varepsilon = \begin{pmatrix} 1.9083 & -0.0629 \\ -0.0629 & 0.0911 \end{pmatrix}.$$

In the lower part of table 4.1 we include the results of the CG test and the criteria FC_1 and FC_2 with parameters and forecasts computed using the full sample.

The results agree with Clark and McCracken (2001) even if the conditions are slightly different. The encompassing tests and our criteria indicate that the unemployment is relevant to forecast inflation based on the out-of-sample forecasts (upper part of table 4.1). On the other hand, the GW and S_1 tests do not reject their respective null hypotheses of conditional and unconditional equal predictive ability.

The analysis without splitting of the series, which is summarized in the lower part of table 4.1, points in the same direction as the encompassing tests. Both criteria FC_1 and FC_2 yield lower values for $I = \{0, 1\}$ and the Granger-Causality test clearly rejects the null.

	test/crit.	statistic	10%-CV	crit. $I=\{0\}$	crit. $I=\{0,1\}$
	FC_1^*			-1.018	-1.045
	FC_2^*			-1.042	-1.092
in-sample estimates	ENC-T	1.656	1.645		
out-of-sample forecasts	ENC-NEW	9.697	1.003		
	S_1	0.5481	1.645		
	GW	3.7862	4.605		
full sample	FC_1			0.7376	0.7083
for estimation	FC_2			0.7163	0.6859
and forecasting	GC	15.115	2.705		

Table 4.1: The first column indicates whether the series is split into in-sample (to estimate) and out-of-sample (to forecast) periods or we use the full-length series to estimate and forecast; the second column is the test or criteria; the third is the value of the test statistic; the fourth is the 10% critical value and the last two are the values of the criteria for the two possible cross sections.

4.A. Lemmas and Proofs

Lemma 4.1. *If x_t is a process*[3] *satisfying assumptions 4.2 and 4.3 and we denote the BLPs of x_{t+h} using respectively $\{x_s : -\infty < s \leq t\}$ and $\{x_s : t-r < s \leq t\}$ by $\mathbb{P}_h = \sum_{k=0}^{\infty} P_{h,k} L^k$ and $\mathbb{P}_{h,r} = \sum_{k=0}^{r} P_{h,r,k} L^k$, then*

(i) $\mathbb{P}_h(z) = z^{-h}(1 - \Psi_{[h-1]}(z)\Psi(z)^{-1})$, where the notation $A_{[j]}(z)$ means $\sum_{k=0}^{j} A_k z^k$ and

(ii) $\|\mathbb{P}_{h,r}\Psi\|_{\ell^2} \leq \|\Psi - \Psi_{[h]}\|_{\ell^2}$, where $\|A(z)\|_{\ell^p}^p := \sum_{k=0}^{\infty} \operatorname{tr}(A_k A_k')^{p/2}$.

Proof. The first part is a straightforward consequence of the Wold representation and the invertibility of $\Psi(z)$. For (ii), let $\ell^2 = \{A = \sum_{k=0}^{j} A_k z^k : \|A\|_{\ell^2} < +\infty\}$ endowed with the inner product $\langle A, B\rangle = \sum_k \operatorname{tr}(A_k B_k')$. We can use that $\mathbb{P}_{h,r}$ is the solution to the problem of minimizing $\|(1 - z^h\mathbb{P})\Psi\|_{\ell^2}^2$,

[3] In lemmas 4.1 and 4.2 we omit the superscript I.

where $\mathbb{P} \in \ell^2$ subject to the constraints $\langle \mathbb{P}, z^j \rangle = 0$ for $j \geq r$. Consequently, the solution has to satisfy the identity $2\langle (1 - z^h \mathbb{P}\Psi), z^h \mathbb{Q}\Psi \rangle = \sum_{j \geq r} \lambda_j \langle z^j, \mathbb{Q} \rangle$, for all $\mathbb{Q} \in \ell^2$ and certain Lagrange multipliers λ_j. Then, letting $\mathbb{Q} = \mathbb{P}$, we get $\langle (1 - z^h \mathbb{P}\Psi), z^h \mathbb{P}\Psi \rangle = 0$ and thus, $\|\mathbb{P}\Psi\|^2_{\ell^2} \leq \|\Psi(z) - \Psi_{[h]}(z)\|_{\ell^2} \|\mathbb{P}\Psi\|_{\ell^2}$. □

Lemma 4.2. *If $\mathbb{P}^*_h$ and $\mathbb{P}^*_{h,r}$ are the BLPs for $y_t = x_{-t}$ as in lemma 4.1, then*

$$\mathbb{P}_h(z) - \mathbb{P}_{h,t}(z) = \sum_{k=t}^{\infty} P_{h,k} z^k \big(1 - \mathbb{P}^*_{k-t,t}(z^{-1})\big). \tag{4.25}$$

Proof. Let $H(r,s)$ be the linear span of $\{x^i_t : i = 1, \ldots, n, r < t \leq s\}$ and $\mathbf{P}_A$ be the orthogonal projection onto the set A according to the scalar product $\langle u, v \rangle = \mathbf{E}[u'v]$. Then, $\mathbb{P}_h(L)x_t = \mathbf{P}_{H(-\infty,t)} x_{t+h}$ and $\mathbb{P}_{h,t}(L)x_t = \mathbf{P}_{H(0,t)} x_{t+h}$, where the projections are applied to the components of the vectors. Hence,

$$\mathbb{P}_h(L)x_t - \mathbb{P}_{h,t}(L)x_t = \mathbf{P}_{H(-\infty,t)} x_{t+h} - \mathbf{P}_{H(0,t)} x_{t+h} =$$
$$= \mathbf{P}_{H(-\infty,t)} x_{t+h} - \mathbf{P}_{H(0,t)} \mathbf{P}_{H(-\infty,t)} x_{t+h} = (1 - \mathbf{P}_{H(0,t)}) \mathbf{P}_{H(-\infty,t)} x_{t+h}.$$

Since $\mathbf{P}_{H(-\infty,t)} x_{t+h} = \sum_k P_{h,k} x_{t-k}$, we have $\mathbb{P}_h(L)x_t - \mathbb{P}_{h,t}(L)x_t = \sum_{k=t}^{\infty} P_{h,k}(1 - \mathbf{P}_{H(0,t)})x_{t-k}$. We get (4.25) by noting that $(1 - \mathbf{P}_{H(0,t)})x_{t-k} = (1 - \mathbb{P}^*_{k-t,t}(F))x_{t-k}$, where $F = L^{-1}$. □

Proof of Proposition 4.1. In order to avoid inessential complications we will assume that $\hat{\Sigma}^I = \Sigma^I = \mathbb{I}$. If we prove the lemma for $I \subset J$, then it is easy to see that it holds for any $I, J \in \mathcal{I}_0$. We just have to apply it in turn to $I \subset I \cup J$ and $J \subset I \cup J$. Thus, with no loss of generality, we assume that $I \subset J = I \cup K$, where $I \cap K = \emptyset$.

The best predictors of x^0_{t+h} using x^I_s respectively with $s \in (1,t)$ and with $s \in (-\infty, t)$ are $\mathbb{P}^{0,I}_{h,t} = e_0 \mathbb{P}_{h,t}$ and $\mathbb{P}^{0,I}_h = e_0 \mathbb{P}_h$ with $e_0 = (1, 0, \ldots, 0)$

and $\mathbb{P}_{h,t}$, $\mathbb{P}_h$ as in lemma 4.1 with $x_t = X_t^I$. We can now write $x_{t+h|t}^{0,I}$ as $\mathbb{P}_{t,h}^{0,I}(L)x_t^I$.

We can decompose the predictor as $\mathbb{P}_h^{0,J}$ as $[\mathbb{P}_h^{0,J,1} : \mathbb{P}_h^{0,J,2}]$ in such way that $\mathbb{P}_h^{0,J}(L)x_t^J = \mathbb{P}_h^{0,J,1}(L)x_t^I + \mathbb{P}_h^{0,J,2}(L)x_t^K$. Since $\mathbb{P}_h^{0,J}$ is the least squares predictor, then the minimum of the quadratic functional,

$$(P, Q) \mapsto q(P, Q) = \mathbf{E}[x_{t+h}^0 - P(L)x_t^I - Q(L)x_t^K]^2, \tag{4.26}$$

is attained at $(\mathbb{P}_h^{0,J,1}, \mathbb{P}_h^{0,J,2})$ and the minimal value is $\sigma_h^2(J)$, but this value is also attained at $(\mathbb{P}_{h,t}^{0,I}, 0)$, because I is also in $\mathcal{I}_0$. If the functional q is strictly convex, then the minimum is unique and $\mathbb{P}_h^{0,J,1} = \mathbb{P}_h^{0,I}, \mathbb{P}_h^{0,J,2} = 0$.

We can see that the strict convexity of q is equivalent to the condition that $P(L)x_t^I + Q(L)x_t^K = 0$ implies $P, Q = 0$, but this property holds because of the uniqueness of the Wold representation when x_t^J is linearly regular and Ψ^I does not have unit modulus roots. This property is stated, for example, in chapter 1 of HD.

Let us turn now to $\hat{\sigma}_h^2(J) - \hat{\sigma}_h^2(I)$. First, we can see that in the strong case, only $O(T^{-1})$ terms are neglected if we replace $\hat{\sigma}_h^2(I)$ by $\dot{\sigma}_h^2(I)$, where $\dot{\sigma}_h^2(I) = T^{-1}\sum_{t=1}^{T-h}(\dot{\varepsilon}_{t,h}^{0,I})^2$ and

$$\dot{\varepsilon}_{t,h}^{0,I} = x_{t+h}^0 - \hat{\mathbb{Q}}_h^{0,I}(L)x_t^I = x_{t+h}^0 - \sum_{k=0}^{\infty}\hat{P}_{h,k}^{0,I}x_{t-k}^I. \tag{4.27}$$

Let us consider the difference

$$\dot{\sigma}_h^2(I) - \hat{\sigma}_h^2(I) = \frac{1}{T}\sum_{t=1}^{T-h}\left[(\dot{\varepsilon}_{t,h}^{0,I})^2 - (\hat{\varepsilon}_{t,h}^{0,I})^2\right]. \tag{4.28}$$

We can write

$$|(\dot{\varepsilon}_{t,h}^{0,I})^2 - (\hat{\varepsilon}_{t,h}^{0,I})^2| = |\dot{\varepsilon}_{t,h}^{0,I} - \hat{\varepsilon}_{t,h}^{0,I}| \cdot |\dot{\varepsilon}_{t,h}^{0,I} + \hat{\varepsilon}_{t,h}^{0,I}|. \tag{4.29}$$

We will analyze separately both factors. First, note that we can use lemma 4.2 with the estimated transfer function $\hat{\Psi}$, so we get $\dot{\varepsilon}_{t,h}^{0,I} - \hat{\varepsilon}_{t,h}^{0,I} =$

$\sum_{k=t}^{\infty} \hat{P}_{h,k}^{0,I} z_{t-k}$, where $z_{t-k} = x_{t-k}^I - \hat{\mathbb{P}}_{k-t,t}^*(F) x_{k-t}^I$ and $F = L^{-1}$. On the other hand, we can write $x_t^I = \Psi'(F)\xi_t$ and then, $z_{t-k} = (1 - \hat{\mathbb{P}}_{k-t,t}^*(F))\Psi'(F)\xi_{k-t}$. Using lemma 4.1 and the inequality $\|AB\|_{\ell^2} \leq \|A\|_{\ell^2} \cdot \|B\|_{\ell^1}$, that is a consequence of Hölder's inequality, we get

$$|z_{t-k}| \leq \|(1 - \hat{\mathbb{P}}_{k-t,t}^*)\Psi'\|_{\ell^2} \Big(\sum_{j=1}^{t} |\xi_j|^2 \Big)^{1/2} \leq$$

$$\leq \Big(\|\Psi'\|_{\ell^2} + \|\hat{\Psi}' - \hat{\Psi}'_{[k-t]}\|_{\ell^2} \cdot \|\hat{\Psi}^{-1}\|_{\ell^1} \cdot \|\Psi'\|_{\ell^1} \Big) \Big(\sum_{j=1}^{t} |\xi_j|^2 \Big)^{1/2}. \quad (4.30)$$

All terms in the first factor of (4.30) are bounded. The first, as a direct consequence of assumption 4.2. For the second, we can use inequality $\|\hat{\Psi}\|_{\ell^2} \leq \|\hat{\Psi} - \Psi\|_{\ell^2} + \|\Psi\|_{\ell^2}$ and the identity $\hat{\Psi} - \Psi = \hat{\Psi}(\Pi - \hat{\Pi})\Psi$ to get $\|\hat{\Psi}\|_{\ell^2}(1 - \|\hat{\Pi} - \Pi\|_{\ell^2} \cdot \|\Psi\|_{\ell^1}) \leq \|\Psi\|_{\ell^2}$. The term $\|\hat{\Psi}^{-1}\|_{\ell^1}$ is bounded from assumption 4.4. Hence, we get $|z_{t-k}| \leq c(\sum_{j=1}^{t} |\xi_j|^2)^{1/2}$ for a certain constant c.

Then, using,

$$|\dot{\varepsilon}_{t,h}^{0,I} - \hat{\varepsilon}_{t,h}^{0,I}| \leq \sum_{k=t}^{\infty} |\hat{P}_{h,k}^{0,I}| \cdot |z_{t-k}|,$$

and dealing with the second factor in (4.29) in a similar way, it follows that, $|(\dot{\varepsilon}_{t,h}^{0,I})^2 - (\hat{\varepsilon}_{t,h}^{0,I})^2| \leq c(\sum_{j=1}^{t} |\xi_j|^2) \sum_{k=t}^{\infty} |\hat{P}_{h,k}^{0,I}|$. Using lemma 4.1, we get that $|(\dot{\varepsilon}_{t,h}^{0,I})^2 - (\hat{\varepsilon}_{t,h}^{0,I})^2| \leq \zeta_t$ with $\sum_t^{\infty} \mathbf{E}\zeta_t < +\infty$. Then, by theorem 2, page 66, in Gihman and Skorohod (1974), we get that $\sum_t[\ldots]$ in (4.28) is bounded with probability 1 and then $\dot{\sigma}_h^2(I) - \hat{\sigma}_h^2(I) = O(T^{-1})$. We can now use the $\dot{\sigma}^2$ terms instead of the $\hat{\sigma}^2$ ones, in the difference $\hat{\sigma}_h^2(J) - \hat{\sigma}_h^2(I)$. Then, we proceed as

$$\frac{1}{T} \sum_{t=1}^{T-h} \Big[(\dot{\varepsilon}_{t,h}^{0,J})^2 - (\dot{\varepsilon}_{t,h}^{0,I})^2 \Big] = \frac{1}{T} \sum_{t=1}^{T-h} \Big[\dot{\varepsilon}_{t,h}^{0,J} - \dot{\varepsilon}_{t,h}^{0,I} \Big] \Big[\dot{\varepsilon}_{t,h}^{0,J} + \dot{\varepsilon}_{t,h}^{0,I} \Big]. \quad (4.31)$$

On the other hand, we can write $\dot{\varepsilon}_{t,h}^{0,I} = x_{t+h}^0 - \hat{\mathbb{Q}}_h^{0,I,*}(L) x_t^J$, with $\hat{\mathbb{Q}}_h^{0,I,*} =$

$[\hat{\mathbb{Q}}_h^{0,I} : 0]$ and $\hat{\mathbb{Q}}_h^{0,I}(L) = \sum_{k=1}^{t-1} P_{h,k}^{0,I} L^k$. Then, $\dot{\sigma}_h^2(J) - \dot{\sigma}_h^2(I)$ equals

$$\frac{1}{T}\sum_{t=1}^{T-h}\left[\sum_{k=1}^{t-1}\left(\hat{P}_{h,k}^{0,J} - \hat{P}_{h,k}^{0,I,*}\right)x_{t-k}^J\right]\left[2x_{t+h}^0 - \sum_{l=1}^{t-1}\left(\hat{P}_{h,l}^{0,J} + \hat{P}_{h,l}^{0,I,*}\right)x_{t-l}^J\right]. \tag{4.32}$$

We will denote the first $[\ldots]$ factor as a_t whereas the second is decomposed as $b_t + c_t$, with

$$\begin{aligned} b_t &= 2\varepsilon_{t,h}^{0,J}, \\ c_t &= \sum_{l=0}^{t-1}\left(2P_{h,l}^{0,J} - \hat{P}_{h,l}^{0,J} - \hat{P}_{h,l}^{0,I,*}\right)x_{t-l}^J, \end{aligned}$$

where $\varepsilon_{t,h}^0 = x_{t+h}^0 - x_{t+h|t}^{0,J}$. Let us deal first with the product $(1/T)\sum_t a_t b_t$,

$$\frac{1}{T}\sum_{t=1}^{T-h} a_t b_t = 2\frac{1}{T}\sum_{t=1}^{T-h}\sum_{k=1}^{t-1}\left(\hat{P}_{h,k}^{0,J} - \hat{P}_{h,k}^{0,I,*}\right)x_{t-k}^J\varepsilon_{t,h}^{0,J}. \tag{4.33}$$

We can swap the order of summation and use that the difference in parentheses does not depend on t. Thus, it becomes

$$2\sum_{k=1}^{T-h-1}\left\{\left(\hat{P}_{h,k}^{0,J} - \hat{P}_{h,k}^{0,I,*}\right)\frac{1}{T}\sum_{t=k+1}^{T-h} x_{t-k}^J\varepsilon_{t,h}^{0,J}\right\} =: 2\sum_{k=1}^{T-h-1}\Delta_k s_{k,T}. \tag{4.34}$$

Let us write now

$$\left|\sum_k \frac{\Delta_k}{Q_T}\frac{s_{k,T}}{Q_T}\right| \leq \left[\sum_{k\leq g(T)}\frac{\Delta_k}{Q_T}\frac{s_{k,T}}{Q_T} + \sum_{k>g(T)}\frac{\Delta_k}{Q_T}\frac{s_{k,T}}{Q_T}\right] \tag{4.35}$$

$$\leq \left[\frac{\|\Delta_{[g(T)]}(z)\|_{\ell^1}}{Q_T}\sup_{k\leq g(T)}\frac{|s_{k,T}|}{Q_T} + \frac{\|\Delta(z) - \Delta_{[g(T)]}(z)\|_{\ell^1}}{Q_T}\sup_{k>g(T)}\frac{|s_{k,T}|}{Q_T}\right], \tag{4.36}$$

where $\Delta(z) = \sum_k \Delta_k z^k$. If $g(T) = (\log T)^a$, then by theorem 5.3.5 in HD, we have that $\sup_{k\leq g(T)}|s_{k,T}| = O(Q_T)$. Thus, using assumption 4.4 the first term inside the brackets in (4.36) is $O(1)$. On the other hand, $\sup_{0\leq k<\infty}|s_{k,T}| = O([\log T/T]^{1/2})$ by theorem 7.4.3, again in HD,

$$\frac{\|\Delta(z) - \Delta_{[g(T)]}(z)\|_{\ell^1}}{Q_T}\sup_{k>g(T)}\frac{|s_{k,T}|}{Q_T} = O((\log T)^{-\alpha a})\left(\frac{\log T}{\log\log T}\right)^{1/2}. \tag{4.37}$$

Thus, if $a > 1/(2\alpha)$ then (4.35) is bounded with probability 1.

We put now $(1/T)\sum_t a_t c_t = \sum_{k,l} \Delta_k G_{k,l,T} \tilde{\Delta}_l'$, where $G_{k,l,T}$ is defined as $(1/T)\sum_t x_{t-k}^J x_{t-l}^{J\,\prime}$ and $\tilde{\Delta}_l := 2P_{h,l}^{0,J} - \hat{P}_{h,l}^{0,J} - \hat{P}_{h,l}^{0,I,*}$. Using that $G_{k,l,T}$ is almost surely uniformly bounded (this is implied by Theorem 7.4.3 in HD), then

$$\left| \sum_{k,l} \frac{\Delta_k}{Q_T} G_{k,l,T} \frac{\tilde{\Delta}_l'}{Q_T} \right| = O\left(\frac{\|\Delta\|_{\ell^1}}{Q_T} \right)^2 = O(1). \tag{4.38}$$

With this, the first part of the lemma is proved. For the order in probability, it is only necessary to replace Q_T by $T^{-1/2}$ and use that $T^{-1/2}\mathbf{E}|s_{k,T}|$ is uniformly bounded. Let us see this.

$$\mathbf{E}|s_{k,T}| = \mathbf{E}\left| \sum_{j=0}^{\infty} \Psi_j^J \frac{1}{T} \sum_{t=k+1}^{T-h} \varepsilon_{t-k-j}^J \varepsilon_{t,h}^{0,J} \right| \leq \tag{4.39}$$

$$\leq \left(\mathbf{E}\left[\sum_{j=0}^{\infty} \Psi_j^J \frac{1}{T} \sum_{t=k+1}^{T-h} \varepsilon_{t-k-j}^J \varepsilon_{t,h}^{0,J} \right]^2 \right)^{1/2} \tag{4.40}$$

The term inside $(\cdot)^{1/2}$ can be written as

$$\frac{1}{T^2} \sum_{j,l} \sum_{t,s} \mathbf{E}\left[\varepsilon_{t-k-j}^{J\,\prime} \Psi_j^{J\prime} \Psi_l^J \varepsilon_{s-k-l}^J \varepsilon_{t,h}^{0,J} \varepsilon_{s,h}^{0,J} \right] \lessapprox \frac{\|\Sigma\|^2}{T} \left(\sum_{\nu=1}^{h} \|\Psi_\nu^J\| \right) \sum_{j}^{\infty} \|\Psi_j^J\|^2, \tag{4.41}$$

because of assumption 4.3 and the fact that $\varepsilon_{t,h}^{0,J}$ is the first component of vector $\sum_{\nu=0}^{h-1} \Psi_\nu^J \varepsilon_{t+h-\nu}^J$. □

Proof of Prop. 4.2. In order to prove strong consistency of $\hat{I}_T$ it suffices to prove that w. p. 1, every convergent subsequence converges to an element of $\mathcal{I}_{00}$. We avoid cumbersome notation by using $\hat{I}_T$ for a convergent subsequence. If $\hat{I}_T \to J$, we will show that necessarily $J \in \mathcal{I}_{00}$. Let us consider first the case $J \notin \mathcal{I}_0$ and then, $J \in \mathcal{I}_0 \setminus \mathcal{I}_{00}$.

If $J \notin \mathcal{I}_0$, then for any $I \in \mathcal{I}_0$, $\sigma_h^2(J) > \sigma_h^2(I)$. For large T, $\hat{I}_T = J$, so

$$\mathrm{FC}(\hat{I}_T) - \mathrm{FC}(I) = \log \hat{\sigma}_h^2(J) - \log \hat{\sigma}_h^2(I) + [\delta(J) - \delta(I)] \frac{S_T}{T}, \tag{4.42}$$

and the first difference in the right hand side converges to a strictly positive value, whereas the last term converges to zero. Thus, w.p. 1, for large T , $\mathrm{FC}(\hat{I}_T) - \mathrm{FC}(I) > 0$.

For the case $J \in \mathcal{I}_0 \setminus \mathcal{I}_{00}$ we need the order of convergence of $\log \hat{\sigma}_h^2(J) - \log \hat{\sigma}_h^2(I)$ established in proposition 4.1 for $I \in \mathcal{I}_0$, $I \subset J$. We can write

$$\log \hat{\sigma}_h^2(J) - \log \hat{\sigma}_h^2(I) = \log \left\{ 1 + \frac{\hat{\sigma}_h^2(J) - \hat{\sigma}_h^2(I)}{\sigma_h^2(I)} \right\}, \tag{4.43}$$

and by a first-order Taylor expansion, we obtain

$$\log \hat{\sigma}_h^2(J) - \log \hat{\sigma}_h^2(I) = [1 + o(1)] \left\{ \frac{\hat{\sigma}_h^2(J) - \hat{\sigma}_h^2(I)}{\sigma_h^2(I)} \right\}. \tag{4.44}$$

Since $\hat{\sigma}_h^2(J) - \hat{\sigma}_h^2(I) = O(Q_T^2)$,

$$\frac{\mathrm{FC}(J) - \mathrm{FC}(I)}{Q_T^2} = O(1) + \frac{S_T}{\log \log T}, \tag{4.45}$$

that diverges to $+\infty$ and then, for large T, $\hat{I}_T \neq J$.

The same arguments can be easily adapted to prove consistency in probability. □

Proof of Prop. 4.3. Let $I, J \in \mathcal{I}_0$. Then, $K = I \cup J \in \mathcal{I}_0$. Let us assume an ordering of the processes such that $x_t^K = [x_t^{I \setminus J'} : x_t^{I \cap J'} : x_t^{J \setminus I'}]'$. Then, by the same uniqueness argument used in the proof of proposition 4.1, we get that $\mathbb{P}_h^{0,K} = [\mathbb{P}_h^{0,I} : 0] = \mathbb{P}_h^{0,K} = [0 : \mathbb{P}_h^{0,J}]$. This means that the BLP using x_t^K only uses effectively the processes in $I \cap J$, so $I \cap J \in \mathcal{I}_0$.

As a consequence, if $I, J \in \mathcal{I}_{00}$, then $I \cap J \in \mathcal{I}_{00}$, but then $I \cap J = I = J$, so the first part of the proposition is proved. For the second, if $I \in \mathcal{I}_{00}$ and $J \in \mathcal{I}_0$, then $I \cap J \in \mathcal{I}_{00}$, but then $I \cap J = I$ and thus, $I \subset J$. □

Proof of Prop. 4.5. With probability 1, there exists T_1 such that for all $T > T_1$, $\underline{\mathcal{I}}_\infty^1 \subset \mathcal{I}_T \subset \bigcup_{p>0} \overline{\mathcal{I}}_\infty^p$.

Let us assume with no loss of generality that $\hat{I}_T \to I_0$. By assumption 4.6, we can discard with probability 1 all elements in $\mathcal{P}(\{0, \ldots, N\}) \setminus \underline{\mathcal{I}}_\infty^1$

as possible limits. On the other hand, any I such that $\sigma_h^2(I) > \sigma_{h,*}^2$ can be ruled out. We conclude by applying proposition 4.1 with $\mathcal{I}_0 = \mathcal{I}_{\infty,0}$.

□

4.B. Tables

0	0.0	0.0	0.0	0.0	0.05	0.05	0.05	0.05	0.05	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2
50	100	200	400	800	50	100	200	400	800	50	100	200	400	800	50	100	200	400	800
4	0.035	0.021	0.014	0.010	0.076	0.062	0.071	0.099	0.165	0.131	0.156	0.252	0.434	0.748	0.358	0.550	0.816	0.982	1.000
1	0.043	0.029	0.020	0.016	0.075	0.064	0.071	0.092	0.148	0.123	0.140	0.206	0.344	0.589	0.277	0.415	0.643	0.887	0.991
3	0.022	0.015	0.009	0.006	0.043	0.030	0.033	0.041	0.067	0.067	0.074	0.100	0.172	0.331	0.155	0.232	0.382	0.618	0.856
2	0.082	0.068	0.061	0.051	0.138	0.131	0.163	0.230	0.372	0.215	0.268	0.426	0.659	0.905	0.485	0.697	0.915	0.996	1.000
3	0.092	0.077	0.073	0.064	0.130	0.123	0.150	0.197	0.309	0.192	0.234	0.347	0.521	0.764	0.377	0.548	0.771	0.944	0.996
9	0.056	0.047	0.039	0.036	0.081	0.074	0.083	0.120	0.188	0.115	0.143	0.210	0.331	0.533	0.241	0.355	0.540	0.757	0.919
3	0.005	0.002	0.001	0.001	0.018	0.011	0.009	0.010	0.017	0.033	0.034	0.045	0.082	0.205	0.112	0.164	0.310	0.557	0.815
6	0.006	0.003	0.001	0.001	0.021	0.012	0.010	0.014	0.017	0.030	0.033	0.045	0.076	0.172	0.088	0.131	0.250	0.441	0.692
7	0.004	0.002	0.001	0.000	0.008	0.004	0.005	0.004	0.005	0.015	0.015	0.018	0.030	0.067	0.041	0.059	0.115	0.232	0.428
7	0.014	0.007	0.005	0.005	0.037	0.028	0.026	0.042	0.071	0.060	0.068	0.108	0.195	0.397	0.177	0.257	0.447	0.694	0.883
0	0.017	0.013	0.009	0.007	0.040	0.028	0.033	0.044	0.075	0.053	0.064	0.101	0.175	0.344	0.147	0.215	0.381	0.583	0.792
7	0.009	0.007	0.004	0.004	0.022	0.013	0.017	0.023	0.034	0.034	0.034	0.053	0.088	0.193	0.077	0.114	0.223	0.381	0.588
3	0.034	0.029	0.034	0.032	0.047	0.052	0.053	0.067	0.105	0.061	0.078	0.098	0.149	0.235	0.110	0.151	0.225	0.340	0.528
4	0.054	0.041	0.040	0.036	0.076	0.067	0.059	0.065	0.089	0.086	0.080	0.097	0.120	0.170	0.126	0.131	0.174	0.244	0.374
7	0.073	0.055	0.049	0.041	0.099	0.081	0.066	0.060	0.075	0.106	0.091	0.090	0.091	0.124	0.134	0.124	0.138	0.163	0.229
2	0.017	0.014	0.017	0.015	0.024	0.028	0.029	0.036	0.058	0.034	0.045	0.056	0.091	0.149	0.060	0.086	0.140	0.234	0.410
8	0.029	0.020	0.020	0.017	0.049	0.038	0.033	0.034	0.051	0.057	0.045	0.055	0.067	0.104	0.083	0.077	0.103	0.152	0.262
2	0.047	0.032	0.025	0.023	0.069	0.053	0.035	0.032	0.041	0.076	0.057	0.052	0.052	0.071	0.097	0.079	0.083	0.091	0.138
2	0.045	0.041	0.040	0.040	0.056	0.047	0.052	0.066	0.083	0.064	0.062	0.085	0.117	0.176	0.098	0.125	0.171	0.241	0.418
5	0.086	0.092	0.151	0.289	0.105	0.084	0.092	0.172	0.352	0.112	0.102	0.115	0.214	0.403	0.129	0.153	0.172	0.286	0.527
3	0.129	0.143	0.190	0.344	0.108	0.139	0.149	0.218	0.374	0.115	0.148	0.154	0.248	0.419	0.128	0.169	0.185	0.265	0.464
5	0.021	0.017	0.018	0.018	0.028	0.021	0.026	0.035	0.045	0.034	0.031	0.041	0.068	0.107	0.053	0.068	0.098	0.153	0.307
7	0.054	0.049	0.082	0.201	0.077	0.055	0.050	0.097	0.249	0.080	0.069	0.065	0.128	0.289	0.098	0.103	0.102	0.167	0.386
1	0.091	0.089	0.120	0.258	0.089	0.100	0.093	0.133	0.283	0.091	0.109	0.099	0.158	0.320	0.102	0.125	0.121	0.160	0.350
0	0.111	0.098	0.104	0.102	0.134	0.142	0.163	0.214	0.320	0.175	0.223	0.319	0.466	0.675	0.318	0.455	0.663	0.877	0.982
4	0.053	0.050	0.052	0.050	0.074	0.080	0.091	0.128	0.203	0.099	0.134	0.200	0.326	0.534	0.195	0.309	0.518	0.781	0.961
0	0.039	0.031	0.033	0.031	0.064	0.063	0.091	0.145	0.260	0.101	0.147	0.268	0.485	0.786	0.278	0.473	0.774	0.965	0.999
4	0.017	0.012	0.011	0.011	0.034	0.032	0.044	0.077	0.155	0.057	0.088	0.165	0.339	0.676	0.190	0.343	0.662	0.935	0.998
0	0.115	0.109	0.107	0.102	0.161	0.162	0.218	0.314	0.495	0.239	0.312	0.502	0.745	0.945	0.506	0.742	0.948	0.998	1.000
9	0.066	0.057	0.054	0.050	0.102	0.105	0.137	0.214	0.371	0.166	0.222	0.381	0.636	0.904	0.403	0.645	0.907	0.996	1.000

lts of the simulations of DGP_1. The value of the parameter b is indicated in the first row and the ries in the second. The leftmost column indicates the test or criteria (with forecasting horizon h heses; if h is not specified, $h = 1$ or it is not applicable). The figures in the remaining places are of of selecting $\{0, 1\}$ for each combination of b, length and test/criteria.

0	0.0	0.0	0.0	0.0	0.05	0.05	0.05	0.05	0.05	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2
50	100	200	400	800	50	100	200	400	800	50	100	200	400	800	50	100	200	400	800
0	0.037	0.021	0.016	0.011	0.080	0.066	0.072	0.100	0.172	0.137	0.175	0.262	0.444	0.744	0.376	0.551	0.800	0.969	0.999
0	0.045	0.028	0.021	0.017	0.075	0.068	0.070	0.095	0.147	0.119	0.147	0.217	0.349	0.583	0.291	0.425	0.642	0.871	0.987
0	0.021	0.013	0.009	0.006	0.040	0.035	0.035	0.046	0.067	0.066	0.078	0.108	0.177	0.329	0.164	0.238	0.394	0.606	0.854
2	0.085	0.069	0.063	0.048	0.145	0.135	0.164	0.230	0.376	0.221	0.293	0.433	0.657	0.895	0.491	0.685	0.899	0.992	1.000
9	0.092	0.079	0.075	0.062	0.134	0.137	0.151	0.207	0.311	0.189	0.249	0.351	0.527	0.755	0.388	0.553	0.765	0.936	0.995
4	0.057	0.044	0.039	0.032	0.079	0.080	0.088	0.122	0.183	0.116	0.149	0.213	0.334	0.528	0.247	0.358	0.541	0.742	0.921
5	0.005	0.003	0.001	0.000	0.020	0.013	0.011	0.011	0.020	0.036	0.038	0.052	0.094	0.214	0.116	0.182	0.314	0.551	0.801
6	0.009	0.004	0.002	0.001	0.018	0.015	0.012	0.013	0.020	0.034	0.038	0.051	0.086	0.173	0.094	0.147	0.248	0.444	0.688
9	0.004	0.002	0.001	0.001	0.012	0.009	0.006	0.006	0.008	0.016	0.017	0.023	0.036	0.068	0.049	0.065	0.118	0.231	0.436
7	0.014	0.009	0.006	0.005	0.039	0.028	0.026	0.041	0.075	0.064	0.075	0.116	0.209	0.399	0.177	0.273	0.448	0.684	0.875
0	0.020	0.012	0.010	0.007	0.038	0.032	0.033	0.046	0.073	0.060	0.073	0.109	0.190	0.337	0.152	0.230	0.374	0.584	0.785
1	0.013	0.008	0.005	0.004	0.023	0.020	0.020	0.023	0.035	0.035	0.038	0.058	0.100	0.190	0.086	0.122	0.216	0.381	0.591
7	0.036	0.029	0.034	0.032	0.044	0.050	0.054	0.071	0.108	0.059	0.077	0.108	0.163	0.235	0.101	0.159	0.232	0.346	0.534
1	0.050	0.040	0.036	0.037	0.070	0.060	0.057	0.068	0.090	0.085	0.082	0.095	0.125	0.169	0.117	0.135	0.173	0.243	0.375
5	0.072	0.054	0.044	0.040	0.097	0.076	0.071	0.065	0.074	0.106	0.095	0.091	0.098	0.120	0.125	0.122	0.129	0.154	0.233
9	0.015	0.015	0.014	0.015	0.021	0.024	0.025	0.037	0.060	0.031	0.040	0.056	0.094	0.152	0.053	0.086	0.139	0.238	0.409
3	0.027	0.018	0.016	0.017	0.044	0.032	0.029	0.035	0.047	0.051	0.047	0.053	0.071	0.101	0.072	0.075	0.099	0.143	0.260
5	0.043	0.031	0.020	0.020	0.066	0.047	0.041	0.033	0.039	0.076	0.056	0.052	0.053	0.068	0.087	0.075	0.073	0.086	0.146
3	0.038	0.040	0.038	0.035	0.050	0.043	0.047	0.063	0.086	0.057	0.062	0.085	0.122	0.178	0.092	0.118	0.170	0.264	0.426
0	0.079	0.077	0.116	0.217	0.089	0.088	0.091	0.149	0.279	0.108	0.102	0.120	0.195	0.332	0.131	0.149	0.165	0.266	0.439
1	0.115	0.125	0.169	0.278	0.099	0.133	0.135	0.189	0.317	0.108	0.143	0.148	0.206	0.351	0.121	0.159	0.183	0.252	0.403
2	0.018	0.017	0.016	0.016	0.024	0.019	0.021	0.030	0.042	0.027	0.031	0.042	0.064	0.104	0.050	0.069	0.094	0.167	0.308
5	0.048	0.041	0.061	0.145	0.068	0.057	0.044	0.083	0.192	0.080	0.067	0.069	0.101	0.221	0.095	0.098	0.095	0.158	0.304
3	0.083	0.078	0.099	0.189	0.079	0.095	0.078	0.113	0.221	0.087	0.101	0.088	0.121	0.247	0.093	0.116	0.112	0.154	0.280
8	0.113	0.103	0.103	0.105	0.135	0.148	0.175	0.231	0.332	0.180	0.242	0.332	0.484	0.679	0.324	0.478	0.669	0.861	0.974
9	0.056	0.050	0.051	0.050	0.071	0.082	0.095	0.132	0.215	0.100	0.140	0.211	0.349	0.546	0.192	0.321	0.517	0.760	0.945
5	0.043	0.036	0.033	0.032	0.071	0.073	0.094	0.147	0.268	0.114	0.169	0.282	0.491	0.779	0.287	0.482	0.750	0.946	0.997
9	0.020	0.014	0.014	0.013	0.039	0.038	0.048	0.083	0.167	0.070	0.100	0.182	0.360	0.673	0.202	0.366	0.640	0.908	0.994
2	0.143	0.126	0.117	0.103	0.192	0.180	0.223	0.306	0.489	0.250	0.331	0.491	0.742	0.946	0.482	0.719	0.940	0.998	1.000
6	0.091	0.078	0.068	0.059	0.138	0.122	0.152	0.213	0.369	0.184	0.233	0.376	0.628	0.900	0.384	0.611	0.892	0.994	1.000

lts of the simulations of DGP_2. The value of the parameter b is indicated in the first row and the ries in the second. The leftmost column indicates the test or criteria (with forecasting horizon h heses; if h is not specified, $h = 1$ or it is not applicable). The figures in the remaining places are of of selecting $\{0, 1\}$ for each combination of b, length and test/criteria.

0	0.0	0.0	0.0	0.0	0.05	0.05	0.05	0.05	0.05	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2
50	100	200	400	800	50	100	200	400	800	50	100	200	400	800	50	100	200	400	800
8	0.045	0.034	0.021	0.015	0.100	0.106	0.138	0.214	0.400	0.210	0.290	0.486	0.774	0.972	0.515	0.774	0.965	0.999	1.000
5	0.033	0.026	0.017	0.012	0.089	0.106	0.165	0.281	0.532	0.222	0.349	0.594	0.881	0.994	0.591	0.870	0.991	1.000	1.000
1	0.021	0.016	0.009	0.007	0.054	0.065	0.092	0.158	0.325	0.139	0.216	0.388	0.695	0.948	0.384	0.656	0.924	0.998	1.000
8	0.103	0.088	0.077	0.069	0.173	0.194	0.261	0.401	0.639	0.314	0.434	0.658	0.898	0.994	0.647	0.872	0.988	1.000	1.000
1	0.077	0.066	0.061	0.057	0.150	0.192	0.294	0.479	0.738	0.316	0.491	0.744	0.954	0.999	0.703	0.929	0.998	1.000	1.000
0	0.053	0.051	0.043	0.040	0.100	0.129	0.195	0.328	0.565	0.215	0.338	0.569	0.850	0.986	0.501	0.781	0.970	1.000	1.000
7	0.007	0.004	0.001	0.001	0.026	0.017	0.020	0.031	0.067	0.057	0.071	0.126	0.248	0.490	0.176	0.308	0.533	0.788	0.952
1	0.006	0.004	0.001	0.001	0.020	0.018	0.027	0.045	0.101	0.052	0.079	0.162	0.349	0.680	0.194	0.383	0.686	0.928	0.994
5	0.002	0.001	0.000	0.000	0.009	0.009	0.006	0.016	0.029	0.022	0.033	0.064	0.158	0.418	0.088	0.186	0.424	0.779	0.967
2	0.018	0.012	0.007	0.006	0.049	0.041	0.054	0.089	0.184	0.099	0.127	0.226	0.420	0.665	0.262	0.421	0.659	0.857	0.970
6	0.015	0.015	0.007	0.007	0.042	0.043	0.068	0.120	0.250	0.093	0.143	0.289	0.533	0.824	0.289	0.509	0.796	0.960	0.997
1	0.007	0.005	0.003	0.002	0.021	0.021	0.028	0.055	0.121	0.046	0.069	0.147	0.333	0.667	0.151	0.295	0.589	0.880	0.985
8	0.043	0.043	0.041	0.037	0.056	0.061	0.080	0.106	0.163	0.091	0.109	0.161	0.231	0.340	0.139	0.204	0.303	0.481	0.728
0	0.064	0.059	0.051	0.047	0.107	0.097	0.121	0.168	0.250	0.147	0.175	0.249	0.366	0.565	0.244	0.350	0.514	0.748	0.945
7	0.080	0.065	0.057	0.052	0.113	0.110	0.111	0.149	0.217	0.152	0.162	0.222	0.318	0.502	0.212	0.283	0.405	0.640	0.878
5	0.023	0.023	0.020	0.020	0.031	0.034	0.047	0.062	0.099	0.052	0.063	0.097	0.150	0.243	0.082	0.129	0.206	0.360	0.612
1	0.039	0.035	0.027	0.023	0.071	0.060	0.075	0.107	0.167	0.102	0.113	0.160	0.258	0.449	0.167	0.241	0.380	0.638	0.901
9	0.054	0.037	0.031	0.028	0.081	0.073	0.065	0.090	0.139	0.110	0.107	0.137	0.211	0.375	0.148	0.184	0.277	0.496	0.796
5	0.052	0.060	0.069	0.071	0.066	0.068	0.075	0.114	0.171	0.083	0.102	0.135	0.206	0.303	0.129	0.156	0.246	0.382	0.648
8	0.089	0.080	0.095	0.176	0.122	0.119	0.102	0.184	0.343	0.140	0.175	0.204	0.319	0.546	0.213	0.292	0.415	0.629	0.876
1	0.124	0.094	0.115	0.189	0.118	0.135	0.126	0.169	0.325	0.136	0.175	0.196	0.285	0.478	0.176	0.254	0.331	0.493	0.770
3	0.025	0.032	0.035	0.040	0.036	0.033	0.037	0.064	0.106	0.045	0.052	0.077	0.120	0.205	0.077	0.089	0.153	0.262	0.521
9	0.058	0.044	0.049	0.111	0.095	0.083	0.061	0.105	0.224	0.109	0.122	0.127	0.211	0.410	0.172	0.220	0.303	0.496	0.810
9	0.089	0.058	0.062	0.116	0.091	0.098	0.079	0.098	0.203	0.107	0.127	0.127	0.182	0.332	0.139	0.190	0.235	0.361	0.667
0	0.119	0.116	0.119	0.113	0.155	0.177	0.224	0.318	0.468	0.236	0.305	0.461	0.654	0.865	0.407	0.604	0.828	0.970	0.999
0	0.066	0.066	0.064	0.058	0.088	0.100	0.137	0.210	0.339	0.144	0.197	0.322	0.519	0.772	0.264	0.444	0.710	0.932	0.998
4	0.047	0.045	0.042	0.039	0.081	0.099	0.153	0.269	0.487	0.156	0.255	0.472	0.764	0.961	0.399	0.679	0.930	0.996	1.000
8	0.020	0.021	0.016	0.015	0.045	0.049	0.084	0.167	0.352	0.097	0.164	0.345	0.652	0.927	0.287	0.558	0.881	0.993	1.000
9	0.139	0.132	0.126	0.124	0.200	0.235	0.335	0.503	0.742	0.343	0.499	0.734	0.943	0.998	0.714	0.923	0.996	1.000	1.000
6	0.082	0.079	0.070	0.069	0.133	0.157	0.234	0.385	0.644	0.257	0.389	0.635	0.900	0.995	0.620	0.880	0.992	1.000	1.000

lts of the simulations of DGP_3. The value of the parameter b is indicated in the first row and the ries in the second. The leftmost column indicates the test or criteria (with forecasting horizon h heses; if h is not specified, $h = 1$ or it is not applicable). The figures in the remaining places are of of selecting $\{0, 1\}$ for each combination of b, length and test/criteria.

0	0.0	0.0	0.0	0.0	0.05	0.05	0.05	0.05	0.05	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2
50	100	200	400	800	50	100	200	400	800	50	100	200	400	800	50	100	200	400	800
6	0.041	0.024	0.016	0.010	0.088	0.076	0.084	0.103	0.191	0.161	0.197	0.285	0.460	0.740	0.404	0.557	0.775	0.947	0.998
6	0.046	0.033	0.022	0.016	0.081	0.073	0.081	0.094	0.152	0.135	0.161	0.224	0.352	0.588	0.300	0.429	0.634	0.845	0.980
8	0.023	0.016	0.010	0.008	0.048	0.038	0.040	0.045	0.071	0.078	0.086	0.119	0.187	0.336	0.170	0.246	0.392	0.609	0.834
3	0.099	0.074	0.062	0.048	0.156	0.153	0.179	0.231	0.385	0.251	0.309	0.440	0.656	0.890	0.514	0.683	0.874	0.980	1.000
7	0.099	0.085	0.065	0.063	0.139	0.141	0.164	0.203	0.307	0.210	0.256	0.356	0.529	0.754	0.397	0.544	0.751	0.914	0.991
0	0.056	0.048	0.040	0.035	0.086	0.083	0.097	0.119	0.188	0.127	0.156	0.226	0.335	0.527	0.243	0.361	0.533	0.737	0.907
3	0.008	0.003	0.002	0.001	0.020	0.014	0.013	0.014	0.024	0.040	0.044	0.067	0.114	0.227	0.139	0.199	0.334	0.538	0.788
6	0.011	0.005	0.004	0.001	0.021	0.017	0.014	0.016	0.025	0.037	0.042	0.062	0.100	0.192	0.107	0.154	0.272	0.442	0.668
0	0.007	0.003	0.002	0.000	0.012	0.009	0.009	0.008	0.011	0.022	0.021	0.029	0.044	0.083	0.049	0.079	0.133	0.246	0.427
9	0.017	0.009	0.006	0.004	0.039	0.031	0.034	0.046	0.082	0.071	0.083	0.129	0.222	0.407	0.206	0.280	0.453	0.661	0.863
0	0.022	0.014	0.010	0.006	0.038	0.033	0.036	0.046	0.075	0.067	0.077	0.118	0.197	0.349	0.158	0.229	0.380	0.572	0.769
3	0.016	0.007	0.007	0.004	0.026	0.023	0.023	0.025	0.043	0.042	0.046	0.065	0.109	0.207	0.087	0.132	0.230	0.384	0.575
0	0.038	0.034	0.033	0.028	0.047	0.046	0.056	0.070	0.106	0.064	0.082	0.113	0.167	0.238	0.115	0.158	0.240	0.358	0.543
2	0.054	0.043	0.038	0.037	0.079	0.059	0.057	0.069	0.086	0.090	0.087	0.098	0.132	0.175	0.119	0.137	0.177	0.255	0.377
4	0.075	0.051	0.048	0.041	0.102	0.080	0.066	0.063	0.075	0.110	0.093	0.089	0.101	0.121	0.126	0.124	0.132	0.165	0.227
1	0.020	0.017	0.017	0.014	0.026	0.025	0.031	0.037	0.063	0.035	0.047	0.067	0.106	0.157	0.067	0.096	0.157	0.259	0.427
5	0.031	0.021	0.018	0.018	0.050	0.034	0.028	0.036	0.048	0.059	0.052	0.056	0.074	0.108	0.078	0.082	0.107	0.163	0.261
6	0.048	0.026	0.025	0.019	0.074	0.051	0.039	0.033	0.038	0.079	0.059	0.050	0.054	0.065	0.088	0.079	0.079	0.096	0.141
0	0.038	0.045	0.038	0.037	0.060	0.049	0.050	0.071	0.088	0.067	0.075	0.088	0.119	0.182	0.099	0.118	0.179	0.269	0.435
2	0.078	0.077	0.108	0.191	0.102	0.089	0.082	0.130	0.251	0.111	0.104	0.117	0.171	0.307	0.141	0.149	0.161	0.243	0.423
6	0.117	0.123	0.150	0.241	0.110	0.122	0.125	0.168	0.268	0.110	0.132	0.139	0.178	0.314	0.130	0.151	0.162	0.228	0.364
2	0.017	0.019	0.016	0.018	0.032	0.023	0.022	0.029	0.044	0.037	0.038	0.041	0.061	0.108	0.055	0.066	0.110	0.177	0.320
4	0.051	0.044	0.057	0.114	0.074	0.057	0.046	0.070	0.156	0.080	0.067	0.069	0.092	0.195	0.105	0.100	0.101	0.150	0.288
3	0.079	0.076	0.086	0.158	0.084	0.092	0.084	0.103	0.174	0.088	0.099	0.090	0.104	0.206	0.103	0.111	0.103	0.133	0.241
8	0.116	0.102	0.105	0.095	0.134	0.142	0.176	0.222	0.328	0.189	0.246	0.335	0.486	0.680	0.341	0.473	0.658	0.834	0.963
3	0.061	0.051	0.053	0.047	0.073	0.079	0.101	0.135	0.215	0.111	0.150	0.222	0.360	0.547	0.221	0.345	0.527	0.747	0.929
8	0.051	0.043	0.039	0.034	0.075	0.079	0.112	0.155	0.276	0.125	0.189	0.292	0.492	0.763	0.317	0.482	0.722	0.913	0.990
1	0.027	0.019	0.018	0.013	0.041	0.046	0.062	0.092	0.177	0.080	0.119	0.203	0.373	0.658	0.231	0.374	0.624	0.869	0.985
1	0.184	0.163	0.135	0.124	0.214	0.211	0.240	0.300	0.476	0.266	0.331	0.469	0.717	0.946	0.474	0.695	0.929	0.996	1.000
0	0.138	0.114	0.090	0.074	0.163	0.158	0.176	0.218	0.361	0.210	0.250	0.369	0.606	0.896	0.386	0.591	0.876	0.993	1.000

lts of the simulations of DGP_4. The value of the parameter b is indicated in the first row and the ries in the second. The leftmost column indicates the test or criteria (with forecasting horizon h heses; if h is not specified, $h = 1$ or it is not applicable). The figures in the remaining places are of of selecting $\{0, 1\}$ for each combination of b, length and test/criteria.

0	0.0	0.0	0.0	0.0	0.05	0.05	0.05	0.05	0.05	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2
50	100	200	400	800	50	100	200	400	800	50	100	200	400	800	50	100	200	400	800
5	0.042	0.030	0.018	0.012	0.064	0.051	0.042	0.048	0.062	0.090	0.087	0.108	0.154	0.270	0.176	0.231	0.373	0.625	0.906
6	0.032	0.021	0.014	0.010	0.042	0.031	0.022	0.016	0.014	0.043	0.031	0.021	0.019	0.013	0.047	0.035	0.026	0.018	0.013
6	0.004	0.001	0.001	0.000	0.007	0.004	0.002	0.001	0.001	0.008	0.005	0.003	0.002	0.003	0.013	0.009	0.007	0.007	0.010
1	0.100	0.083	0.065	0.060	0.121	0.113	0.111	0.134	0.176	0.160	0.162	0.213	0.318	0.513	0.267	0.367	0.556	0.803	0.973
0	0.077	0.064	0.053	0.045	0.078	0.075	0.062	0.054	0.048	0.087	0.069	0.061	0.055	0.046	0.086	0.072	0.063	0.045	0.029
0	0.013	0.010	0.007	0.004	0.020	0.015	0.012	0.008	0.010	0.022	0.016	0.017	0.017	0.022	0.030	0.026	0.034	0.040	0.051
4	0.006	0.003	0.001	0.000	0.014	0.008	0.005	0.006	0.006	0.020	0.013	0.013	0.021	0.038	0.049	0.052	0.080	0.162	0.348
2	0.004	0.002	0.001	0.001	0.009	0.006	0.003	0.002	0.001	0.010	0.005	0.004	0.003	0.003	0.016	0.011	0.010	0.010	0.014
3	0.001	0.000	0.000	0.000	0.003	0.001	0.000	0.000	0.000	0.004	0.001	0.001	0.000	0.000	0.003	0.002	0.002	0.001	0.003
0	0.016	0.010	0.007	0.004	0.030	0.020	0.016	0.021	0.029	0.038	0.032	0.042	0.063	0.121	0.082	0.097	0.162	0.309	0.547
6	0.013	0.008	0.006	0.004	0.021	0.012	0.010	0.009	0.010	0.025	0.016	0.013	0.014	0.018	0.033	0.028	0.028	0.036	0.048
9	0.003	0.002	0.001	0.001	0.007	0.003	0.002	0.001	0.002	0.008	0.005	0.003	0.004	0.005	0.011	0.009	0.009	0.014	0.025
5	0.040	0.039	0.033	0.035	0.043	0.043	0.048	0.051	0.068	0.052	0.050	0.068	0.093	0.129	0.068	0.096	0.132	0.195	0.285
7	0.063	0.047	0.043	0.039	0.086	0.067	0.049	0.042	0.041	0.078	0.058	0.044	0.036	0.029	0.069	0.055	0.033	0.024	0.016
2	0.080	0.062	0.054	0.046	0.099	0.084	0.061	0.051	0.048	0.106	0.082	0.063	0.048	0.040	0.097	0.081	0.049	0.039	0.031
3	0.020	0.019	0.015	0.017	0.023	0.021	0.025	0.025	0.039	0.028	0.028	0.037	0.053	0.079	0.039	0.053	0.079	0.125	0.194
7	0.036	0.025	0.024	0.021	0.055	0.040	0.026	0.022	0.021	0.053	0.036	0.023	0.018	0.014	0.045	0.029	0.016	0.011	0.007
3	0.052	0.037	0.028	0.022	0.071	0.054	0.034	0.026	0.026	0.077	0.054	0.037	0.026	0.022	0.073	0.053	0.029	0.020	0.015
5	0.044	0.038	0.047	0.038	0.047	0.045	0.040	0.048	0.064	0.058	0.048	0.059	0.075	0.114	0.067	0.079	0.112	0.162	0.234
3	0.091	0.084	0.114	0.192	0.099	0.094	0.080	0.110	0.187	0.103	0.089	0.080	0.095	0.171	0.104	0.081	0.063	0.073	0.128
5	0.124	0.123	0.148	0.259	0.102	0.119	0.115	0.134	0.236	0.112	0.120	0.124	0.142	0.228	0.105	0.125	0.106	0.120	0.195
1	0.020	0.018	0.020	0.018	0.022	0.021	0.015	0.022	0.027	0.033	0.025	0.028	0.038	0.062	0.035	0.041	0.061	0.096	0.152
5	0.057	0.045	0.062	0.118	0.075	0.063	0.044	0.060	0.118	0.082	0.062	0.045	0.051	0.109	0.085	0.058	0.032	0.035	0.078
3	0.089	0.082	0.089	0.178	0.079	0.090	0.073	0.082	0.162	0.088	0.085	0.080	0.081	0.152	0.085	0.095	0.069	0.069	0.128
6	0.116	0.118	0.109	0.105	0.130	0.121	0.137	0.149	0.193	0.144	0.158	0.202	0.268	0.392	0.200	0.266	0.388	0.570	0.784
5	0.064	0.062	0.053	0.056	0.068	0.067	0.075	0.087	0.115	0.081	0.084	0.119	0.170	0.267	0.117	0.167	0.263	0.427	0.663
5	0.044	0.040	0.036	0.036	0.053	0.049	0.056	0.078	0.117	0.068	0.082	0.124	0.204	0.377	0.135	0.207	0.369	0.643	0.900
7	0.019	0.017	0.014	0.013	0.026	0.024	0.024	0.038	0.062	0.035	0.040	0.067	0.117	0.254	0.081	0.124	0.246	0.512	0.834
3	0.130	0.125	0.112	0.113	0.141	0.146	0.158	0.198	0.262	0.180	0.197	0.275	0.411	0.627	0.293	0.421	0.634	0.871	0.988
0	0.074	0.070	0.061	0.061	0.087	0.085	0.092	0.120	0.173	0.119	0.128	0.188	0.296	0.513	0.209	0.313	0.523	0.801	0.975

lts of the simulations of DGP_5. The value of the parameter b is indicated in the first row and the ries in the second. The leftmost column indicates the test or criteria (with forecasting horizon h heses; if h is not specified, $h = 1$ or it is not applicable). The figures in the remaining places are of of selecting $\{0, 1\}$ for each combination of b, length and test/criteria.

Bibliography

[1] Akaike H., 1973, Information Theory and an extension of the Maximum Likelihood Principle, in: B. N. Petrov and F. Csaki, (eds.), Second international symposium on information theory. Academiai Kiado: Budapest, pp. 267-281.

[2] Akaike, H., 1974, A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716-723.

[3] Clark T. E. and McCracken M. W., 2001. Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics 105, 85-110.

[4] Clark T. E. and McCracken M. W., 2007. Approximately normal tests for equal predictive accuracy in nested models. Journal of Econometrics 138, 291-311.

[5] Diebold F. and Mariano R., 1995. Comparing Predictive Accuracy. Journal of Business and Economics Statistics 13, 252-263.

[6] Forni M., Hallin M., Lippi F. and Reichlin L., 2005. The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting. Journal of the American Statistical Association 100, 830-840.

[7] Giacomini R. and White H., 2006. Tests of conditional predictive ability. Econometrica 74(6), 1545-1578.

[8] Gihman I. I. and Skorohod A. V., 1974. The theory of Stochastic Processes. Springer, New York.

[9] Granger C. W. J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424-438.

[10] Hannan E. J. and Deistler M., 1988. The Statistical Theory of Linear Systems. John Wiley and Sons, New York.

[11] Hannan E. J. and Quinn B. G., 1979. The determination of the order of an autoregression. Journal of the Royal Statistical Society B 41, 190-195.

[12] Harville, D. A., 1997. Matrix Algebra From a Statistician's Perspective. Springer, New York.

[13] Nishii, R., 1988. Maximum Likelihood Principle and Model Selection when the True Model Is Unspecified. Journal of Multivariate Analysis 27, 392-403.

[14] Peña D. and Sánchez I., 2007. Measuring the Advantages of Multivariate versus Univariate Forecasts. Journal of Time Series Analysis 28, 886-909.

[15] Schwarz G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461-464.

[16] Shibata, R., 1980. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. The Annals of Statistics 8, 147-164.

[17] Sin Ch.-Y. and White H., 1996. Information criteria for selecting possibly misspecified parametric models. Journal of Econometrics 71, 207-225.

[18] Stock J. H. and Watson, M. W., 2002. Forecasting Using Principal Components From a Large Number of Predictors. Journal of the American Statistical Association 97, 1167-1179.