Université de Montréal

Problems in Time Series and Financial Econometrics: Linear Methods for VARMA Modelling, Multivariate Volatility Analysis, Causality and Value-at-Risk

par

Denis Pelletier

Département de sciences économiques

Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures

en vue de l'obtention du grade de

Philosophiae Doctor (Ph.D.)

en sciences économiques

Mai 2004

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée :

Problems in Time Series and Financial Econometrics: Linear Methods for VARMA Modelling, Multivariate Volatility Analysis, Causality and Value-at-Risk

présentée par:

Denis Pelletier

a été évaluée par un jury composé des personnes suivantes :

Éric Renault:	président-rapporteur
Jean-Marie Dufour:	directeur de recherche
Nour Meddahi:	codirecteur de recherche
Benoit Perron:	membre du jury
Johh Galbraith:	examinateur externe (McGill University)
Roch Roy:	représentant du doyen de la FES

Sommaire

L'objectif de cette thèse est d'étudier divers problèmes d'économétrie des séries chronologiques et de la finance. Le thème qui relie les différents essais est la malédiction de la dimension qui est intrinsèque de l'étude des séries chronologiques multivariées.

Dans le premier essai, nous considérons le problème de la modélisation des modèles VARMA par des méthodes simples qui ne requièrent que des régressions linéaires. Dans ce but, nous utilisons une méthode d'estimation proposée par Hannan et Rissanen (1982, *Biometrika*) pour les modèles ARMA univariés. Nous dérivons les propriétés asymptotiques de ces estimateurs sous des hypothèses faibles à propos des innovations (non corrélées et mélangeantes fortes) afin d'élargir la classe de modèles auxquels ils peuvent être appliqués.

Pour faciliter l'utilisation des modèles VARMA, nous présentons des nouvelles représentations identifiées, la *forme équation diagonale MA* et la *forme équation finale MA*, où les opérateurs MA sont respectivement diagonaux et scalaires. Nous présentons également un critère d'information modifié qui donne des estimations convergentes des ordres de ces différentes représentations. Pour démontrer l'importance des modèles VARMA dans l'étude des séries chronologiques multivariées, nous comparons les coefficients d'impulsion générés par des modèles VARMA et VAR.

Dans le deuxième essai, nous proposons un nouveau modèle pour la variance entre plusieurs séries chronologiques, le modèle *Regime Switching Dynamic Correlation*. Nous décomposons les covariances en corrélations et écarts types. La matrice de corrélation suit un modèle à changement de régime : elle est constante à l'intérieur d'un régime mais différente entre les régimes. Les transitions entre ceux-ci sont déterminées par une chaîne de Markov. Ce modèle ne souffre pas d'une malédiction de la dimension et permet le calcul analytique d'espérances conditionnelles sur plusieurs horizons de la matrice de variance. Nous présentons également une application empirique qui illustre que notre modèle peut obtenir une meilleure performance interéchantillon que le modèle *Dynamic Conditional Correlation* proposé par Engle (2002, *JBES*).

Dans le troisième essai, nous examinons des méthodes pour tester des hypothèses de non-causalité à différents horizons, tel qu'ils sont définis dans Dufour et Renault

(1998, *Econometrica*). Nous étudions en détail le cas des modèles VAR et nous proposons des méthodes linéaires basées sur l'estimation d'autorégressions vectorielles à différents horizons. Même si les hypothèses considérées sont non linéaires, ces méthodes ne requièrent que des techniques de régression linéaire de même que la théorie distributionnelle asymptotique gaussienne habituelle. Dans le cas des processus intégrés, nous avons recours des méthodes de régression étendue qui n'exigent pas de théorie asymptotique non standard. Les méthodes sont appliquées à un modèle VAR de l'économie américaine.

Dans le quatrième essai, nous proposons des nouveaux tests statistiques pour l'évaluation des modèles de risque financier utilisés pour le calcul des Valeurs-à-Risque (VaR), tel que le modèle dont il est question dans le deuxième essai. Ces tests sont basés sur la durée en jours entre les violations de la VaR. Les résultats de nos simulations Monte Carlo montrent que pour des situations réalistes, les tests basés sur les durées donnent de meilleures propriétés en matière de puissance que ceux précédemment avancés.

Mots clés : équation forme finale, critère d'information, représentation faible, coefficients d'impulsion, corrélation dynamique, chaîne de Markov, causalité indirecte, autorégression vectorielle, GARCH, évaluation de modèle de risque.

Summary

The objective of this thesis is to study various problems in time series and financial econometrics. The common thread of the various parts is the intrinsic curse of dimensionality underlying the study of multivariate time series.

In the first essay, we consider the problem of modelling VARMA models by relatively simple methods which require linear regressions. For that purpose, we consider the regression-based estimation method proposed by Hannan and Rissanen (1982, *Biometrika*) for univariate ARMA models. The asymptotic properties of the estimator are derived under weak hypotheses for the innovations (uncorrelated and strong mixing) so as to broaden the class of models to which it can be applied.

To further ease the use of VARMA models we present new identified VARMA representations, *diagonal MA equation form* and *final MA equation form*, where the MA operators are diagonal and scalar respectively. We also present a modified information criterion which gives consistent estimates of the orders of these representations. To demonstrate the importance of using VARMA models to study multivariate time series we compare the impulse-response functions generated by VARMA and VAR models.

In the second essay, we propose a new model for the variance between multiple time series, the Regime Switching Dynamic Correlation model. In this model, we decompose the covariances into correlations and standard deviations. The correlation matrix follows a regime switching model: it is constant within a regime but different across regimes. The transitions between the regimes are governed by a Markov chain. This model does not suffer from a curse of dimensionality and it allows analytic computation of multi-step ahead conditional expectations of the variance matrix. We also present an empirical application which illustrates that our model can have a better in-sample fit of the data than the Dynamic Conditional Correlation model proposed by Engle (2002, *JBES*).

In the third essay, we discuss methods for testing hypothesis of non-causality at various horizons, as defined in Dufour and Renault (1998, *Econometrica*). We study in detail the case of VAR models and we propose linear methods based on running vector autoregressions at different horizons. While the hypotheses considered are nonlinear, the proposed methods only require linear regression techniques as well as standard

Gaussian asymptotic distributional theory. For the case of integrated processes, we propose extended regression methods that avoid nonstandard asymptotics. The methods are applied to a VAR model of the U.S. economy.

In the fourth essay, we propose new statistical tests for backtesting financial risk models used for computing Value-at-Risk (VaR), like the model we proposed in the second essay. These tests are based on the duration in days between the violations of the VaR. Our Monte Carlo results show that in realistic situations, the new duration-based tests have considerably better power properties than the previously suggested tests.

Key words: final equation form, information criterion, weak representation, impulse-response functions, dynamic correlation, Markov chain, indirect causality, vector autoregression, GARCH, risk model evaluation.

Contents

List of Tables

List of Figures

Remerciements

J'ai de nombreuses personnes et institutions à remercier. Tout d'abord mes directeurs de recherche, Jean-Marie Dufour et Nour Meddahi. Je m'estime chanceux d'avoir côtoyé des scientifiques et des personnes de leur calibre. J'ai très apprécié leur support moral sans lequel j'aurais sûrement abandonné en cours de route.

Je voudrais également remercier mes autres co-auteurs pour les essais de cette thèse, Peter Christoffersen et Éric Renault[1]. Tous deux m'ont également beaucoup appris et je suis heureux de m'être fait deux amis par le coup même.

Je voudrais également remercier les différents organismes qui m'ont accordé bourses et financements sans lesquels je n'aurais pu écrire cette thèse : conseil de recherches en sciences humaines du Canada (CRSH), Fonds québécois de la recherche sur la nature et les technologies, centre interuniversitaire de recherche en économie quantitative (CIREQ), le centre interuniversitaire de recherche en analyse des organisations (CIRANO). Je voudrais remercier North Carolina State University pour leur support lors de la rédaction finale de cette thèse.

Je veux remercier mon épouse Paule de m'avoir non seulement enduré, mais encouragé dans les périodes les plus exigeantes de mes études.

[1]Le premier essai a été écrit en collaboration avec Jean-Marie Dufour, le troisième avec Jean-Marie Dufour et Éric Renault, et le quatrième avec Peter Christoffersen.

Introduction

Un des problèmes intrinsèques de l'étude des séries chronologiques multivariées est la malédiction de la dimension. Bien souvent, la complexité et le nombre de paramètres des modèles que l'on tente d'utiliser augmentent avec le nombre de séries chronologiques, ce qui rend l'analyse de telles séries très difficile, voire impossible. La ligne directrice de cette thèse est l'étude de méthodes permettant de contourner cette malédiction de la dimension, pour les séries tant macroéconomiques que financières.

Pour étudier la dynamique des séries chronologiques macroéconomiques, les économistes se servent la plupart du temps des modèles VAR. Le grand attrait de ces modèles est que leur estimation ne requièrt que des régressions linéaires, ce qui les rend très faciles d'utilisation.

En revanche, l'utilisation des modèles VAR a deux grands défauts. Le premier est le manque de parcimonie. Tout comme il est admis que les modèles ARMA sont plus parcimonieux que les modèles AR pour les séries univariées, les modèles VARMA ont le potentiel d'être plus parcimonieux que les modèles VAR, surtout lorsqu'on remarque que des modèles VAR avec des ordres très élevés sont nécessaires pour de nombreuses séries macroéconomiques.

Le deuxième défaut est que la spécification d'un modèle VAR est très arbitraire puisque cette classe de modèles n'est pas robuste à l'agrégation temporelle et à la marginalisation. Si un vecteur suit un processus VAR, des sous-vecteurs ne suivent pas typiquement des modèles VAR (mais des processus VARMA). De la même façon, si un processus VAR est observé à une fréquence différente, alors la série obtenue ne suit pas un modèle VAR mais un processus VARMA. Par opposition, l'agrégation temporelle ou la marginalisation d'un processus VARMA demeure un processus VARMA.

Les économistes persistent tout de même à utiliser seulement les modèles VAR au lieu d'envisager les modèles VARMA, ce qu'on peut expliquer par deux raisons. La première est que la représentation VARMA identifiée privilégiée par la littérature économétrique, i.e. la forme échelon, est difficile à manipuler. L'utilisateur doit spécifier les indices de Kronecker (le nombre d'indices est égal au nombre de séries), et les ordres des polynômes composant les opérateurs AR et MA sont fonction de leur po-

sition relativement à la diagonale. La seconde raison est que la méthode d'estimation habituellement proposée pour les modèles VARMA est le maximum de la vraisemblance. Les modèles VARMA sont plus parcimonieux que les modèles VAR mais le nombre de paramètres peut être élevé, ce qui rend très compliquée la maximisation de la vraisemblance.

Dans le premier essai de cette thèse, nous présentons une méthode pour la modélisation des modèles VARMA qui franchit ces deux obstacles. Dans un premier temps, nous introduisons deux nouvelles représentations VARMA identifiées, la *forme équation diagonale MA* et la *forme équation finale MA*, où les opérateurs MA sont respectivement diagonaux et scalaires. Ces représentations ont de nombreux avantages. Elles peuvent être interprétées comme de simples extensions du modèle VAR. Contrairement à la forme échelon, elles imposent une forme très simple sur la partie MA, celle qui complexifie l'utilisation des modèles VARMA. Les ordres des polynômes qui compose la partie MA ne sont pas reliés entre eux, contrairement à la forme échelon.

Dans un second temps, nous proposons une méthode d'estimation qui ne requiert que trois régressions linéaires. Cette méthode est une généralisation de celle proposée par Hannan et Rissanen (1982) pour les modèles ARMA. Les estimateurs de la troisième régression ont les mêmes propriétés asymptotiques que ceux obtenus par maximum de vraisemblance sous l'hypothèse que les innovations sont gaussiennes. Avec cette méthode d'estimation, nous combinons un critère d'information qui donne des estimations convergentes des ordres des polynômes AR et MA.

Pour l'étude des séries financières, la malédiction de la dimension force les économistes à utiliser des modèles aux dynamiques très simples. Les généralisations multivariées directes des modèles GARCH univariés, tel que le modèle BEKK de Engle et Kroner (1995), ne peuvent être appliquées à plus de quatre ou cinq séries sans quoi la maximisation de la vraisemblance devient prohibitive [voir Ding et Engle (2001)]. Une avenue intéressante pour la spécification des modèles de volatilité multivariés est la décomposition des covariances en corrélations et écarts types. Le chercheur spécifie ensuite des modèles pour les écarts types et un modèle pour la matrice de corrélation. On se débarrasse ainsi de la malédiction de la dimension puisqu'on peut estimer le modèle deux étapes : d'abord pour les écarts types puis ensuite pour la matrice de cor-

rélation en utilisant les résidus standardisés. Le premier à utiliser cette décomposition a été Bollerslev (1990), en posant l'hypothèse que les corrélations sont constantes.

L'hypothèse selon laquelle la matrice de corrélation est constante n'étant pas toujours appuyée par les données, de nouveaux modèles ont été proposés au cours des dernières années. Les modèles *Dynamic Conditional Correlations* de Engle (2002) et *Multivariate GARCH* de Tse et Tsui (2002) avancent plutôt une dynamique de type GARCH pour la matrice de corrélation : la matrice de corrélation est aujourd'hui une fonction des matrices de corrélation passées et des produits croisés des innovations standardisées passées.

On préfère ces modèles à ceux qui ont une matrice de corrélation constante, mais une dynamique de type GARCH pour la matrice de corrélation n'est pas entièrement satisfaisante. Le fait que les produits croisés des innovations standardisées ne soit pas borné par -1 et 1 est un problème, puisque cela implique qu'aucun élément de la matrice de corrélation n'est borné par -1 et 1. Par conséquent, on doit remettre à l'échelle les matrices obtenues afin de vraiment aboutir à des matrices de corrélation, mais ces mises à l'échelle introduisent des non-linéarités qui ont pour effet d'empêcher les calculs analytiques d'espérance conditionnelle pour les covariances et corrélations. On s'aperçoit qu'un modèle qui ne tient pas directement compte des caractéristiques d'une matrice de corrélation n'est pas satisfaisant.

Dans le deuxième essai, nous proposons un nouveau modèle de volatilité multivarié, le modèle *Regime Switching Dynamic Correlation*. Nous décomposons également les covariances en corrélations et écarts types, mais la matrice de corrélations suit un modèle à changement de régime : elle est constante à l'intérieur d'un régime mais différente entre régimes. Les transitions entre régimes sont déterminées par une chaîne de Markov. Ce modèle ne souffre pas d'une malédiction de la dimension puisqu'on peut l'estimer en deux étapes, tout comme les modèles de Bollerslev (1990), Engle (2002), Tse et Tsui (2002). Notre modèle a aussi l'avantage de permettre le calcul analytique d'espérance conditionnelle sur plusieurs horizons de la matrice de corrélation, et de la matrice de variance si un modèle approprié pour les écarts types est employé [le modèle ARMACH de Taylor (1986)]. Nous présentons également une application empirique qui montre que notre modèle peut avoir une meilleure performance inter-échantillon

que celui d'Engle (2002).

Les tests de causalité à plusieurs horizons, tel que définis dans Dufour et Renault (1998), présentent également des problèmes associés à l'étude des séries macroéconomiques multivariées. Même dans les modèles VAR, les hypothèses de causalité à plusieurs horizons sont non linéaires et prennent la forme de contraintes sur des transformations multilinéaires des paramètres du modèle VAR. L'application des tests statistiques habituels, de type Wald, par exemple, pourrait générer des matrices de covariance asymptotiquement singulières, avec comme résultat que la théorie asymptotique standard ne s'appliquerait pas à ces statistiques.

C'est pourquoi nous présentons, dans le troisième essai, des méthodes de test simples pour tester les hypothèses de non-causalité à plusieurs horizons dans les modèles VAR d'ordre fini qui ne requièrent que des méthodes de régression linéaire. Celles-ci méthodes sont basées sur des autorégressions vectorielles à multiples horizons où on peut estimer les paramètres au moyen de méthodes linéaires. En utilisant cette approche, on peut tester les restrictions de non-causalité à divers horizons en utilisant des critères de type Wald ou Fisher, une fois que l'on tient compte de la structure moyenne mobile des erreurs (qui sont orthogonales aux régresseurs).

Une des raisons d'être des modèles de volatilité multivariés tels que celui que nous présentons dans le deuxième essai est de prédire la distribution de rendements futurs d'un portefeuille. Ces prédictions sont nécessaires pour le calcul de la Valeur-à-Risque (VaR) d'un portefeuille d'actifs financiers. La VaR d'un portefeuille est tout simplement un quantile de la distribution des rendements futurs du portefeuille. C'est une mesure du risque d'un portefeuille : plus ses rendements sont volatils, plus la variance est élevée, et plus les petits quantiles sont éloignés de la moyenne. Les institutions financières sont maintenant tenues de calculer ces VaR par, notamment, les Accords de Basle.

Dans le quatrième essai, nous présentons de nouveaux tests statistiques pour évaluer si le modèle utilisé pour calculer la VaR est correctement spécifié. Si aujourd'hui la VaR pour demain et pour un niveau de couverture de 1 % est 10 000$, cela signifie que demain, la probabilité que ce portefeuille perde plus que 10 000$ est égale à 1 %. L'évaluation des modèles utilisés pour calculer les VaR est basée sur la comparaison

des VaR (ex-ante) et des pertes effectives (ex-post). On crée ainsi une séquence binaire I_t : on marque un 1 pour les jours où les pertes excèdent la VaR et un 0 pour les jours où la VaR n'excède pas les pertes.

Si la VaR est calculée de façon optimale, il devrait être impossible de prévoir à quel moment elle sera violée (quand les pertes vont excéder la VaR), ce qui implique que la séquence I_t devrait être indépendante. Si on calcule une VaR avec un niveau de couverture de $p\%$, alors on devrait excéder la VaR p % des jours. Donc, ce qui nous intéresse, c'est de vérifier si la séquence I_t est i.i.d. Bernoulli(p). Des tests basés sur l'hypothèse alternative d'une chaîne de Markov pour décrire la séquence I_t ont été avancés par Christoffersen (1998).

Dans cet essai, nous proposons des nouveaux tests statistiques qui sont basés sur la durée en nombre de jours entre les violations de la VaR. Si le modèle utilisé pour calculer la VaR est optimale, alors ces durées devraient être i.i.d. exponentielles de moyenne $1/p$. S'il est impossible de prévoir quand la VaR sera violée, il ne peut y avoir d'effet de mémoire et si la VaR est excédée p % du temps, on devra attendre $1/p$ jours en moyenne entre les violations. Pour tester cette hypothèse, nous proposons deux alternatives qui englobent le cas i.i.d. exponentiel : la distribution Weibull et le modèle EACD de Engle et Russel (1998). À l'aide de simulations Monte Carlo, nous montrons que pour des situations réalistes, ces tests ont plus de puissance que ceux proposés précédemment.

Chapter 1: Linear estimation of weak VARMA models with a macroeconomic application

1. Introduction

In time series analysis and econometrics, VARMA models are scarcely used to represent multivariate time series. VAR models are much more widely employed because they are easier to implement: the latter models can be estimated by least squares methods, while VARMA models typically require nonlinear methods (such as maximum likelihood).

VAR models, however, have important drawbacks. First, they are typically much less parsimonious than VARMA models. Second, the family of VAR models is not closed under marginalization and temporal aggregation. If a vector satisfies a VAR model, subvectors do not typically satisfy VAR models (but VARMA models). Similarly, if the variables of a VAR process are observed at a different frequency, the resulting process is not a VAR process. In contrast, the class of (weak) VARMA models is closed under such operations. We say that a VARMA model is strong if the innovations are independent, and it is weak if they are merely uncorrelated.

It follows that VARMA models appear to be preferable from a theoretical viewpoint, but their adoption is complicated by identification issues and estimation difficulties. The direct multivariate generalization of ARMA models does not give an identified representation. It follows that a one has to decide on a set of constraints to impose so as to gain identification. Standard estimation methods for VARMA models (maximum likelihood, nonlinear least squares) require nonlinear optimization which may not be feasible as soon as the model involves a few time series, because the number of parameters can increase quickly.

In this paper, we consider the problem of estimating VARMA models by relatively simple methods which only require linear regressions. For that purpose, we consider

a generalization by Hannan and Kavalieris (1984a) of the regression-based estimation method proposed by Hannan and Rissanen (1982) for univariate ARMA models. Their method is performed in three steps. In a first step a long autoregression is fitted to the data. In the second step, the lagged innovations in the ARMA model are replaced by the corresponding residuals from the long autoregression and a regression is performed. In a third step, the data from the second step are filtered so as to give estimates that have the same asymptotic covariance matrix than one would get with the maximum likelihood [claimed in Hannan and Rissanen (1982), proven in Zhao-Guo (1985)]. Extension of this innovation-substitution method to VARMA models was also proposed by Koreisha and Pukkila (1989), but these authors did not provide a detailed asymptotic theory for their proposed extension.

Here, we first provide such a theory by showing that the linear regression-based estimators are consistent under weak hypotheses on the innovations and how filtering in the third step gives estimators that have the same asymptotic distribution as their nonlinear counterparts (maximum likelihood if the innovations are independent and identically distributed (i.i.d.), or nonlinear least squares if they are merely uncorrelated). In the non i.i.d. case, we consider strong mixing conditions [Doukhan (1995), Bosq (1998)], rather than the usual martingale difference sequence (m.d.s.) assumption. By using weaker assumptions for the process of the innovations we broaden the class of models to which our method can be applied.[1]

Second, in order to avoid identification problems and to further ease the use of VARMA models, we introduce three new identified VARMA representations, the *diagonal MA equation form*, the *final MA equation form* and the *diagonal AR equation form*. Under the diagonal MA equation form (diagonal AR equation form) representation, the MA (AR) operator is diagonal and each lag operator may have a different order q_i (p_i). Under the final MA equation form representation the MA operator is scalar, i.e. the the operators are equal across equations. The diagonal and final MA equation form representations can be interpreted as simple extensions of the VAR model, which should be appealing to practitioners who prefer to employ VAR models due to their ease of use.

[1] For univariate ARMA models Francq and Zakoïan (1998) presents numerous cases where the representation is only weak.

The identified VARMA representation that is the most widely employed in the literature is the *echelon form*. Specification of VARMA models in echelon form does not amount to specifying the order p and q as with ARMA models. Under this representation, VARMA models are specified by as many parameters, called Kronecker indices, as the number of time series studied. These indices determine the order of the elements of the AR and MA operators in a non trivial way. The complicated nature of the echelon form representation might be a reason why practitioners are not using VARMA models, so the introduction of a simpler identified representation is interesting. The proposed representations may be less parsimonious than the echelon form but since our estimation method only involve regressions we can afford it.

Thirdly, we suggest a modified information criterion to choose the orders of VARMA models under these representations. This criterion is to be minimized in the second step of the estimation method over the orders of the AR and MA operators and gives consistent estimates of these orders. Our criterion is a generalization of the information criterion proposed by Hannan and Rissanen (1982), which was corrected later on in Hannan and Rissanen (1983, 1984b), for choosing the orders p and q in ARMA models. The idea of generalizing this information criterion is mentioned in Koreisha and Pukkila (1989) but a specific generalization and theoretical properties are not presented.

Fourth, the method is applied to U.S. macroeconomic data previously studied by Bernanke and Mihov (1998) and McMillin (2001). To illustrate the impact of using VARMA models instead of VAR models to study multivariate time series we compare the impulse-response functions generated by each model. We show that we can obtain much more precise estimates of the impulse-response function by using VARMA models instead of VAR models.

The rest of the paper is organized as follows. Our framework and notation are introduced in section 2. The new identified representations are presented in section 3. In section 4, we present the estimation method. In section 5, we describe the information criterion used for choosing the orders of VARMA models under the representation proposed in our work. Section 6 contains results of Monte Carlo simulations which illustrate the properties of our method. Section 7 presents the macroeconomic

application where we compare the impulse-response functions from a VAR model and VARMA models. Section 8 contains a few concluding remarks. Finally, proofs are in the appendix.

2. Framework

Consider the following K-variate zero mean VARMA(p,q) model in standard representation:

$$Y_t = \sum_{i=1}^{p} A_i Y_{t-i} + U_t - \sum_{j=1}^{q} B_j U_{t-j} \tag{2.1}$$

where U_t is a sequence of uncorrelated random variables defined on some probability space (Ω, $\mathcal{A}$, $\mathcal{P}$). The vectors Y_t and U_t contain the K univariate time series: $Y_t = [y_t(1), y_t(2), \ldots, y_t(K)]'$ and $U_t = [u_t(1), u_t(2), \ldots, u_t(K)]'$. We can also write the previous equation with lag operators:

$$A(L)Y_t = B(L)U_t \tag{2.2}$$

where

$$A(L) = I_K - A_1 L - \cdots - A_p L^p, \tag{2.3}$$

$$B(L) = I_K - B_1 L - \cdots - B_q L^q. \tag{2.4}$$

Let H_t be the Hilbert space generated by $(Y_s,\ s < t)$. The process (U_t) can be interpreted as the linear innovation of Y_t:

$$U_t = Y_t - EL[Y_t|H_t]. \tag{2.5}$$

Also assume that Y_t is a strictly stationary and ergodic sequence and that the process $\{U_t\}$ has common variance ($Var[U_t] = \Sigma$) and finite fourth moment ($E[|u_t(i)|^{4+2\delta}] < \infty$ for some $\delta > 0$). We make the zero mean-mean hypothesis only to simplify the

notation.

Assuming that the process $\{Y_t\}$ is stable,

$$\det[A(z)] \neq 0 \text{ for } |z| \leq 1, \tag{2.6}$$

and invertible,

$$\det[B(z)] \neq 0 \text{ for } |z| \leq 1, \tag{2.7}$$

it, can be represented as an infinite VAR

$$\Pi(L)Y_t = U_t,$$

where

$$\Pi(L) = B(L)^{-1}A(L) = I_K - \sum_{i=1}^{\infty} \Pi_i L^i,$$

or an infinite VMA

$$Y_t = \Psi(L)U_t,$$

where

$$\Psi(L) = A(L)^{-1}B(L) = I_K - \sum_{j=1}^{\infty} \Psi_j L^j.$$

The matrices Π_i and Ψ_j could be zero past a finite order if $\det[B(L)]$ or $\det[A(L)]$ respectively is a non-zero constant. We will denote by $a_{ij}(L)$ the polynomial in row i and column j of $A(L)$, and the row i or column j of $A(L)$ by

$$A_{i\bullet}(L) = [a_{i1}(L), \ldots, a_{iK}(L)], \tag{2.8}$$

$$A_{\bullet j}(L) = [a_{1j}(L), \ldots, a_{Kj}(L)]'. \tag{2.9}$$

The *diag* operator creates a diagonal matrix,

$$\begin{aligned} diag[a_{ii}(L)] &= diag[a_{11}(L), \ldots, a_{KK}(L)] \\ &= \begin{bmatrix} a_{11}(L) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{KK}(L) \end{bmatrix}, \end{aligned} \tag{2.10}$$

where

$$a_{ii}(L) = 1 - a_{ii,1}L - \cdots - a_{ii,p}L^p. \tag{2.11}$$

The function $deg[a(L)]$ returns the degree of the polynomial $a(L)$ and the function $dim(\gamma)$ gives the dimension of the vector γ.

We need to impose a minimum of structure on the process $\{U_t\}$ because saying that it is uncorrelated is not enough to get any significant results. The typical hypothesis that is imposed in the time series literature is that the U_t's are either independent and identically distributed (i.i.d.) or a martingale difference sequence (m.d.s.). In this work we do not impose such strong assumptions because we want to broaden the class of models to which it can be applied. We only assume that it satisfies a strong mixing condition [Doukhan (1995), Bosq (1998)]. Let $\{U_t\}$ be a strictly stationary process, then its α-mixing coefficient of order h is defined as

$$\alpha(h) = \sup_{\substack{B \in \sigma(U_s, s \leq t) \\ C \in \sigma(U_s, s \geq t+h)}} |\Pr(B \cap C) - \Pr(B)\Pr(C)| \ , \ h \geq 1. \tag{2.12}$$

The strong mixing condition that we impose is

$$\sum_{h=1}^{\infty} \alpha(h)^{\delta/(2+\delta)} < \infty \quad \text{for some} \quad \delta > 0. \tag{2.13}$$

This is a fairly minimal condition that will be satisfied by many processes of interest.

3. Diagonal VARMA representations

It is important to note that we cannot work with the standard representation (2.1) because it is not identified. To help us gain intuition on the identification of VARMA models we can consider a more general representation where A_0 and B_0 are not identity matrices:

$$A_0 Y_t = A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + B_0 \bar{U}_t - B_1 \bar{U}_{t-1} + \cdots + B_q \bar{U}_{t-q}. \tag{3.1}$$

By this specification, we mean the well-defined process

$$Y_t \;=\; (A_0 - A_1 L - \cdots - A_p L^p)^{-1}(B_0 + B_1 L + \cdots + B_q L^q)\bar{U}(t).$$

But we can see that such process has a standard representation if A_0 and B_0 are non-singular. To see this we left-multiply (3.1) by ${A_0}^{-1}$ and define $U_t = {A_0}^{-1} B_0 \bar{U}_t$:

$$\begin{aligned} Y_t \;=\; & {A_0}^{-1} A_1 Y_{t-1} + \cdots + {A_0}^{-1} A_p Y_{t-p} + U(t) - \\ & {A_0}^{-1} B_1 {B_0}^{-1} A_0 U_{t-1} - \cdots - {A_0}^{-1} B_q {B_0}^{-1} A_0 U_{t-q}. \end{aligned}$$

Redefining the matrices we get a representation of the type (2.1). With this example we see that as long as A_0 and B_0 are non-singular we can transform a non-standard VARMA into a standard one.

We say that two VARMA representations are equivalent if $A(L)^{-1}B(L)$ results in the same operator $\Psi(L)$. Thus, to insure uniqueness of a VARMA representation we must impose restrictions on the AR and MA operators such that for a given $\Psi(L)$ there is one and only one set of operators $A(L)$ and $B(L)$ that can generate this infinite MA representation.

A first restriction that we impose is a multivariate equivalent of the coprime property in the univariate case. We don't want elements of $A(L)$ and $B(L)$ to "cancel out" when we take $A(L)^{-1}B(L)$. We call this the *left-coprime* property [see Hannan (1969), Lütkepohl (1993a)]. It may be defined by calling the matrix operator $\Psi[A(L), B(L)] = A(L)^{-1}B(L)$ left-coprime if the existence of operators $D(L)$, $\bar{A}(L)$,

and $\bar{B}(L)$ satisfying

$$D(L)\Psi[\bar{A}(L), \bar{B}(L)] = \Psi[A(L), B(L)] \tag{3.2}$$

implies that $D(L)$ is unimodular, that is $\det D(L)$ is a nonzero constant. To obtain uniqueness of left-coprime operators we have to impose restrictions ensuring that the only feasible unimodular operator $D(L)$ in (3.2) is $D(L) = I_K$. There exist more than one representation which guarantee the uniqueness of the left-coprime operators. The predominant representation in the literature is the *echelon form* [see Deistler and Hannan (1981), Hannan and Kavalieris (1984b), Lütkepohl (1993a), Lütkepohl and Poskitt (1996a)].

Definition 3.1 (Echelon form) *The VARMA representation in (2.1) is said to be in echelon form if the AR and MA operators $A(L) = [a_{ij}(L)]_{i,j=1,\ldots,K}$ and $B(L) = [b_{ij}(L)]_{i,j=1,\ldots,K}$ satisfy the following conditions: all operators $a_{ij}(L)$ and $b_{ij}(L)$ in the i-th row of $A(L)$ and $B(L)$ have the same degree p_i and have the form*

$$\begin{aligned}
a_{ii}(L) &= 1 - \sum_{m=1}^{p_i} a_{ii,m} L^m, \quad \text{for } i = 1, \ldots, K \\
a_{ij}(L) &= - \sum_{m=p_i - p_{ij}+1}^{p_i} a_{ij,m} L^m, \quad \text{for } j \neq i \\
b_{ij}(L) &= \sum_{m=0}^{p_i} b_{ij,m} L^m \quad \text{for } i, j = 1, \ldots, K, \quad \text{with } B_0 = A_0.
\end{aligned}$$

Further, in the VAR operator $a_{ij}(L)$,

$$p_{ij} = \begin{cases} \min(p_i + 1, p_j) & \text{for } i \geq j \\ \min(p_i, p_j) & \text{for } i < j \end{cases} \qquad i, j = 1, \ldots, K.$$

i.e., p_{ij} specifies the number of free coefficients in the operator $a_{ij}(L)$ for $j \neq i$. The row orders $(p_1, \ldots, p_K)$ are the Kronecker indices and their sum $\sum_{i=1}^{K} p_i$ is the McMillan degree. For the VARMA orders we have in general $p = q = max(p_1, \ldots, p_K)$.

We see that dealing with VARMA models in echelon form is not as easy as dealing with univariate ARMA models where everything is specified by choosing the value of

p and q. The number of Kronecker indices is bigger than two (if K is bigger than two) and when choosing p_{ij} we have to consider if we are above or below the diagonal. Having a summation subscript in the operator a_{ij}, $m = p_i - p_{ij} + 1$, different across rows and columns also complicates the use of this representation. The task is far from being impossible but it is more complicated than for ARMA models. Specification of VARMA models in echelon form is discussed in Hannan and Kavalieris (1984b), Lütkepohl and Claessen (1997), Poskitt (1992), Nsiri and Roy (1992), Nsiri and Roy (1996), Lütkepohl and Poskitt (1996b), Bartel and Lütkepohl (1998). This might be a reason why practitioners are reluctant to employ VARMA models. Who could blame them for sticking with VAR models when they probably need to refer to a textbook to simply write down an identified VARMA representation?

In this work, to ease the use of VARMA models we present new VARMA representations which can be seen as a simple extensions of the VAR model. To introduce them, we first review another identified representation, the *final equation form*, which will refer to as the *final AR equation form*, under which the AR operator is scalar [see Zellner and Palm (1974), Hannan (1976), Wallis (1977), Lütkepohl (1993a)].

Definition 3.2 (Final AR equation form) *The VARMA representation (2.1) is said to be in final AR equation form if $A(L) = a(L)I_K$, where $a(L) = 1 - a_1 L - \cdots - a_p L^p$ is a scalar polynomial with $a_p \neq 0$.*

To see how we can obtain a VARMA model with a final AR equation form representation, we can proceed as follows. By standard linear algebra, we have

$$A^*(L)A(L) = A(L)A(L)^* = \det\left[A(L)\right] I_K$$

where $A^*(L)$ is the adjoint matrix of $A(L)$. On multiplying both sides of (2.2) by $A^*(L)$, we get:

$$\det\left[A(L)\right] Y_t \;=\; A(L)^* B(L) U_t.$$

This representation may not be attractive for several reasons. First, it is quite far from usual VAR models by excluding lagged values of other variables in each equation

[e.g., the AR part of the first equation include lagged values of $y_t(1)$ but no lagged values of $y_t(2), \ldots, y_t(K)$]. Further, the AR coefficients are the same in all the equations, which will require a polynomial of higher order ($p\,K$). Second, the interaction between the different variables are modeled through the MA part of the model, which may have to be quite complex.

We can obtain our new representations with analogous manipulations. Upon multiplying both sides of (2.2) by $B^{\star}(L)$, we get:

$$B(L)^{\star}A(L)Y_t = \det\left[B(L)\right]U_t \tag{3.3}$$

where $B(L)^{\star}$ is the adjoint matrix of $B(L)$. We refer to VARMA models in (3.3) as being in *final MA equation form.*

Definition 3.3 (Final MA equation form) *The VARMA representation (2.1) is said to be in final MA equation form if $B(L) = b(L)I_K$, where $b(L) = 1 - b_1 L - \cdots - b_q L^q$ is a scalar operator with $b_q \neq 0$.*

The same criticism regarding the parsimony of the final equation form would apply but it is possible to get a more parsimonious representation by looking at common structures across equations. Suppose there are common roots across rows for some columns of $B(L)$, so that starting from (2.1) we can write

$$\begin{aligned} A(L)Y_t &= \bar{B}(L)D(L)U_t \\ \bar{B}(L)^{\star}A(L)Y_t &= \det\left[\bar{B}(L)\right]D(L)U_t \end{aligned} \tag{3.4}$$

where $D(L) = diag[d_1(L), \ldots, d_K(L)]$ and $d_j(L)$ is a polynomial common to $b_{ij}(L)$, $\forall i = 1, \ldots, K$. We see that allowing non-equal diagonal polynomials in the moving average as in equation (3.4) may give a more parsimonious representation than in (3.3). We will call the representation (3.4) *diagonal MA equation form* representation.

Definition 3.4 (Diagonal MA equation form) *The VARMA representation (2.1) is said to be in diagonal MA equation form if $B(L) = diag[b_{ii}(L)] = I_K - B_1 L - \cdots - B_q L^q$ where $b_{ii}(L) = 1 - b_{ii,1}L - \cdots - b_{ii,q_i}L^{q_i}$, $b_{ii,q_i} \neq 0$, and $q = \max_{1 \leq i \leq K}(q_i)$.*

This representation is interesting because contrary to the echelon form it is easy to specify. We don't have to deal with rules for the orders of the off-diagonal elements in the AR and MA operators. The fact that it can be seen as a simple extension of the VAR model is appealing. Practitioners are comfortable using VAR models, so simply adding lags of $u_i(t)$ to equation i is a natural extension of the VAR model which could give a more parsimonious representation. It also has the advantage of putting the simple structure on the MA part, the part which complicates the estimation, instead of on the AR part as in the final AR equation form. Notice that in VARMA models, it is not necessary to include lags of all the innovations $u_1(t), \cdots, u_K(t)$ in every equations. This could entice practitioners to consider VARMA models if it is combined with a simple regression-based estimation method. For this representation to be useful, it needs to be identified. This is demonstrated in Theorem **3.11** below under the following assumptions and using Lemma **3.8** below

Assumption 3.5 *The matrices $A(z)$ and $B(z)$ have the following form:*

$$\begin{aligned} A(z) &= I_K - A_1 z - \cdots - A_p z^p \\ B(z) &= I_K - B_1 z - \cdots - B_q z^q \end{aligned}$$

Assumption 3.6 *$B(z)$ is diagonal:*

$$B(z) = diag\left[b_{ii}(z)\right]$$

with $b_{ii}(z) = 1 - b_{ii,1} z - \cdots - b_{ii,q_i} z^{q_i}$, $b_{ii,q_i} \neq 0$.

Assumption 3.7 *For each $i = 1, \ldots, K$, there are no roots common to $A_{i\bullet}(z)$ and $b_{ii}(z)$, i.e. there is no value $z^\star$ such that $A_{i\bullet}(z^\star) = 0$ and $b_{ii}(z^\star) = 0$.*

Lemma 3.8 *Let $[A(z), B(z)]$ and $\left[\bar{A}(z), \bar{B}(z)\right]$ be two pairs of polynomial matrices which satisfy the assumptions* **3.5** *to* **3.7**. *If R_0 is a positive constant such that*

$$\begin{aligned} \det\left[A(z)\right] &\neq 0, \det\left[B(z)\right] \neq 0 \\ \det\left[\bar{A}(z)\right] &\neq 0, \det\left[\bar{B}(z)\right] \neq 0 \end{aligned}$$

for $0 \leq |z| < R_0$, *and*

$$A(z)^{-1}B(z) = \bar{A}(z)^{-1}\bar{B}(z)$$

for $0 \leq |z| < R_0$, *then*

$$A(z) = \bar{A}(z) \text{ and } B(z) = \bar{B}(z), \forall z$$

Remark 3.9 In Lemma **3.8**, the condition

$$A(z)^{-1}B(z) = \bar{A}(z)^{-1}\bar{B}(z)$$

could be replaced by

$$B(z)^{-1}A(z) = \bar{B}(z)^{-1}\bar{A}(z)$$

since by assumption the inverse of $B(z)$ and $\bar{B}(z)$ exist.

Remark 3.10 The assumptions **3.5** to **3.7** and conditions in Lemma **3.8** allow $\det[A(z)]$ and $\det[B(z)]$ to have roots on or inside the unit circle $|z| = 1$.

PROOF OF LEMMA **3.8** Clearly, $A(0) = B(0) = I_K$ and $\det[A(0)] = \det[B(0)] = 1 \neq 0$. The polynomials $\det[A(z)]$ and $\det[B(z)]$ are different from zero in a neighborhood of zero. In particular, we can choose $R_0 > 0$ such that $\det[A(z)] \neq 0$ and $\det[B(z)] \neq 0$ for $0 \leq |z| < R_0$. It follows that the matrices $A(z)$ and $B(z)$ are invertible for $0 \leq |z| < R_0$.

Let

$$C_0 = \{z \in \mathbb{C} \mid 0 \leq |z| < R_0\}$$

and

$$\Psi(z) = A(z)^{-1}B(z)$$

for $z \in C_0$. Since

$$\begin{aligned} A(z)^{-1} &= \frac{1}{\det[A(z)]} A^{\star}(z), \\ B(z)^{-1} &= \frac{1}{\det[B(z)]} B^{\star}(z), \end{aligned}$$

where $A^{\star}(z)$ and $B^{\star}(z)$ are matrices of polynomials, it follows that, for $z \in C_0$, each element of $A(z)^{-1}$ and $B(z)^{-1}$ is a rational function whose denominator is different from zero. Thus, for $z \in C_0$, $A(z)^{-1}$ and $B(z)^{-1}$ are matrices of analytic functions. It follows that the function

$$\Psi(z) = A(z)^{-1} B(z)$$

is analytic in the circle $0 \leq |z| < R_0$. Hence, it has a unique representation of the form

$$\Psi(z) = \sum_{k=0}^{\infty} \Psi_k z^k, \quad z \in C_0.$$

By assumption,

$$\Psi(z) = A(z)^{-1} B(z) = \bar{A}(z)^{-1} \bar{B}(z)$$

for $z \in C_0$. Hence, for $z \in C_0$,

$$\begin{aligned} \bar{A}(z) A(z)^{-1} B(z) &= \bar{B}(z) \\ \bar{A}(z) A(z)^{-1} &= \bar{B}(z) B(z)^{-1} \equiv \Delta(z) \end{aligned} \tag{3.5}$$

where $\Delta(z)$ is a diagonal matrix because $B(z)$ and $\bar{B}(z)$ are both diagonal,

$$\Delta(z) = diag\left[\delta_{ii}(z)\right],$$

where

$$\delta_{ii}(z) = \frac{\bar{b}_{ii}(z)}{b_{ii}(z)}, \; b_{ii}(0) = 1, \; \delta_{ii}(0) = \bar{b}_{ii}(0), \; i = 1, \ldots, K. \tag{3.6}$$

From (3.6), it follows that each $\delta_{ii}(z)$ is rational with no pole in C_0 such that $\delta_{ii}(0) = 1$, so it can be written in the form

$$\delta_{ii}(z) = \frac{e_i(z)}{f_i(z)}$$

where $e_i(z)$ and $f_i(z)$ have no common roots, $f_i(z) \neq 0$ for $z \in C_0$ and $\delta_{ii}(0) = e_{ii}(0) = 1$.

From (3.5), it follows that for $i = 1, \ldots, K$,

$$\begin{aligned} \bar{b}_{ii}(z) &= \delta_{ii}(z) b_{ii}(z), \\ \bar{a}_{ij}(z) &= \delta_{ii}(z) a_{ij}(z), \quad j = 1, \ldots, K, \end{aligned}$$

for $z \in C_0$.

We first show that $\delta_{ii}(z)$ must be a polynomial. If $f_i(z) \neq 1$, then its order cannot be greater than the order $q_i \equiv deg[b_{ii}(z)]$ for otherwise $\bar{b}_{ii}(z)$ would not be a polynomial. Similarly, if $f_i(z) \neq 1$ and is a polynomial of order less or equal to q_i, then all its roots must be roots of $b_{ii}(z)$ and $a_{ij}(z)$, for otherwise $\bar{b}_{ii}(z)$ or $\bar{a}_{ij}(z)$ would be a rational function. If $q_i \geq 1$, these roots are then common to $b_{ii}(z)$ and $a_{ij}(z)$, $j = 1, \ldots, K$, which is in contradiction with Assumption **3.7**. Thus the degree of $f_i(z)$ must be zero, and $\delta_{ii}(z)$ is a polynomial.

If $\delta_{ii}(z)$ is a polynomial of degree greater than zero, this would entail that $\bar{b}_{ii}(z)$ and $\bar{a}_{ij}(z)$ have roots in common, in contradiction with Assumption **3.7**. Thus $\delta_{ii}(z)$ must be a constant. Further, $\delta_{ii}(0) = 1$ so that for $i = 1, \ldots, K$,

$$\begin{aligned} \bar{b}_{ii}(z) &= b_{ii}(z), \\ \bar{a}_{ij}(z) &= a_{ij}(z), \quad j = 1 \ldots, K, \end{aligned}$$

and

$$\begin{aligned} \bar{B}(z) &= B(z), \\ \bar{A}(z) &= A(z). \end{aligned}$$

□

It should be noted that Assumption **3.7** is weaker than the hypothesis that $\det[A(L)]$ and $\det[B(L)]$ have no common roots, which would be a generalization of the usual identification condition for ARMA models.

Theorem 3.11 IDENTIFICATION OF DIAGONAL MA EQUATION FORM REPRESENTATION. *Under Assumptions* **3.5**, **3.6**, **3.7**, *and the assumption that the VARMA process is invertible, VARMA models in diagonal MA equation form are identified.*

PROOF OF THEOREM **3.11** Under the assumption that the VARMA process is invertible, we can write

$$B(L)^{-1}A(L)Y_t = U_t$$

Now suppose by contradiction that there exist operators $\bar{A}(L)$ and $\bar{B}(L)$, with $\bar{B}(L)$ diagonal and invertible, and $\bar{A}(L) \neq A(L)$ or $\bar{B}(L) \neq B(L)$, such that

$$\bar{B}(L)^{-1}\bar{A}(L) = B(L)^{-1}A(L)$$

If the above equality hold, then it must also be the case that

$$\bar{B}(z)^{-1}\bar{A}(z) = B(z)^{-1}A(z), \quad \forall z \in C_0$$

where $C_0 = \{z \in \mathbb{C} \mid 0 \leq |z| < R_0\}$ and $R_0 > 0$. By Lemma **3.8**, it follows that

$$\begin{aligned} \bar{A}(z) &= A(z) \\ \bar{B}(z) &= B(z) \quad \forall z. \end{aligned}$$

Hence, the representation is unique. □

Similarly, we can demonstrate that the final MA equation form representation is identified under the following assumption.

Assumption 3.12 *There are no roots common to $A(z)$ and $b(z)$, i.e. there is no value $z^\star$ such that $A(z^\star) = 0$ and $b(z^\star) = 0$.*

Theorem 3.13 IDENTIFICATION OF FINAL MA EQUATION FORM REPRESENTATION. *Under Assumption* **3.12**, *VARMA models in final MA equation form are identified.*

PROOF OF THEOREM **3.13** The proof can be easily adapted from the proof of Theorem **3.11** once we replace Assumption **3.7** by Assumption **3.12**.

□

Looking at equation (3.3), we see that it is always possible to obtain a diagonal MA equation form representation starting from any VARMA representation. One case where we would obtain a diagonal and not final MA representation is when there are common factors across rows of columns of $B(L)$ as in (3.4).

One strong appeal of the diagonal and final MA equation form representations is that it is really easy to get the equivalent (in term of autocovariances) invertible MA representation of a non-invertible representation. With ARMA models, we simply have to invert the roots of the MA polynomial which are inside the unit circle and adjust the standard deviation of the innovations (divide it by the square of these roots), see Hamilton (1994, Section 3.7). The same procedure could be applied to VARMA models in diagonal or final MA equation form.

For VARMA representations where no particular simple structure is imposed on the MA part, at the moment we are not aware of an algorithm to go from the non-invertible to the invertible representation tough theoretically this invertible representation exist and is unique as long as $\det[B(z)] \neq 0$ for $|z| = 1$ [see Hannan and Deistler (1988, chapter 1, section 3)]. So it might be troublesome to use a nonlinear optimization with these VARMA representations since we don't know how to go from the non-invertible to the invertible representation.

We can also consider the following natural generalization of the final AR equation form, where we simply replace the scalar AR operator by a diagonal operator.

Definition 3.14 (Diagonal AR equation form) *The VARMA representation (2.1) is*

said to be in diagonal AR equation form if $A(L) = diag[a_{ii}(L)] = I_K - A_1 L - \cdots - A_p L^p$ *where* $a_{ii}(L) = 1 - a_{ii,1}L - \cdots - a_{ii,p_i}L^{p_i}$ *and* $p = \max_{1 \leq i \leq K}(p_i)$.

Assumption 3.15 *For each* $i = 1, \ldots, K$, *there are no roots common to* $a_{ii}(z)$ *and* $B_{i\bullet}(z)$, *i.e. there is no value* $z^\star$ *such that* $a_{ii}(z^\star) = 0$ *and* $B_{i\bullet}(z^\star) = 0$.

Theorem 3.16 IDENTIFICATION OF DIAGONAL AR EQUATION FORM REPRESENTATION. *Under Assumption* **3.15**, *VARMA models in diagonal AR equation form are identified.*

From Theorem **3.11** we can see that one way to ensure identification is to impose constraints on the MA operator. This is an alternative approach to the ones developed for example in Hannan (1971, 1976) where the identification is obtained by restricting the autoregressive part to be lower triangular with $deg[a_{ij}(L)] \leq deg[a_{ii}(L)]$ for $j > i$, or in the final AR equation form where $A(L)$ is scalar. It may be more interesting to impose constraints on the moving average part instead because it is this part which causes problems in the estimation of VARMA models. Other identified representations which do not have a simple MA operator include the reversed echelon canonical form [see Poskitt (1992)] where we permute the rows of the VARMA model in echelon form so that the Kronecker indices are ordered from smallest to largest, and the scalar component model [see Tiao and Tsay (1989)] where we study contemporaneous linear transformations of the vector process. A general treatment of algebraic and topological structure underlying VARMA models is given in Hannan and Kavalieris (1984b).

4. Estimation Method

We next introduce elements of notation for the parameters of our model. First, irrespective of the VARMA representation employed we split the whole vector of parameters γ in two parts γ_1 (the parameters for the AR part) and γ_2 (MA part):

$$\gamma = [\gamma_1, \gamma_2]'.$$

For a VARMA model in diagonal MA equation form, γ_1 and γ_2 are

$$\gamma_1 = [a_{1\bullet,1}, \ldots, a_{1\bullet,p}, \ldots, a_{K\bullet,1}, \ldots, a_{K\bullet,p}], \tag{4.1}$$

$$\gamma_2 = [b_{11,1}, \ldots, b_{11,q_1}, \ldots, b_{KK,1}, \ldots, b_{KK,q_K}], \tag{4.2}$$

while for a VARMA model in final MA equation form, γ_2 is

$$\gamma_2 = [b_1, \ldots, b_q]. \tag{4.3}$$

For VARMA models in diagonal AR equation form, we simply invert γ_1 and γ_2:

$$\gamma_1 = [a_{11,1}, \ldots, a_{11,p_1}, \ldots, a_{KK,1}, \ldots, a_{KK,p_K}], \tag{4.4}$$

$$\gamma_2 = [b_{1\bullet,1}, \ldots, b_{1\bullet,q}, \ldots, b_{K\bullet,1}, \ldots, b_{K\bullet,q}], \tag{4.5}$$

while for a VARMA model in final AR equation form, γ_1 is

$$\gamma_1 = [a_1, \ldots, a_p]. \tag{4.6}$$

The estimation method is in three steps.

Step 1. Estimate a VAR(n_T) to approximate the VARMA(p,q) and recuperate the residuals that we will call $\hat{U}_t$:

$$\hat{U}_t = Y_t - \sum_{l=1}^{n_T} \hat{\Pi}_l(n_T) Y_{t-l} \tag{4.7}$$

with $T > 2 \times K \times n_T$.

Step 2. With the residuals from step 1, compute an estimate of the variance matrix of U_t, $\hat{\Sigma} = \sum_{t=n_T+1}^{T} \hat{U}_t \hat{U}_t'/T$ and estimate by GLS the following multivariate regression:

$$A(L)Y_t = [B(L) - I_K]\hat{U}_t + e_t$$

to get estimates $\tilde{A}(L)$ and $\tilde{B}(L)$ of $A(L)$ and $B(L)$. The regression is

$$\tilde{\gamma} = \left[\sum_{t=l}^{T} \hat{Z}'_{t-1}\hat{\Sigma}^{-1}\hat{Z}_{t-1}\right]^{-1}\left[\sum_{t=l}^{T} \hat{Z}'_{t-1}\hat{\Sigma}^{-1}Y_t\right] \tag{4.8}$$

with $l = n_T + \max(p,q) + 1$. If we define the following vectors

$$\begin{aligned}
\mathbf{Y}_{t-1} &= [y_{t-1}(1), \ldots, y_{t-1}(K), \ldots, y_{t-p}(1), \ldots, y_{t-p}(K)] \\
\hat{\mathbf{U}}_{t-1} &= [\hat{u}_{t-1}(1), \ldots, \hat{u}_{t-1}(K), \ldots, \hat{u}_{t-q}(1), \ldots, \hat{u}_{t-q}(K)] \\
\mathbf{y}_{t-1}(k) &= [y_{t-1}(k), \ldots, y_{t-p_k}(k)] \\
\hat{\mathbf{u}}_{t-1}(k) &= [\hat{u}_{t-1}(k), \ldots, \hat{u}_{t-q_k}(k)]
\end{aligned}$$

then the matrix $\hat{Z}_{t-1}$ for the various representations is

$$\hat{Z}^{DMA}_{t-1} = \begin{bmatrix} \mathbf{Y}_{t-1} & \cdots & 0 & \hat{\mathbf{u}}_{t-1}(1) & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{Y}_{t-1} & 0 & \cdots & \hat{\mathbf{u}}_{t-1}(K) \end{bmatrix}$$

$$\hat{Z}^{FMA}_{t-1} = \begin{bmatrix} \mathbf{Y}_{t-1} & \cdots & 0 & \hat{\mathbf{u}}_{t-1}(1) \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \mathbf{Y}_{t-1} & \hat{\mathbf{u}}_{t-1}(K) \end{bmatrix}$$

$$\hat{Z}^{DAR}_{t-1} = \begin{bmatrix} \mathbf{y}_{t-1}(1) & \cdots & 0 & \hat{\mathbf{U}}_{t-1} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{y}_{t-1}(K) & 0 & 0 & \hat{\mathbf{U}}_{t-1} \end{bmatrix}$$

$$\hat{Z}^{FAR}_{t-1} = \begin{bmatrix} \mathbf{y}_{t-1}(1) & \hat{\mathbf{U}}_{t-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_{t-1}(K) & 0 & 0 & \hat{\mathbf{U}}_{t-1} \end{bmatrix}$$

where DMA, FMA, DAR and FAR respectively stands for Diagonal MA, Final MA, Diagonal AR and Final AR equation form.

Step 3. Using the second step estimates, we first form new residuals

$$\tilde{U}_t = Y_t - \sum_{i=1}^{p} \tilde{A}_i Y_{t-i} + \sum_{j=1}^{q} \tilde{B}_j \tilde{U}_{t-j}$$

initiating with $\tilde{U}_t = 0$, $t \leq \max(p, q)$, and we define

$$\begin{aligned} X_t &= \sum_{j=1}^{q} \tilde{B}_j X_{t-j} + Y_t \\ W_t &= \sum_{j=1}^{q} \tilde{B}_j W_{t-j} + \tilde{U}_t \end{aligned}$$

initiating with $X_t = W_t = 0$ for $t \leq \max(p, q)$. We also compute a new estimate of Σ, $\tilde{\Sigma} = \sum_{t=\max(p,q)+1}^{T} \tilde{U}_t \tilde{U}_t'/T$. Then we regress by GLS $\tilde{U}_t + X_t - W_t$ on $\tilde{V}_{t-1}$ with

$$\tilde{V}_t = \sum_{j=1}^{q} \tilde{B}_j \tilde{V}_{t-j} + \tilde{Z}_t$$

where $\tilde{Z}_t$ is just like $\hat{Z}_t$ from step 2 except that it is computed with $\tilde{U}_t$ instead of $\hat{U}_t$ to obtain regression coefficients that we call $\hat{A}_i$ and $\hat{B}_j$:

$$\hat{\gamma} = \left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}_{t-1}' \tilde{\Sigma}^{-1} \tilde{V}_{t-1} \right]^{-1} \left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}_{t-1}' \tilde{\Sigma}^{-1} [\tilde{U}_t + X_t - W_t] \right]. \quad (4.9)$$

The properties of the estimation method are summarized in the following three theorems. Theorem **4.1** is a generalization of results from Lewis and Reinsel (1985) where convergence is demonstrated for i.i.d. innovations. We denote the Euclidean norm by $\|B\|^2 = \text{tr}(B'B)$.

Theorem 4.1 VARMA FIRST STEP ESTIMATES. *Under the above hypothesis on the process* $\{Y_t\}$ *and if* n_T *grows at a rate faster than* $\log T$ *with* ${n_T}^2/T \to 0$ *then for the first stage estimates* $\sum_{l=1}^{n_T} \|\hat{\Pi}_l(n_T) - \Pi_l\| \xrightarrow{q.m.} 0$.

Theorem 4.2 VARMA SECOND STEP ESTIMATES. *Under the above hypothesis on the process* $\{Y_t\}$ *and if* n_T *grows at a rate faster than* $\log T$ *with* ${n_T}^2/T \to 0$ *then the second stage estimates converge in quadratic mean to their true value.*

Their asymptotic distribution is given by

$$\sqrt{T}\left(\tilde{\gamma}-\gamma\right) \xrightarrow{d} \mathcal{N}\left(0, \tilde{J}^{-1}\tilde{I}\tilde{J}^{-1}\right) \tag{4.10}$$

with

$$\begin{aligned}\tilde{I} &= \sum_{j=-\infty}^{\infty} E\left[\left\{Z'_{t-1}\Sigma^{-1}U_t\right\}\left\{Z'_{t-1-j}\Sigma^{-1}U_{t-j}\right\}'\right] \\ \tilde{J} &= E\left[Z'_{t-1}\Sigma^{-1}Z_{t-1}\right]\end{aligned}$$

and Z_{t-1} is equal to the matrix $\hat{Z}_{t-1}$ where $\hat{U}_t$ is replaced by U_t.

Also, if $m_T^4/T \to 0$ with $m_T \to \infty$ then the matrix $\tilde{I}$ and $\tilde{J}$ can be consistently estimated in probability respectively by

$$\tilde{I}_T = \frac{1}{T}\sum_{j=-m_T}^{m_T}\omega(j,m_T)\sum_{t=l+|j|}^{T}\left\{\hat{Z}'_{t-1}\hat{\Sigma}^{-1}\tilde{U}_t\right\}\left\{\hat{Z}'_{t-1-j}\hat{\Sigma}^{-1}\tilde{U}_{t-j}\right\}' \tag{4.11}$$

$$\tilde{J}_T = \frac{1}{T}\sum_{t=l}^{T}\hat{Z}'_{t-1}\hat{\Sigma}^{-1}\hat{Z}_{t-1} \tag{4.12}$$

with $\omega(j,m_T) = 1-|j|/(m_T+1)$.

Theorem 4.3 VARMA THIRD STEP ESTIMATES. *Under the above hypothesis on the process $\{Y_t\}$ and if n_T grows at a rate faster than $\log T$ with ${n_T}^2/T \to 0$ then the third stage estimates converge in quadratic mean to their true value.*

Their asymptotic distribution is given by

$$\sqrt{T}\left(\hat{\gamma}-\gamma\right) \xrightarrow{d} \mathcal{N}\left(0, \hat{J}^{-1}\hat{I}\hat{J}^{-1}\right) \tag{4.13}$$

with

$$\begin{aligned}\hat{I} &= \sum_{j=-\infty}^{\infty} E\left[\left\{V'_{t-1}\Sigma^{-1}U_t\right\}\left\{V'_{t-1-j}\Sigma^{-1}U_{t-j}\right\}'\right] \\ \hat{J} &= E\left[V'_{t-1}\Sigma^{-1}V_{t-1}\right]\end{aligned}$$

and V_{t-1} is equal to the matrix $\tilde{V}_{t-1}$ where $\tilde{U}_t$ is replaced by U_t.

Also, if $m_T^4/T \to 0$ *with* $m_T \to \infty$ *then the matrix* $\hat{I}$ *and* $\hat{J}$ *can be consistently estimated in probability respectively by*

$$\hat{I}_T = \frac{1}{T}\sum_{j=-m_T}^{m_T} \omega(j, m_T) \sum_{t=l'+|j|}^{T} \left\{\tilde{V}'_{t-1}\tilde{\Sigma}^{-1}\bar{U}_t\right\}\left\{\tilde{V}'_{t-1-j}\tilde{\Sigma}^{-1}\bar{U}_{t-j}\right\}' \quad (4.14)$$

$$\hat{J}_T = \frac{1}{T}\sum_{t=\max(p,q)+1}^{T} \tilde{V}'_{t-1}\tilde{\Sigma}^{-1}\tilde{V}_{t-1} \quad (4.15)$$

with $l' = \max(p, q) + 1$ *and* $\bar{U}_t$ *are the filtered residuals computed with* $\hat{\gamma}$.

Notice the simplicity of this estimation method. Only three regressions are needed so we can avoid all the caveats associated with nonlinear optimizations. This is an important problem with VARMA models where we typically have to deal with a high number of parameters and numerical convergence might be hard to obtain. This is especially important when we consider the fact that the asymptotic distribution of our estimators, on which we would base our inference, may be a bad approximation to the finite-sample distribution in high-dimensional dynamic models. See for example Dufour, Pelletier, and Renault (2002) where we see that even for VAR models the asymptotic approximation may be unreliable. Because of this, an estimation procedure which only requires linear methods is interesting since it suggest that simulation-based procedures – bootstrap techniques for example – should be used, something that would be impractical if the estimation is based on non-linear optimizations.

It is also important to mention that this procedure is not specific to the representations considered in this work. The expressions can be easily adapted to other identified representation, e.g. the echelon form. Since our estimation method is only based on regressions we can afford to use a less parsimonious representation whereas for nonlinear method it is highly important to keep the number of parameters to a minimum.

For the estimation of VARMA models the emphasis has been on maximizing the likelihood (minimizing the nonlinear least squares) quickly. There are two ways of doing this. The first is having quick and efficient algorithm to evaluate the likelihood[2] [e.g.

[2]Expressions for the exact and approximate likelihood of VARMA model are presented in Hillmer and Tiao (1979) and theoretical properties of maximum likelihood estimation of VARMA models under the hypothesis that the innovations follow a m.d.s. is presented in Hannan, Dunsmuir, and Deistler (1980).

Luceño (1994) and the reference therein, Mauricio (2002), Shea (1989)]. The second is to find preliminary consistent estimates that can be computed quickly to initialize the optimization algorithm.

We are not the first to present a generalization to VARMA models of the Hannan and Rissanen (1982) estimation procedure for ARMA models [whose asymptotic properties are further studied in Zhao-Guo (1985) and Saikkonen (1986)]. A similar method in three steps is also presented in Hannan and Kavalieris (1984a) where the third step is presented as a correction to the second step estimates. Further relevant results concerning the approximation of a VARMA process by a long VAR are given in Lewis and Reinsel (1985), Hannan and Kavalieris (1986), Paparoditis (1996), Huang and Guo (1990), Wahlberg (1989). A third step to improve the efficiency of the estimators is rarely employed, surely because these procedures are often seen as a way to get initial values to startup a nonlinear optimization [e.g. see Poskitt (1992), Koreisha and Pukkila (1989), Lütkepohl and Claessen (1997)].

There are many variations around the innovation-substitution approach for the estimation of VARMA models. In some of them, we replace the lagged and current innovations by the corresponding residuals and we do a GLS estimation [Koreisha and Pukkila (1989) which is a multivariate generalization of Koreisha and Pukkila (1990a) and Koreisha and Pukkila (1990b), Flores de Frutos and Serrano (2002)]. Another is Spliid (1983), where in the first step a VAR of fixed length (for example $p+q$) is fitted. We then have to iterate the second step of the estimation to get consistent estimates.

Another approach to get estimators for VARMA models that do not require nonlinear estimation is use the link that exist between the VARMA parameters and the infinite VAR or VMA representation. This is an extension of a procedure proposed by Durbin (1959, 1960a, 1960b). With this approach, using a VAR we can estimate VMA models [see Galbraith, Ullah, and Zinde-Walsh (2000), which generalizes Galbraith and Zinde-Walsh (1994) and Galbraith and Zinde-Walsh (1997)] and VARMA models [Koreisha and Pukkila (1989)].

4.1. Asymptotic efficiency

We can ask ourselves what is the cost of not doing the nonlinear estimation. For a given sample size we will certainly lose some efficiency because of the first step estimation. We can none the less compare the asymptotic variance matrix of our estimates with the corresponding nonlinear estimates. We first can see that if the innovations are a m.d.s., then the asymptotic variance of our linear estimates is the same as the variance of maximum likelihood estimates under Gaussianity. The variance of maximum likelihood estimates for i.i.d. Gaussian innovations is given in Lütkepohl (1993a):

$$I = plim \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\partial U_t'}{\partial \gamma} \Sigma^{-1} \frac{\partial U_t}{\partial \gamma'} \right]$$

We can transform this expression so as to obtain an equation more closely related to our previous results. First, we split γ in the same two vectors γ_1 (the AR parameters) and γ_2 (the MA parameters), then we compute $\partial U_t / \partial \gamma_1'$ and $\partial U_t / \partial \gamma_2'$. We know that

$$U_t = Y_t - A_1 Y_{t-1} - \cdots - A_p Y_{t-p} + B_1 U_{t-1} + \cdots + B_q U_{t-q}.$$

So taking the derivative with respect to γ_1':

$$\begin{aligned} \frac{\partial U_t}{\partial \gamma_1'} &= Z_{\bullet 1:dim(\gamma_1),t-1} + B_1 \frac{\partial U_{t-1}}{\partial \gamma_1'} + \cdots + B_q \frac{\partial U_{t-q}}{\partial \gamma_1'} \\ B(L) \frac{\partial U_t}{\partial \gamma_1'} &= Z_{\bullet 1:dim(\gamma_1),t-1} \\ \frac{\partial U_t}{\partial \gamma_1'} &= B(L)^{-1} Z_{\bullet 1:dim(\gamma_1),t-1} \end{aligned}$$

where $Z_{\bullet 1:dim(\gamma_1),t-1}$ is the first $dim(\gamma_1)$ columns of Z_{t-1}. Similarly the derivative with respect to γ_2' is

$$\begin{aligned} \frac{\partial U_t}{\partial \gamma_2'} &= Z_{\bullet dim(\gamma_1)+1:dim(\gamma),t-1} + B_1 \frac{\partial U_{t-1}}{\partial \gamma_2'} + \cdots + B_q \frac{\partial U_{t-q}}{\partial \gamma_2'}, \\ &= B(L)^{-1} Z_{\bullet dim(\gamma_1)+1:dim(\gamma),t-1} \end{aligned}$$

Combining the two expressions we see that

$$\frac{\partial U_t}{\partial \gamma'} = V_{t-1}$$

so the variance matrix for maximum likelihood estimates I is equal to the matrix $\hat{J}$ from the third step estimation. Moreover if U_t is a m.d.s. we see that we have the equality $\hat{J} = \hat{I}$ so that the asymptotic variance matrix that we get in the third step of our method is the same as one would get by doing the maximum likelihood.

If the innovations are merely uncorrelated then we can generalize the results of Francq and Zakoïan (1998) who prove the consistency of nonlinear least squares for univariate weak ARMA model. The authors show that the asymptotic distribution of the estimates are

$$\sqrt{T}\left(\hat{\gamma} - \gamma_0\right) \xrightarrow{d} \mathcal{N}\left(0, J^{-1} I J^{-1}\right)$$

with

$$\begin{aligned} I &= 4 \sum_{k=-\infty}^{\infty} Cov\left[u_t \frac{\partial u_t}{\partial \gamma} \; ; \; u_{t-k} \frac{\partial u_{t-k}}{\partial \gamma}\right] \\ J &= 2E\left[\frac{\partial u_t}{\partial \gamma} \frac{\partial u_t}{\partial \gamma'}\right] \\ \gamma &= \{a_1, \ldots, a_p, b_1, \ldots, b_q\}. \end{aligned}$$

Without formally proving it we can generalize these expressions for the multivariate case. Writing the multivariate nonlinear least squares problem and doing a first order expansion of the first order condition we find that the expression for the asymptotic covariance matrix of the estimates would again be $J^{-1} I J^{-1}$ with

$$\begin{aligned} I &= 4 \sum_{k=-\infty}^{\infty} Cov\left[U_t \, \Sigma^{-1} \frac{\partial U_t}{\partial \gamma'} \; ; \; U_{t-k} \, \Sigma^{-1} \frac{\partial U_{t-k}}{\partial \gamma'}\right] \\ J &= 2E\left[\frac{\partial U_t'}{\partial \gamma} \, \Sigma^{-1} \frac{\partial U_t}{\partial \gamma'}\right] \end{aligned}$$

In our previous results we saw that $\partial U_t / \partial \gamma' = V_{t-1}$. From this we see that $J = 2\hat{J}$,

$I = 4\hat{I}$ and our third-step estimator have the same asymptotic variance-covariance matrix as maximum likelihood or non-linear least squares estimators depending on the properties of the innovations. To get a feel for the loss of efficiency in finite samples due to replacing the true innovations by residuals from a long VAR we perform Monte Carlo simulations and report the results in section 6.

5. Estimation of orders in VARMA models

We still have unknowns in our model, the orders of the AR and MA operators. If no theory specifies these parameters, we have to use a statistical procedure to choose them. We propose the following information criterion method to choose the orders for VARMA models in the different identified representations proposed in Section 3. In the second step of the estimation we compute for all $p_i \leq P$ and $q_i \leq Q$ the following information criterion:

$$\log(\det \tilde{\Sigma}) + \dim(\gamma)\frac{(\log T)^{1+\delta}}{T}, \quad \delta > 0. \tag{5.1}$$

We then choose $\hat{p}_i$ and $\hat{q}_i$ as the set which minimizes the information criterion. We assume that the upper bound P and Q on the order of the AR and MA part are bigger than the true values of p_i and q_i (or that they slowly grow with the sample size). The properties of $\hat{p}_i$ and $\hat{q}_i$ are summarized in the following theorem.

Theorem 5.1 ESTIMATION OF THE ORDER p AND q IN VARMA MODELS. *Under the above hypothesis on the process $\{Y_t\}$ and if n_T grows at a rate faster than $\log T$ with ${n_T}^2/T \to 0$ then $\hat{p}_i$ and $\hat{q}_i$, $i = 1, \ldots, K$, converge in probability to their true value.*

This criterion is a generalization of the information criterion proposed by Hannan and Rissanen (1982) which the authors acknowledged that it must in fact be modified to provide consistent estimates of the order p and q. The original criterion was

$$\log \tilde{\sigma}^2 + (p+q)\frac{(\log T)^{\delta}}{T}$$

with $\delta > 0$. But in Hannan and Rissanen (1983) they acknowledged that $\tilde{\sigma}^2 - \sigma^2$ is $O(n_T T^{-1})$ and not $O(T^{-1})$ so the penalty $(\log T)^\delta / T$ is not strong enough. The authors proposed two possible modifications to their procedure. The simpler is to take $(\log T)^{1+\delta}$ instead of $(\log T)^\delta$ in the information criterion so that the penalty on $p+q$ will dominate $\log \hat{\sigma}^2$ in the criterion. The second, which they favored and was used in latter work [see Hannan and Kavalieris (1984b)], is to modify the first step of the procedure. Instead of taking $n_T = O(\log T)$ they used another information criterion to choose the order of the long autoregression and they iterated the whole procedure picking a potentially different p and q at every iteration. A similar approach is also proposed in Poskitt (1987). In this work we prefer the first solution so as to keep the procedure as simple as possible.

The literature on information criterion to choose the order p and q in univariate ARMA models is vast. The best known criterion are certainly the AIC [Akaike (1973)], AICc [Sugiura (1978), Hurvich and Tsai (1989)], FPE [Akaike (1973)], Mallow's Cp [Mallows (1973)], SIC [Schwarz (1978)] and HQ [Hannan and Quinn (1979)]. McQuarrie and Tsai (1998) would be a good starting point for interested readers. Another approach for choosing p and q is to check if the residuals are uncorrelated [see, e.g., Pukkila, Koreisha, and Kallinen (1990), Koreisha and Pukkila (1995)].

Much work has also been done on information criterion to choose the order of VAR models. A good summary of the work in this field is Lütkepohl (1985) where he studied the performance of nine different procedures. Methods based on testing for uncorrelated residuals have also been developed [e.g., Koreisha and Pukkila (1999)].

For the identification of the order of VARMA models, it all depends on the representation that is used. Although it was one of the first representation studied, not much work has been done with the final AR equation form. People felt that this representation gives VARMA models with too many parameters. A complete procedure to fit VARMA models under this representation is given in Lütkepohl (1993a): One would first fit an ARMA(p_i, q_i) model to every univariate time series, using maybe the procedure of Hannan and Rissanen (1982). To build the VARMA representation in final AR equation form, knowing that the VAR operator is the same for every equation we would take it to be the product of all the univariate AR polynomials. This would give

a VAR operator of order $p = \sum_{i=1}^{K} p_i$. Accordingly, for the VMA part we would take $q = \max_k[q_k + \sum_{i=1, i\neq k}^{K} p_i]$. It is no wonder that people feel that the final equation form uses to many parameters.

There has been a lot more work done on the identification of Kronecker indices for VARMA models in echelon form. The problem has been studied by, among others, Hannan and Kavalieris (1984b), Poskitt (1992) and Lütkepohl and Poskitt (1996b). Non-stationary or co-integrated systems are studied by Huang and Guo (1990), Bartel and Lütkepohl (1998) and Lütkepohl and Claessen (1997). Additional references are given in Lütkepohl (1993a, Chapter 8).

A complementing approach to specify VARMA models, which is based on Cooper and Wood (1982), aims at finding simplifying structures via some combinations of the different series to obtain more parsimonious models. It includes Tiao and Tsay (1989), Tsay (1989a), Tsay (1989b), Tsay (1991), Nsiri and Roy (1992), Nsiri and Roy (1996).

6. Monte Carlo Simulations

To illustrate the performance of our estimation method we ran two types of simulations. For the first type, strong VARMA models were simulated (VARMA models with i.i.d. Gaussian innovations). The second type of simulations involves weak representations where the innovations are not independent nor a m.d.s but merely uncorrelated. This is done by time-aggregating a strong VARMA process with non-Gaussian innovations or an ARCH process. All the simulated models are bivariate so the results are easier to analyze. The results are generated using Ox version 3.30 on Linux [see Doornik (1999)]. We performed 1000 simulations for each model.

6.1. Strong VARMA

For the simulations with a strong representation we report results for a sample size of 250 which represent about 20 years of monthly data, a reasonable sample size for macroeconomic data. Tables 1 and 2 gives results for VARMA models in final MA equation form (VARMA(1,1) and VARMA(2,1) respectively), while results for VARMA models in diagonal MA equation form are given in Tables 3 and 4

(VARMA(1,1) with $q = (1,1)$ and VARMA(2,1) with $q = (1,1)$ respectively). We present the results (mean, standard deviations, root mean square error, 5% quantile, %95 quantile and median) for the second (when the number of parameters does not exceed five) and third step estimates, and the maximum likelihood estimates. We employed the likelihood conditional on the initial observations, and maximized the likelihood using the true value of the parameters as initial values. Samples for which the optimization algorithm did not converge were dropped (this happened for less than 1% of the simulations).

Looking at the RMSE, a first thing to notice is that there can be sizable improvement in doing the third step. Some of the third step RMSEs in Tables 1 and 3 are more than 50% smaller than for the second step. This is an interesting observation considering that the third step basically involve only one extra regression. Comparing the third step RMSEs and the RMSEs for the maximum likelihood estimates, we see that the former are usually no more than 15% bigger. This is also an interesting observation. The cost of avoiding a numerical optimization, which can become quite challenging as the number of time series studied or order of the operators increases, appears to be small.

In the top part of these Tables we also present the results for the selection of the order of the operators using our proposed information criterion. For models in final MA equation form, we have to select the orders p and q, and for models in diagonal MA equation, the selection is over p, q_1 and q_2. Looking at Table 1 and 2, we see the most frequently chosen orders are the true ones, and the criterion will tend to pick a higher value for q than for p. This result might partially be skewed by the fact that the simulated models have a highly persistent moving average ($b_1 = 0.9$). For VARMA models in diagonal equation form (Tables 3 and 4), we get similar results. The orders which are selected with the highest frequency are the true ones, but for some models we pick the wrong orders more than 50% of the time.

These results for the information criterion are fairly sensitive to the value of δ and c_0, more so for the model with a diagonal representation. This can be compared to non-parametric regressions and the selection of the bandwidth parameter. The performance of the information criterion with respect to these two parameters should be investigated further.

The results for models in final AR or diagonal AR equation form are presented in Tables 5 and 6, and Tables 7 and 8 respectively. The results are similar. We can see big improvements between the second step and third step estimates, and the RMSEs for the maximum likelihood estimates are usually 15% smaller than for the third step estimates. The performance of the information criterion is less satisfactory. We could have expected the information criterion to give similar results since the simulated AR equation form models are very much related to the MA equation form models (we simply interchanged the AR and MA operators). What we instead see is that the true order is chosen less often and in one case (Table 8) the true order is not the one selected the most often. One symmetrical result we do observe is now a higher AR order tends to be chosen more often than a higher MA order.

6.2. Weak VARMA

In this work we simulate weak VARMA models, where the innovations are uncorrelated but are not a m.d.s., by two different methods. Both methods are based on time-aggregation of a strong process. The first approach is to simulate directly weak innovations, from which we will build the simulated series Y_t. The second approach is to simulate a strong VARMA process and then time-aggregate it to obtain the series Y_t.

From the results in Drost and Nijman (1993), we know that the temporal aggregation of a strong GARCH process (where the standardized innovations are i.i.d.) will give a weak process. Suppose $\tilde{U}_t$ is given by the following bivariate ARCH model:

$$\begin{aligned} \tilde{U}_t &= H_t^{1/2}\varepsilon_t \\ H_t &= \Omega + \alpha\tilde{U}_{t-1}\tilde{U}'_{t-1} \end{aligned}$$

where ε_t is i.i.d. $N(0, I_2)$, $H_t^{1/2}$ is the Cholesky decomposition of H_t and α is a scalar. If we consider $\tilde{U}_t$ as a stock variable, then temporal aggregation of $\tilde{U}_t$ over two periods, i.e.

$$U_t = \tilde{U}_{2t}$$

will give a weak process. The series U_t will be uncorrelated but not a m.d.s., its mean will be zero and the variance will be $\Omega(1-\alpha^2)/(1-\alpha)$.

In these examples, because the innovations are not a m.d.s., we cannot do maximum likelihood. We instead employ nonlinear generalized least-squares (GLS), i.e. we minimize the nonlinear least squares, compute an estimate of the variance matrix of the innovations and then do nonlinear GLS. We did not operate this procedure, partly to reduce the estimation time in our Monte Carlo study, partly because there is no asymptotic gain in iterating.

Using this method, we simulated weak version of the previously simulated VARMA models in final MA (Tables 9 and 10), diagonal MA (Tables 11 and 12) and diagonal AR equation form (Tables 13 and 14). We kept the same values for the orders p and q, and the same values for the AR and MA parameters. The goal of these experiments is to confirm that the properties of our method does not rely on having i.i.d. innovations. In our simulations, we took

$$\begin{aligned} \Omega &= \begin{bmatrix} 1.0 & 0.7 \\ 0.7 & 1.0 \end{bmatrix} \\ \alpha &= 0.3. \end{aligned}$$

As expected, we get the same results as for the cases where the innovations were i.i.d. Gaussian. We can get big reductions of the RMSEs by doing the third step of the regression-based estimation method and the RMSEs of the third step estimates are often slightly bigger than those obtained by nonlinear GLS. It appears that we don't lose much by doing only three regressions instead of doing the full non-linear minimization. The performance of the information criterion with weak VARMA models is also similar to cases where the VARMA models were strong.

Another easy and relevant way to simulate a weak VARMA model is by time aggregating a strong one. For univariate time series it has been shown that if y_t is a strong ARMA(p,q) then y_{mt} (we observe the process y_t every m periods) will be an ARMA(p,$p+[(q-p)/m]$) where the brackets represent the integer part. But the innovations of the aggregated process even tough they are uncorrelated they are not i.i.d. or a m.d.s. anymore. The temporal aggregation of ARMA processes has been extensively

studied [Palm and Nijman (1984), Nijman and Palm (1990), Amemiya and Wu (1972), Drost (1993)].

The generalization of these results to multivariate time series is straightforward. The time aggregation of a strong VARMA will give a weak VARMA. For example, take Y_t a bivariate VARMA(1,1) in final equation form with i.i.d. $(0, \Sigma)$ innovations at the monthly frequency. If we only have quarterly data then for the process $Y_{3(t)}$ we can write:

$$Y_{3(t)} = A_1^3 Y_{3(t-1)} + U_{3(t)} - (B_1 - A_1)U_{3(t)-1} - (A_1 B_1 - A_1^2)U_{3(t)-2} - A_1^2 B_1 U_{3(t-1)}$$

We can compute the first autocorrelation which give

$$\begin{aligned} \Gamma(1) &= E\left[Y_{3(t+1)} Y'_{3(t)}\right] \\ &= E\left[\left\{A_1^3 Y_{3(t)} + V_{3(t+1)}\right\} Y'_{3(t)}\right] \\ &= A_1^3 \Gamma(0) - A_1^2 B_1 \Sigma \end{aligned}$$

where $V_{3(t+1)} = U_{3(t+1)} - (B_1 - A_1)U_{3(t+1)-1} - (A_1 B_1 - A_1^2)U_{3(t+1)-2} - A_1^2 B_1 U_{3(t+1)-3}$ and $\Gamma(0) = E\left[Y_{3(t)} Y'_{3(t)}\right]$. For the second autocorrelation we get

$$\begin{aligned} \Gamma(2) &= E\left[Y_{3(t+2)} Y'_{3(t)}\right] \\ &= E\left[\left\{A_1^3 Y_{3(t+1)} + V_{3(t+2)}\right\} Y'_{3(t)}\right] \\ &= E\left[A_1^3 Y_{3(t+1)} Y'_{3(t)}\right] + E\left[V_{3(t+2)} Y_{3(t)}\right] \\ &= A_1^3 \Gamma(1) \end{aligned}$$

where the last term is zero because $V_{3(t+2)}$ is a linear combination of $U_{3(t+2)}$, $U_{3(t+2)-1}$, $U_{3(t+2)-2}$ and $U_{3(t+2)-3}$ and hence uncorrelated with $Y_{3(t)}$. We then see that in general, we have $\Gamma(h) - A_1^3 \Gamma(h-1) = 0$ for $h > 1$ which imply that $Y_{3(t)}$ is a VARMA(1,1) and the AR coefficient is A_1^3. We can then write

$$Y_{3(t+1)} = A_1^3 Y_{3t} + \varepsilon_{3(t+1)} - \Theta \varepsilon_{3t}$$

where the ε_{3t}'s are uncorrelated and $E[\varepsilon_{3t}\varepsilon'_{3t}] = \Sigma_\varepsilon$. To find the value of Θ and Σ_ε we

have to solve the following equations:

$$\begin{aligned} E[\{\varepsilon_{3t} - \Theta\varepsilon_{3(t-1)}\}\{\varepsilon_{3t} - \Theta\varepsilon_{3(t-1)}\}'] &= E[v_{3t}v_{3t}'] \\ E[\{\varepsilon_{3t} - \Theta\varepsilon_{3(t-1)}\}\{\varepsilon_{3(t-1)} - \Theta\varepsilon_{3(t-2)}\}] &= E[v_{3t}v_{3(t-1)}'] \end{aligned}$$

which give the following system of equations

$$\begin{aligned} \Sigma_\varepsilon + \Theta\Sigma_\varepsilon\Theta' &= \Sigma_U + (B_1 - A_1)\Sigma(B_1 - A_1)' + (A_1B_1 - {A_1}^2)\Sigma(A_1B_1 - {A_1}^2)' \\ &\quad +({A_1}^2B_1)\Sigma({A_1}^2B_1)' \\ -\Theta\Sigma_\varepsilon &= -{A_1}^2B_1\Sigma. \end{aligned}$$

Unlike in the univariate case we are not aware of any algorithm to solve this system of equations for the general case. What we can do is solve these equations numerically for a given value of A_1, B_1 and Σ.

In our example, we took the model described in Table 5 (an VARMA model in final AR equation form) and assumed that Y_t were monthly stock data and we time aggregated them to the quarterly frequency. Instead of taking the innovations to be Gaussian (which is a special case where the aggregated VARMA would be strong) we take them to be a mixture of two Gaussian distributions with different means (but with mean zero unconditionally). This will give skewed marginal distributions and we can appeal to the results of Francq and Zakoïan (1998, Section 2.2.1) to claim that the resulting VARMA is only weak. We take U_t to be

$$U_t = p\, u_t(1) + (1-p)\, u_t(2)$$

with

$$\begin{aligned} u_t(1) &\sim N\left(\begin{bmatrix} -1 \\ -1 \end{bmatrix}; \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}\right) \\ u_t(2) &\sim N\left(\begin{bmatrix} 0.25 \\ 0.25 \end{bmatrix}; \begin{bmatrix} 1.75 & 1.3 \\ 1.3 & 1.75 \end{bmatrix}\right) \end{aligned}$$

and $\Pr[p = 1] = 0.2$. The different parameters are chosen such that the unconditional mean of U_t is zero and the variance is the same as with the examples for the strong VARMA models.

Solving the system of equations for Θ and Σ_ϵ we have four solutions, two complex and two real. For the two real solutions, one gave a non-invertible MA operator and the other an invertible MA operator. The invertible solution is:

$$\Theta = \begin{bmatrix} 0.0593618 & -0.14134 \\ 0.20598 & 0.296472 \end{bmatrix}$$

$$\Sigma_\epsilon = \begin{bmatrix} 2.64155 & 0.650962 \\ 0.650962 & 1.70611 \end{bmatrix}.$$

Using these results we report one set of simulations. Table 15 contains results for a sample size after aggregation of 200. For the selection of the order p and q with a sample size of 200 we get results similar to those in table 5. For the estimates of the AR and MA parameters we get results similar to the cases of VARMA models with i.i.d innovations. Again the RMSEs are smaller for the third step estimates compared to the second step and just like in Table 6 the third step estimates and the nonlinear least-squares have the same RMSEs. This is again an indication that we don't necessarily loose efficiency by not doing the nonlinear estimation.

7. Application to macroeconomics time series

To illustrate our estimation method and the gains that can be obtain by using a more parsimonious representation, we fit VARMA and VAR models to six macroeconomic time series and compute the impulse-response functions generated by each model. What people typically do to get the impulse-response functions is first fit a VAR to their multiple time series and then get the implied infinite VMA representation. If the time series are not stationary this representation can't be interpreted as a VMA representation but its coefficients can be computed none the less. The order of the VAR required for macro series is usually high. For example, Bernanke and Mihov (1998) uses a VAR(13) to model six monthly macroeconomic time series when about 30 years of

data are available. The resulting standard errors for the impulse-response functions are very large, like in most macroeconomic study. We can ask ourselves how much of this is due to the fact that so many parameters are estimated. To try to answer this we will study the impulse-response functions generated by VARMA models estimated on the same data.

For this exercise we take the time series from Bernanke and Mihov (1998). They consist of the log of the real GDP (*gdpm*), total bank reserves (*trl*), nonborrowed reserves (*nbrecl*), federal funds rate (*fyff*), log of the GDP deflator (*pgdpm*), log of the Dow-Jones index of spot commodity prices (*psccom*). These are monthly data and cover the period January 1962 to December 1996. The monthly data for real GDP and the GDP deflator were constructed by state space methods, using a list of monthly interpolator variables and assuming that the interpolation error is describable as an AR(1) process. Both total reserves and nonborrowed reserves are normalized by a 36-month moving average of total reserves. The series are plotted in Figure 1.

Our example is based on McMillin (2001) who compare numerous identification restrictions for the structural effects of monetary policy shocks using the same dataset as Bernanke and Mihov (1998). One of the model studied is a VAR applied to the first difference of the series, in order, *gdpm*, (*psscom-pgdpm*), *fyff*, *nbrecl*, *trl*, *psscom*. With an argument based on Keating (2002), the author state that using this ordering of the variables the Cholesky decomposition, based on long-run macroeconomic restrictions, which are described in an appendix, of the variance matrix of the innovations will identify the structural effects of the policy variable *nbrecl* without imposing any contemporaneous restrictions among the variables. Since the model is in first difference, the impulse-response at a given order is the cumulative shocks up to that order.

By fitting a VAR(12) to these series we get sensibly the same impulse-response functions and confidence band as in McMillin (2001) [3]. They are plotted in figure 2. The impulse-response function for the output and federal funds rate tends to zero as the order increase which is consistent with the notion that a monetary variable does not have a long term impact on real variables. The impulse response of the price level

[3]The magnitude of the IRF for *gdpm* is smaller and the confidence band for *gdpm* and *fyff* are tighter than in McMillin (2001).

increase as we let the order grow and does not revert to zero.

We next estimate VARMA models for the four representations proposed in this work. Selection of the orders for the two diagonal representations is more complicated now that we are studying six time series. If we take the maximum order of the diagonal operators to be ten, then we would have to perform the second step of the estimation method more than one million times. We instead impose that the order of the diagonal operators are equal, so we only have to minimize the information criterion over two parameters.

The information criterion picked the following orders for the different representations: VARMA(1,7) for the final MA equation form ($\delta = 0.1, c_0 = 1$), VARMA(3,3) for the diagonal MA equation form ($\delta = 0.1, c_0 = 1/2$), VARMA(12,1) for the final AR equation form ($\delta = 0.1, c_0 = 2/3$), VARMA(16,4) for the diagonal AR equation form ($\delta = 0.1, c_0 = 1/3$). Looking at the impulse responses for VARMA models with parameters p and q close to the above values for the respective representations, the following orders give impulse-response functions closer to the ones generated by the VAR(12): VARMA(5,5) for the final MA (Figure 3), VARMA(5,1) for the diagonal MA (Figure 4), VARMA(12,5) for the final AR (Figure 5), VARMA(12,5) for the diagonal AR (Figure 6).

The behavior of the impulse-response function for the federal funds rate from the VARMA models are similar to what we obtained with a VAR. The closest match is given by the VARMA in final MA representation. We see an initial decrease in the federal funds rate, followed by a return to the initial level. The VARMA models are generating a smaller initial decrease -0.13 to -0.2 percentage point versus -0.32 for the VAR.

For the price level the VARMA models are giving the same pattern as the VAR model but the amplitude of the impact is smaller. The VARMA model in final AR equation form gives the smallest impact.

The shape of impulse-response functions generated by the VARMA models for the output variable are also similar to the one from the VAR model, except that the amplitude is smaller. The initial impact is negative, then output goes up, and return to it's original level after reaching a peak. Of the four VARMA representations, the final

MA equation form gives the result closest to the VAR model but the initial reduction is more important and the impulse-response is peaking a few months earlier.

What is the most interesting is the behavior of the confidence bands for the VARMA's impulse-response functions. For the output and the federal funds rate series, we see that the bands are much smaller for the VARMA models and they shrink more quickly as the horizon increases compared to the VAR model. This result should not be so surprising since we expect that there should be no long-term effect of the policy variable on these two variables so the uncertainty about the long term effect should decrease as the horizon increases. The situation is different for the price level. For this variable the confidence band grows with the order. Again this is not so surprising because we expect that a change in the non-borrowed reserves should have a long-term impact on the price level. With a non-dying impact it is natural that the uncertainty about this impact can grow as time passes.

From this example, we see that VARMA models in final MA or diagonal MA equation form are giving results the closest to what we would obtain with a VAR model. This result could be expected since these models are simple extensions of the VAR approach. The introduction of a simple MA operator allows the reduction of the required AR order so we can get more precise estimates, which translate into more precise impulse-response functions.

8. Conclusion

In this paper we propose a modeling and estimation method which ease the use of VARMA model. We first propose new identified VARMA representations, the final MA equation form and the diagonal MA equation form. These two representations are simple extensions of the class of VAR models where we add a simple MA operator, either a scalar or diagonal operator. The addition of a MA part can give more parsimonious representations, yet the simple form of the MA operators does not introduce undue complications.

To ease the estimation we consider the problem of estimating VARMA models by relatively simple methods which only require linear regressions. For that purpose,

we consider a generalization of the regression-based estimation method proposed by Hannan and Rissanen (1982) for univariate ARMA models. Our method is in three steps. In a first step a long VAR is fitted to the data. In the second step, the lagged innovations in the VARMA model are replaced by the corresponding lagged residuals from the first step and a regression is performed. In a third step, the data from the second step are filtered and another regression is performed. We show that the third step estimators have the same asymptotic variance as their nonlinear counterpart (Gaussian maximum likelihood if the innovations are i.i.d., or generalized nonlinear least squares if they are merely uncorrelated). In the non i.i.d. case, we consider strong mixing conditions, rather than the usual martingale difference sequence assumption. We make these minimal assumptions on the innovations to broaden the class of models to which this method can be applied.

We also propose a modified information criterion that gives consistent estimates of the orders of the AR and MA operators of the proposed VARMA representations. This criterion is to be minimized in the second step of the estimation method over a set of possible values for the different orders.

Monte Carlo simulation results indicates that the estimation method works well for small sample sizes and the information criterion picks the true value of the order p and q most of the time. These results holds for sample sizes commonly used in macroeconomics, i.e. 20 years of monthly data or 250 sample points. To demonstrate the importance of using VARMA models to study multivariate time series we compare the impulse-response functions generated by VARMA and VAR models when these models are applied to the dataset of macroeconomic time series used by Bernanke and Mihov (1998).

9. Appendix: Proofs

Lemma 9.1 DAVYDOV (1968). *Let U and V be random variables measurable with respect to $\mathcal{F}_{-\infty}^{0}$ and $\mathcal{F}_{n}^{\infty}$, respectively. Let r_1, r_2, r_3 be positive numbers. Assume that $\|U\|_{r_1} < \infty$ and $\|V\|_{r_2} < \infty$ where $\|U\|_r = (E[|U|]^r)^{1/r}$. If $r_1^{-1} + r_2^{-1} + r_3^{-1} = 1$, then there exists a positive constant C independent of U, V and n, such that*

$$|E[UV] - E[U]E[V]| \leq C\|U\|_{r_1}\|V\|_{r_2}(\alpha(n))^{1/r_3}.$$

Lemma 9.2 IBRAGIMOV (1962). *If the random process (y_t) is strictly stationary and satisfies the strong mixing condition (2.12), with*

$$E|y_t|^{2+\delta} < \infty$$

for some $\delta > 0$, and if

$$\sum_{j=1}^{\infty} \alpha(j)^{\delta/(2+\delta)} < \infty,$$

then

$$\sigma^2 = E\left[(y_t - E[y_t])^2\right] + 2\sum_{j=1}^{\infty} E\left[(y_t - E[y_t])(y_{t+j} - E[y_{t+j}])\right].$$

Moreover, if $\sigma \neq 0$ and $E[y_t] = 0$, then

$$\Pr\left[\frac{y_1 + \cdots + y_t}{\sigma\sqrt{t}} < z\right] \xrightarrow[T\to\infty]{} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du$$

Lemma 9.3 INFINITE VAR. *If the VARMA model is invertible and if n_T grows at a rate faster then $\log T$, then $\sum_{j=1}^{K} \sum_{m=n_T+1}^{\infty} |\pi_{ij,m}| = o(T^{-1})$ for $i = 1, \ldots, K$.*

PROOF OF LEMMA **9.3** The matrix $B(L)^{-1}$ can be seen has its adjoint matrix divided by its determinant. Since Y_t is invertible, the roots of $\det B(L)$ are outside the unit

circle and so the elements of $\Pi(L) = B(L)^{-1}A(L)$ decrease exponentially:

$$|\pi_{ij,m}| \leq c\rho^m, \ \forall i, j.$$

with $c > 0$ and $0 < \rho < 1$. From this

$$\begin{aligned} \sum_{j=1}^{K} \sum_{l=n_T+1}^{T} |\pi_{ij,m}| &\leq \sum_{j=1}^{K} \sum_{l=n_T+1}^{T} c\rho^m \\ &= cK \frac{\rho^{n_T+1}}{1-\rho}. \end{aligned}$$

If n_T grows at a rate faster than $\log T$ then $T\rho^{n_T}$ will tend to zero since $|\rho| < 1$. □

Lemma 9.4 COVARIANCE ESTIMATION. *If the process $\{Y_t\}$ is a strictly stationary VARMA process with (U_t) uncorrelated, $E[|u_t(i)|^{4+2\delta}] < \infty$ for some $\delta > 0$, α-mixing with $\sum_{h=1}^{\infty} \alpha(h)^{2/(2+\delta)} < \infty$ and if $n_T/T \to 0$ then*

$$\frac{1}{T} \sum_{t=1}^{T} y_{t-r}(k) y_{t-s}(k') - E[y_{t-r}(k) y_{t-s}(k')] \xrightarrow{q.m.} 0, \ \forall k, k'.$$

PROOF OF LEMMA **9.4**

First notice that by stationarity,

$$E\left[\frac{1}{T} \sum_{t=n_T+1}^{T} y_{t-r}(k) y_{t-s}(k')\right] - E[y_{t-r}(k) y_{t-s}(k')] = 0.$$

Now taking the variance,

$$\begin{aligned} & Var\left[\frac{1}{T} \sum_{t=1}^{T} y_{t-r}(k) y_{t-s}(k')\right] \\ = & \frac{1}{T^2} \sum_{t=1}^{T} \sum_{t'=1}^{T} Cov\left[y_{t-r}(k) y_{t-s}(k') \, ; \, y_{t'-r}(k) y_{t'-s}(k')\right] \\ = & \frac{1}{T^2} \sum_{t=1}^{T} \sum_{t'=t+s-r+1}^{T} Cov\left[y_{t-r}(k) y_{t-s}(k') \, ; \, y_{t'-r}(k) y_{t'-s}(k')\right] \end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{T^2}\sum_{t'=1}^{T}\sum_{t=t'+s-r+1}^{T} Cov\left[y_{t-r}(k)y_{t-s}(k')\,;\, y_{t'-r}(k)y_{t'-s}(k')\right] \\
&+ \frac{1}{T^2}\sum_{t'=1+(s-r)}^{T-(s-r)}\sum_{t=t'-(s-r)}^{t'+(s-r)} Cov\left[y_{t-r}(k)y_{t-s}(k')\,;\, y_{t'-r}(k)y_{t'-s}(k')\right] \\
&+ \frac{1}{T^2}\sum_{t'=1}^{(s-r+1)}\sum_{t=1}^{t'+(s-r)} Cov\left[y_{t-r}(k)y_{t-s}(k')\,;\, y_{t'-r}(k)y_{t'-s}(k')\right] \\
&+ \frac{1}{T^2}\sum_{t=T-(s-r)}^{T}\sum_{t'=t-(s-r)}^{t+(s-r)} Cov\left[y_{t-r}(k)y_{t-s}(k')\,;\, y_{t'-r}(k)y_{t'-s}(k')\right] \qquad (9.1)
\end{aligned}
$$

For the first two terms of equation (9.1), Using Davydov's inequality (lemma **9.1**), the strong mixing hypothesis and the finite fourth moment we know that

$$
\begin{aligned}
&\sum_{t'=t+s-r+1}^{T} \left|Cov\left[y_{t-r}(k)y_{t-s}(k')\,;\, y_{t'-r}(k)y_{t'-s}(k')\right]\right| \\
\leq\ &\sum_{t'=t+s-r+1}^{T} \Big[C\,\|(y_{t-r}(k)y_{t-s}(k')\|_{2+\delta}\|y_{t'-r}(k)y_{t'-s}(k')\|_{2+\delta} \\
&\qquad \times\alpha(t'-t-s+r-1)^{\delta/(2+\delta)}\Big] \\
<\ &\infty
\end{aligned}
$$

from which we conclude that the first two terms converge to zero at rate $1/T$. For the other three terms, since these covariances are finite, the sums divided by T^2 will also converge to zero.

□

PROOF OF THEOREM **4.1**

We first introduce some additional matrix norms:

$$
\|B\|_2^2 = \sup_{l\neq 0}\frac{l'B'Bl}{l'l} \qquad (9.2)
$$

$$
\|B\|_1 = \max_{i\leq j\leq n}\sum_{i=1}^{n}|b_{ij}| \qquad (9.3)
$$

$$
\|B\|_\infty = \max_{1\leq i\leq n}\sum_{j=1}^{n}|b_{ij}| \qquad (9.4)
$$

where (9.2) is the largest eigenvalue of $B'B$. Useful inequalities relating these norms are given in Horn and Johnson (1985, p. 313)

$$\|AB\|^2 \leq \|A\|_2^2\|B\|^2, \tag{9.5}$$

$$\|AB\|^2 \leq \|A\|^2\|B\|_2^2, \tag{9.6}$$

$$\|B\|_2^2 \leq \|B\|_1\|B\|_\infty. \tag{9.7}$$

In the first step estimation, we regress

$$y_t(i) = \sum_{l=1}^{n_T}\sum_{j=1}^{K}\pi_{ij,l}y_{t-l}(j) + e_t(i) \tag{9.8}$$

when in fact

$$y_t(i) = \sum_{l=1}^{\infty}\sum_{j=1}^{K}\pi_{ij,l}y_{t-l}(j) + u_t(i). \tag{9.9}$$

Let

$$\hat{B}(n_T) = \sum_{t=n_T+1}^{T}\frac{\mathbf{Y}'_{t-1}\mathbf{Y}_{t-1}}{T-n_T}.$$

Then OLS applied to (9.8) gives:

$$\begin{aligned}
\hat{\Pi}_{i\bullet}(n_T) &= [\hat{\pi}_{i\bullet,1},\ldots,\hat{\pi}_{i\bullet,n_T}]' \\
&= \hat{B}(n_T)^{-1}\sum_{t=n_T+1}^{T}\frac{\mathbf{Y}'_{t-1}y_t(i)}{T-n_T} \\
&= \hat{B}(n_T)^{-1}\sum_{t=n_T+1}^{T}\frac{\mathbf{Y}'_{t-1}}{T-n_T}\left\{\sum_{l=1}^{\infty}\pi_{i\bullet,l}Y_{t-l} + u_t(i)\right\} \\
&= \Pi_{i\bullet}(n_T) + \hat{B}(n_T)^{-1}\sum_{t=n_T+1}^{T}\frac{\mathbf{Y}'_{t-1}}{T-n_T}\left\{\sum_{l=n_T+1}^{\infty}\pi_{i\bullet,l}Y_{t-l} + u_t(i)\right\}.
\end{aligned}$$

Rearranging the elements,

$$\hat{\Pi}_{i\bullet}(n_T) - \Pi_{i\bullet}(n_T) = \hat{B}(n_T)^{-1}\sum_{t=n_T+1}^{T}\frac{\mathbf{Y}'_{t-1}}{T-n_T}\left\{\sum_{l=n_T+1}^{\infty}\pi_{i\bullet,l}Y_{t-l}\right\} +$$

$$\hat{B}(n_T)^{-1} \sum_{t=n_T+1}^{T} \frac{\mathbf{Y}'_{t-1} u_t(i)}{T - n_T}$$

From which we get, using inequalities (9.5) to (9.7) and the fact that $\hat{B}(n_T)$ is symmetric,

$$\begin{aligned} \|\hat{\Pi}_{i\bullet}(n_T) - \Pi_{i\bullet}(n_T)\| &\leq \|\hat{B}(n_T)^{-1}\|_2 \|V_{1T}\| + \|\hat{B}(n_T)^{-1}\|_2 \|V_{2T}\| \\ &\leq \|\hat{B}(n_T)^{-1}\|_1 \|V_{1T}\| + \|\hat{B}(n_T)^{-1}\|_1 \|V_{2T}\| \end{aligned} \quad (9.10)$$

where

$$\begin{aligned} V_{1T} &= \frac{1}{T - n_T} \sum_{t=n_T+1}^{T} \mathbf{Y}'_{t-1} \sum_{l=n_T+1}^{\infty} \pi_{i\bullet,l} Y_{t-l} \\ &= \frac{1}{T - n_T} \sum_{t=n_T+1}^{T} \mathbf{Y}'_{t-1} [u_t(i) - e_t(i)] \\ V_{2T} &= \frac{1}{T - n_T} \sum_{t=n_T+1}^{T} \mathbf{Y}'_{t-1} u_t(i) \end{aligned}$$

Firstly, $\|V_{2T}\|^2$ can be expanded into

$$\begin{aligned} \|V_{2T}\|^2 &= tr\left(V'_{2T} V_{2T}\right) \\ &= \frac{1}{(T - n_T)^2} \sum_{t=n_T+1}^{T} \sum_{t'=n_T+1}^{T} u_t(i) u_{t'}(i) \mathbf{Y}_{t-1} \mathbf{Y}'_{t'-1} \\ &= \frac{1}{(T - n_T)^2} \sum_{t=n_T+1}^{T} \sum_{t'=n_T+1}^{T} u_t(i) u_{t'}(i) \left(\sum_{k=1}^{K} \sum_{l=1}^{n_T} y_{t-l}(k) y_{t'-l}(k) \right) \\ &= \frac{1}{(T - n_T)^2} \sum_{k=1}^{K} \sum_{l=1}^{n_T} \sum_{t=n_T+1}^{T} \sum_{t'=n_T+1}^{T} u_t(i) u_{t'}(i) y_{t-l}(k) y_{t'-l}(k). \end{aligned}$$

Taking the expectation,

$$E\left[\|V_{2T}\|^2\right] = \frac{1}{(T - n_T)^2} \sum_{k=1}^{K} \sum_{l=1}^{n_T} \sum_{t=n_T+1}^{T} \sum_{t'=n_T+1}^{T} Cov\left(u_t(i) y_{t-l}(k); u_{t'}(i) y_{t'-l}(k)\right)$$

As in the proof of Lemma **9.4**, using Davydov's inequality, the strong mixing hypothesis and the finite fourth moment, it can be shown that

$\sum_{t'=n_T+1}^{T} Cov\left(u_t(i)y_{t-l}(k); u_{t'}(i)y_{t'-l}(k)\right)$ is bounded by c_0, with $0 < c_0 < \infty$, so that

$$\begin{aligned} E\left[\|V_{2T}\|^2\right] &\leq \frac{Kn_T}{T-n_T}c_0 \\ &\to 0 \end{aligned}$$

since n_T is chosen such that $n_T/T \to 0$.

Similarly, for $\|V_{1T}\|^2$

$$\begin{aligned} &\|V_{1T}\|^2 \\ &= \frac{1}{(T-n_T)^2}\left[\sum_{t=n_T+1}^{T}\mathbf{Y}'_{t-1}[u_t(i)-e_t(i)]\right]'\left[\sum_{t=n_T+1}^{T}\mathbf{Y}'_{t-1}[u_t(i)-e_t(i)]\right] \\ &= \frac{1}{(T-n_T)^2}\sum_{t=n_T+1}^{T}\sum_{t'=n_T+1}^{T}\sum_{k=1}^{K}\sum_{l=1}^{n_T} y_{t-l}(k)\left[u_t(i)-e_t(i)\right]y_{t'-l}(k)\left[u_{t'}(i)-e_{t'}(i)\right] \\ &= \frac{1}{(T-n_T)^2}\sum_{t=n_T+1}^{T}\sum_{t'=n_T+1}^{T}\sum_{k=1}^{K}\sum_{l=1}^{n_T} y_{t-l}(k)\left[\sum_{m=n_T+1}^{\infty}\sum_{j=1}^{K}\pi_{ij,m}y_{t-m}(j)\right]\times \\ &\qquad y_{t'-l}(k)\left[\sum_{m'=n_T+1}^{\infty}\sum_{j'=1}^{K}\pi_{ij',m'}y_{t'-m'}(j')\right] \\ &= \frac{1}{(T-n_T)^2}\sum_{t=n_T+1}^{T}\sum_{t'=n_T+1}^{T}\sum_{k=1}^{K}\sum_{l=1}^{n_T}\sum_{m=n_T+1}^{\infty}\sum_{j=1}^{K}\sum_{m'=n_T+1}^{\infty}\sum_{j'=1}^{K} \\ &\qquad \pi_{ij,m}\pi_{ij',m'}y_{t-l}(k)y_{t-m}(j)y_{t'-l}(k)y_{t'-m'}(j') \end{aligned}$$

Again is in the proof of Lemma **9.4**, when taking the expectations of $\|V_{1T}\|^2$ using Davydov's inequality and the strong mixing hypothesis, we know that the sum over t' is bounded so that for some c_0, $0 < c_0 < \infty$,

$$\begin{aligned} E\left[\|V_{1T}\|\right] &\leq \frac{n_T}{T-n_T}\sum_{m=n_T+1}^{\infty}\sum_{m'=n_T+1}^{\infty}\sum_{j=1}^{K}\sum_{j'=1}^{K}c_0\pi_{ij,m}\pi_{ij',m'} \\ &\to 0 \end{aligned}$$

using the result from Lemma **9.3**.

For $\|\hat{B}(n_T)^{-1}\|_1$ and $\|\hat{B}(n_T)^{-1}\|_\infty$, the existence of $\hat{B}(n_T)^{-1}$ is guaranteed by a lemma that can be found in Tiao and Tsay (1983). The argument is the following. It is

clear that $\hat{B}(n_T)$ is a symmetric non-negative definite matrix. To show that it is positive definite take $c = [c_1, \ldots c_{k\,n_T}]'$ be any arbitrary vector and consider

$$c'\hat{B}(n_T)c = \frac{1}{T^2}\sum_{t=n_T+1}^{T}\left(\sum_{j=1}^{n_T}\sum_{k=1}^{K} c_{(j-1)K+k}y_{t-j}(k)\right)^2$$

If $c'\hat{B}(n_T)c = 0$, then

$$\sum_{j=1}^{n_T}\sum_{k=1}^{K} c_{(j-1)K+k}y_{t-j}(k) = 0 \quad \text{for} \quad t = n_T+1, \ldots, T$$

which since $T > 2Kn_T$, is a system of linear equations of Kn_T unknowns and at least Kn_T equations. Since Y_t is continuous and non deterministic, this implies that $c = 0$. This proves that $\hat{B}(n_T)$ is positive definite. Denoting by $B(n_T)$ the $(Kn_T \times Kn_T)$ matrix of the corresponding covariances instead of the empirical covariances, we can use a similar argument to show that $B(n_T)$ is also positive definite.

We next show that the sum of the elements along a row of B_{n_T} is uniformly bounded in n_T and the row number. If we take the sum along the row $h = (i-1)K + j$

$$\begin{aligned}
&\sum_{l=1}^{Kn_T} B_{hl}(n_T) \\
&= \sum_{k=1}^{K}\sum_{l=1}^{n_T} E[y_{t-j}(i)y_{t-l}(k)] \\
&= \sum_{k=1}^{K}\sum_{l=1}^{n_T} E\left[\left(\sum_{u=0}^{\infty}\sum_{v=1}^{K}\psi_{iv,u}u_{t-j-u}(v)\right)\left(\sum_{u'=0}^{\infty}\sum_{v'=1}^{K}\psi_{kv',u'}u_{t-l-u'}(v')\right)\right] \\
&= \sum_{k=1}^{K}\sum_{l=1}^{n_T}\sum_{u=0}^{\infty}\sum_{v=1}^{K}\sum_{u'=0}^{\infty}\sum_{v'=1}^{K}\psi_{iv,u}\psi_{kv',u'}E\left[u_{t-j-u}(v)u_{t-l-u'}(v')\right] \\
&= \sum_{k=1}^{K}\sum_{l=1}^{n_T}\sum_{v=1}^{K}\sum_{u'=0}^{\infty}\sum_{v'=1}^{K}\psi_{iv,l+u'-j}\psi_{kv',u'}\Sigma_{vv'} \\
&= \sum_{k=1}^{K}\sum_{u'=0}^{\infty}\sum_{v'=1}^{K}\psi_{kv',u'}\sum_{l=1}^{n_T}\sum_{v=1}^{K}\psi_{iv,l+u'-j}\Sigma_{vv'}
\end{aligned}$$

which is bounded because the Ψ's decreases exponentially. This property also holds for

$B(n_T)^{-1}$. If it was not the case we would have

$$\begin{aligned} B(n_T)^{-1}B(n_T) &= I_{n_T} \\ B(n_T)^{-1}B(n_T)\mathbf{i}_{n_T} &= \mathbf{i}_{n_T} \end{aligned}$$

with $\mathbf{i}_{n_T}$ an $(K\ n_T \times 1)$ vector of ones. Since $B(n_T)\mathbf{i}_{n_T}$ gives a vector of bounded elements, $B(n_T)^{-1}$ must have the sum of elements bounded along any row.

From lemma **9.4** we know that each element of $\hat{B}(n_T) - B_{n_T}$ converges in quadratic mean to zero and that $\|\hat{B}(n_T) - B_{n_T}\| \leq c_0(Kn_T)^2/T$. Hence, $\|\hat{B}(n_T) - B_{n_T}\| \xrightarrow{T\to\infty} 0$ and the sum of the elements along a row of $B_{n_T}^{-1}$ is bounded. It follows that the two terms on the right-hand side of equation (9.10) converge in quadratic mean to zero.

□

PROOF OF THEOREM **4.2** If we denote by Z_{t-1} the equivalent of $\hat{Z}_{t-1}$ which contains the true innovations $u_t(k)$ instead of the residuals $\hat{u}_t(k)$,

$$\begin{aligned} \hat{\gamma} &= \left[\sum_{t=l}^{T} \hat{Z}'_{t-1}\hat{\Sigma}^{-1}\hat{Z}_{t-1}\right]^{-1}\left[\sum_{t=l}^{T} \hat{Z}_{t-1}\hat{\Sigma}^{-1}\left(Z_{t-1}\gamma + U_t\right)\right] \\ &= \left[\sum_{t=l}^{T} \hat{Z}'_{t-1}\hat{\Sigma}^{-1}\hat{Z}_{t-1}\right]^{-1}\left[\sum_{t=l}^{T} \hat{Z}'_{t-1}\hat{\Sigma}^{-1}Z_{t-1}\right]\gamma + \\ &\quad \left[\sum_{t=l}^{T} \hat{Z}'_{t-1}\hat{\Sigma}^{-1}\hat{Z}_{t-1}\right]^{-1}\left[\sum_{t=l}^{T} \hat{Z}_{t-1}\hat{\Sigma}^{-1}U_t\right] \end{aligned}$$

Firstly, we show that $\hat{\Sigma} \to \Sigma$:

$$\begin{aligned} & \frac{1}{T-n_T}\sum_{t=n_T+1}^{T} \hat{U}_t\hat{U}'_t \\ = & \frac{1}{T-n_T}\sum_{t=n_T+1}^{T} \left(\hat{\Pi}^{n_T}(L)Y_t\right)\left(\cdots\right)' \\ = & \frac{1}{T-n_T}\sum_{t=n_T+1}^{T} \left(\hat{\Pi}^{n_T}(L)\Psi(L)U_t\right)\left(\cdots\right)' \\ = & \frac{1}{T-n_T}\sum_{t=n_T+1}^{T} \left([I_K + (\hat{\Pi}^{n_T}(L)\Psi(L) - I_K)]U_t\right)\left(\cdots\right)' \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{T-n_T}\sum_{t=n_T+1}^{T}\left([I_K+(\hat{\Pi}^{n_T}(L)-\Pi(L))\Psi(L)]U_t\right)\left(\cdots\right)' \\
&= \frac{1}{T-n_T}\sum_{t=n_T+1}^{T}\left([U_t+(\hat{\Pi}^{n_T}(L)-\Pi(L))Y_t]\right)\left(\cdots\right)' \\
&\xrightarrow{p} \Sigma
\end{aligned}
$$

using the results from Theorem **4.1** where we showed that $\|\hat{\Pi}^{n_T}(L)-\Pi(L)\| \xrightarrow{q.m.} 0$.

To show that $\sum_{t=l}^{T}\hat{Z}'_{t-1}\hat{\Sigma}^{-1}\hat{Z}_{t-1}/T$ converge to $\tilde{J}=E[Z'_{t-1}\Sigma^{-1}Z_{t-1}]$ in probability, since $\hat{\Sigma}\to\Sigma$ we only have to show that

- $\sum_{t=l}^{T} y_{t-k}(i)y_{t-l}(j)/T \to E[y_{t-k}(i)y_{t-l}(j)]$,
- $\sum_{t=l}^{T} \hat{u}_{t-k}(i)\hat{u}_{t-l}(j)/T \to E[u_{t-k}(i)u_{t-l}(j)]$,
- $\sum_{t=l}^{T} y_{t-k}(i)\hat{u}_{t-l}(j) \to E[y_{t-k}(i)u_{t-l}(j)]$.

The first is proved in lemma **9.4** and the second can be proved in a similar manner. We can easily prove the third by using results for the previous two and Theorem **4.1**. Similarly, $\sum_{t=l}^{T}\hat{Z}'_{t-1}\hat{\Sigma}^{-1}Z_{t-1}/T$ converge also to $\tilde{J}$.

We next study $\sum_{t=l}^{T}\hat{Z}_{t-1}\hat{\Sigma}^{-1}U_t/T$. Using similar calculus we see that it converge in probability to zero. Combining all these results we can conclude that $\tilde{\gamma}-\gamma \xrightarrow{p} 0$.

For the asymptotic distribution, using Ibragimov's central limit theorem we can conclude that

$$
\frac{1}{\sqrt{T}}\sum_{t=l}^{T}\hat{Z}'_{t-1}\hat{\Sigma}^{-1}U_t \xrightarrow{d} \mathcal{N}\left(0,\hat{I}\right)
$$

with

$$
\tilde{I}=\sum_{j=-\infty}^{\infty}E\left[\left\{Z'_{t-1}\Sigma^{-1}U_t\right\}\left\{Z'_{t-1-j}\Sigma^{-1}U_{t-j}\right\}\right].
$$

From this,

$$
\sqrt{T}\left(\tilde{\gamma}-\gamma\right) \xrightarrow{d} \mathcal{N}\left(0,\tilde{J}^{-1}\tilde{I}\tilde{J}^{-1}\right).
$$

From the preceding results, it is obvious that $\tilde{J}$ can be consistently estimated by

$$\tilde{J}_T = \frac{1}{T}\sum_{t=n_T+1}^{T} \hat{Z}'_{t-1}\hat{\Sigma}^{-1}\hat{Z}_{t-1}$$

and using theorem 2 of Newey and West (1987b), we know that $\tilde{I}_T \xrightarrow{P} \tilde{I}$ if we take $m_T^4/T \longrightarrow 0$ with $m_T \longrightarrow \infty$. □

PROOF OF THEOREM **4.3**

First we can rewrite X_t, W_t and $\tilde{V}_t$ as

$$\begin{aligned} X_t &= \hat{B}(L)^{-1}Y_t \\ W_t &= \hat{B}(L)^{-1}\tilde{U}_t \\ \tilde{V}_t &= \hat{B}(L)^{-1}\tilde{Z}_t. \end{aligned}$$

We can also rewrite $\tilde{U}_t + X_t - W_t$ as

$$\begin{aligned} \tilde{U}_t + X_t - W_t &= \tilde{U}_t + \hat{B}(L)^{-1}\left[Y_t - \tilde{U}_t\right] \\ &= \hat{B}(L)^{-1}Y_t + \tilde{U}_t - \hat{B}(L)^{-1}\tilde{U}_t \\ &= \hat{B}(L)^{-1}Z_{t-1}\gamma + \hat{B}(L)^{-1}U_t + \tilde{U}_t - \hat{B}(L)^{-1}\tilde{U}_t \\ &= V_{t-1}\gamma + \left[I_K + O(T^{-1/2})\right]U_t. \end{aligned}$$

With this, the regression becomes

$$\begin{aligned} \hat{\gamma} &= \left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}'_{t-1}\tilde{\Sigma}^{-1}\tilde{V}_{t-1}\right]^{-1}\left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}'_{t-1}\tilde{\Sigma}^{-1}\left(\tilde{U}_t + X_t - W_t\right)\right] \\ &= \left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}'_{t-1}\tilde{\Sigma}^{-1}\tilde{V}_{t-1}\right]^{-1}\left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}'_{t-1}\tilde{\Sigma}^{-1}V_{t-1}\right]\gamma + \\ &\quad \left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}'_{t-1}\tilde{\Sigma}^{-1}\tilde{V}_{t-1}\right]^{-1}\left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}'_{t-1}\tilde{\Sigma}^{-1}U_t\right] + O(T^{-1}) \end{aligned}$$

just like in the proof of theorem **4.2** we see that $\hat{\gamma} - \gamma = O(T^{-1/2})$. Using Ibragimov's

central limit theorem we conclude that

$$\sqrt{T}(\hat{\gamma}-\gamma) \xrightarrow{d} \mathcal{N}\left(0, \hat{J}^{-1}\hat{I}\hat{J}\right)$$

with

$$\begin{aligned}\hat{I} &= \sum_{j=-\infty}^{\infty} E\left[\left\{V_{t-1}\Sigma^{-1}U_t\right\}\left\{V_{t-1-j}\Sigma^{-1}U_{t-j}\right\}\right] \\ \hat{J} &= E\left[V'_{t-1}\Sigma^{-1}V_{t-1}\right]\end{aligned}$$

and just like in the proof of theorem **4.2** the matrix $\hat{I}$ and $\hat{J}$ can be consistently estimated respectively by

$$\begin{aligned}\hat{I}_T &= \frac{1}{T}\sum_{j=-m_T}^{m_T}\omega(j,m_T)\sum_{t=\max(p,q)+1+|j|}^{T}\left\{\tilde{V}_{t-1}\tilde{\Sigma}^{-1}\tilde{U}_t\right\}\left\{\tilde{V}_{t-1-j}\tilde{\Sigma}^{-1}\tilde{U}_{t-j}\right\} \\ \hat{J}_T &= \frac{1}{T}\sum_{t=\max(p,q)+1}^{T}\tilde{V}'_{t-1}\tilde{\Sigma}^{-1}\tilde{V}_{t-1}.\end{aligned}$$

□

PROOF OF THEOREM **5.1**

Take the difference between the information criterion for given values of the orders p and q, and its true value (for the true values p_0, q_0 and Σ_0)

$$\log(\det\tilde{\Sigma}) - \log\left(\det\Sigma_0\right) + \left[\dim\gamma(p,q) - \dim\gamma(p_0,q_0)\right]\frac{(\log T)^{1+\delta}}{T}.$$

First, consider the case where $p < p_0$ or $q < q_0$. In this case, as T grows to infinity, eventually $\log(\det\tilde{\Sigma}) > \log(\det\Sigma_0)$ because of the left-coprime property. So eventually we must have $p \geq p_0$ and $q \geq q_0$ because the difference of the two criterions will be positive. Second, consider the case where $p \geq p_0$ or $q \geq q_0$. We will first study the behavior of $\tilde{\Sigma}$ for the case $p = p_0$ and $q = q_0$. Dropping the inf after the first

equality and using the subscript 0 for true values of a parameter we get:

$$\begin{aligned}
\tilde{\Sigma} &= \inf_{A,B} \frac{1}{T} \sum_{t=n_T+1}^{T} \tilde{U}_t \tilde{U}_t' \\
&= \frac{1}{T} \sum_{t=n_T+1}^{T} \left[A(L)Y_t - (B(L) - I_K)\hat{U}_t \right] \left[\cdots \right]' \\
&= \frac{1}{T} \sum_{t=n_T+1}^{T} \left[A(L)Y_t - (B(L) - I_K)\hat{\Pi}(L)Y_t \right] \left[\cdots \right]' \\
&= \frac{1}{T} \sum_{t=n_T+1}^{T} \left[A(L) - (B(L) - I_K)\hat{\Pi}(L) \right] Y_t Y_t' \left[\cdots \right]' \\
&= \frac{1}{T} \sum_{t=n_T+1}^{T} \left[A(L) - (B(L) - I_K)B_0(L)^{-1}B_0(L)\hat{\Pi}(L) \right] Y_t Y_t' \left[\cdots \right]' \\
&= \frac{1}{T} \sum_{t=n_T+1}^{T} \left[\chi(L) - (B(L) - I_K)B_0(L)^{-1}C_T(L) + B_0(L)^{-1}A_0(L) \right] Y_t Y_t' \left[\cdots \right]'.
\end{aligned}$$

where $C_T(L) = B_0(L)\hat{\Pi}(L) - A_0(L)$ and $\chi(L) = A(L) - B(L)B_0(L)^{-1}A_0(L)$. From previous calculus we know that

$$\begin{aligned}
&\frac{1}{T} \sum_{t=n_T+1}^{T} \left\{ B_0(L)^{-1}A_0(L)Y_t \right\} \left\{ B_0(L)^{-1}A_0(L)Y_t \right\}' \\
&= \frac{1}{T} \sum_{t=n_T+1}^{T} U_t U_t' \\
&= \Sigma_0 + O(T^{-1/2}).
\end{aligned}$$

For the cross-product involving $\chi(L)$ we see that

$$\begin{aligned}
&\frac{1}{T} \sum_{t=n_T+1}^{T} \left\{ \chi(L)Y_t \right\} \left\{ \chi(L)Y_t \right\}' \\
&= \frac{1}{T} \sum_{t=n_T+1}^{T} \left\{ [A(L) - B(L)B_0(L)^{-1}A_0(L)]Y_t \right\} \left\{ [A(L) - B(L)B_0(L)^{-1}A_0(L)]Y_t \right\}' \\
&= \frac{1}{T} \sum_{t=n_T+1}^{T} \left\{ [A(L)A_0(L)^{-1}B_0(L) - B(L)]U_t \right\} \left\{ [A(L)A_0(L)^{-1}B_0(L) - B(L)]U_t \right\}' \\
&= O(T^{-1})
\end{aligned}$$

because $A(L)A_0(L)^{-1}B_0(L) - B(L)$ is $O(T^{-1/2})$. We have similar results for all the remaining cross-product except for

$$\frac{1}{T}\sum_{t=n_T+1}^{T}\left\{[B(L)-I_K]B_0(L)^{-1}C_{n_T}(L)Y_y\right\}\left\{[B(L)-I_K]B_0(L)^{-1}C_{n_T}(L)Y_y\right\}'.$$

We saw previously that $\sum_{l=1}^{n_T}\|\hat{\Pi}_l(n_T)-\Pi_l\| = O(n_T T^{-1/2})$ so we have the same result for $C_{n_T}(L)$, i.e. $\sum_{l=1}^{n_T+q}\sum_{j=1}^{K} C_l^{n_T}(i,j) = O(n_T T^{-1/2})$. Combining this with the fact that $\frac{1}{T}\sum_{t=n_T+1}^{T} Y_t Y_t' - E[Y_t Y_t'] = O(T^{-1/2})$ we can conclude that

$$\frac{1}{T}\sum_{t=n_T+1}^{T}\left\{[B(L)-I_K]B_0(L)^{-1}C_{n_T}(L)Y_t\right\}\left\{\cdots\right\}' = O(n_T T^{-1}).$$

Combining these results we see that for $p = p_0$ and $q = q_0$, $\tilde{\Sigma} = \Sigma_0 + O(n_T T^{-1})$ and equivalently

$$\det\tilde{\Sigma} = \det\Sigma_0 + O(n_T T^{-1}).$$

For the case where $p \geq p_0$, $q \geq q_0$ with either p or q greater then their true value, even though the model might not be identified in this case, for the minimization of $\det\tilde{\Sigma}$ we can not do any worse than in the case where $p = p_0$, $q = q_0$ so the infimum will yield the same result than for the case $p = p_0$ and $q = q_0$. So eventually

$$\det\tilde{\Sigma} - \det\Sigma_0 + [\dim\gamma(p,q) - \dim\gamma(p_0,q_0)]\frac{(\log T)^{1+\delta}}{T} \geq 0$$

because the penalty on the number of parameters will dominate. So if to select the order p and q we use an information criterion such as

$$\log(\det\hat{\Sigma}) + (\dim\gamma)\frac{(\log T)^{1+\delta}}{T} \tag{9.11}$$

where $\delta > 0$, we will get $\hat{p} \to p_0$, $\hat{q} \to q_0$ since $\log(\det\hat{\Sigma}) - \log(\det\Sigma_0) = O(n_T T^{-1})$.

□

Table 1: Strong final MA equation form VARMA(1,1). The simulated model is a strong VARMA(1,1) in final MA equation form with $a_1(1,1) = 0.5$, $a_1(1,2) = -0.6$, $a_1(2,1) = 0.7$, $a_1(2,2) = 0.3$ and $b_1 = 0.9$, . The variance of the innovations is 1.0 and the correlation is 0.7. Sample size is 250, the length of the long AR is $n_T = 20$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.3$.

$p \setminus q$	0	1	2	3	4	5	6
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.565	0.109	0.054	0.011	0.001	0.000
2	0.000	0.060	0.190	0.040	0.013	0.004	0.002
3	0.000	0.000	0.015	0.050	0.000	0.000	0.000
4	0.002	0.000	0.000	0.000	0.000	0.000	0.000
5	0.001	0.000	0.000	0.000	0.000	0.000	0.000
6	0.001	0.000	0.000	0.000	0.000	0.000	0.000

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Second step							
$a_1(1,1)$	0.5	0.4255	0.0596	0.0954	0.3282	0.5221	0.4287
$a_1(1,2)$	-0.6	-0.6385	0.0520	0.0647	-0.7247	-0.5539	-0.6397
$a_1(2,1)$	0.7	0.6686	0.0561	0.0643	0.5733	0.7592	0.6709
$a_1(2,2)$	0.3	0.2120	0.0555	0.1040	0.1201	0.3066	0.2131
b_1	0.9	0.8127	0.0566	0.1041	0.7225	0.9046	0.8141
Third step							
$a_1(1,1)$	0.5	0.4985	0.0502	0.0502	0.4122	0.5810	0.4997
$a_1(1,2)$	-0.6	-0.5883	0.0471	0.0486	-0.6624	-0.5097	-0.5899
$a_1(2,1)$	0.7	0.6825	0.0549	0.0576	0.5945	0.7657	0.6844
$a_1(2,2)$	0.3	0.3130	0.0558	0.0573	0.2322	0.3921	0.3100
b_1	0.9	0.8964	0.0327	0.0329	0.8438	0.9480	0.8968
MLE							
$a_1(1,1)$	0.5	0.4945	0.0486	0.0489	0.4105	0.5718	0.4959
$a_1(1,2)$	-0.6	-0.6084	0.0440	0.0448	-0.6793	-0.5363	-0.6096
$a_1(2,1)$	0.7	0.7002	0.0494	0.0494	0.6173	0.7787	0.7009
$a_1(2,2)$	0.3	0.2887	0.0451	0.0465	0.2135	0.3630	0.2886
b_1	0.9	0.8868	0.0253	0.0285	0.8442	0.9272	0.8880

Table 2: Strong final MA equation form VARMA(2,1). The simulated model is a strong VARMA(2,1) in final MA equation form with $a_1(1,1) = 0.9$, $a_1(1,2) = -0.5$, $a_1(2,1) = 0.3$, $a_1(2,2) = 0.1$, $a_2(1,1) = -0.1$, $a_2(1,2) = -0.2$, $a_2(2,1) = 0.1$, $a_2(2,2) = -0.15$ and $b_1 = 0.9$. The variance of the innovations is 1.0 and the correlation is 0.7. Sample size is 250, the length of the long AR is $n_T = 20$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.3$.

$p \setminus q$	0	1	2	3	4	5	6
0	0.000	0.001	0.059	0.131	0.027	0.001	0.001
1	0.000	0.012	0.075	0.033	0.010	0.001	0.000
2	0.000	0.343	0.166	0.104	0.019	0.001	0.001
3	0.000	0.001	0.011	0.002	0.001	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.000	0.000	0.000	0.000

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Third step							
$a_1(1,1)$	0.90	0.9186	0.1001	0.1018	0.7546	1.0774	0.9210
$a_1(1,2)$	-0.50	-0.4826	0.0905	0.0922	-0.6327	-0.3279	-0.4829
$a_1(2,1)$	0.30	0.2843	0.0873	0.0887	0.1415	0.4247	0.2860
$a_1(2,2)$	0.10	0.1461	0.1119	0.1210	-0.0391	0.3279	0.1458
$a_2(1,1)$	-0.10	-0.0815	0.0896	0.0914	-0.2281	0.0634	-0.0794
$a_2(1,2)$	-0.20	-0.1773	0.1103	0.1126	-0.3522	0.0099	-0.1833
$a_2(2,1)$	0.10	0.0766	0.1020	0.1046	-0.0917	0.2401	0.0794
$a_2(2,2)$	-0.15	-0.0939	0.1235	0.1356	-0.2825	0.1182	-0.1005
b_1	0.90	0.9094	0.0759	0.0765	0.7814	1.0272	0.9116
MLE							
$a_1(1,1)$	0.90	0.8756	0.0843	0.0877	0.7382	1.0130	0.8777
$a_1(1,2)$	-0.50	-0.4972	0.0876	0.0876	-0.6409	-0.3518	-0.4980
$a_1(2,1)$	0.30	0.2964	0.0858	0.0859	0.1496	0.4376	0.2985
$a_1(2,2)$	0.10	0.0792	0.0939	0.0962	-0.0749	0.2361	0.0810
$a_2(1,1)$	-0.10	-0.0904	0.0836	0.0842	-0.2328	0.0452	-0.0893
$a_2(1,2)$	-0.20	-0.2208	0.0961	0.0983	-0.3815	-0.0585	-0.2260
$a_2(2,1)$	0.10	0.1209	0.0893	0.0917	-0.0300	0.2678	0.1230
$a_2(2,2)$	-0.15	-0.1744	0.0997	0.1026	-0.3359	-0.0060	-0.1739
b_1	0.90	0.8811	0.0341	0.0390	0.8202	0.9312	0.8838

Table 3: Strong diagonal MA equation form VARMA(1,1). The simulated model is a strong VARMA(1,1) in diagonal MA equation form with $a_1(1,1) = 0.5$, $a_1(1,2) = -0.6$, $a_1(2,1) = 0.7$, $a_1(2,2) = 0.3$ and $b_1(1) = 0.9$, $b_1(2) = 0.7$. The variance of the innovations is 1.0 and the correlation is 0.7. Sample size is 250, the length of the long AR is $n_T = 20$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.3$.

(p, q_1, q_2)	Frequency	(p, q_1, q_2)	Frequency
1,1,1	0.579	2,2,2	0.022
1,2,1	0.124	1,4,1	0.018
1,1,2	0.076	2,2,1	0.014
1,3,1	0.060	1,5,1	0.013
1,1,3	0.026	2,1,1	0.011

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Second step							
$a_1(1,1)$	0.5	0.4282	0.0573	0.0919	0.3339	0.5172	0.4300
$a_1(1,2)$	-0.6	-0.6433	0.0507	0.0667	-0.7275	-0.5606	-0.6451
$a_1(2,1)$	0.7	0.6734	0.0491	0.0559	0.5911	0.7524	0.6742
$a_1(2,2)$	0.3	0.2312	0.0536	0.0872	0.1447	0.3204	0.2311
$b_1(1)$	0.9	0.8146	0.0681	0.1092	0.6985	0.9272	0.8142
$b_1(2)$	0.7	0.6358	0.0695	0.0946	0.5184	0.7447	0.6375
Third step							
$a_1(1,1)$	0.5	0.5069	0.0627	0.0630	0.4267	0.5819	0.5074
$a_1(1,2)$	-0.6	-0.5952	0.0490	0.0492	-0.6716	-0.5135	-0.5958
$a_1(2,1)$	0.7	0.6967	0.0421	0.0422	0.6246	0.7663	0.6980
$a_1(2,2)$	0.3	0.3017	0.0459	0.0459	0.2281	0.3798	0.2992
$b_1(1)$	0.9	0.8882	0.0416	0.0433	0.8201	0.9526	0.8895
$b_1(2)$	0.7	0.6937	0.0520	0.0523	0.6107	0.7813	0.6936
MLE							
$a_1(1,1)$	0.5	0.4967	0.0433	0.0434	0.4263	0.5638	0.5004
$a_1(1,2)$	-0.6	-0.6114	0.0446	0.0460	-0.6848	-0.5378	-0.6125
$a_1(2,1)$	0.7	0.6994	0.0419	0.0419	0.6301	0.7675	0.7006
$a_1(2,2)$	0.3	0.2894	0.0432	0.0445	0.2159	0.3602	0.2875
$b_1(1)$	0.9	0.8878	0.0316	0.0339	0.8325	0.9378	0.8893
$b_1(2)$	0.7	0.6937	0.0452	0.0457	0.6200	0.7652	0.6952

Table 4: Strong diagonal MA equation form VARMA(2,1). The simulated model is a strong VARMA(2,1) in diagonal MA equation form with $a_1(1,1) = 0.9$, $a_1(1,2) = -0.5$, $a_1(2,1) = 0.3$, $a_1(2,2) = 0.1$, $a_2(1,1) = -0.1$, $a_2(1,2) = -0.2$, $a_2(2,1) = 0.1$, $a_2(2,2) = -0.15$ and $b_1(1) = 0.9$, $b_1(2) = 0.7$. The variance of the innovations is 1.0 and the correlation is 0.7. Sample size is 250, the length of the long AR is $n_T = 20$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.2$.

(p, q_1, q_2)	Frequency	(p, q_1, q_2)	Frequency
2,1,1	0.263	2,1,0	0.047
1,2,1	0.224	2,4,0	0.031
2,3,0	0.068	0,3,1	0.025
1,3,1	0.055	1,4,1	0.018
2,3,1	0.054	2,3,2	0.018

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Third step							
$a_1(1,1)$	0.90	0.9194	0.0997	0.1016	0.7548	1.0842	0.9185
$a_1(1,2)$	-0.50	-0.5094	0.0893	0.0898	-0.6603	-0.3665	-0.5104
$a_1(2,1)$	0.30	0.3033	0.0801	0.0801	0.1685	0.4306	0.3047
$a_1(2,2)$	0.10	0.1122	0.1652	0.1657	-0.1433	0.3959	0.1080
$a_2(1,1)$	-0.10	-0.0724	0.0887	0.0930	-0.2180	0.0707	-0.0712
$a_2(1,2)$	-0.20	-0.1927	0.1242	0.1244	-0.3875	0.0220	-0.1989
$a_2(2,1)$	0.10	0.0961	0.1140	0.1141	-0.1122	0.2644	0.1095
$a_2(2,2)$	-0.15	-0.1231	0.1426	0.1451	-0.3298	0.1417	-0.1365
$b_1(1)$	0.90	0.8934	0.0784	0.0787	0.7620	1.0148	0.8982
$b_1(2)$	0.70	0.7100	0.1455	0.1458	0.4807	0.9689	0.7063
MLE							
$a_1(1,1)$	0.90	0.8789	0.0841	0.0867	0.7432	1.0168	0.8804
$a_1(1,2)$	-0.50	-0.4997	0.0879	0.0879	-0.6476	-0.3555	-0.4990
$a_1(2,1)$	0.30	0.2960	0.0796	0.0797	0.1617	0.4238	0.2988
$a_1(2,2)$	0.10	0.0739	0.1199	0.1227	-0.1341	0.2619	0.0768
$a_2(1,1)$	-0.10	-0.0827	0.0827	0.0845	-0.2225	0.0556	-0.0810
$a_2(1,2)$	-0.20	-0.2278	0.1084	0.1119	-0.4039	-0.0492	-0.2289
$a_2(2,1)$	0.10	0.1206	0.0869	0.0893	-0.0303	0.2580	0.1225
$a_2(2,2)$	-0.15	-0.1754	0.1059	0.1089	-0.3471	0.0078	-0.1764
$b_1(1)$	0.90	0.8875	0.0402	0.0421	0.8177	0.9468	0.8908
$b_1(2)$	0.70	0.6794	0.0869	0.0893	0.5248	0.8079	0.6892

Table 5: Strong final AR equation form VARMA(1,1). The simulated model is a strong VARMA(1,1) in final AR equation form with $a_1 = 0.9$, $b(1,1) = 0.5$, $b(1,2) = -0.6$, $b(2,1) = 0.7$, $b(2,2) = 0.3$. The variance of the innovations is 1.0 and the correlation is 0.7. Sample size is 250, the length of the long AR is $n_T = 15$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.2$.

$p \setminus q$	0	1	2	3	4
0	0.000	0.000	0.000	0.000	0.000
1	0.000	0.484	0.002	0.000	0.000
2	0.000	0.178	0.007	0.001	0.000
3	0.000	0.113	0.001	0.000	0.000
4	0.000	0.089	0.003	0.000	0.001
5	0.000	0.042	0.001	0.000	0.000
6	0.000	0.076	0.002	0.000	0.000

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Second step							
a_1	0.9	0.8861	0.0298	0.0329	0.8326	0.9279	0.8907
$b_1(1,1)$	0.5	0.4956	0.0979	0.0980	0.3413	0.6583	0.4938
$b_1(1,2)$	-0.6	-0.6051	0.0976	0.0977	-0.7693	-0.4425	-0.6045
$b_1(2,1)$	0.7	0.7030	0.0980	0.0980	0.5481	0.8730	0.7005
$b_1(2,2)$	0.3	0.2890	0.1002	0.1008	0.1302	0.4499	0.2877
Third step							
a_1	0.9	0.8932	0.0234	0.0244	0.8507	0.9261	0.8957
$b_1(1,1)$	0.5	0.4978	0.0570	0.0570	0.4010	0.5878	0.4986
$b_1(1,2)$	-0.6	-0.5945	0.0584	0.0587	-0.6853	-0.4999	-0.5963
$b_1(2,1)$	0.7	0.6988	0.0574	0.0574	0.6052	0.7913	0.7007
$b_1(2,2)$	0.3	0.2998	0.0598	0.0598	0.2027	0.3947	0.2995
MLE							
a_1	0.9	0.8963	0.0220	0.0223	0.8570	0.9284	0.8990
$b_1(1,1)$	0.5	0.4998	0.0496	0.0496	0.4165	0.5794	0.5006
$b_1(1,2)$	-0.6	-0.5998	0.0496	0.0497	-0.6766	-0.5218	-0.6012
$b_1(2,1)$	0.7	0.7028	0.0484	0.0485	0.6224	0.7785	0.7054
$b_1(2,2)$	0.3	0.2988	0.0495	0.0495	0.2193	0.3806	0.2961

Table 6: Strong final AR equation form VARMA(1,2). The simulated model is a strong VARMA(1,2) in final AR equation form with $a_1 = 0.9$, $b_1(1,1) = 0.9$, $b_1(1,2) = -0.5$, $b_1(2,1) = 0.3$, $b_1(2,2) = 0.1$, $b_2(1,1) = -0.1$, $b_2(1,2) = -0.2$, $b_2(2,1) = 0.1$, $b_2(2,2) = -0.15$. The variance of the innovations is 1.0 and the correlation is 0.7. Sample size is 250, the length of the long AR is $n_T = 15$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.5$.

$p \setminus q$	0	1	2	3	4	5
0	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.160	0.161	0.000	0.001	0.000
2	0.090	0.236	0.142	0.001	0.000	0.000
3	0.039	0.051	0.041	0.000	0.000	0.000
4	0.012	0.018	0.016	0.001	0.000	0.000
5	0.007	0.011	0.003	0.000	0.000	0.000
6	0.005	0.005	0.001	0.000	0.000	0.000

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Third step							
a_1	0.90	0.8959	0.0262	0.0265	0.8456	0.9331	0.8994
$b_1(1,1)$	0.90	0.8955	0.0910	0.0911	0.7501	1.0495	0.8949
$b_1(1,2)$	-0.50	-0.5004	0.0860	0.0860	-0.6472	-0.3608	-0.5004
$b_1(2,1)$	0.30	0.3000	0.0874	0.0874	0.1596	0.4458	0.2993
$b_1(2,2)$	0.10	0.0950	0.0870	0.0871	-0.0448	0.2373	0.0944
$b_2(1,1)$	-0.10	-0.0953	0.0949	0.0950	-0.2477	0.0716	-0.0946
$b_2(1,2)$	-0.20	-0.2035	0.0995	0.0995	-0.3704	-0.0393	-0.1994
$b_2(2,1)$	0.10	0.1016	0.0870	0.0870	-0.0315	0.2441	0.1012
$b_2(2,2)$	-0.15	-0.1502	0.0917	0.0917	-0.3014	-0.0069	-0.1463
MLE							
a_1	0.90	0.8953	0.0258	0.0263	0.8484	0.9326	0.8987
$b_1(1,1)$	0.90	0.9049	0.0908	0.0909	0.7620	1.0572	0.9010
$b_1(1,2)$	-0.50	-0.5042	0.0856	0.0857	-0.6510	-0.3646	-0.5033
$b_1(2,1)$	0.30	0.3032	0.0873	0.0874	0.1635	0.4501	0.3042
$b_1(2,2)$	0.10	0.0972	0.0868	0.0869	-0.0424	0.2395	0.0961
$b_2(1,1)$	-0.10	-0.0991	0.0953	0.0953	-0.2536	0.0657	-0.1005
$b_2(1,2)$	-0.20	-0.2071	0.0988	0.0991	-0.3725	-0.0450	-0.2040
$b_2(2,1)$	0.10	0.1050	0.0866	0.0868	-0.0286	0.2459	0.1050
$b_2(2,2)$	-0.15	-0.1568	0.0911	0.0914	-0.3139	-0.0162	-0.1507

Table 7: Strong diagonal AR equation form VARMA(1,1). The simulated model is a strong VARMA(1,1) in diagonal AR equation form with $a_1(1) = 0.9$, $a_1(2) = 0.7$, $b(1,1) = 0.5$, $b(1,2) = -0.6$, $b(2,1) = 0.7$, $b(2,2) = 0.3$. The variance of the innovations is 1.0 and the correlation is 0.7. Sample size is 250, the length of the long AR is $n_T = 15$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.5$.

(p_1, p_2, q)	Frequency	(p_1, p_2, q)	Frequency
1,1,1	0.414	1,4,1	0.038
1,2,1	0.116	1,6,1	0.026
2,1,1	0.098	4,1,1	0.025
1,3,1	0.057	1,2,2	0.023
3,1,1	0.040	1,5,1	0.016

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Second step							
$a_1(1)$	0.9	0.9243	0.0298	0.0385	0.8720	0.9666	0.9283
$a_1(2)$	0.7	0.4898	0.0934	0.2300	0.3352	0.6347	0.4915
$b_1(1,1)$	0.5	0.5327	0.1031	0.1082	0.3663	0.7023	0.5324
$b_1(1,2)$	-0.6	-0.6053	0.0975	0.0977	-0.7696	-0.4468	-0.6051
$b_1(2,1)$	0.7	0.7028	0.0976	0.0977	0.5469	0.8701	0.7010
$b_1(2,2)$	0.3	0.0916	0.1467	0.2549	-0.1585	0.3282	0.0917
Third step							
$a_1(1)$	0.9	0.8964	0.0320	0.0322	0.8391	0.9414	0.9017
$a_1(2)$	0.7	0.7077	0.0851	0.0854	0.5608	0.8377	0.7117
$b_1(1,1)$	0.5	0.4886	0.0875	0.0883	0.3557	0.6426	0.4821
$b_1(1,2)$	-0.6	-0.5893	0.0695	0.0703	-0.7054	-0.4817	-0.5853
$b_1(2,1)$	0.7	0.6834	0.0676	0.0696	0.5718	0.7922	0.6873
$b_1(2,2)$	0.3	0.3109	0.1303	0.1307	0.0983	0.5167	0.3207
MLE							
$a_1(1)$	0.9	0.8967	0.0301	0.0302	0.8404	0.9384	0.9012
$a_1(2)$	0.7	0.6772	0.0663	0.0701	0.5574	0.7691	0.6829
$b_1(1,1)$	0.5	0.5167	0.0752	0.0770	0.4089	0.6444	0.5102
$b_1(1,2)$	-0.6	-0.6099	0.0597	0.0606	-0.7127	-0.5202	-0.6067
$b_1(2,1)$	0.7	0.7119	0.0550	0.0563	0.6233	0.8000	0.7132
$b_1(2,2)$	0.3	0.2745	0.1004	0.1036	0.1025	0.4177	0.2884

Table 8: Strong diagonal AR equation form VARMA(1,2). The simulated model is a strong VARMA(1,2) in diagonal AR equation form with $a_1(1) = 0.9$, $a_1(2) = 0.7$, $b_1(1,1) = 0.9$, $b_1(1,2) = -0.5$, $b_1(2,1) = 0.3$, $b_1(2,2) = 0.1$, $b_2(1,1) = -0.1$, $b_2(1,2) = -0.2$, $b_2(2,1) = 0.1$, $b_2(2,2) = -0.15$. The variance of the innovations is 1.0 and the correlation is 0.7. Sample size is 250, the length of the long AR is $n_T = 15$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.1$.

(p_1, p_2, q)	Frequency	(p_1, p_2, q)	Frequency
2,1,1	0.316	1,1,1	0.043
3,1,1	0.136	2,1,2	0.034
4,1,1	0.106	6,1,1	0.025
1,1,2	0.083	2,1,0	0.020
5,1,1	0.047	3,1,2	0.016

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Third step							
$a_1(1)$	0.90	0.8946	0.0242	0.0248	0.8491	0.9312	0.8971
$a_1(2)$	0.70	0.6878	0.0773	0.0783	0.5534	0.8047	0.6897
$b_1(1,1)$	0.90	0.8976	0.0907	0.0908	0.7546	1.0525	0.8947
$b_1(1,2)$	-0.50	-0.5048	0.0871	0.0873	-0.6503	-0.3681	-0.5045
$b_1(2,1)$	0.30	0.3001	0.0883	0.0883	0.1598	0.4538	0.2999
$b_1(2,2)$	0.10	0.0854	0.1204	0.1213	-0.1177	0.2855	0.0811
$b_2(1,1)$	-0.10	-0.0933	0.0964	0.0966	-0.2514	0.0674	-0.0939
$b_2(1,2)$	-0.20	-0.2087	0.1001	0.1005	-0.3765	-0.0466	-0.2038
$b_2(2,1)$	0.10	0.1056	0.0974	0.0976	-0.0527	0.2639	0.1028
$b_2(2,2)$	-0.15	-0.1595	0.1181	0.1185	-0.3518	0.0415	-0.1620
MLE							
$a_1(1)$	0.90	0.8944	0.0242	0.0249	0.8512	0.9302	0.8964
$a_1(2)$	0.70	0.6819	0.0703	0.0726	0.5630	0.7882	0.6818
$b_1(1,1)$	0.90	0.9104	0.0904	0.0910	0.7720	1.0623	0.9083
$b_1(1,2)$	-0.50	-0.5103	0.0866	0.0872	-0.6544	-0.3712	-0.5076
$b_1(2,1)$	0.30	0.3051	0.0880	0.0882	0.1671	0.4515	0.3062
$b_1(2,2)$	0.10	0.0810	0.1131	0.1146	-0.1080	0.2663	0.0787
$b_2(1,1)$	-0.10	-0.0974	0.0972	0.0973	-0.2515	0.0743	-0.0998
$b_2(1,2)$	-0.20	-0.2104	0.1015	0.1021	-0.3824	-0.0408	-0.2072
$b_2(2,1)$	0.10	0.1116	0.0964	0.0971	-0.0439	0.2689	0.1098
$b_2(2,2)$	-0.15	-0.1702	0.1153	0.1170	-0.3592	0.0186	-0.1722

Table 9: Weak final MA equation form VARMA(1,1). The simulated model is a weak VARMA(1,1) in final MA equation form with $a(1,1) = 0.5$, $a(1,2) = -0.6$, $a(2,1) = 0.7$, $a(2,2) = 0.3$ and $b_1 = 0.9$. The variance of the innovations is 1.3 and the covariance is 0.91. Sample size is 250, the length of the long AR is $n_T = 20$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.3$.

$p \setminus q$	0	1	2	3	4	5
0	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.736	0.101	0.024	0.007	0.003
2	0.000	0.002	0.107	0.013	0.003	0.003
3	0.000	0.000	0.000	0.001	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Second step							
$a_1(1,1)$	0.5	0.4255	0.0629	0.0975	0.3243	0.5307	0.4281
$a_1(1,2)$	-0.6	-0.6390	0.0515	0.0646	-0.7239	-0.5541	-0.6401
$a_1(2,1)$	0.7	0.6682	0.0586	0.0666	0.5677	0.7599	0.6686
$a_1(2,2)$	0.3	0.2117	0.0551	0.1041	0.1195	0.3043	0.2129
b_1	0.9	0.8128	0.0593	0.1054	0.7148	0.9079	0.8139
Third step							
$a_1(1,1)$	0.5	0.5001	0.0505	0.0505	0.4174	0.5857	0.5006
$a_1(1,2)$	-0.6	-0.5896	0.0469	0.0481	-0.6685	-0.5154	-0.5899
$a_1(2,1)$	0.7	0.6859	0.0524	0.0543	0.6018	0.7682	0.6852
$a_1(2,2)$	0.3	0.3111	0.0494	0.0507	0.2341	0.3911	0.3101
b_1	0.9	0.8978	0.0348	0.0349	0.8368	0.9494	0.9000
NLLS							
$a_1(1,1)$	0.5	0.4952	0.0504	0.0507	0.4120	0.5789	0.4962
$a_1(1,2)$	-0.6	-0.6089	0.0432	0.0441	-0.6813	-0.5402	-0.6094
$a_1(2,1)$	0.7	0.7017	0.0494	0.0494	0.6209	0.7810	0.7023
$a_1(2,2)$	0.3	0.2875	0.0460	0.0476	0.2138	0.3660	0.2868
b_1	0.9	0.8866	0.0282	0.0312	0.8378	0.9294	0.8884

Table 10: Weak final MA equation form VARMA(2,1). The simulated model is a weak VARMA(2,1) in final MA equation form with $a_1(1,1) = 0.9$, $a_1(1,2) = -0.5$, $a_1(2,1) = 0.3$, $a_1(2,2) = 0.1$, $a_2(1,1) = -0.1$, $a_2(1,2) = -0.2$, $a_2(2,1) = 0.1$, $a_2(2,2) = -0.15$ and $b_1 = 0.9$, . The variance of the innovations is 1.3 and the covariance is 0.91. Sample size is 250, the length of the long AR is $n_T = 20$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.3$.

$p \setminus q$	0	1	2	3	4	5	6
0	0.000	0.001	0.050	0.114	0.022	0.007	0.003
1	0.000	0.013	0.084	0.035	0.006	0.002	0.001
2	0.000	0.318	0.180	0.116	0.021	0.003	0.002
3	0.000	0.003	0.015	0.002	0.002	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.000	0.000	0.000	0.000

	Value	Average	Std. dev.	RMSE	5%	95%	Median
third step							
$a_1(1,1)$	0.90	0.9219	0.1021	0.1044	0.7565	1.0866	0.9229
$a_1(1,2)$	-0.50	-0.4865	0.0930	0.0940	-0.6407	-0.3358	-0.4880
$a_1(2,1)$	0.30	0.2855	0.0871	0.0883	0.1396	0.4333	0.2872
$a_1(2,2)$	0.10	0.1425	0.1153	0.1229	-0.0514	0.3292	0.1442
$a_2(1,1)$	-0.10	-0.0807	0.0956	0.0975	-0.2389	0.0726	-0.0799
$a_2(1,2)$	-0.20	-0.1803	0.1125	0.1142	-0.3645	0.0157	-0.1762
$a_2(2,1)$	0.10	0.0802	0.1027	0.1046	-0.0924	0.2525	0.0825
$a_2(2,2)$	-0.15	-0.1010	0.1197	0.1293	-0.2976	0.0964	-0.1025
b_1	0.90	0.9101	0.0737	0.0744	0.7836	1.0231	0.9125
NLLS							
$a_1(1,1)$	0.90	0.8768	0.0912	0.0941	0.7300	1.0322	0.8774
$a_1(1,2)$	-0.50	-0.4996	0.0907	0.0907	-0.6474	-0.3518	-0.4981
$a_1(2,1)$	0.30	0.2969	0.0857	0.0857	0.1564	0.4349	0.2925
$a_1(2,2)$	0.10	0.0762	0.1012	0.1039	-0.0953	0.2380	0.0824
$a_2(1,1)$	-0.10	-0.0879	0.0894	0.0903	-0.2360	0.0591	-0.0879
$a_2(1,2)$	-0.20	-0.2261	0.1000	0.1034	-0.3938	-0.0663	-0.2234
$a_2(2,1)$	0.10	0.1250	0.0907	0.0941	-0.0215	0.2737	0.1230
$a_2(2,2)$	-0.15	-0.1826	0.0987	0.1039	-0.3428	-0.0224	-0.1814
b_1	0.90	0.8811	0.0340	0.0389	0.8226	0.9294	0.8840

Table 11: Weak diagonal MA equation form VARMA(1,1). The simulated model is a weak VARMA(1,1) in diagonal MA equation form with $a(1,1) = 0.5$, $a(1,2) = -0.6$, $a(2,1) = 0.7$, $a(2,2) = 0.3$, $b_1(1) = 0.9$ and $b_1(1) = 0.7$. The variance of the innovations is 1.3 and the covariance is 0.91. Sample size is 250, the length of the long AR is $n_T = 20$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.3$.

(p, q_1, q_2)	Frequency	(p, q_1, q_2)	Frequency
1,1,1	0.588	1,1,3	0.026
1,2,1	0.123	2,1,1	0.014
1,1,2	0.062	1,4,1	0.014
1,3,1	0.045	1,5,1	0.012
2,2,2	0.043	1,1,5	0.010

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Second step							
$a_1(1,1)$	0.5	0.4277	0.0601	0.0940	0.3284	0.5233	0.4303
$a_1(1,2)$	-0.6	-0.6439	0.0507	0.0671	-0.7291	-0.5594	-0.6444
$a_1(2,1)$	0.7	0.6732	0.0514	0.0579	0.5863	0.7550	0.6729
$a_1(2,2)$	0.3	0.2314	0.0526	0.0865	0.1446	0.3193	0.2309
$b_1(1)$	0.9	0.8130	0.0707	0.1122	0.6976	0.9266	0.8150
$b_1(2)$	0.7	0.6364	0.0708	0.0952	0.5185	0.7476	0.6393
Third step							
$a_1(1,1)$	0.5	0.5064	0.0469	0.0473	0.4324	0.5845	0.5062
$a_1(1,2)$	-0.6	-0.5960	0.0552	0.0554	-0.6762	-0.5183	-0.5969
$a_1(2,1)$	0.7	0.6988	0.0418	0.0418	0.6314	0.7659	0.6997
$a_1(2,2)$	0.3	0.3021	0.0469	0.0469	0.2272	0.3830	0.3032
$b_1(1)$	0.9	0.8885	0.0442	0.0456	0.8100	0.9531	0.8910
$b_1(2)$	0.7	0.6967	0.0522	0.0523	0.6092	0.7843	0.6969
NLLS							
$a_1(1,1)$	0.5	0.4973	0.0453	0.0453	0.4222	0.5703	0.4972
$a_1(1,2)$	-0.6	-0.6116	0.0443	0.0458	-0.6864	-0.5371	-0.6114
$a_1(2,1)$	0.7	0.7009	0.0411	0.0411	0.6334	0.7683	0.7006
$a_1(2,2)$	0.3	0.2897	0.0441	0.0453	0.2185	0.3645	0.2893
$b_1(1)$	0.9	0.8874	0.0349	0.0371	0.8260	0.9385	0.8894
$b_1(2)$	0.7	0.6950	0.0446	0.0449	0.6198	0.7673	0.6955

Table 12: Weak diagonal MA equation form weak VARMA(2,1). The simulated model is a weak VARMA(2,1) in diagonal MA equation form with $a_1(1,1) = 0.9$, $a_1(1,2) = -0.5$, $a_1(2,1) = 0.3$, $a_1(2,2) = 0.1$, $a_2(1,1) = -0.1$, $a_2(1,2) = -0.2$, $a_2(2,1) = 0.1$, $a_2(2,2) = -0.15$, $b_1(1) = 0.9$, and $b_1(2) = 0.7$. The variance of the innovations is 1.3 and the covariance is 0.91. Sample size is 250, the length of the long AR is $n_T = 20$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.2$.

(p, q_1, q_2)	Frequency	(p, q_1, q_2)	Frequency
2,1,1	0.267	2,1,0	0.047
1,2,1	0.204	2,2,1	0.031
2,3,0	0.057	2,4,0	0.029
1,3,1	0.051	2,2,2	0.020
2,3,1	0.050	0,3,1	0.020

	Value	Average	Std. dev.	RMSE	5%	95%	Median
third step							
$a_1(1,1)$	0.90	0.9205	0.1016	0.1036	0.7554	1.0882	0.9204
$a_1(1,2)$	-0.50	-0.5137	0.0922	0.0932	-0.6643	-0.3620	-0.5142
$a_1(2,1)$	0.30	0.3036	0.0802	0.0802	0.1737	0.4326	0.3020
$a_1(2,2)$	0.10	0.1071	0.1666	0.1668	-0.1533	0.3802	0.1037
$a_2(1,1)$	-0.10	-0.0716	0.0937	0.0979	-0.2302	0.0781	-0.0715
$a_2(1,2)$	-0.20	-0.1976	0.1262	0.1262	-0.3995	0.0159	-0.1995
$a_2(2,1)$	0.10	0.1014	0.1127	0.1127	-0.0969	0.2749	0.1111
$a_2(2,2)$	-0.15	-0.1326	0.1363	0.1374	-0.3462	0.1156	-0.1440
$b_1(1)$	0.90	0.8917	0.0774	0.0778	0.7654	1.0079	0.8973
$b_1(2)$	0.70	0.7084	0.1423	0.1426	0.4724	0.9397	0.7112
NLLS							
$a_1(1,1)$	0.90	0.8787	0.0914	0.0939	0.7254	1.0347	0.8799
$a_1(1,2)$	-0.50	-0.5015	0.0918	0.0918	-0.6523	-0.3517	-0.5008
$a_1(2,1)$	0.30	0.2957	0.0801	0.0802	0.1665	0.4244	0.2927
$a_1(2,2)$	0.10	0.0715	0.1252	0.1284	-0.1412	0.2643	0.0748
$a_2(1,1)$	-0.10	-0.0815	0.0887	0.0906	-0.2250	0.0622	-0.0814
$a_2(1,2)$	-0.20	-0.2328	0.1144	0.1190	-0.4268	-0.0491	-0.2332
$a_2(2,1)$	0.10	0.1243	0.0901	0.0933	-0.0216	0.2639	0.1278
$a_2(2,2)$	-0.15	-0.1831	0.1074	0.1124	-0.3522	-0.0050	-0.1826
$b_1(1)$	0.90	0.8861	0.0404	0.0427	0.8133	0.9448	0.8892
$b_1(2)$	0.70	0.6789	0.0883	0.0908	0.5215	0.8065	0.6889

Table 13: Weak diagonal AR equation form VARMA(1,1). The simulated model is a weak VARMA(1,1) in diagonal AR equation form with $a_1(1) = 0.9$, $a_1(2) = 0.7$, $b_1(1,1) = 0.5$, $b_1(1,2) = -0.6$, $b_1(2,1) = 0.7$, $b_1(2,2) = 0.3$. The variance of the innovations is 1.3 and the covariance is 0.91. Sample size is 250, the length of the long AR is $n_T = 15$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.2$.

(p_1, p_2, q)	Frequency	(p_1, p_2, q)	Frequency
1,1,1	0.428	1,2,2	0.039
1,2,1	0.091	1,4,1	0.034
2,1,1	0.088	1,5,1	0.023
1,3,1	0.064	4,1,1	0.019
3,1,1	0.044	2,2,2	0.019

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Second step							
$a_1(1)$	0.9	0.9244	0.0318	0.0401	0.8662	0.9709	0.9268
$a_1(2)$	0.7	0.4850	0.0894	0.2328	0.3211	0.6286	0.4853
$b_1(1,1)$	0.5	0.5309	0.1093	0.1136	0.3520	0.7219	0.5294
$b_1(1,2)$	-0.6	-0.6036	0.1008	0.1009	-0.7738	-0.4489	-0.6007
$b_1(2,1)$	0.7	0.7036	0.1004	0.1004	0.5417	0.8661	0.7031
$b_1(2,2)$	0.3	0.0880	0.1477	0.2584	-0.1642	0.3324	0.0875
Third step							
$a_1(1)$	0.9	0.8964	0.0332	0.0334	0.8395	0.9459	0.9000
$a_1(2)$	0.7	0.7073	0.0960	0.0963	0.5444	0.8503	0.7096
$b_1(1,1)$	0.5	0.4864	0.0930	0.0940	0.3347	0.6391	0.4836
$b_1(1,2)$	-0.6	-0.5847	0.0696	0.0712	-0.6988	-0.4690	-0.5871
$b_1(2,1)$	0.7	0.6813	0.0698	0.0723	0.5659	0.7928	0.6815
$b_1(2,2)$	0.3	0.3136	0.1401	0.1407	0.0809	0.5505	0.3179
NLLS							
$a_1(1)$	0.9	0.8961	0.0319	0.0322	0.8422	0.9415	0.9000
$a_1(2)$	0.7	0.6762	0.0676	0.0717	0.5438	0.7735	0.6835
$b_1(1,1)$	0.5	0.5152	0.0767	0.0782	0.3996	0.6510	0.5117
$b_1(1,2)$	-0.6	-0.6070	0.0582	0.0586	-0.7070	-0.5143	-0.6032
$b_1(2,1)$	0.7	0.7112	0.0544	0.0555	0.6265	0.8020	0.7080
$b_1(2,2)$	0.3	0.2759	0.0977	0.1006	0.0968	0.4252	0.2843

Table 14: Weak diagonal AR equation form VARMA(1,2). The simulated model is a weak VARMA(1,2) in diagonal AR equation form with $a_1(1) = 0.9$, $a_1(2) = 0.7$, $b_1(1,1) = 0.9$, $b_1(1,2) = -0.5$, $b_1(2,1) = 0.3$, $b_1(2,2) = 0.1$, $b_2(1,1) = -0.1$, $b_2(1,2) = -0.2$, $b_2(2,1) = 0.1$, $b_2(2,2) = -0.15$. The variance of the innovations is 1.3 and the covariance is 0.91. Sample size is 250, the length of the long AR is $n_T = 15$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.1$.

(p_1, p_2, q)	Frequency	(p_1, p_2, q)	Frequency
2,1,1	0.306	5,1,1	0.045
3,1,1	0.133	6,1,1	0.031
4,1,1	0.089	2,1,2	0.026
1,1,2	0.083	2,1,0	0.024
1,1,1	0.050	2,2,0	0.023

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Third step							
$a_1(1)$	0.9	0.8953	0.0238	0.0242	0.8538	0.9297	0.8977
$a_1(2)$	0.7	0.6901	0.0771	0.0777	0.5653	0.8075	0.6935
$b_1(1,1)$	0.5	0.8982	0.0976	0.0976	0.7368	1.0596	0.8973
$b_1(1,2)$	-0.6	-0.5040	0.0907	0.0908	-0.6568	-0.3589	-0.5043
$b_1(2,1)$	0.7	0.3019	0.0894	0.0894	0.1604	0.4479	0.3006
$b_1(2,2)$	0.3	0.0890	0.1244	0.1249	-0.1113	0.2933	0.0868
$b_2(1,1)$	-0.1	-0.0930	0.0983	0.0985	-0.2466	0.0827	-0.0973
$b_2(1,2)$	-0.2	-0.2049	0.1002	0.1004	-0.3729	-0.0342	-0.2072
$b_2(2,1)$	0.1	0.1011	0.0981	0.0982	-0.0563	0.2579	0.0995
$b_2(2,2)$	-0.15	-0.1518	0.1135	0.1135	-0.3355	0.0315	-0.1533
NLLS							
$a_1(1)$	0.9	0.8950	0.0240	0.0245	0.8534	0.9292	0.8973
$a_1(2)$	0.7	0.6822	0.0708	0.0730	0.5650	0.7901	0.6857
$b_1(1,1)$	0.5	0.9110	0.0967	0.0973	0.7517	1.0691	0.9121
$b_1(1,2)$	-0.6	-0.5094	0.0898	0.0903	-0.6639	-0.3673	-0.5090
$b_1(2,1)$	0.7	0.3067	0.0894	0.0896	0.1631	0.4474	0.3065
$b_1(2,2)$	0.3	0.0828	0.1192	0.1205	-0.1117	0.2787	0.0772
$b_2(1,1)$	-0.1	-0.0978	0.0998	0.0998	-0.2538	0.0725	-0.0985
$b_2(1,2)$	-0.2	-0.2065	0.1024	0.1026	-0.3753	-0.0278	-0.2081
$b_2(2,1)$	0.1	0.1081	0.0991	0.0994	-0.0586	0.2645	0.1074
$b_2(2,2)$	-0.15	-0.1646	0.1103	0.1113	-0.3415	0.0278	-0.1657

Table 15: Weak final AR equation form VARMA(1,1). The simulated model is a weak VARMA(1,1) in final AR equation form with $a_1 = 0.729$, $b(1,1) = 0.0593618$, $b(1,2) = -0.14134$, $b(2,1) = 0.20598$, $b(2,2) = 0.296472$. The variance of the innovations is 2.64155 and 1.70611 and the covariance is 0.650962. Sample size is 200, the length of the long AR is $n_T = 15$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.2$.

$p \setminus q$	0	1	2	3	4	5
0	0.000	0.000	0.000	0.000	0.000	0.000
1	0.129	0.574	0.000	0.000	0.000	0.000
2	0.092	0.113	0.000	0.000	0.000	0.000
3	0.018	0.034	0.000	0.000	0.000	0.000
4	0.009	0.015	0.000	0.000	0.000	0.000
5	0.005	0.004	0.000	0.000	0.000	0.000
5	0.003	0.004	0.000	0.000	0.000	0.000

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Second step							
a_1	0.7290	0.6962	0.0579	0.0665	0.6006	0.7827	0.7000
$b_1(1,1)$	0.0594	0.0254	0.0932	0.0992	-0.1250	0.1795	0.0215
$b_1(1,2)$	-0.1413	-0.1272	0.1041	0.1051	-0.2969	0.0390	-0.1258
$b_1(2,1)$	0.2060	0.1986	0.0679	0.0683	0.0836	0.3064	0.2004
$b_1(2,2)$	0.2965	0.3005	0.1091	0.1091	0.1083	0.4721	0.3076
Third step							
a_1	0.7290	0.7205	0.0559	0.0566	0.6203	0.8024	0.7236
$b_1(1,1)$	0.0594	0.0451	0.0829	0.0841	-0.0899	0.1820	0.0445
$b_1(1,2)$	-0.1413	-0.1265	0.0927	0.0938	-0.2827	0.0206	-0.1256
$b_1(2,1)$	0.2060	0.1984	0.0567	0.0572	0.1026	0.2887	0.1983
$b_1(2,2)$	0.2965	0.3178	0.0996	0.1018	0.1483	0.4760	0.3220
NLLS							
a_1	0.7290	0.7217	0.0557	0.0562	0.6268	0.8033	0.7242
$b_1(1,1)$	0.0594	0.0488	0.0832	0.0839	-0.0883	0.1869	0.0508
$b_1(1,2)$	-0.1413	-0.1291	0.0930	0.0938	-0.2921	0.0201	-0.1284
$b_1(2,1)$	0.2060	0.2000	0.0570	0.0573	0.1065	0.2901	0.2002
$b_1(2,2)$	0.2965	0.3206	0.0997	0.1026	0.1456	0.4757	0.3234

Figure 1: Macroeconomic series.

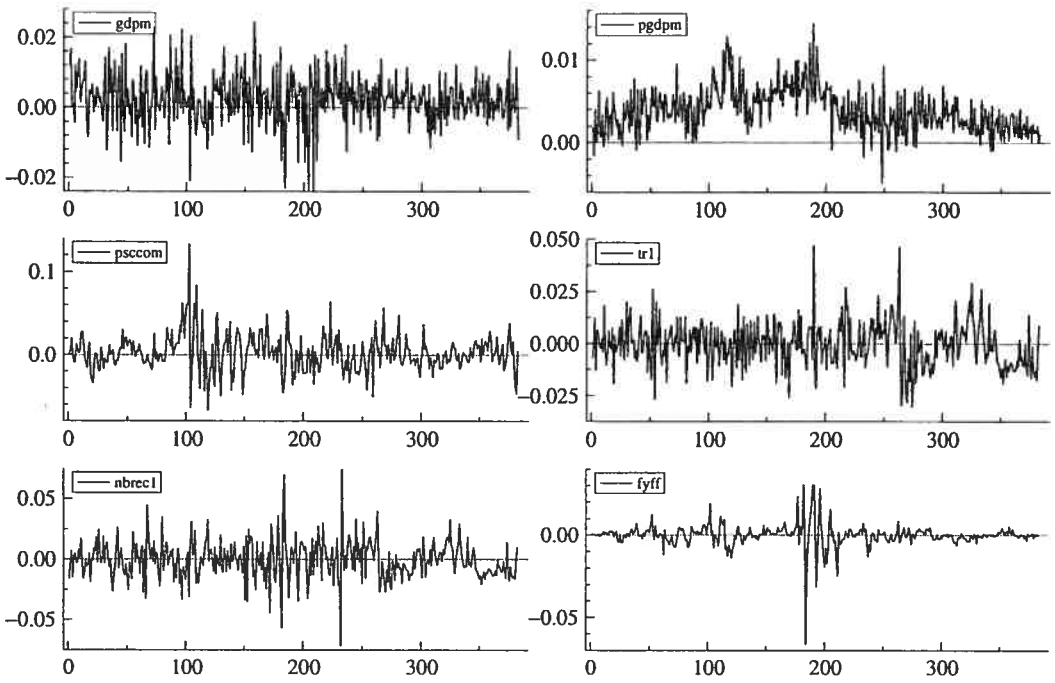


Figure 2: Impulse-response functions for VAR model. A VAR(12) is fitted to the first difference of the six time series. The confidence band represent a one standard deviation. The standard deviations are derived from Monte Carlo simulations with 1000 draws.

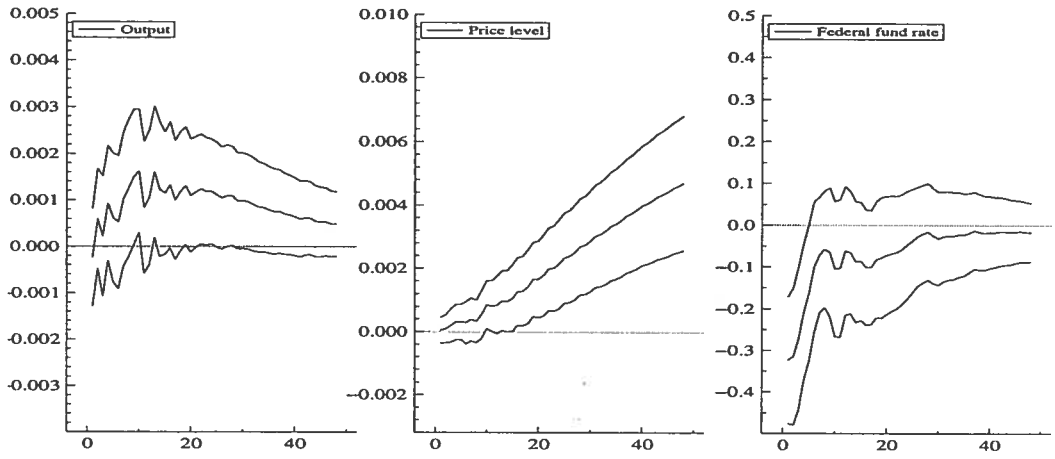


Figure 3: Impulse-response functions for VARMA model in final MA equation form. A VARMA(5,5) is fitted to the first difference of the six time series. The confidence band represent a one standard deviation. The standard deviations are derived from Monte Carlo simulations with 1000 draws.

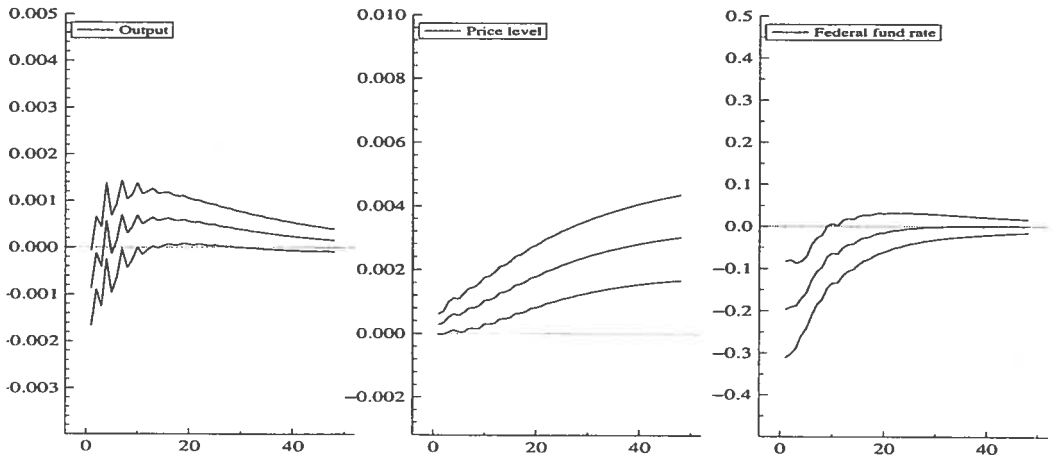


Figure 4: Impulse-response functions for VARMA model in diagonal MA equation form. A VARMA(5,1) with $q = (1, 1, 1, 1, 1, 1)$ is fitted to the first difference of the six time series. The confidence band represent a one standard deviation. The standard deviations are derived from Monte Carlo simulations with 1000 draws.

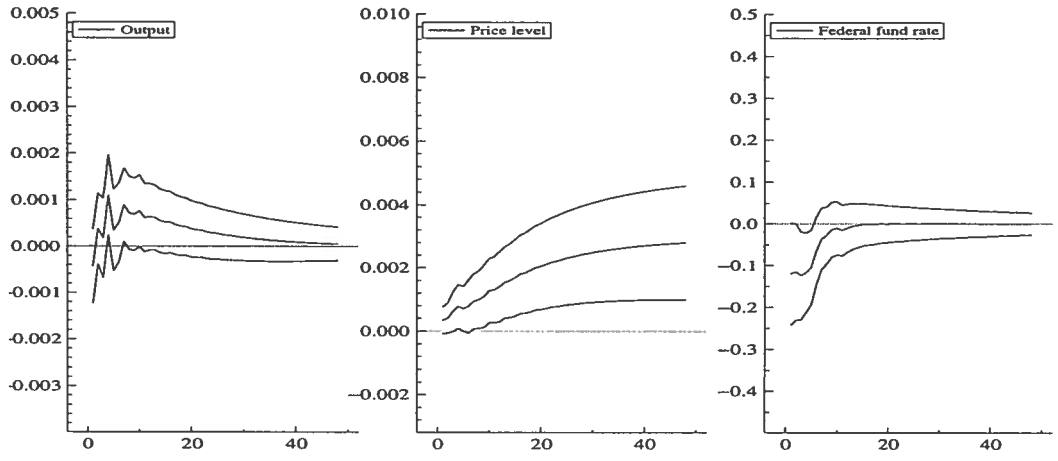


Figure 5: Impulse-response functions for VARMA model in final AR equation form. A VARMA(12,5) is fitted to the first difference of the six time series. The confidence band represent a one standard deviation. The standard deviations are derived from Monte Carlo simulations with 1000 draws.

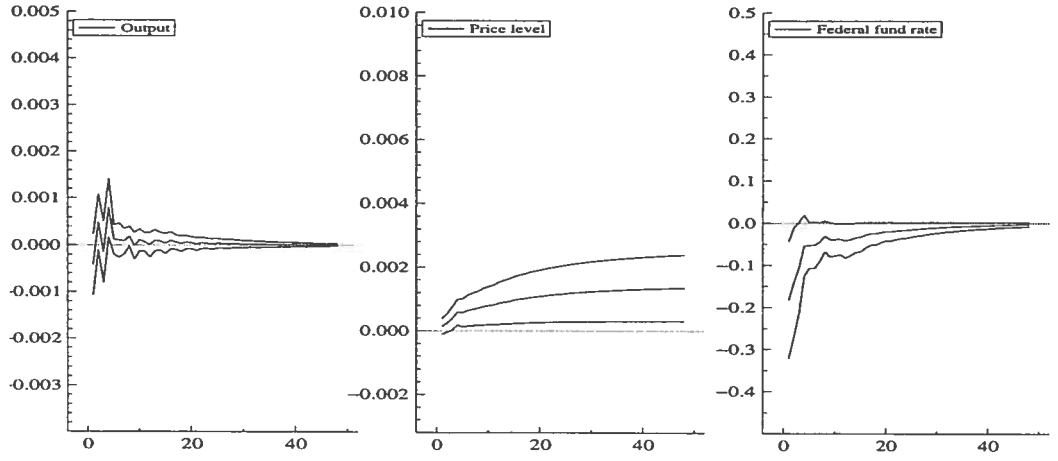
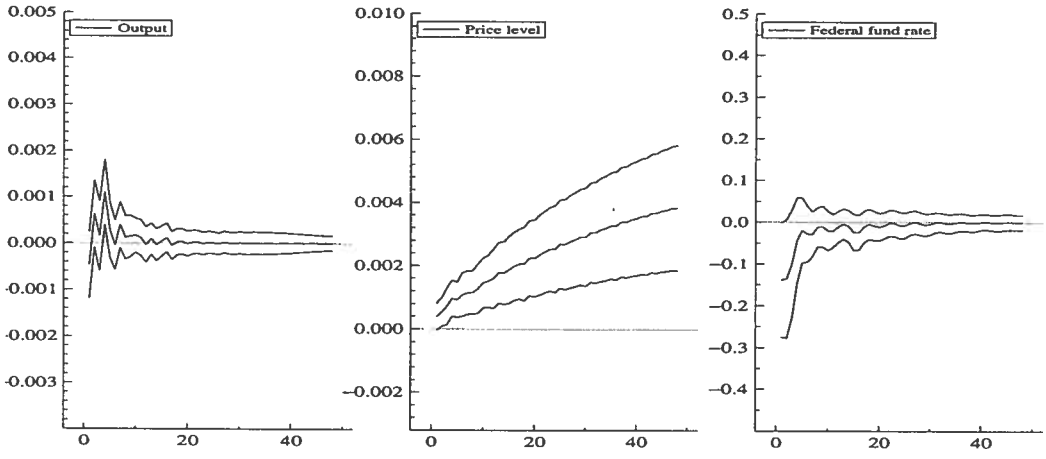


Figure 6: Impulse-response functions for VARMA model in diagonal AR equation form. A VARMA(12,5) with $p = (12, 12, 12, 12, 12, 12)$ is fitted to the first difference of the six time series. The confidence band represent a one standard deviation. The standard deviations are derived from Monte Carlo simulations with 1000 draws.



Chapter 2: Regime switching for dynamic correlations

1. Introduction

It is a well known fact that the variance and covariance of most financial time series are time-varying. Modeling time-varying variance is not just a statistical exercise where someone tries to increase the value of the likelihood; it has important impacts in terms of asset allocation, asset pricing, computation of Value-at-Risk (VaR). A lot of work has been done to model univariate financial time series since the introduction of the ARCH model by Engle (1982). However, we face additional problems when we try to write a multivariate model of volatility. Not only must the variances be positive, the variance matrix must also be positive semi-definite (PSD) at every point in time. Another important problem is the curse of dimensionality. We want models that can be applied to more than a few time series. This rules out the direct generalizations of univariate GARCH models such as the BEKK model of Engle and Kroner (1995).

The most popular multivariate volatility model so far is certainly the Constant Conditional Correlation (CCC) model of Bollerslev (1990). In this model, the covariances of a vector of returns are decomposed into standard deviations and correlations. The major hypothesis in this model is that the conditional correlations are constant through time. With this hypothesis, it is easy to get PSD variance matrices because we only have to ensure that the correlation matrix is PSD and that the standard deviations are non-negative. It also breaks the curse of dimensionality because the likelihood can be seen as a set of SURE equations, i.e. a two-step estimation procedure where univariate volatility models are estimated in a first step that will yield consistent estimates. However, the hypothesis of constant correlations is not always supported by the data [e.g. Engle and Sheppard (2001)].

In this work, we present a new multivariate volatility model, the Regime Switching Dynamic Correlation (RSDC) model. We also decompose the covariances into standard deviations and correlations, but these correlations are dynamic. The correlation matrix

follows a regime switching model; it is constant within a regime but different across regimes. The transitions between the regimes are governed by a Markov chain. The CCC model is a special case of ours where we take the number of regimes to be one.

The RSDC model has many interesting properties. First, it is easy to impose that the variance matrices are PSD. Second, it does not suffer from a curse of dimensionality because it can be estimated with a two-step procedure. Third, when combined with the ARMACH model [see Taylor (1986) and Schwert (1989)] for the standard deviations, this correlation model allows analytic computation of multi-step ahead conditional expectations of the whole variance matrix. Fourth, it can produce smooth patterns for the correlations. We also present an empirical application to exchange rate time series which illustrates that it can have a better in-sample fit of the data than the Dynamic Conditional Correlation (DCC) model recently proposed in Engle (2002).

The model of Engle (2002) and the model proposed in Tse and Tsui (2002) use the same decomposition for the variance matrix as in Bollerslev (1990), but instead of taking constant correlations they propose a GARCH-type dynamic. Because a correlation must lie between -1 and 1, these models must include a rescaling that introduces non-linearities. One side effect of this rescaling is that we can't analytically compute multi-step ahead conditional expectations of the correlation and variance matrices. We can also ask ourselves if a GARCH-type model is appropriate for the correlations because the dynamic of a correlation can be intrinsically different than the behavior of a covariance, e.g. a correlation is bounded from below and above while a covariance is not.

Another approach for breaking the curse of dimensionality of the multivariate GARCH is Ledoit, Santa Clara, and Wolf (2003)'s that proposes a flexible estimation procedure for the Diagonal-Vech model of Bollerslev, Engle, and Wooldridge (1988). The maximization of the likelihood of this model is not computationally feasible if the number of time series is greater than five [see Ding and Engle (2001)]. They propose a way to combine the estimates from univariate and bivariate model so as to get consistent estimates of the parameters of the full multivariate Diagonal-Vech and insure that the variance matrices are PSD. This procedure is only valid for the somewhat restrictive Diagonal-Vech model.

The paper is organized as follows. The second section presents the RSDC model and its properties. Section three describes the estimation of this model and the theoretical properties of the estimates. Section four outlines the computation of one-step and multi-step ahead conditional expectations of the variance matrix. Section five presents an application of the model to multiple exchange rates series. Section six contains a few concluding remarks. Finally, proofs are in the appendix.

2. The RSDC model

In this section we present the Regime Switching Dynamic Correlation (RSDC) model. Assume that the K-variate process Y_t has the form:

$$Y_t = H_t^{1/2} U_t \tag{2.1}$$

where U_t is an i.i.d. $(0, I_K)$ process. The time varying covariance matrix H_t can be decomposed into:

$$H_t \equiv S_t \Gamma_t S_t \tag{2.2}$$

where S_t is a diagonal matrix composed of the standard deviations $s_{k,t}$, $k = 1, \ldots, K$ and the matrix Γ_t contains the correlations. Both S_t and Γ_t are time varying. This decomposition of the covariance matrix has previously been used by Bollerslev (1990), Tse and Tsui (2002), Engle (2002) and Barnard, McCulloch, and Meng (2000). The series Y_t could be a filtered process.

With this decomposition the log-likelihood can be written

$$\begin{aligned} L &= -\frac{1}{2}\sum_{t=1}^{T}\left(K\log(2\pi) + \log(|H_t|) + Y_t' H_t^{-1} Y_t\right) \\ &= -\frac{1}{2}\sum_{t=1}^{T}\left(K\log(2\pi) + \log(|S_t\Gamma_t S_t|) + Y_t' S_t^{-1}\Gamma_t S_t^{-1} Y_t\right) \\ &= -\frac{1}{2}\sum_{t=1}^{T}\left(K\log(2\pi) + 2\log(|S_t|) + \log(|\Gamma_t|) + \tilde{U}_t'\Gamma_T^{-1}\tilde{U}_t\right) \end{aligned} \tag{2.3}$$

where $\tilde{U}_t = [\tilde{u}_{1,t}, \ldots, \tilde{u}_{K,t}]'$ is a zero-mean process with covariance matrix Γ_t and $|H_t| = \det(H_t)$. This is the first building block of our RSDC model: to model the full covariance matrix we model the variances and the correlations separately.

2.1. Regime switching for the correlations

In this work we will argue for a regime switching model for the correlations. This can be seen as a midpoint between the CCC model of Bollerslev (1990) and models such as the DCC of Engle (2002) where the correlations change every period. This model will have the appealing property of constant correlations within a regime but will still have dynamic correlations because of the regime switching. More specifically, the time-varying correlation matrix Γ_t follows:

$$\Gamma_t = \sum_{n=1}^{N} \mathbb{1}_{\{\Delta_t = n\}} \Gamma_n \tag{2.4}$$

with Δ_t an unobserved Markov chain process independent of U_t which can take N possible values ($\Delta_t = 1, 2, \ldots, N$). The symbol $\mathbb{1}$ is the indicator function. The $K \times K$ matrices Γ_n are correlation matrices (symmetric, PSD, ones on the diagonal, off-diagonal elements between -1 and 1) with $\Gamma_n \neq \Gamma_{n'}$ for $n \neq n'$. The probability law governing Δ_t is defined by its transition probability matrix, denoted by Π. The probability of going from regime i in period t to regime j in period $t+1$ is denoted by $\pi_{i,j}$ and the limiting probability of being in regime n is π_n. The element on row j and column i of Π is $\pi_{i,j}$. We make the standard assumptions on the Markov chain [aperiodic, irreducible and ergodic. See Ross (1993, Chapter 4)].

Beside its very intuitive interpretation, this model has many appealing properties. It is easy to impose that Γ_t is a correlation matrix because we only have to impose it for every Γ_n. Imposing that the diagonal elements are equal to one and that the off-diagonal elements are in $[-1, 1]$ does not guarantee that Γ_n is PSD. One way to impose that Γ_n will be a correlation matrix is to take its Choleski decomposition, i.e. $\Gamma_n = P_n P_n'$ where P_n is a lower triangular matrix, and to impose constraints on P_n so that we get ones on the diagonal. These constraints will automatically give off-diagonal elements

between -1 and 1. Consider a trivariate example:

$$\begin{aligned}\Gamma &= \begin{bmatrix} p_{1,1} & 0 & 0 \\ p_{2,1} & p_{2,2} & 0 \\ p_{3,1} & p_{3,2} & p_{3,3} \end{bmatrix} \begin{bmatrix} p_{1,1} & p_{2,1} & p_{3,1} \\ 0 & p_{2,2} & p_{3,2} \\ 0 & 0 & p_{3,3} \end{bmatrix} \\ &= \begin{bmatrix} p_{1,1}^2 & p_{1,1}p_{2,1} & p_{1,1}p_{3,1} \\ p_{1,1}p_{2,1} & p_{2,1}^2 + p_{2,2}^2 & p_{2,1}p_{3,1} + p_{2,2}p_{3,2} \\ p_{1,1}p_{3,1} & p_{2,1}p_{3,1} + p_{2,2}p_{3,2} & p_{3,1}^2 + p_{3,2}^2 + p_{3,3}^2 \end{bmatrix}.\end{aligned}$$

Imposing the additional constraint that the elements on the diagonal P_n are positive, the restrictions becomes

$$p_{j,j} = \sqrt{1 - \sum_{i=1}^{j-1} p_{j,i}^2}, \quad (j = 1, \ldots, K) \tag{2.5}$$

where the sum is zero for $j = 1$. Equation (2.5) is restricting elements $p_{j,i}$, $i = 1, \ldots, j-1$ to be inside a sphere of unit radius and these restrictions are easy to impose.

We could think that estimation of the RSDC model would be complicated by the possibly high number of parameters coming from each Γ_n. Fortunately we will see later on that we can use the EM algorithm [Dempster, Laird, and Rubin (1977)] as presented in Hamilton (1994, chapter 22) so that increasing the number of time series, to which the model is applied will not complicate the estimation.

This specification has three additional interesting properties. The first is that because this model for the correlations is basically linear due to the Markov chain we are able to compute multi-step ahead conditional expectations of the correlation matrix. Also, if we use an appropriate model for the standard deviations, we will also be able to perform these computations for the whole variance matrix. We present such a model in Section 2.3. This is in contrast to the models of Engle (2002) and Tse and Tsui (2002) where the rescaling that is used to keep the correlations between -1 and 1 introduces non-linearities that forbid the computation of multi-step ahead conditional expectations. The second property comes from the Markov chain. If there is some general form of persistence in the chain (high probability of staying in a given regime

for more than one period), then this will lead to smooth time-varying correlations. This could have important impacts namely for the computation of VaR and dynamic portfolio allocation because the benefits of portfolio diversification would be less volatile. The third is that by having a regime switching for the correlations, the variances and covariances are not bounded which is the case when they are the ones following a regime switching [e.g. see Geweke and Amisano (2001)]

2.2. A parsimonious model

We next present a restricted version of the general regime switching model which will have a reduced number of parameters and will remain easy to estimate. For the matrix Γ_t we propose the following form:

$$\Gamma_t = \Gamma\lambda(\Delta_t) + I_K(1 - \lambda(\Delta_t)) \tag{2.6}$$

where Γ is a fixed correlation matrix, I_K is a $K \times K$ identity matrix, $\lambda(\Delta_t) \in [0, 1]$ is a univariate random process governed by an unobserved Markov chain process Δ_t that can take N possible values ($\Delta_t = 1, 2, \ldots, N$) and is independent of U_t. The probability law governing Δ_t is defined by its transition probability matrix, denoted by Π.

The correlation matrix at time t is a weighted average of two extreme states of the world. In one state, the returns are uncorrelated [$\lambda(\Delta_t) = 0$]and in the other the returns are (highly) correlated [$\lambda(\Delta_t) = 1$]. We then have regimes of generally higher or lower correlations and the changes across correlations in a given regime are proportional. The variable $\lambda(\Delta_t)$ can be related to the notion of common features and factor models [Engle and Susmel (1993), Bollerslev and Engle (1993), King, Sentana, and Wadhwani (1994), Diebold and Nerlove (1989), Engle, Ng, and Rothschild (1990), Ng, Engle, and Rothschild (1992)] where the factor affects the variance matrix instead of the correlation matrix.

Note that for the off-diagonal elements only the product of Γ and λ can be identified (by construction the diagonal elements of Γ_t are equal to 1). To solve this identification

problem we can consider two natural sets of constraints. The first is:

$$\lambda(1) = 1,\ \lambda(1) > \lambda(2),\ \ldots,\ \lambda(N-1) > \lambda(N), \tag{2.7}$$

In this case, fixing one of the $\lambda(n)$ to be one identifies the product of Γ and λ. We also restrict the $\lambda(n)$s to be a decreasing sequence to remove the possibility of relabelling regime i as regime j and vice versa. An alternative identification assumption is:

$$\max_{i \neq j} |\Gamma_{i,j}| = 1 \quad \text{with} \quad 1 > \lambda(1), \lambda(1) > \lambda(2), \ldots, \lambda(N-1) > \lambda(N). \tag{2.8}$$

In this case, instead of fixing the highest value of $\lambda(n)$ to be one, we impose this restriction on an off-diagonal element of Γ. The second identification scheme does not impose that one correlation is equal to 1 or -1 because we multiply Γ by $\lambda(\Delta_t)$. Depending on the estimation scheme that we use, one of the two sets of constraints will be more appropriate. We can prove that the matrix H_t is positive semi-definite with probability one for all t.

Proposition 2.1 PSD VARIANCE MATRIX. *If the standard deviations $s_{k,t}$ are non-negative with probability one for all t, $\lambda(n) \in [0,1]$ for $n = 1, \ldots, N$ and Γ is a PSD correlation matrix then the variance matrix H_t will be PSD with probability one for all t.*

It is tempting to allow $\lambda(\Delta_t)$ to take negative values to allow the correlations to change sign, however we don't have a result for a lower bound on $\lambda(\Delta_t)$ that would guarantee that Γ_t is PSD. To understand the problem, consider the correlation matrix of a trivariate time series. If all the correlations are 0.99 then the correlation matrix is PSD; if all the correlations are -0.99, then it will not be PSD.

2.3. Univariate volatility models

To complete the RSDC model we have to specify the dynamic for the standard deviations. The most common one for the volatility of univariate processes is certainly the GARCH model of Bollerslev (1986) where the conditional variance at time t, $s_{k,t}^2$, is a

linear function of past squared innovations and past conditional variances:

$$s_{k,t}^2 = \omega + \sum_{i=1}^{q} \alpha_i y_{k,t-i}^2 + \sum_{j=1}^{p} \beta_j s_{k,t-j}^2. \tag{2.9}$$

where the k subscript on the GARCH parameters is removed.

We should notice that our RSDC model is not written in terms of variances but in terms of standard deviations; a covariance is a correlation times the standard deviations. By using a model such as the GARCH for the variance, the covariance becomes the product of a correlation and the square-root of the product of two variances. The square-root introduces non-linearities that will prohibit analytic computation of conditional expectations.

One model for the volatility of univariate time series that would not have this problem is the GARCH in absolute innovations of Taylor (1986) and Schwert (1989). This class of model is also referred to as ARMACH process in Taylor (1986). In these models the conditional standard deviations follows:

$$s_{k,t} = \omega + \sum_{i=1}^{q} \tilde{\alpha}_i |y_{k,t-i}| + \sum_{j=1}^{p} \beta_j s_{k,t-j} \tag{2.10}$$

with $\tilde{\alpha}_i = \alpha_i / E|\tilde{u}_{k,t}|$. The conditional standard deviations (instead of the conditional variance) are a recursive function of absolute value of past innovations (instead of squared innovations).

There are numerous reasons why a volatility model based on absolute values instead of squared innovations could be a good thing. One reason can be linked to the least absolute deviations versus least squares approach. As argued by Davidian and Carroll (1987), the model could be more robust if we use the absolute value instead of the squared innovation. However, we must reckon that the interpretation of an outlier in a volatility model is not as straightforward as in a regression context. It could also be that the absolute return is a better measure of risk than the squared return. This question is studied by Granger and Ding (1993).

Using the ARMACH model for the volatility of univariate time series is not a prerequisite of our model. We consider this model because it allows the computation of

multi-step ahead conditional expectations of the variance matrix. If conditional expectations are not a point of interest or if the ARMACH gives a clearly inferior fit of the data then another model could be used.

2.4. Review of multivariate GARCH models

To motivate why further work on multivariate volatility model is relevant, we can review some of the existing models. The most straightforward multivariate generalization of the univariate GARCH model can be written in the following way:

$$vech(H_t) = C + \sum_{i=1}^{p} A_i vech(Y_{t-i}Y_{t-i}') + \sum_{j=1}^{q} B_j vec(H_{t-j}). \tag{2.11}$$

where the operator $vech$ stacks in a vector the elements on and below the diagonal of each column of a matrix. This model is not really useful because it is very hard to impose that the matrices H_t are PSD, it is not parsimonious and it is hard to estimate because of the high number of parameters. Engle and Kroner (1995) propose the BEKK representation which guarantees that (2.11) will generate PSD variance matrices, but the problem of simultaneous estimation of a high number of parameters is not solved.

The most popular multivariate variance model is certainly the CCC model of Bollerslev (1990). As its name states, the correlations are constant, i.e. in equation (2.3) we have $\Gamma_t = \Gamma, \ \forall t$. Standard univariate GARCH models are used for the conditional variances. This model has many attractive properties. Interpretation of the parameters is easy because of the correlations and standard deviations decomposition. We only have to take Γ to be PSD to obtain a variance matrix which is PSD. The model is also easy to work with because we can perform the estimation in two steps: firstly, estimate univariate GARCH models and secondly, compute a correlation matrix with the standardized residuals.

To test the hypothesis that the conditional correlations are constant, Bollerslev computed Portmanteau test statistics with the standardized residuals from the univariate GARCH estimations. Under the null hypothesis of constant conditional correlations, the cross-product of the standardized innovations from the univariate GARCH should be i.i.d. Given the low value of these (Ljung-Box) tests he did not reject the null hy-

pothesis. Since then, questions have been raised about the power of these tests [e.g. see Hong (1996)]. To illustrate the lack of power, we repeated the work of Bollerslev by fitting a GARCH(1,1) to the four exchange rate series that we will later use in section 5 and computed the autocorrelation function for each cross-product of the standardized residuals. These are plotted in Figure 7 with the two standard deviations confidence band (under the i.i.d. assumption). Looking at these we are tempted to conclude that there is no dynamic in the cross-product of the standardized innovations, just as Bollerslev (1990) did, and assume that the correlations are constant. The same argument is used by Baillie and Bollerslev (1990). Another paper which favors constant correlations is Schwert and Seguin (1990) who tried several specifications of the multivariate GARCH model (2.11) for monthly stock returns and they could not find one that obviously dominated the constant conditional correlations model. They don't mention which model they tried.

We can run simple Monte Carlo simulations to illustrate that the conclusion of constant correlations could be erroneous. We simulate a multivariate volatility model with a strong dynamic in the correlations, estimate the univariate volatility model and plot the ACF of the cross-product of the standardized residuals. In our example we simulated two models: our restricted model (Figure 8) and the DCC-GARCH (Figure 9) of Engle (2002) which we review below. The parameter values are the estimates obtained with the same exchange rates dataset. Looking at the two figures, we see that the results from the simulated sample and the true data are similar. This is certainly a reason that would explain why there is little evidence in the literature that the conditional correlations are not constant. One evidence is Andersen, Bollerslev, Diebold, and Labys (2001) who gives strong proofs of important dynamics in the correlations by studying realized volatilities computed with high frequency observations of exchange rates data.

More recently, Engle (2002) and Tse and Tsui (2002) introduced multivariate GARCH models with dynamic correlations. Both of them employ the $S_t \Gamma_t S_t$ decomposition of the variance matrix H_t. In Engle (2002) the conditional correlation matrix

Γ_t follows

$$\tilde{\Gamma}_t = (1 - \sum_{i=1}^{q} a_i - \sum_{j=1}^{p} b_j)\Gamma + \sum_{i=1}^{q} a_i(\tilde{U}_{t-i}\tilde{U}'_{t-i}) + \sum_{j=1}^{p} b_j\tilde{\Gamma}_{t-j}, \quad (2.12)$$

$$\Gamma_t = D_t^{-1}\tilde{\Gamma}_t D_t^{-1} \quad (2.13)$$

where D_t is a diagonal matrix with $\sqrt{\tilde{\Gamma}_t(i,i)}$ on row i and column i, and a_i and b_j are scalars. The intuition behind this model is to impose a GARCH-type dynamic for the correlations. Since a correlation matrix must have ones on the diagonal and off-diagonal elements between -1 and 1, we must rescale the correlation matrix [equation (2.13)] because $\tilde{U}_{t-i}\tilde{U}'_{t-i}$ is not constrained to have elements between -1 and 1. The theoretical and empirical properties of this model are developed in Engle and Sheppard (2001).

The model of Tse and Tsui (2002) is similar to the one of Engle (2002) but the rescaling is done differently:

$$\Gamma_t = (1 - \theta_1 - \theta_2)\Gamma + \theta_1\Gamma_{t-1} + \theta_2\Psi_{t-1}, \quad (2.14)$$

$$\Psi_{i,j,t-1} = \frac{\sum_{h=1}^{M} \tilde{u}_{i,t-h}\tilde{u}_{j,t-h}}{\sqrt{\left(\sum_{h=1}^{M} \tilde{u}_{i,t-h}^2\right)\left(\sum_{h=1}^{M} \tilde{u}_{j,t-h}^2\right)}} \quad (2.15)$$

with $M \geq K$. We can see that both rescaling forbid even the analytic computation of multi-step ahead conditional expectation of the correlation matrix. It is unfortunate because one reason why we study volatility is to be able to forecast it.

3. Estimation

The estimation of the RSDC model can in theory be done in one step but if we have more than a few time series the high number of parameters will prohibit us from doing so. Fortunately, we can use a two-step estimation procedure as in Engle (2002). In a first step, we can estimate the univariate volatility models and in a second step, we can estimate the parameters in the correlation matrix conditional on the first step estimates.

In the first subsection we review the theoretical properties of the one-step estimates

and explain how the likelihood can be evaluated. In the following subsection we present estimation methods which can greatly ease the estimation problem due to the high number of parameters.

3.1. One-step estimation

To maximize the likelihood we need to evaluate

$$QL(\boldsymbol{\theta}; \boldsymbol{Y}) = \sum_{t=1}^{T} \log f(Y_t | \underline{Y_{t-1}}), \tag{3.1}$$

where $\underline{Y_{t-1}} = \{Y_{t-1}, Y_{t-2}, \ldots\}$ and $\boldsymbol{\theta}$ is the vector of parameter values. Since the variable Δ_t which drives the correlation matrix is unobserved it is not straightforward. To do this we use Hamilton's filter [Hamilton (1989), Hamilton (1994, chapter 22)] which we adapt to our setup. Inference on the state of the Markov chain is given by the following equations:

$$\hat{\xi}_{t|t} = \frac{(\hat{\xi}_{t|t-1} \odot \eta_t)}{\mathbf{1}'(\hat{\xi}_{t|t-1} \odot \eta_t)}, \tag{3.2}$$

$$\hat{\xi}_{t+1|t} = \Pi \hat{\xi}_{t|t}, \tag{3.3}$$

$$\eta_t = \begin{bmatrix} f(Y_t | \underline{Y_{t-1}}, \Delta_t = 1; \boldsymbol{\theta}) \\ \vdots \\ f(Y_t | \underline{Y_{t-1}}, \Delta_t = N; \boldsymbol{\theta}) \end{bmatrix}, \tag{3.4}$$

where $\hat{\xi}_{t|t}$ is an $(N \times 1)$ vector which contains the probability of being in each regime at time t conditional on the observations up to time t. The $(N \times 1)$ vector $\hat{\xi}_{t+1|t}$ gives these probabilities at time $t+1$ conditional on observations up to time t. The n-th element of the $(N \times 1)$ vector η_t is the density of Y_t conditional on past observations and being in regime n at time t, $\mathbf{1}$ is an $(N \times 1)$ vector of 1s, and $\odot$ denotes elements-by-elements multiplication. Given a starting value $\hat{\xi}_{1|0}$ and parameter values $\boldsymbol{\theta}$, one can iterate over (3.2) and (3.3) for $t = 1, \ldots, T$. The likelihood is obtained as a by-product of this

algorithm:

$$QL(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log\left(1'(\hat{\xi}_{t|t-1} \odot \eta_t)\right). \tag{3.5}$$

Smoothing inference on the state of the Markov chain can also be computed using an algorithm developed by Kim (1994). The probability of being in each regime at time t conditional on observations up to time T is given by the following equation:

$$\hat{\xi}_{t|T} = \hat{\xi}_{t|t} \odot \left\{ \Pi' \left[\hat{\xi}_{t+1|T} \,(\div)\, \hat{\xi}_{t+1|t} \right] \right\} \tag{3.6}$$

where $(\div)$ denotes element-by-element division. One would start iterating over (3.6) with $t = T$, where $\hat{\xi}_{T|T}$ is given by (3.2).

What remains is deciding how to start up the algorithm, i.e. specifying $\hat{\xi}_{1|0}$. One approach would be to add this vector to the parameter space and estimate these initial probabilities. This would add N parameters, $p_1, \ldots, p_N \geq 0$ with $p_1 + \cdots + p_N = 1$. Another approach would be to use the limiting probabilities $(\pi_1, \pi_2, \ldots, \pi_N)$ of the Markov process [Ross (1993, Chapter 4)]. These probabilities are the solution of the following system of equations:

$$\begin{aligned} \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_N \end{bmatrix} &= \Pi \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_N \end{bmatrix}, \\ \sum_{n=1}^{N} \pi_n &= 1. \end{aligned}$$

In the two-regime case the solution is

$$\pi_1 = \frac{1 - \pi_{2,2}}{(1 - \pi_{1,1}) + (1 - \pi_{2,2})}; \quad \pi_2 = \frac{1 - \pi_{1,1}}{(1 - \pi_{1,1}) + (1 - \pi_{2,2})}.$$

In this work both approaches will be used, depending on the estimation method. As we will see below, when using the EM algorithm there is an advantage in treating $\xi_{1|0}$ as unknown parameters. If we are not using the EM algorithm then we will use the limiting probabilities of the Markov chain because in this case these extra parameters

would complicate the estimation.

In the evaluation of the likelihood, notice that the correlation matrix can take N possible values in our model so we only have to invert N times a $K \times K$ matrix. When the number of time series is large this can be a computational advantage over models such as Engle (2002) and Tse and Tsui (2002) where a different correlation matrix has to be inverted for every observation. We are now ready to state the properties of the maximum likelihood estimates.

Theorem 3.1 ONE-STEP MAXIMUM LIKELIHOOD ESTIMATION. *If the assumptions of proposition* **2.1** *and if the usual regularity assumptions for the validity of the QMLE are satisfied then the maximum likelihood estimates are consistent and their asymptotic distribution is given by:*

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \longrightarrow N\left(0; J^{-1} I J^{-1}\right)$$

with

$$J = E[\frac{\partial^2 \log f}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}], \quad I = E[\frac{\partial \log f}{\partial \boldsymbol{\theta}} \frac{\partial \log f}{\partial \boldsymbol{\theta}'}].$$

The matrices I and J can be consistently estimated by their plug-in estimates:

$$\begin{aligned} \hat{I} &= \frac{1}{T} \sum_{t=1}^{T} \left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(Y_t | \underline{Y_{t-1}}; \hat{\boldsymbol{\theta}}) \right) \left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(Y_t | \underline{Y_{t-1}}; \hat{\boldsymbol{\theta}}) \right)', \\ \hat{J} &= \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \left(\frac{1}{T} \sum_{t=1}^{T} \log f(Y_t | \underline{Y_{t-1}}; \hat{\boldsymbol{\theta}}) \right). \end{aligned}$$

PROOF OF THEOREM **3.1** See Newey and McFadden (1994). □

One-step estimation is not really practicable if the number of time series is more than a few because of a curse of dimensionality. In this case, we need an estimation method which only requires non-linear optimization of $O(1)$ parameters at a time. This is what we present in the next subsection.

3.2. Two-step estimation

By splitting the model in two parts, standard deviations and correlations, we can estimate the model in two steps as in Engle (2002). The first step involves the parameters of the univariate volatility models and the second step involves the parameters of the correlation model. We first begin by introducing elements of notation. The complete parameter space $\boldsymbol{\theta}$ is split into $\boldsymbol{\theta_1}$ for the parameters in the univariate volatility model and $\boldsymbol{\theta_2}$ for the parameters in the correlation model. We denote by QL_1 the likelihood where the correlation matrix is taken to be an identity matrix:

$$QL_1(\boldsymbol{\theta_1}; \boldsymbol{Y}) = -\frac{1}{2}\sum_{t=1}^{T}\left(K\log(2\pi) + 2\log(|S_t|) + U_t'U_t\right). \tag{3.7}$$

We denote by QL_2 the likelihood given $\boldsymbol{\theta_1}$ where we have concentrate out S_t:

$$QL_2(\boldsymbol{\theta_2}; \boldsymbol{Y}, \boldsymbol{\theta_1}) = -\frac{1}{2}\sum_{t=1}^{T}\left(K\log(2\pi) + \log(|\Gamma_t|) + U_t'\Gamma_T^{-1}U_t\right). \tag{3.8}$$

Notice two important features of QL_1. Firstly, it is the sum of K univariate log-likelihood so maximizing it is equivalent to maximizing each univariate log-likelihood separately. Secondly, the evaluation of these log-likelihood is straightforward since it does not involve the use of Hamilton's filter. To maximize QL_2 we again have to use Hamilton's filter since Δ_t is unobserved. The procedure is the same as the one-step case because the correlations are not a function of the standard deviations.

Because the number of parameters in the correlation model grows at a quadratic rate with the number of time series, direct maximization of the likelihood is not practicable if we analyze more than a few series. To bypass this problem, we present two estimation methods, one for the non-restricted model and one for the restricted model, which do not rely on the simultaneous non-linear maximization of all the parameters.

For the non-restricted model, it turns out that maximization of the likelihood QL_2 for the correlation model can be done with the EM algorithm. Using the results of Hamilton (1994, chapter 22) we know that the MLE estimates of the transition probabilities and the correlation matrices satisfy the following equations if the initial proba-

bilities $\hat{\xi}_{1|0}$ are not a function of Π and Γ_i:

$$\hat{\pi}_{i,j} = \frac{\sum_{t=2}^{T} P[\Delta_t = j, \Delta_{t-1} = i|\underline{\hat{U}_T}; \hat{\boldsymbol{\theta}}_2]}{\sum_{t=2}^{T} P\left[\Delta_{t-1} = i|\underline{\hat{U}_T}; \hat{\boldsymbol{\theta}}_2\right]}, \tag{3.9}$$

$$\hat{\Gamma}_n = \frac{\sum_{t=1}^{T} (\hat{U}_t \hat{U}_t') P[\Delta_t = n|\underline{\hat{U}_T}; \hat{\boldsymbol{\theta}}_2]}{\sum_{t=1}^{T} P[\Delta_t = n|\underline{\hat{U}_T}; \hat{\boldsymbol{\theta}}_2]}. \tag{3.10}$$

Starting with an initial value $\hat{\boldsymbol{\theta}}_2^{(0)}$ for the vector $\boldsymbol{\theta}_2$, we can compute a new vector $\hat{\boldsymbol{\theta}}_2^{(1)}$ using equations (3.9) and (3.10). We then continue the iteration until the difference between successive vectors $\hat{\boldsymbol{\theta}}_2^{(m)}$ and $\hat{\boldsymbol{\theta}}_2^{(m+1)}$ is small. This estimation method is more efficient than blindly maximizing the likelihood with Newton-type algorithms because more information on the structure of the problem is used. Notice also that the dimension of Γ_n (i.e. the number of time series) does not affect the complexity of the estimation because we only have to take weighted sums of outer-products. We should also mention that equation (3.10) cannot be used directly because typically it does not provide correlation matrices, i.e. the elements on the diagonal of $\hat{\Gamma}_n$ are not imposed to be one. One should rescale these matrices as in equation (2.13) so they are correlation matrices. By doing this transformation, the estimates obtained with these equations will not exactly be the numerical maximum of the likelihood, but very close to it. From our experience, a limited number of Newton-type iterations are necessary to obtain the exact numerical maximum. For the vector of initial probabilities $\hat{\xi}_{1|0}$, it is also shown that their MLE estimates are given by the smoothed probabilities of the first observation.

For the restricted model we can estimate the matrix Γ, up to a scale factor, by doing correlation targeting. This leaves $O(1)$ parameters to be non-linearly estimated. To do the correlation targeting notice that

$$E[\Gamma_t] = \Gamma \sum_{n=1}^{N} \lambda(n)\pi_n + I_K \sum_{n=1}^{N} (1 - \lambda(n))\pi_n.$$

Therefore a correlation matrix computed with the standardized residuals from the first step estimation will provide an estimate $\hat{\Gamma}$ of Γ up to the scale factor $\sum_{n=1}^{N} \lambda(n)\pi_n$ for the off-diagonal elements. The scale indetermination can be solved by using the con-

straints on Γ and $\lambda(n)$ described in equation (2.8). We would divide the off-diagonal elements of $\hat{\Gamma}$ by the highest in absolute value, so as to get a 1 or -1 off the diagonal, and we would take $\lambda(1) > 1$. This leaves a number of parameters to be non-linearly estimated which increase with the number of regimes, not with the number of time series. The properties of the two-step estimation are described in the following theorem.

Theorem 3.2 TWO-STEP MAXIMUM LIKELIHOOD ESTIMATION. *If the assumptions of Theorem* **3.1** *are satisfied then the two-step estimates are consistent and their asymptotic distribution is:*

$$\sqrt{T}\left(\begin{bmatrix}\hat{\boldsymbol{\theta}}_1\\ \hat{\boldsymbol{\theta}}_2\end{bmatrix} - \begin{bmatrix}\boldsymbol{\theta}_1\\ \boldsymbol{\theta}_2\end{bmatrix}\right) \longrightarrow N(0;V)$$

with

$$V = \begin{bmatrix} G_{\boldsymbol{\theta}_1}^{-1} & -G_{\boldsymbol{\theta}_1}^{-1}G_{\boldsymbol{\theta}_2}M^{-1} \\ 0 & M^{-1}\end{bmatrix} E\left[\frac{\partial \ln f}{\partial \boldsymbol{\theta}}\frac{\partial \ln f}{\partial \boldsymbol{\theta}'}\right] \begin{bmatrix} G_{\boldsymbol{\theta}_1}^{-1} & -G_{\boldsymbol{\theta}_1}^{-1}G_{\boldsymbol{\theta}_2}M^{-1} \\ 0 & M^{-1}\end{bmatrix}'$$

where

$$G_{\boldsymbol{\theta}_1} = E\left[\frac{\partial g(Y,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_1{}'}\right]; \quad G_{\boldsymbol{\theta}_2} = E\left[\frac{\partial g(Y,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2{}'}\right]; \quad M = E\left[\frac{\partial m(Y,\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2{}'}\right]$$

$$g(Y,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \frac{\partial \ln f(Y_t|\underline{Y_{t-1}})}{\partial \boldsymbol{\theta}_1}; \quad m(Y,\boldsymbol{\theta}_2) = \frac{\partial \ln f(Y_t|\underline{Y_{t-1}})}{\partial \boldsymbol{\theta}_2}$$

The matrix V can be consistently estimated by their plug-in estimate:

$$\hat{V} = \begin{bmatrix} \hat{G}_{\boldsymbol{\theta}_1}^{-1} & -\hat{G}_{\boldsymbol{\theta}_1}^{-1}\hat{G}_{\boldsymbol{\theta}_2}\hat{M}^{-1} \\ 0 & \hat{M}^{-1}\end{bmatrix} \hat{I} \begin{bmatrix} \hat{G}_{\boldsymbol{\theta}_1}^{-1} & -\hat{G}_{\boldsymbol{\theta}_1}^{-1}\hat{G}_{\boldsymbol{\theta}_2}\hat{M}^{-1} \\ 0 & \hat{M}^{-1}\end{bmatrix}'$$

where

$$\hat{G}_{\boldsymbol{\theta}_1} = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2}{\partial \boldsymbol{\theta}_1\, \partial \boldsymbol{\theta}_1{}'} \ln f(Y_t|\underline{Y_{t-1}}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2),$$

$$\hat{G}_{\boldsymbol{\theta}_2} = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2}{\partial \boldsymbol{\theta}_1\, \partial \boldsymbol{\theta}_2{}'} \ln f(Y_t|\underline{Y_{t-1}}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2),$$

$$\hat{M} = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2}{\partial \boldsymbol{\theta_2}\, \partial \boldsymbol{\theta_2}'} \ln f(Y_t|\underline{Y_{t-1}}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2),$$

$$\hat{I} = \frac{1}{T}\sum_{t=1}^{T} \left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(Y_t|\underline{Y_{t-1}}, \ldots, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\right) \left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(Y_t|\underline{Y_{t-1}}, \ldots, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\right)'.$$

The proof is in the appendix.

Using the general results summarized in Pagan (1986) on two-step estimation we can compute efficient estimates from the two-step estimates by doing one step of a Newton-Raphson estimation of the full likelihood using our two-step estimates as the starting point. The properties of the estimates resulting from this procedure are described in the following theorem.

Theorem 3.3 TWO-STEP EFFICIENT MAXIMUM LIKELIHOOD ESTIMATION. *If the assumptions of theorem* **3.1** *are satisfied then efficient estimates can be obtained by doing one step of a Newton-Raphson estimation of the full likelihood using the two-step estimates* $\hat{\boldsymbol{\theta}}$*:*

$$\bar{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - \left[\frac{\partial^2 QL}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1}_{\hat{\boldsymbol{\theta}}} \frac{\partial QL}{\partial \boldsymbol{\theta}}\Big|_{\hat{\boldsymbol{\theta}}},$$

$$\sqrt{T}\left(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \longrightarrow N\left(0; J^{-1} I J^{-1}\right).$$

The matrices I *and* J *can be constantly estimated by their plug-in estimate given in Theorem* **3.1**.

Proof: See Pagan (1986). Notice that the computation of these estimates could be costly in computing time when dealing with very large systems because of the need to compute the matrix of second derivatives.

The remaining problem in this work is to specify the number of regimes in the Markov chain. It is well known that testing for the number of regimes in a Markov chain is a hard problem to tackle. We leave this problem for further work. The asymptotic theory of an LR test of $N + 1$ versus N regimes is complicated by the fact that some parameters are not identified under the null hypothesis and we are testing parameter values that are on the boundary of the maintained hypothesis [see Andrews (1999, 2001)]. The asymptotic properties of this test are unknown for the moment. A solution

could be the use of Monte Carlo test procedures [see Dufour (2002)]. An alternative procedure could be the specification tests presented in Hamilton (1996).

4. Multi-step ahead conditional expectations

In this section we study one-step and multi-step ahead conditional expectations of the variance matrix. To compute these we must take the conditional expectations of the product of a correlation and two standard deviations. We begin by introducing a notation for the matrix Γ_t that covers both the restricted and unrestricted model. We will denote by $\Gamma(\Delta_t = n)$ the value taken by Γ_t when the chain is in regime n at time t. All the calculus will be presented for the case where the univariate volatility model is an ARMACH(1,1). Extension to a more general ARMACH(p,q) would not introduce new difficulties.

One-step ahead conditional expectations are straightforward. Using the fact that tomorrow's conditional standard deviations are known, to compute $E_t[H_{t+1}]$ we have to compute

$$\begin{aligned} E_t\left[s_{i,t+1}s_{j,t+1}\Gamma_{i,j}(\Delta_{t+1})\right] &= s_{i,t+1}s_{j,t+1}E_t\left[\Gamma_{i,j}(\Delta_{t+1})\right] \\ &= s_{i,t+1}s_{j,t+1}\Gamma_{i,j,t+1|t} \end{aligned}$$

for $i, j = 1, 2, \ldots, K$, where $\Gamma_{i,j,t+1|t} = E_t\left[\Gamma_{i,j}(\Delta_{t+1})\right]$. To compute this expectation we use the fact that the Markov chain Δ_t is independent of U_t. Given the information up to time t, the probability of being in each regime at time $t+1$ is

$$\xi_{t+1|t} = \Pi\xi_{t|t}.$$

From this we deduce that,

$$\Gamma_{t+1|t} = \sum_{n=1}^{N} \Gamma(\Delta_{t+1} = n)\xi_{n,t+1|t}.$$

We see that for the one-step ahead conditional expectations the choice of the model for the standard deviations does not play a role when a GARCH-type model is used

because tomorrow's standard deviations are known.

To compute the d-step ahead conditional expectations $E_t[H_{t+d}]$ we have to compute elements of the following form, for $i, j = 1, 2, \ldots, K$,

$$E_t[s_{i,t+d} s_{j,t+d} \Gamma_{i,j}(\Delta_{t+d})].$$

In the following, we assume that $i \neq j$. If $i = j$, the correlation is always equal to 1 and the Markov chain does not play a role. At this point we see why we cannot analytically compute multi-step ahead conditional expectations with a GARCH model for the standard deviations. We would have to take conditional expectations of the square root of a linear expression.

The ARMACH model described in equation (2.10) can be rewritten in an ARMA-type representation and for an ARMACH(1,1) we get:

$$s_{k,t} = \omega_k + (\alpha_k + \beta_k) s_{k,t-1} + \alpha_k s_{k,t-1} \tilde{v}_{k,t-1} \tag{4.1}$$

where

$$\tilde{v}_{k,t-1} = \left(\frac{|\tilde{u}_{k,t-1}|}{E|\tilde{u}_{k,t-1}|} - 1 \right) \tag{4.2}$$

is a martingale difference sequence. Using the fact that the Markov chain is independent of the process U_t, we can first compute the expectation conditional on the Markov chain and then integrate it out:

$$E_t^{\Delta} \left[E_t^{U} [s_{i,t+d} s_{j,t+d} \Gamma_{i,j}(\Delta_{t+d}) | \Delta] \right] = E_t^{\Delta} \left[\Gamma_{i,j}(\Delta_{t+d}) E_t^{U} [s_{i,t+d} s_{j,t+d} | \Delta] \right]$$

where $E_t^U[\cdots|\Delta]$ is the expectation with respect to the innovations U_t conditional on the present and future values of Δ_t, and $E_t^{\Delta}[\cdots]$ is the expectation with respect to the process Δ_t. We can now treat the correlations as known for the computation of $E_t^U[\cdots|\Delta]$. Before proceeding, we define the following elements:

$$f_{i,j,t+d}(n_d) \equiv E_t^U[\tilde{v}_{i,t+d} \tilde{v}_{j,t+d} | \Delta_{t+d} = n_d]$$

$$= \frac{1}{(E|\tilde{u}_{i,t}|E|\tilde{u}_{j,t}|)} E\left[|\tilde{u}_{i,t+d}||\tilde{u}_{j,t+d}|\middle|\Delta_{t+d} = n_d\right] \tag{4.3}$$

and

$$\begin{aligned} a_{k,t+d} &\equiv E_t^U\left[s_{k,t+d}|\Delta\right] \\ &= E_t^U\left[\omega_k \sum_{l=0}^{d-2}(\alpha_k+\beta_k)^l + (\alpha_k+\beta_k)^{d-1} s_{k,t+1}\right] \\ &\quad + E_t^U\left[\sum_{l=1}^{d-1}\alpha_k(\alpha_k+\beta_k)^{l-1} s_{k,t+d-l}\tilde{v}_{k,t+d-l}\middle|\Delta\right] \\ &= \omega_k \frac{1-(\alpha_k+\beta_k)^{d-1}}{1-(\alpha_k+\beta_k)} + (\alpha_k+\beta_k)^{d-1} s_{k,t+1}. \end{aligned} \tag{4.4}$$

For the expectation in (4.3), if we assume that the U_t's are jointly Gaussian then, it has a closed-form solution which involves a hyper-geometric function with the correlation between $\tilde{u}_{i,t+d}$ and $\tilde{u}_{j,t+d}$, which is known, as an argument[4]:

$$\begin{aligned} f_{i,j,t+d}(n_d) &= \frac{2\left((1-\Gamma_{i,j}(n_d)^2)^2 + 2\Gamma_{i,j}(n_d)^2 HG\left(\frac{1}{2}, 2, \frac{3}{2}, \frac{-\Gamma_{i,j}(n_d)^2}{1-\Gamma_{i,j}(n_d)^2}\right)\right)}{2\sqrt{1-\Gamma_{i,j}(n_d)^2}}, \\ HG(a,b,c,z) &= \sum_{k=0}^{\infty} \frac{(a)_k (b)_k z^k}{(c)_k\, k!} \end{aligned}$$

where $(x)_k = x(x+1)\cdots(x+k)$.

In the case where U_t is not Gaussian and a closed-form solution cannot be found, $f_{i,j,t+d}(n_d)$ could be evaluated by numerical integration. However this would have to be done only N times because n_d can take only N possible values. In any case, for the form of the distribution of U_t, a stronger stand must be taken than only saying that it has mean zero and an identity matrix for the variance.

Using these expressions the d-step ahead conditional expectation becomes:

$$\begin{aligned} & E_t^U\left[s_{i,t+d}s_{j,t+d}|\Delta\right] \\ = \; & E_t^U\big[\left(\omega_i + (\alpha_i+\beta_i)s_{i,t+d-1} + \alpha_i s_{i,t+d-1}\tilde{v}_{i,t+d-1}\right) \times \\ & \quad \left(\omega_j + (\alpha_j+\beta_j)s_{j,t+d-1} + \alpha_j s_{j,t+d-1}\tilde{v}_{j,t+d-1}\right)|\Delta\big] \end{aligned}$$

[4]Computed with Mathematica.

$$
\begin{aligned}
&= \omega_i\omega_j + \omega_i(\alpha_j+\beta_j)a_{j,t+d-1} + \omega_j(\alpha_i+\beta_i)a_{i,t+d-1} \\
&\quad + \left\{(\alpha_i+\beta_i)(\alpha_j+\beta_j) + \alpha_i\alpha_j f_{i,j,t+d-1}(n_{d-1})\right\} E_t^U\left[s_{i,t+d-1}s_{j,t+d-1}|\Delta\right] \\
&= a_{i,j,t+d-1} + b_{i,j,t+d-1}(n_{d-1})E_t^U\left[s_{i,t+d-1}s_{j,t+d-1}|\Delta\right]
\end{aligned}
$$

where

$$
\begin{aligned}
a_{i,j,t+d-1} &= \omega_i\omega_j + \omega_i(\alpha_j+\beta_j)a_{j,t+d-1} + \omega_j(\alpha_i+\beta_i)a_{i,t+d-1}, \\
b_{i,j,t+d-1}(n_{d-1}) &= (\alpha_i+\beta_i)(\alpha_j+\beta_j) + \alpha_i\alpha_j f_{i,j,t+d-1}(n_{d-1}).
\end{aligned}
$$

We can solve this expression recursively to get

$$
\begin{aligned}
E_t^U\left[s_{i,t+d}s_{j,t+d}|\Delta\right] &= \sum_{l=1}^{d-1} a_{i,j,t+d-l}\left(\prod_{m=1}^{l-1} b_{i,j,t+d-m}(n_{d-m})\right) \\
&\quad + \prod_{m=1}^{d-1} b_{i,j,t+d-m}(n_{d-m})s_{i,t+1}s_{j,t+1}
\end{aligned}
$$

where $\prod_{m=1}^{l-1} b_{i,j,t+d-m}(n_{d-m})$ is equal to one when $l=1$.

Keeping in mind that $b_{i,j,t+d-m}(n_{d-m})$ depends on the state of the Markov chain at time $t+d-m$ we next integrate out the Markov chain. Doing so we get

$$
\begin{aligned}
&E_t[s_{i,t+d}s_{j,t+d}\Gamma_{i,j}(\Delta_{t+d})] \\
&= \sum_{n_d=1}^{N}\cdots\sum_{n_0=1}^{N} E_t^U\left[s_{i,t+d}s_{j,t+d}|\Delta\right]\Gamma_{i,j}(n_d)\xi_{n_0,t|t}\pi_{n_0,n_1}\cdots\pi_{n_{d-1},n_d} \\
&= \sum_{n_d=1}^{N}\cdots\sum_{n_0=1}^{N}\sum_{l=1}^{d-1} a_{i,j,t+d-l}\left[\prod_{m=1}^{l-1} b_{i,j,t+d-m}(n_{d-m})\right]\Gamma_{i,j}(n_d) \\
&\qquad \times\xi_{n_0,t|t}\pi_{n_0,n_1}\cdots\pi_{n_{d-1},n_d} + \\
&\quad \sum_{n_d=1}^{N}\cdots\sum_{n_0=1}^{N} s_{i,t+1}s_{j,t+1}\left[\prod_{m=1}^{d-1} b_{i,j,t+d-m}(n_{d-m})\right]\Gamma_{i,j}(n_d) \\
&\qquad \times\xi_{n_0,t|t}\pi_{n_0,n_1}\cdots\pi_{n_{d-1},n_d}. \qquad (4.5)
\end{aligned}
$$

The summations in the last equality can be rearranged so as to obtain

$$E_t[s_{i,t+d}s_{j,t+d}\Gamma_{i,j}(\Delta_{t+d})]$$

$$
\begin{aligned}
= & \sum_{l=1}^{d-1} a_{i,j,t+d-l} \sum_{n_d=1}^{N} \Gamma_{i,j}(n_d) \sum_{n_{d-1}=1}^{N} b_{i,j,t+d-1}(n_{d-1})^{\mathbb{1}_{\{l>1\}}} \pi_{n_{d-1},n_d} \cdots \times \\
& \sum_{n_1=1}^{N} b_{i,j,t+1}(n_1)^{\mathbb{1}_{\{l>d-1\}}} \pi_{n_1,n_2} \sum_{n_0=1}^{N} \xi_{n_0,t|t} \pi_{n_0,n_1} \\
+ & \; s_{i,t+1} s_{j,t+1} \sum_{n_d=1}^{N} \Gamma_{i,j}(n_d) \sum_{n_{d-1}=1}^{N} b_{i,j,t+d-1}(n_{d-1}) \pi_{n_{d-1},n_d} \cdots \times \\
& \sum_{n_1=1}^{N} b_{i,j,t+1}(n_1) \pi_{n_1,n_2} \sum_{n_0=1}^{N} \xi_{n_0,t|t} \pi_{n_0,n_1}. \qquad (4.6)
\end{aligned}
$$

We see that the sums over N^{d+1} terms in equation (4.5) can be written as a sum over $(d+1)N$ terms.

We are able to compute multi-step ahead conditional expectations of the whole variance matrix for two reasons. The first is that since our model for the correlation matrix is linear, the conditional expectations of the correlation matrix are given by the summation of a constant times a probability which is linearly updated. The second is the use of a model for the conditional standard deviation (ARMACH) instead of the variance. Note that the use of the ARMACH model is not required. If another univariate model for the conditional volatility is obviously better and if analytic computation of multi-step ahead conditional expectations are not of interest then this model should be used.

It is not easy to design a multivariate volatility model that has a rich enough dynamic but allows these analytic computations of multi-step ahead conditional expectations of the variance matrix. For example, in the DCC model of Engle (2002) it is not even possible to compute multi-step ahead conditional expectations of the correlation matrix because the rescaling performed in equation (2.13) introduces non-linearities.

5. Application to exchange rate data

In this section we apply both the unrestricted and restricted version of the RSDC model to the exchange rate dataset used by Harvey, Ruiz, and Shephard (1994) and Kim, Shephard, and Chib (1998). This dataset contains four weekdays close exchange rates (Pound, Deutschmark, Yen, Swiss-Franc all against the U.S. dollar) over the period

1/10/81 to 28/6/85. The number of observation is 946. We first take 100 times the first difference of the logarithm of each series, minus the sample mean, before applying directly our variance model (these are our filtered series). We employ this dataset because Harvey, Ruiz, and Shephard (1994) use it to present a multivariate stochastic volatility model where they assume that correlations are constant through time. Using our model we can check if their assumption was reasonable.

The results are generated using Ox version 3.30 on Linux [see Doornik (1999)]. The estimation results that we present in the various tables are for full one-step maximum likelihood estimation. We first do the two-step estimation (EM algorithm or correlation targeting) and then use these values to initialize the full maximization. We can do it because we have a limited number of time series in our example.

5.1. RSDC model with two regimes

We first present results for the models with two regimes. Models with three regimes are studied in the following subsection. The results for the unrestricted models are presented in Tables 16 [ARMACH(1,1) for the standard deviations] and 17 [GARCH(1,1) for the standard deviations]. The outputs for the restricted version of the model are in Tables 18 [ARMACH(1,1)] and 19 [GARCH(1,1)]. For the restricted model we present the correlation matrix in each regime and their standard deviations computed with the Delta method instead of the matrix Γ and the value of $\lambda(2)$ [we use the identification scheme of equation (2.7) when doing the one-step estimation] so that the results are directly comparable to those of the unrestricted model.

The results for the univariate volatility models are similar to the usual findings with this type of financial series. The level of persistence for the univariate GARCH models ($\alpha + \beta$) are high but strictly lower than one. For an ARMACH(1,1) the degree of persistence of the standard deviations are given by $\alpha + \beta$, not $\tilde{\alpha} + \beta$. We also find that the persistences are high but strictly lower than one. The impact on the likelihood of replacing the ARMACH model by the GARCH model is an increase of about 15 points.

For the estimation of the regime switching model the first thing to notice is that the results do not depend on the univariate model for the standard deviations. The likelihood may be higher with the GARCH model but the parameters of the correlation

model are basically the same in both cases. It is an indication that we can replace the traditional GARCH by the ARMACH or that the correlation model is robust to the specification of the univariate standard deviations.

Looking at the tables and the Figures 11 and 12 where we have plotted, for the unrestricted and the restricted model, the smoothed probabilities of being in regime one and the smoothed correlations at each point in time, we see that the correlations appear to be dynamic. Figure 11 shows that we frequently move between both regimes and there is little uncertainty about the regime we are in at each point in time. The process is spending more time in regime one and spells in regime two are shorter on average than in regime one. This is explained by the estimate of the transition probability matrix, which is very similar across the various models with two regimes. The probability of being in regime one at time $t+1$ conditional on being in regime one at time t, $\pi_{1,1}$, is around 0.93. That means a high level of persistence in the Markov chain because the probability of spending the next five days in regime one is $0.93^5 = 0.70$. In comparison, for regime two this probability is $0.67^5 = 0.14$. This illustrates that 0.93 and 0.67, although both high probabilities, are very different.

As for the value of the correlations in each regime, the results for the restricted model are similar to those of the unrestricted model. Under the unrestricted model, the magnitude of all the correlations in regime two is smaller than in regime one. So the hypothesis of the restricted version of the model that there is an ordering in the magnitude of the correlations across the different regimes seems plausible. The hypothesis that they all decrease in the same proportion is less supported by the data. In the unrestricted model, the implied value for $\lambda(2)$ for each correlation is as low as 0.243 and as high as 0.592 (for the ARMACH case). Since these two models are nested, we can use an LR test for this hypothesis. Under the null hypothesis that the restricted model is the reality, twice the difference in the log-likelihood should follow a Chi-square with five degrees of freedom. The value of the test statistic is 27.2 and the 1% critical values is 15.09. We would reject the restricted version of the model at the 1% level.

We mentioned at the end of Section 3 that a LR test of one regime versus two does not asymptotically follow a Chi-square distribution with degrees of freedom equal to the number of extra parameters. Nonetheless, the increase in the likelihood by going

from one regime [which is the CCC model of Bollerslev (1990)] to two regimes is so high, more than 250 points, that we don't need to perform a formal test to reject the model with one regime. Table 26 contains the likelihood and the number of parameters of all the models estimated in this work.

5.2. RSDC model with three regimes

We next allow a third regime in the Markov chain. The estimation results for the various models are presented in Tables 20 to 23. As expected, the estimates of univariate volatility models are not affected by the addition of an additional regime. The increase of the log-likelihood is about 40 points for the unrestricted model and 50 points for the restricted model, while the third regime adds respectively eleven and five parameters. Again, there is no impact on the estimates of the correlation model when going from the GARCH to the ARMACH model. If we have in mind a likelihood ratio test to gauge the increase in the likelihood we would compare 80 or 100 (twice the increase) to the critical values of a Chi-square with eleven or five degrees of freedom (24.73 and 15.09 respectively), although it is not a valid procedure because the LR test is probably not asymptotically Chi-square with these degrees of freedom.

The addition of a third regime now allows the data to identify two regimes with high correlations and one regime of very low correlations. Again, we have in general the same ordering of the magnitude of the correlations across the regimes with the unrestricted model. The magnitude of the correlations in regime one is smaller than in regime two, which is smaller than in regime three. We can again test the restricted model versus the unrestricted. In this case, we compare twice the difference of the likelihood, i.e. 8 for the ARMACH, to a Chi-square with ten degrees of freedom and doing so we don't reject the restricted model.

Looking at Figure 13, we see that the Markov chain is spending most of its time in regimes of high correlations (regime two and three for the unrestricted model, regime one and two for the restricted model). Very rarely does the chain goes in the regime of low correlation. Again, we see that most of the time we have a strong idea about which regime we are in at every point in time as the smoothed probabilities are close to either zero or one most of the time. Examining more closely the correlation matrix for

each regime, the smoothed probabilities and the smoothed correlations in Figure 14, we see that with a third regime, the Markov chain is beginning to identify what could be outliers. The chain is going very rarely in a regime which is very different from the others. This could be seen as an indicator that three regimes is enough.

5.3. DCC

To evaluate the relative performance of our model to fit the data we estimate the DCC-GARCH(1,1) of Engle (2002). To isolate the impact of not using the same model for the standard deviations we also estimate a DCC-ARMACH(1,1). The results for the DCC-GARCH are in Table 24 and the results for the DCC-ARMACH(1,1) are in Table 25. With both of these univariate volatility models we get similar estimates for the matrix Γ and for the parameters α and β; again an indication that the correlation models are robust to the univariate volatility model employed. The full maximum likelihood estimates are reported.

What is interesting is to compare the log-likelihood of the different models. The GARCH(1,1) appears to fit the data a bit better than the ARMACH(1,1) because the likelihood increases by 28 points when we use the first of the two models. We also get a similar increase in our regime switching model. But there is a big difference in the level of the log-likelihood when we compare the RSDC model and the DCC model. For our restricted model with two regimes (and GARCH model) the log-likelihood is 100 points higher than the DCC-GARCH while the regime switching model has only one more parameter than the DCC-GARCH. The difference in the log-likelihood is 114.5 points between the unrestricted RSDC model with two regimes and the DCC-GARCH at the cost of seven additional parameters.

Because our regime switching model and the DCC model are not nested we cannot perform a likelihood ratio test to verify if the increase in the likelihood is significant. One valid test for testing non-nested models is proposed by Rivers and Vuong (2002, Section 4). With this test, we reject at the 10% level[5] the hypothesis that the DCC model

[5]No parameter is treated as a nuisance parameter. We use the suggested Newey and West (1987b) estimator for the variance. We tried a wide range of values for the truncation lag in the computation of the variance.

is as close to the true model as the RSDC model. Another approach for choosing one model over the other could be the use of information criteria. Ultimately, we are not interested in rejecting a model. A better solution would be to combine the forecasts from these different models.

Another interesting comparison is the correlations extracted from both models. If we compare the smoothed correlations from the unrestricted RSDC model with AR-MACH models for the standard deviations (Figure 14) with the correlations from the DCC-ARMACH (Figure 15), we see that the correlations are generally smoother with the switching regime model. This is even more apparent when we take the smoothed correlations from the restricted model for the comparison. The exception would be the correlation between the Deutschmark and the Swiss-Franc where there is almost no movement for the DCC-GARCH while the single factor imposes changes in this correlation. One interesting implication of smoother patterns for the correlations is for the computation of VaR and portfolio allocation. If the time-varying correlations are smoother, then the gain from portfolio diversification will also be smoother which might imply a smoother pattern for the VaR and portfolio weights.

It might be intriguing that the regime switching gives a higher value for the likelihood than the DCC because both models imply a VARMA dynamic for the outer-product of the standardized innovations. The DCC equation (2.12) can be rewritten as

$$vech(\tilde{U}_t\tilde{U}_t') = \bar{\Gamma}_1 + \sum_{i=1}^{\max(p,q)} (a_i + b_i)vech(\tilde{U}_{t-i}\tilde{U}_{t-i}') + V_t - \sum_{j=1}^{p} b_j V_{t-j}$$

where $V_t = vech(\tilde{U}_t\tilde{U}_t') - vech(\tilde{\Gamma}_t)$. From this equation we see that both the AR and MA operators are scalar.

The VARMA representation of the regime switching model for the correlations presented in this work is derived in Dufour and Pelletier (2003):

$$\prod_{n=1}^{N-1}(1 - e_n\, L)vech(\tilde{U}_t\tilde{U}_t') = \bar{\Gamma}_2 + V_t + \sum_{n=1}^{N-1} B_n V_{t-n}$$

with V_t a white noise process and the e_ns are the eigenvalues of the transition matrix

different than 1. The matrices of parameters B_n are function of the correlation matrices and the transition matrix. From this, we see that one reason why the regime switching model can be doing better is because the MA operator is not restricted to be scalar.

5.4. Series associated to the Markov chain

An interesting exercise with regime switching models is identifying what is driving the latent process Δ_t. Our model is for the standardized innovations but we can nonetheless check if periods of high correlations correspond to a particular pattern for the standard deviations. We plot in Figure 16 the smoothed probabilities of being in the regime of high correlations for the restricted model with two regimes and the standard deviations from an ARMACH(1,1) for each series. At first glance we cannot discern a pattern. We can also regress the smoothed probability on a constant and the standard deviations. Doing so, we get a low R^2 coefficient (0.11) and, contrary to the prevailing intuition, the regression coefficients are not all positive. The coefficient and t-stat for the pound (-0.0824 and -1.3809), Yen (-0.3475 and -4.2725) and Swiss-Franc (-0.47776 and -6.9474) are negative. Only the Deutschmark is positive (0.8013 and 7.8502). We get similar results with the model with three regimes or with the filtered probabilities.

We can also look at series other than the standard deviations of each return. One process which could drive the correlations of the various currencies is the return on the stock market. Since all the currencies are expressed in term of U.S. dollars we can look at the return on the Dow Jones index. The conditional variance from a GARCH(1,1) fitted on this series over the same period as our exchange rates is plotted in Figure 17. Compared to the smoothed probabilities in Figure 13 for the unrestricted model, we see that the increase in the volatility after observation number 200 of the index corresponds to a period where the process is in regime 1 (highest correlations) for a prolonged period. This is far from a complete explanation because we cannot really discern a link between this conditional variance and the rest of the smoothed probabilities.

If we believe that adding a third regime is equivalent to chasing outliers we can try to see if something special happened in the days when the process went into that third and infrequent regime. Looking again at the smoothed probabilities for the unrestricted model in Figure 13, we see that around observation number 450 the process is spending

five days in regime 3. These observations correspond to the July 11, 1983 to July 15, 1983 period. Reading newspapers from this period we see that over this week there was a lot of uncertainty about what the Fed would do with the interest rates. At the beginning of the week, Volcker sent a strong but noisy signal that something might or might not happen to the interest rates (the process enters the regime of very low correlations). Throughout the week, the Fed keeps sending this strong and noisy signal (the process stays in this regime). Then at the end of the week, on July 15, Volcker announces that the interest rates will go up. The uncertainty is resolved. The process leaves the regime of very low correlations. Again, this is not a complete explanation because similar event studies for the other periods where the process goes into the regime of low correlations are not as satisfactory.

6. Conclusion

In this work we propose a new model for the variance between multiple time series, the Regime Switching Dynamic Correlation (RSDC) model. We decompose the covariances into correlations and standard deviations and both the correlations and the standard deviations are dynamic. For the correlation matrix, we propose a regime switching model. It is constant within a regime but different across regimes. The transitions between the regimes are governed by a first order Markov chain. This property of constant correlation could have important impacts, namely for the computation of Value-at-Risk and for dynamic portfolio allocation. We also present a restricted version of our model where the changes across correlations in a given regime are proportional. This regime switching model can be seen as a mid-point between the CCC model of Bollerslev (1990) where the correlations are constant and models such as the DCC model of Engle (2002) where the correlation matrix change at every point in time.

One appealing feature of this model for the correlations is that when combined with the ARMACH model [Taylor (1986) and Schwert (1989)] for the conditional standard deviations, it allows analytic computation of multi-step ahead conditional expectations of the whole variance matrix. The ARMACH model is a GARCH-type model for the conditional standard deviations instead of the conditional variance.

The evaluation of the likelihood is done with Hamilton's filter because of the unobserved Markov chain. By decomposing the variance matrix into a diagonal matrix of standard deviations and a correlation matrix, we can use a two-step estimation procedure as in Engle (2002). Combining this two-step estimation procedure with either correlation targeting (for the restricted model) or the EM algorithm (for the unrestricted model) breaks the curse of dimensionality, i.e. the number of parameters in every nonlinear estimation is not a function of the number of time series.

An application of this model to four major exchange rate series illustrates its good behavior. A comparison of our regime switching model with the DCC model of Engle (2002) shows that our model has a better in-sample fit. An interesting aspect of our regime switching model is that we find strong persistence in the Markov chain, which produces smoother time-varying correlations than the DCC model.

Possible extensions in future work includes the addition of relations between correlations and standard deviations as the work of Andersen, Bollerslev, Diebold, and Labys (2001) seems to indicate. Identification of the number of regimes in the Markov chain is also an ongoing research project.

7. Appendix: Proofs

PROOF OF PROPOSITION **2.1** We can first state that Γ_t is positive semi-definite for all t. To prove this, consider a vector $c = [c_1, \ldots, c_K]' \in \mathbb{R}^K$:

$$c'\Gamma_t c = c'\Gamma c\, \lambda(\Delta_t) + c'c\,(1 - \lambda(\Delta_t)) \geq 0$$

because Γ is PSD and $\lambda(\Delta_t) \in [0, 1]$. If the standard deviations are non-negative then the product $S_t\Gamma_t S_t$, i.e. the variance matrix H_t, will also be PSD. □

PROOF OF THEOREM **3.2**

Scaling (3.7) by $1/T$, the uniform strong law of large numbers implies that a.s. we get

$$\mathcal{L}_1 = -\frac{1}{2}E_{\theta_0}\left[\sum_{k=1}^{K}\left(\log 2\pi + 2\log s_{k,t} + \frac{y_{k,t}^2}{s_{k,t}^2}\right)\right] \tag{7.1}$$

where E_{θ_0} is the expectation with respect to the true density. Similarly, scaling (2.3) by $1/T$, a.s. we get

$$\mathcal{L} = -\frac{1}{2}E_{\theta_0}\left[K\log 2\pi + \log|\Gamma_t| + 2\sum_{k=1}^{K}\log s_{k,t} + \tilde{U}_t'\Gamma_t^{-1}\tilde{U}_t\right] \tag{7.2}$$

If we can show that both sets of first order conditions with respect to $\boldsymbol{\theta_1}$ are satisfied for the same vector of parameters then we can conclude that the estimates from (3.7) will converge to their true value.

Denoting by $\theta_{k,j}$ one of the parameters in $\boldsymbol{\theta_1}$ that appears in the expression of $s_{k,t}$, we can write the first order conditions for $\mathcal{L}_1$ as

$$\frac{\partial \mathcal{L}_1}{\partial \theta_{k,j}} = E_{\theta_0}\left[-\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \theta_{k,j}} + \tilde{u}_{k,t}^2\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \theta_{k,j}}\right] = 0. \tag{7.3}$$

While the first order conditions for $\mathcal{L}$ are

$$\frac{\partial \mathcal{L}}{\partial \theta_{k,j}} = E_{\theta_0}\left[-\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \theta_{k,j}} + \tilde{U}_t' \Gamma_t^{-1} \begin{bmatrix} 0 \\ \vdots \\ \tilde{u}_{k,t} \\ \vdots \\ 0 \end{bmatrix} \frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \theta_{k,j}} \right] = 0 \qquad (7.4)$$

Using the trace operator we can easily see that $\tilde{U}_t' \Gamma_t^{-1}[0, \ldots, \tilde{u}_{k,t}, \ldots, 0]'$ is a random variable with unit mean, just like $\tilde{u}_{k,t}^2$. From this we see that the value of $\theta_{k,j}$ that will solve equation (7.4) will also solve equation (7.3). For the rest of the proof see Newey and McFadden (1994).

□

Figure 7: ACF of the cross-product of the standardized residuals from a AR-MACH(1,1).

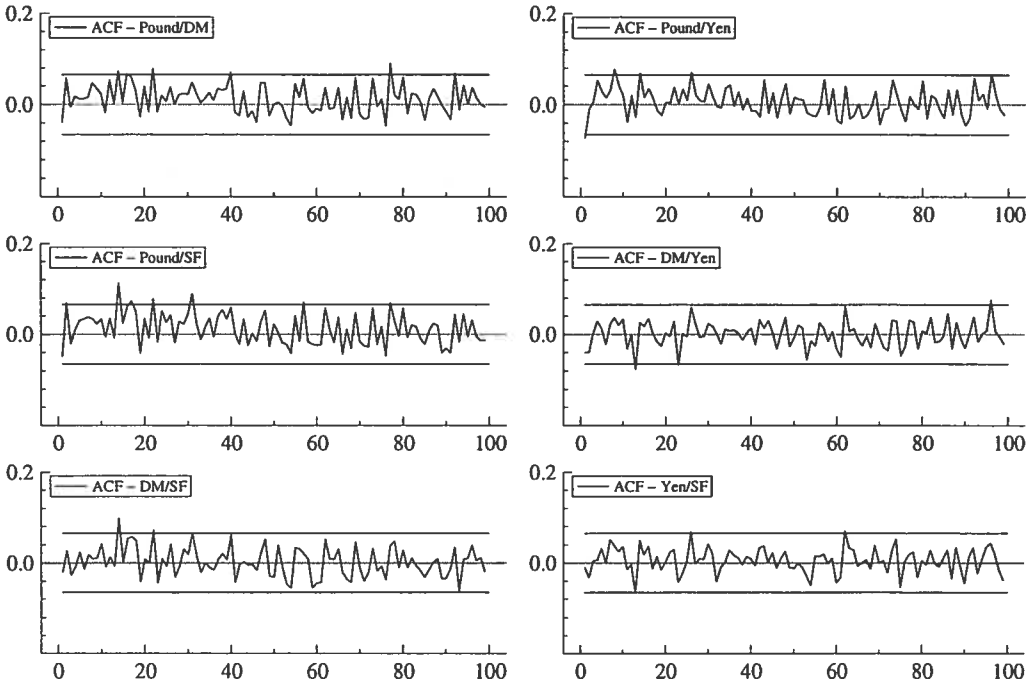


Figure 8: ACF of the cross-product of the standardized residuals with data simulated from a regime switching model. Sample size is 1000.

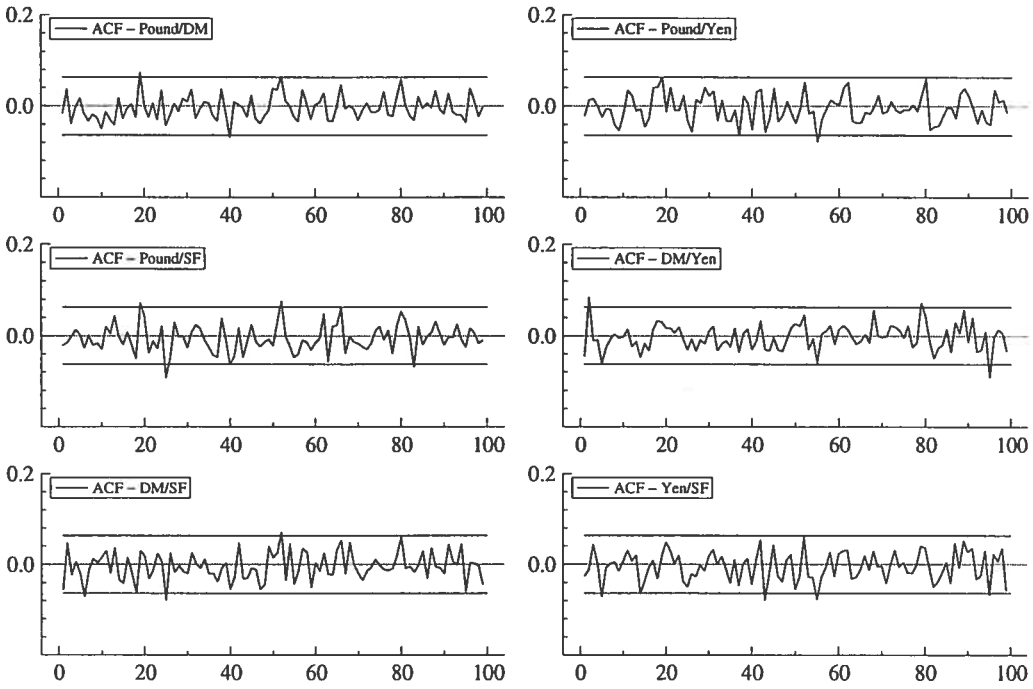


Figure 9: ACF of the cross-product of the standardized residuals with data simulated from a DCC-GARCH. Sample size is 1000.

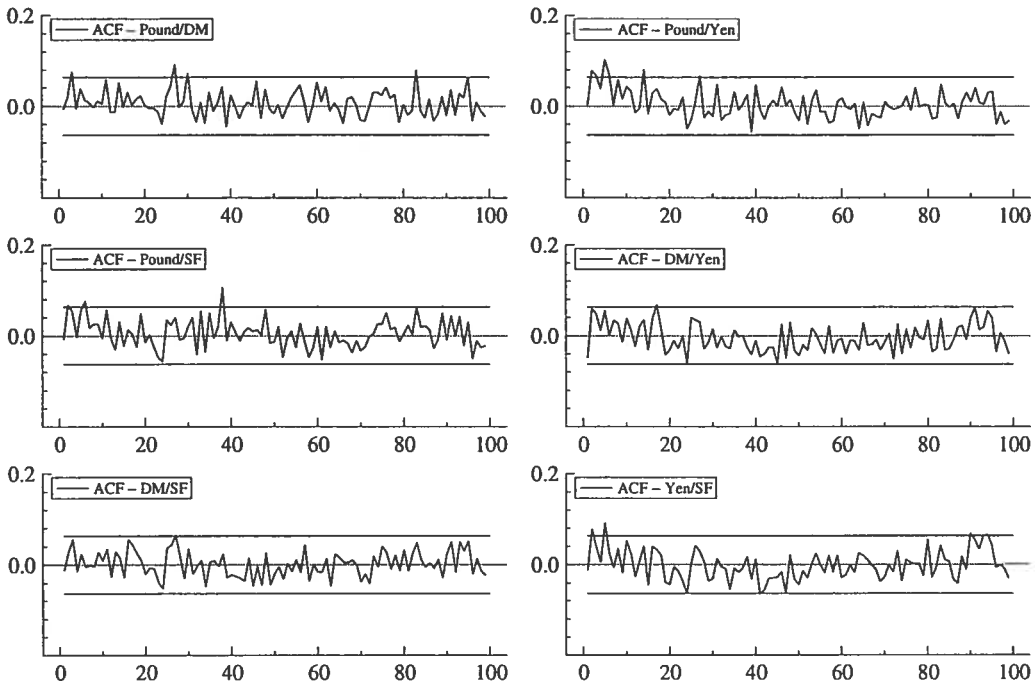


Figure 10: Exchange rate series. The top and bottom figures are respectively the level and the growth rate of each series.

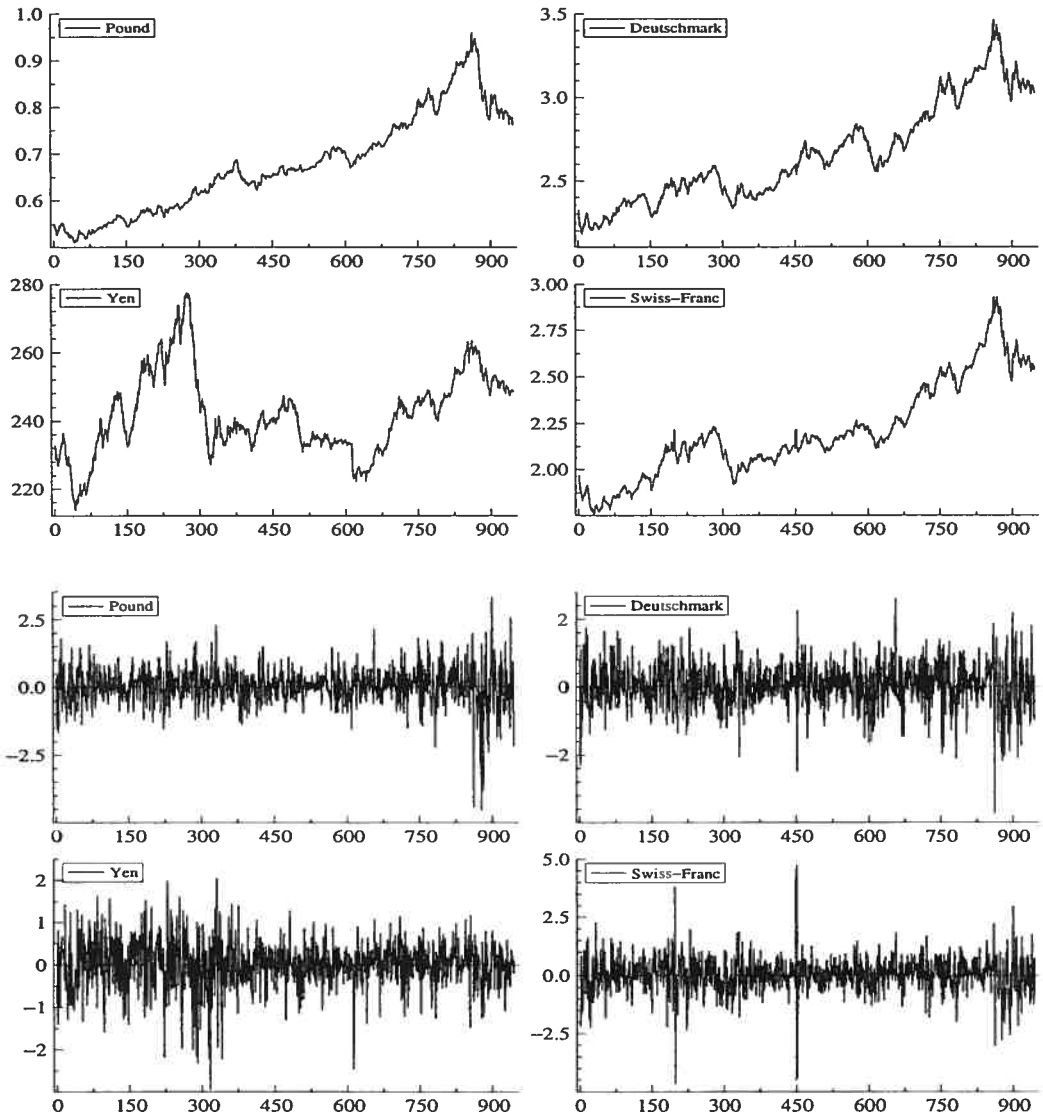


Table 16: Estimation results for the unrestricted model with two regimes and AR-MACH. Standard errors are in parenthesis. The log-likelihood value is -2011.6.

	$\Gamma_{1,2}$	$\Gamma_{1,3}$	$\Gamma_{1,4}$	$\Gamma_{2,3}$	$\Gamma_{2,4}$	$\Gamma_{3,4}$
Regime 1	0.8754	0.7656	0.8569	0.8471	0.9510	0.8617
	(0.0292)	(0.0363)	(0.0283)	(0.0181)	(0.0061)	(0.0184)
Regime 2	0.4011	0.1859	0.3255	0.4739	0.5626	0.3250
	(0.0958)	(0.0996)	(0.1275)	(0.0843)	(0.1871)	(0.1666)

Π	Regime 1	Regime 2
Regime 1	0.9291	0.3334
	(0.0356)	
Regime 2	0.0709	0.6666
		(0.0605)

Series	ω	$\tilde{\alpha}$	α	β
Pound	0.0245	0.1028	0.0795	0.8895
	(0.0094)	(0.0249)	(0.0194)	(0.0263)
Deutschmark	0.0710	0.1286	0.1014	0.8078
	(0.0197)	(0.0295)	(0.0179)	(0.0330)
Yen	0.0049	0.0295	0.0225	0.9705
	(0.0044)	(0.0113)	(0.0086)	(0.0136)
Swiss-Franc	0.0874	0.1225	0.0928	0.7975
	(0.0452)	(0.0430)	(0.0327)	(0.0821)

Table 17: Estimation results for the unrestricted model with two regimes and GARCH. Standard errors are in parenthesis. The log-likelihood value is -1994.7.

	$\Gamma_{1,2}$	$\Gamma_{1,3}$	$\Gamma_{1,4}$	$\Gamma_{2,3}$	$\Gamma_{2,4}$	$\Gamma_{3,4}$
Regime 1	0.8842	0.7805	0.8648	0.8567	0.9536	0.8696
	(0.0264)	(0.0321)	(0.0275)	(0.0191)	(0.0091)	(0.0169)
Regime 2	0.4636	0.2484	0.3930	0.5217	0.6222	0.3953
	(0.1015)	(0.1149)	(0.1329)	(0.0833)	(0.1748)	(0.1792)

Π	Regime 1	Regime 2
Regime 1	0.9131	0.3206
	(0.0392)	
Regime 2	0.0869	0.6794
		(0.0757)

Series	ω	α	β
Pound	0.0193	0.0895	0.8789
	(0.0102)	(0.0344)	(0.0449)
Deutschmark	0.0450	0.1136	0.8160
	(0.0159)	(0.0296)	(0.0418)
Yen	0.0011	0.0181	0.9802
	(0.0018)	(0.0089)	(0.0117)
Swiss-Franc	0.0798	0.1143	0.7646
	(0.0486)	(0.0592)	(0.1209)

Table 18: Estimation results for the restricted model with two regimes and ARMACH. Standard errors are in parenthesis. The log-likelihood value is -2025.2.

	$\Gamma_{1,2}$	$\Gamma_{1,3}$	$\Gamma_{1,4}$	$\Gamma_{2,3}$	$\Gamma_{2,4}$	$\Gamma_{3,4}$
Regime 1	0.8549	0.7274	0.8347	0.8334	0.9479	0.8477
	(0.0233)	(0.0400)	(0.0241)	(0.0227)	(0.0069)	(0.0221)
Regime 2	0.3362	0.2861	0.3283	0.3278	0.3728	0.3334
	(0.1327)	(0.1138)	(0.1296)	(0.1294)	(0.1469)	(0.1316)

Π	Regime 1	Regime 2
Regime 1	0.9473	0.3318
	(0.0254)	
Regime 2	0.0527	0.6682
		(0.0635)

Series	ω	$\tilde{\alpha}$	α	β
Pound	0.0271	0.1068	0.0827	0.8826
	(0.0102)	(0.0265)	(0.0207)	(0.0285)
Deutschmark	0.0739	0.1282	0.1010	0.8037
	(0.0194)	(0.0220)	(0.0175)	(0.0318)
Yen	0.0040	0.0276	0.0211	0.9731
	(0.0038)	(0.0097)	(0.0074)	(0.0115)
Swiss-Franc	0.0866	0.1206	0.0913	0.7983
	(0.0439)	(0.0406)	(0.0406)	(0.0798)

Table 19: Estimation results for the restricted model with two regimes and GARCH. Standard errors are in parenthesis. The log-likelihood value is -2009.0.

	$\Gamma_{1,2}$	$\Gamma_{1,3}$	$\Gamma_{1,4}$	$\Gamma_{2,3}$	$\Gamma_{2,4}$	$\Gamma_{3,4}$
Regime 1	0.8602	0.7377	0.8373	0.8420	0.9500	0.8545
	(0.0273)	(0.0354)	(0.0257)	(0.0194)	(0.0087)	(0.0166)
Regime 2	0.4052	0.3475	0.3944	0.3966	0.4475	0.4025
	(0.1382)	(0.1192)	(0.1345)	(0.1350)	(0.1521)	(0.1369)

Π	Regime 1	Regime 2
Regime 1	0.9381	0.3196
	(0.0332)	
Regime 2	0.0619	0.6804
		(0.0583)

Series	ω	α	β
Pound	0.0219	0.0933	0.8697
	(0.0105)	(0.0354)	(0.0454)
Deutschmark	0.0477	0.1098	0.8145
	(0.0149)	(0.0268)	(0.0352)
Yen	0.0010	0.0176	0.9805
	(0.0016)	(0.0078)	(0.0103)
Swiss-Franc	0.0817	0.1204	0.7521
	(0.0388)	(0.0566)	(0.1015)

Figure 11: Smoothed probabilities for the models with two regimes and ARMACH. The top and bottom figures represent the smoothed probabilities of being in regime 1 for the unrestricted and restricted model respectively.

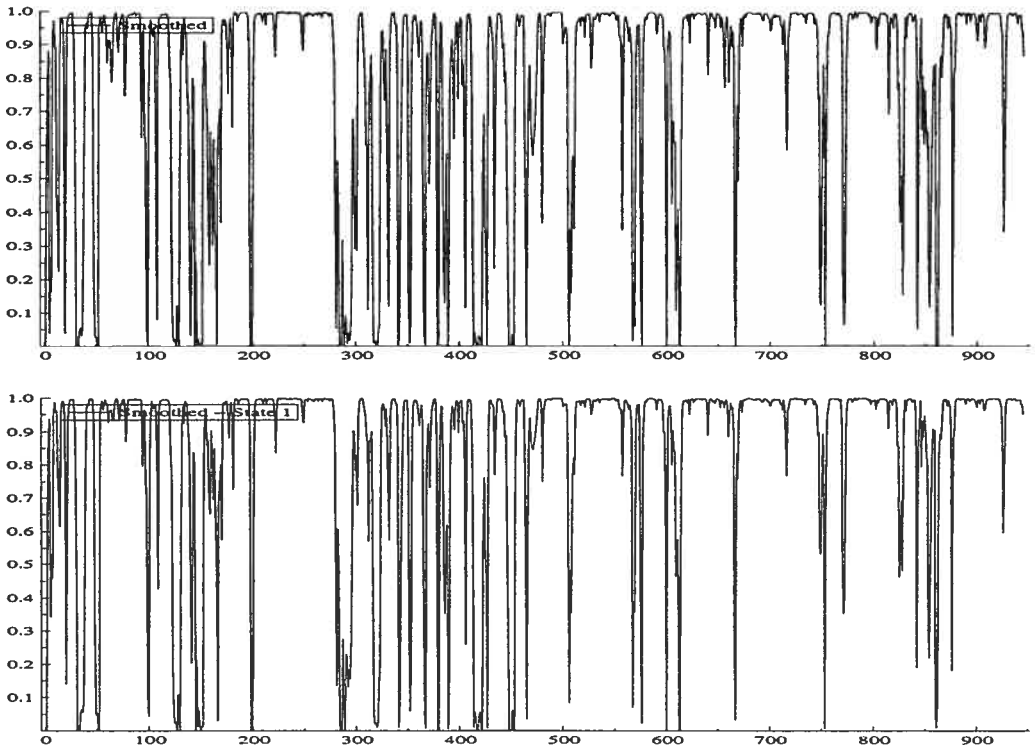


Figure 12: Smoothed correlations for the models with two regimes case and ARMACH. The top and bottom panel are for the unrestricted and restricted version of the model respectively.

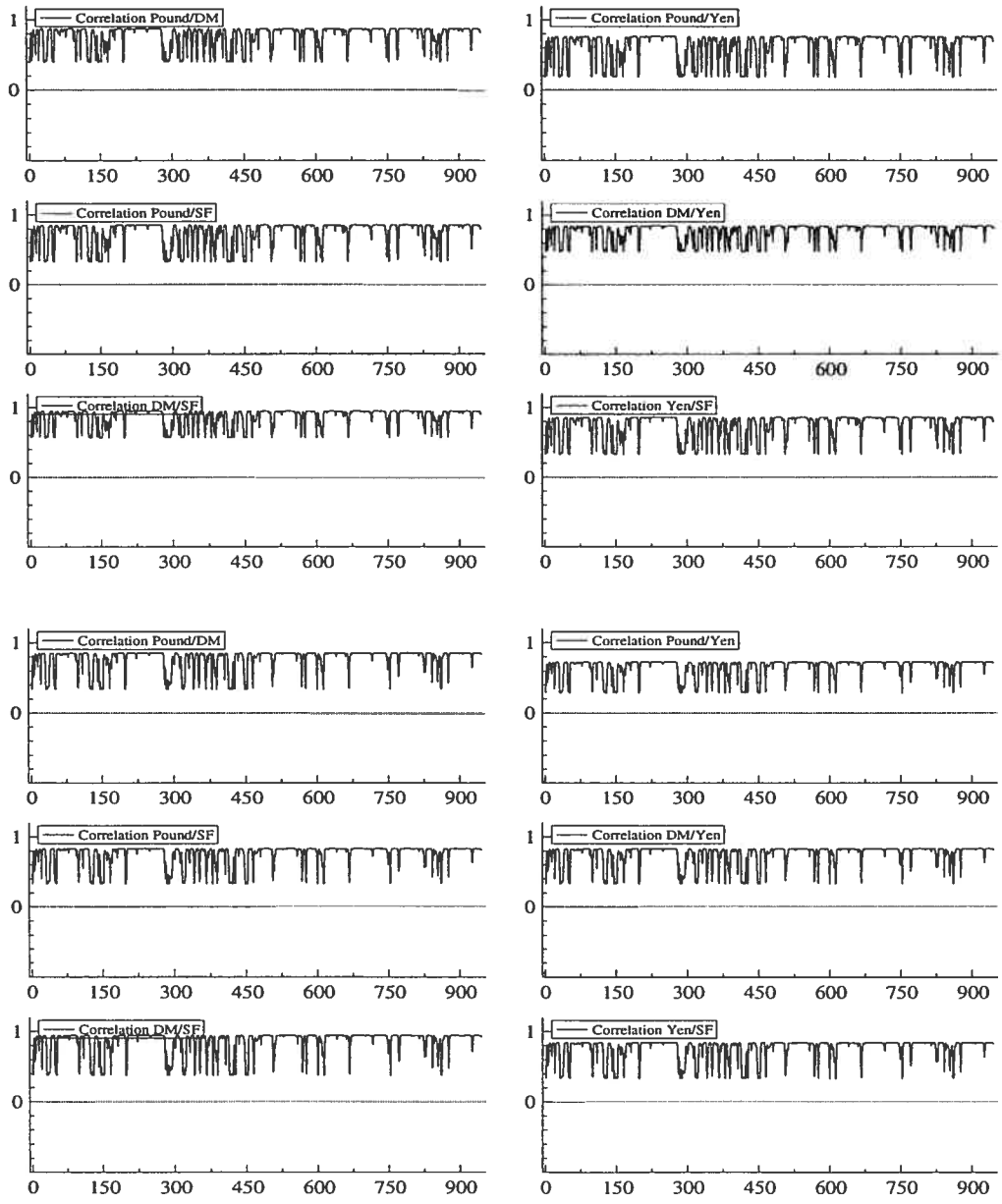


Table 20: Estimation results for the unrestricted model with three regimes and AR-MACH. Standard errors are in parenthesis. The log-likelihood value is -1971.7.

	$\Gamma_{1,2}$	$\Gamma_{1,3}$	$\Gamma_{1,4}$	$\Gamma_{2,3}$	$\Gamma_{2,4}$	$\Gamma_{3,4}$
Regime 1	0.1850	0.0855	0.0730	0.2048	0.2199	0.0620
	(0.2592)	(0.0819)	(0.1263)	(0.0410)	(0.0989)	(0.0830)
Regime 2	0.6039	0.4189	0.5598	0.7222	0.8853	0.7238
	(0.0697)	(0.0831)	(0.1307)	(0.0381)	(0.1341)	(0.0590)
Regime 3	0.9491	0.8497	0.9298	0.8568	0.9251	0.8705
	(0.0101)	(0.0347)	(0.0667)	(0.0672)	(0.2257)	(0.0894)

Π	Regime 1	Regime 2	Regime 3
Regime 1	0.6250	0.0000	0.0177
		(0.0326)	(0.0502)
Regime 2	0.2479	0.8847	0.1248
	(0.2045)		(0.1168)
Regime 3	0.1271	0.1153	0.8575
	(0.2189)	(0.0746)	

Series	ω	$\tilde{\alpha}$	α	β
Pound	0.0332	0.1135	0.0878	0.8615
	(0.0445)	(0.0469)	(0.0364)	(0.1205)
Deutschmark	0.0543	0.1151	0.0907	0.8311
	(0.0480)	(0.0226)	(0.0179)	(0.0905)
Yen	0.0040	0.0313	0.0239	0.9694
	(0.0053)	(0.0253)	(0.0193)	(0.0251)
Swiss-Franc	0.0718	0.1200	0.0908	0.8108
	(0.0965)	(0.0538)	(0.0408)	(0.1892)

Table 21: Estimation results for the unrestricted model with three regimes and GARCH. Standard errors are in parenthesis. The log-likelihood value is -1955.3.

	$\Gamma_{1,2}$	$\Gamma_{1,3}$	$\Gamma_{1,4}$	$\Gamma_{2,3}$	$\Gamma_{2,4}$	$\Gamma_{3,4}$
Regime 1	0.3074	0.1132	0.1101	0.2169	0.1582	-0.0474
	(0.1106)	(0.1491)	(0.1561)	(0.0949)	(0.1728)	(0.1383)
Regime 2	0.5992	0.4174	0.5559	0.7196	0.8863	0.7177
	(0.0930)	(0.0688)	(0.0810)	(0.0587)	(0.1985)	(0.0921)
Regime 3	0.9487	0.8524	0.9297	0.8567	0.9249	0.8718
	(0.0260)	(0.1190)	(0.1433)	(0.2012)	(0.3846)	(0.2039)

Π	Regime 1	Regime 2	Regime 3
Regime 1	0.6759	0.0000	0.0172
		(0.0537)	(0.0678)
Regime 2	0.2702	0.8835	0.1218
	(0.4512)		(0.0566)
Regime 3	0.0539	0.1165	0.8610
	(0.2730)	(0.0698)	

Series	ω	α	β
Pound	0.0290	0.1128	0.8203
	(0.0470)	(0.0742)	(0.2500)
Deutschmark	0.0344	0.1054	0.8258
	(0.0419)	(0.0333)	(0.1277)
Yen	0.0015	0.0234	0.9721
	(0.0044)	(0.0381)	(0.0407)
Swiss-Franc	0.0545	0.1038	0.7973
	(0.0662)	(0.0341)	(0.2056)

Table 22: Estimation results for the restricted model with three regimes and ARMACH. Standard errors are in parenthesis. The log-likelihood value is -1975.7.

	$\Gamma_{1,2}$	$\Gamma_{1,3}$	$\Gamma_{1,4}$	$\Gamma_{2,3}$	$\Gamma_{2,4}$	$\Gamma_{3,4}$
Regime 1	0.8775	0.7348	0.8567	0.8550	0.9723	0.8649
	(0.0160)	(0.0275)	(0.0183)	(0.0143)	(0.0038)	(0.0141)
Regime 2	0.7835	0.6561	0.7649	0.7634	0.8682	0.7723
	(0.0225)	(0.0285)	(0.0236)	(0.0212)	(0.0195)	(0.0172)
Regime 3	0.1508	0.1262	0.1472	0.1469	0.1670	0.1486
	(0.0692)	(0.0581)	(0.0676)	(0.0674)	(0.0766)	(0.0682)

Π	Regime 1	Regime 2	Regime 3
Regime 1	0.9260	0.0797	0.0305
		(0.0288)	(0.1151)
Regime 2	0.0686	0.8787	0.4365
	(0.0200)		(0.1849)
Regime 3	0.0054	0.0416	0.5330
	(0.0097)	(0.0203)	

Series	ω	$\tilde{\alpha}$	α	β
Pound	0.0311	0.1035	0.0801	0.8800
	(0.0123)	(0.0281)	(0.0219)	(0.0333)
Deutschmark	0.0659	0.1101	0.0868	0.8270
	(0.0187)	(0.0212)	(0.0169)	(0.0339)
Yen	0.0043	0.0299	0.0228	0.9709
	(0.0034)	(0.0105)	(0.0080)	(0.0116)
Swiss-Franc	0.0964	0.1281	0.0970	0.7814
	(0.0460)	(0.0421)	(0.0320)	(0.0827)

Table 23: Estimation results for the restricted model with three regimes and GARCH. Standard errors are in parenthesis. The log-likelihood value is -1961.3.

	$\Gamma_{1,2}$	$\Gamma_{1,3}$	$\Gamma_{1,4}$	$\Gamma_{2,3}$	$\Gamma_{2,4}$	$\Gamma_{3,4}$
Regime 1	0.8776	0.7347	0.8558	0.8545	0.9718	0.8645
	(0.0166)	(0.0282)	(0.0191)	(0.0148)	(0.0036)	(0.0144)
Regime 2	0.7759	0.6495	0.7566	0.7555	0.8592	0.7643
	(0.0245)	(0.0299)	(0.0255)	(0.0232)	(0.0220)	(0.0232)
Regime 3	0.1444	0.1209	0.1408	0.1406	0.1599	0.1422
	(0.0832)	(0.0698)	(0.0812)	(0.0810)	(0.0921)	(0.0820)

Π	Regime 1	Regime 2	Regime 3
Regime 1	0.9134	0.1006	0.0138
		(0.0396)	(0.0921)
Regime 2	0.0809	0.8555	0.4823
	(0.0232)		(0.1947)
Regime 3	0.0057	0.0439	0.5039
	(0.0098)	(0.0226)	

Series	ω	α	β
Pound	0.0256	0.0941	0.8621
	(0.0127)	(0.0375)	(0.0521)
Deutschmark	0.0464	0.1058	0.8163
	(0.0140)	(0.0254)	(0.0369)
Yen	0.0017	0.0226	0.9739
	(0.0018)	(0.0091)	(0.0120)
Swiss-Franc	0.0789	0.1222	0.7555
	(0.0346)	(0.0351)	(0.0764)

Figure 13: Smoothed probabilities for the three-regime case with ARMACH. The top and bottom figures represent the smoothed probabilities of being in each regime for the unrestricted and restricted model respectively.

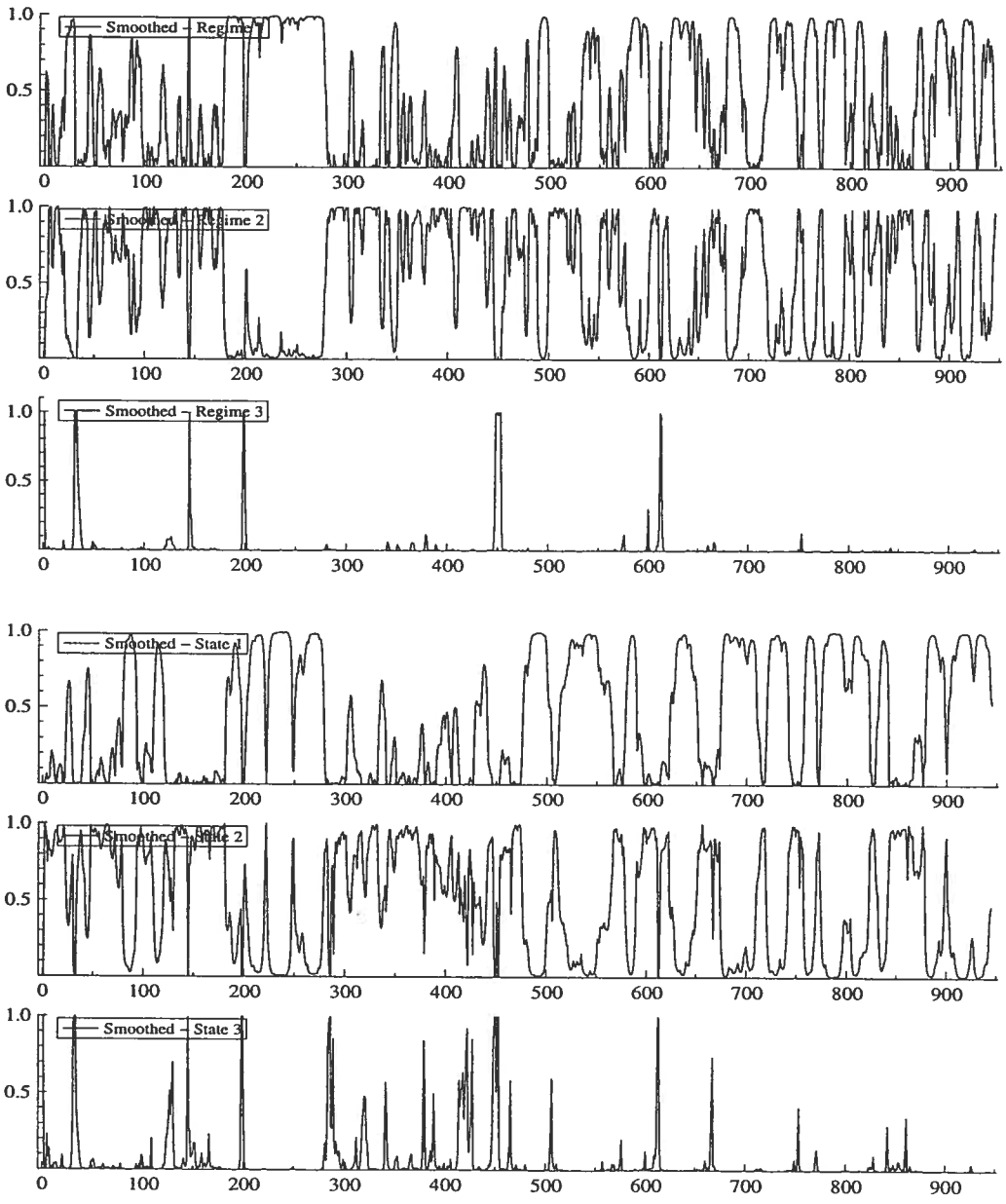


Figure 14: Smoothed correlations for the three-regime case with ARMACH. The top and bottom panel are for the unrestricted and restricted version of the model respectively.

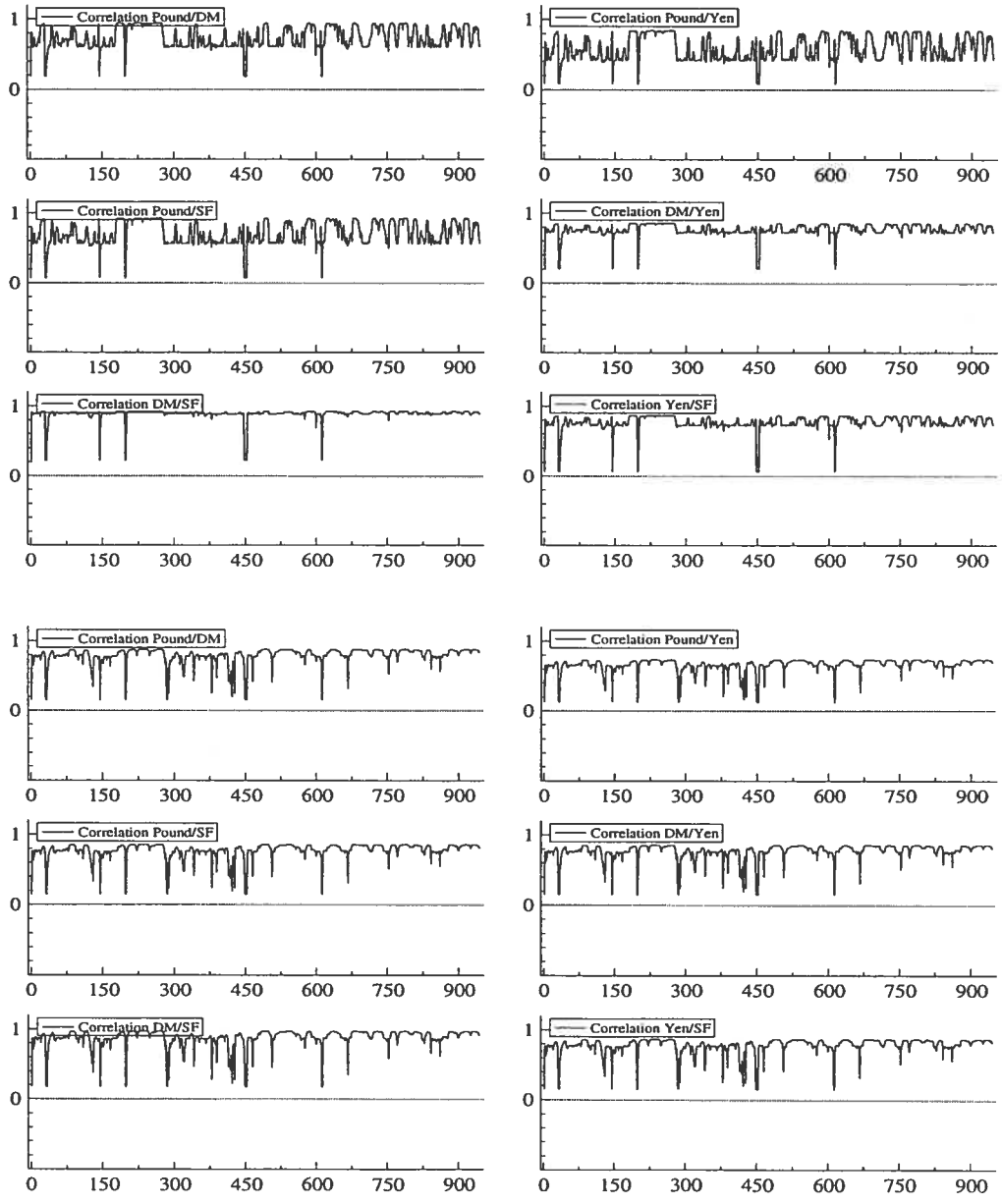


Table 24: Estimation results for the DCC-GARCH(1,1). Standard errors are in parenthesis. The log-likelihood value is -2109.2

Series	$\hat{\omega}$	$\hat{\alpha}$	$\hat{\beta}$	
Pound	0.0282	0.1338	0.8145	
	(0.0158)	(0.0518)	(0.0716)	
Deutschmark	0.0472	0.1689	0.7415	
	(0.0186)	(0.0446)	(0.0629)	
Yen	0.0031	0.0306	0.9615	
	(0.0027)	(0.0128)	(0.0171)	
Swiss-Franc	0.0646	0.1823	0.7189	
	(0.0539)	(0.1085)	(0.1636)	
	1	0.7657	0.5896	0.7093
		(0.0440)	(0.0636)	(0.0483)
		1	0.7219	0.8639
$\hat{\Gamma}$			(0.0467)	(0.0307)
			1	0.6814
				(0.0564)
				1
	$\hat{a}$	$\hat{b}$		
	0.1235	0.7822		
	(0.0451)	(0.0893)		

Table 25: Estimation results for the DCC-ARMACH(1,1). Standard errors are in parenthesis. The log-likelihood value is -2137.8

Series	$\hat{\omega}$	$\hat{\tilde{\alpha}}$	$\hat{\alpha}$	$\hat{\beta}$
Pound	0.0204	0.1120	0.0867	0.8841
	(0.0082)	(0.0288)	(0.0224)	(0.0294)
Deutschmark	0.0562	0.1524	0.1201	0.7932
	(0.0163)	(0.0252)	(0.0201)	(0.0377)
Yen	0.0045	0.0325	0.0248	0.9677
	(0.0031)	(0.0153)	(0.0117)	(0.0152)
Swiss-Franc	0.0882	0.1876	0.1420	0.7377
	(0.0636)	(0.0759)	(0.0553)	(0.1263)
	1	0.7554	0.6036	0.6849
		(0.0486)	(0.0524)	(0.0596)
		1	0.7255	0.8682
$\hat{\Gamma}$			(0.0383)	(0.0291)
			1	0.6843
				(0.0515)
				1

$\hat{a}$	$\hat{b}$
0.1088	0.8083
(0.0344)	(0.0571)

Figure 15: Correlations for the DCC-ARMACH(1,1).

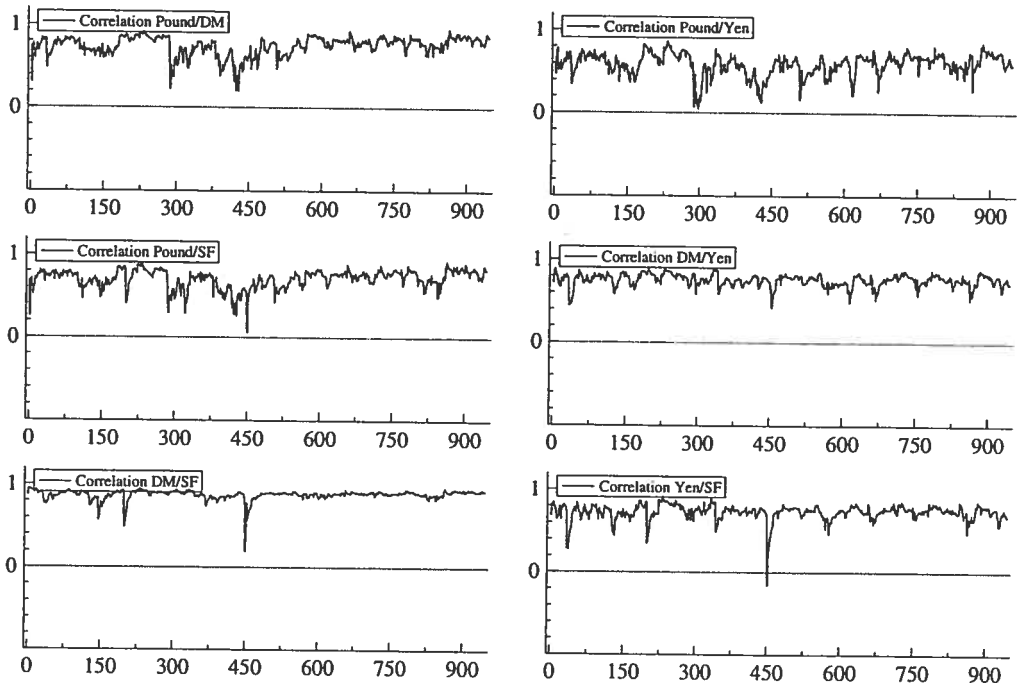


Table 26: Likelihood value and number of parameters for various models.

	Log-likelihood	Nb. par.
Unrestricted 3-regime GARCH	-1955.3	38
Restricted 3-regime GARCH	-1961.3	26
Unrestricted 3-regime ARMACH	-1971.7	38
Restricted 3-regime ARMACH	-1975.7	26
Unrestricted 2-regime GARCH	-1994.7	27
Restricted 2-regime GARCH	-2009.0	21
Unrestricted 2-regime ARMACH	-2011.6	27
Restricted 2-regime ARMACH	-2025.2	21
DCC-GARCH(1,1)	-2109.2	20
DCC-ARMACH(1,1)	-2137.8	20
CCC-GARCH(1,1)	-2272.1	18
CCC-ARMACH(1,1)	-2301.8	18

Figure 16: Smoothed probabilities of being in the regime of high correlations for the restricted model with two regimes and standard deviations from an ARMACH(1,1) for each series.

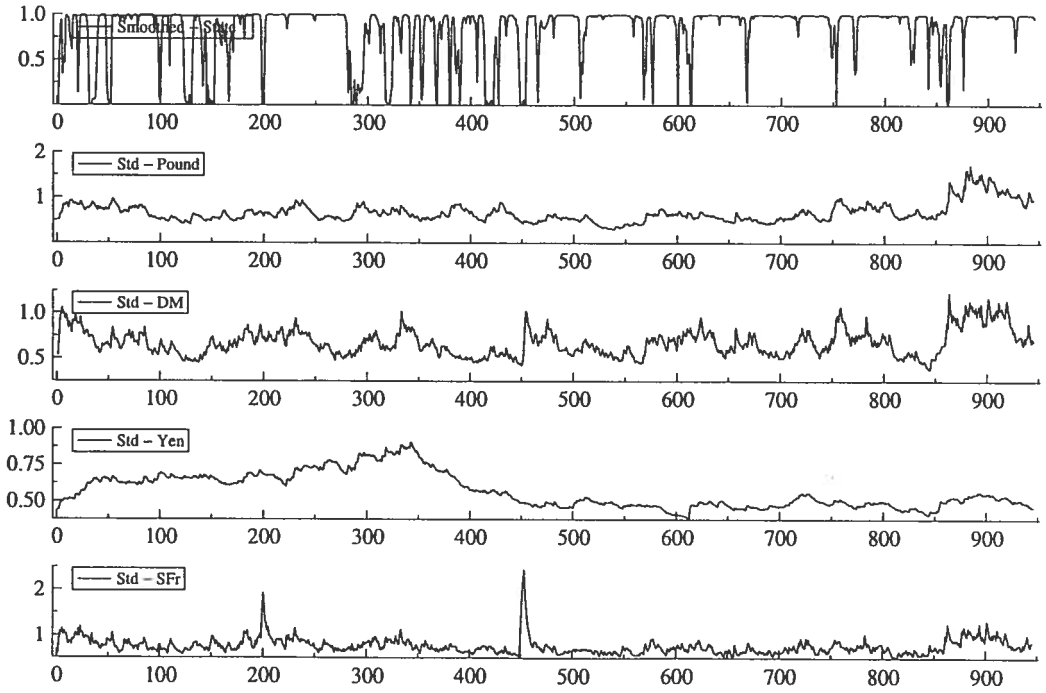
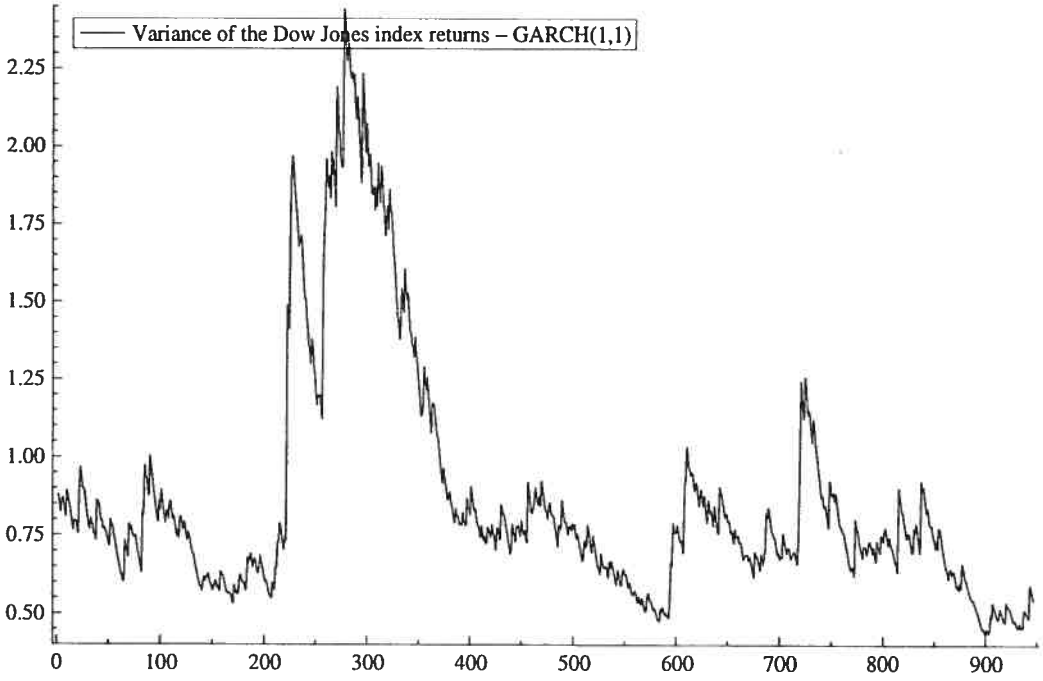


Figure 17: Conditional variance from a GARCH(1,1) for the return on the Dow Jones index.



Chapter 3: Short run and long run causality in time series: inference

1. Introduction

The concept of causality introduced by Wiener (1956) and Granger (1969) is now a basic notion for studying dynamic relationships between time series. The literature on this topic is considerable; see, for example, the reviews of Pierce and Haugh (1977), Newbold (1982), Geweke (1984), Lütkepohl (1991) and Gouriéroux and Monfort (1997, Chapter 10). The original definition of Granger (1969), which is used or adapted by most authors on this topic, refers to the predictability of a variable $X(t)$, where t is an integer, from its own past, the one of another variable $Y(t)$ and possibly a vector $Z(t)$ of auxiliary variables, **one period ahead**: more precisely, we say that Y causes X in the sense of Granger if the observation of Y up to time t $\quad (Y(\tau) : \tau \leq t)$ can help one to predict $X(t+1)$ when the corresponding observations on X and Z are available $(X(\tau),\ Z(\tau) : \tau \leq t)$; a more formal definition will be given below.

Recently, however, Lütkepohl (1993b) and Dufour and Renault (1998) have noted that, for multivariate models where a vector of auxiliary variables Z is used in addition to the variables of interest X and Y, it is possible that Y does not cause X in this sense, but can still help to predict X **several periods ahead**; on this issue, see also Sims (1980) and Renault, Sekkat, and Szafarz (1998). For example, the values $Y(\tau)$ up to time t may help to predict $X(t+2)$, even though they are useless to predict $X(t+1)$. This is due to the fact that Y may help to predict Z one period ahead, which in turn has an effect on X at a subsequent period. It is clear that studying such indirect effects can have a great interest for analyzing the relationships between time series. In particular, one can distinguish in this way properties of "short-run (non-)causality" and "long-run (non-)causality".

In this paper, we study the problem of testing non-causality at various horizons as defined in Dufour and Renault (1998) for finite-order vector autoregressive (VAR) models. In such models, the non-causality restriction at horizon one takes the form

of relatively simple zero restrictions on the coefficients of the VAR [see Boudjellaba, Dufour, and Roy (1992) and Dufour and Renault (1998)]. However non-causality restrictions at higher horizons (greater than or equal to 2) are generally nonlinear, taking the form of zero restrictions on multilinear forms in the coefficients of the VAR. When applying standard test statistics such as Wald-type test criteria, such forms can easily lead to asymptotically singular covariance matrices, so that standard asymptotic theory would not apply to such statistics. Further, calculation of the relevant covariance matrices _ which involve the derivatives of potentially large numbers of restrictions _ can become quite awkward.

Consequently, we propose simple tests for non-causality restrictions at various horizons [as defined in Dufour and Renault (1998)] which can be implemented only through linear regression methods and do not involve the use of artificial simulations [e.g., as in Lütkepohl and Burda (1997)]. This will be done, in particular, by considering multiple horizon vector autoregressions [called (p, h)-*autoregressions*] where the parameters of interest can be estimated by linear methods. Restrictions of non-causality at different horizons may then be tested through simple Wald-type (or Fisher-type) criteria after taking into account the fact that such autoregressions involve autocorrelated errors [following simple moving average processes] which are orthogonal to the regressors. The correction for the presence of autocorrelation in the errors may then be performed by using an autocorrelation consistent [or heteroskedasticity-autocorrelation-consistent (HAC)] covariance matrix estimator. Further, we distinguish between the case where the VAR process considered is stable (*i.e.*, the roots of the determinant of the associated AR polynomial are all outside the unit circle) and the one where the process may be integrated of an unknown order (although not explosive). In the first case, the test statistics follow standard chi-square distributions while, in the second case, they may follow nonstandard asymptotic distributions involving nuisance parameters, as already observed by several authors for the case of causality tests at horizon one [see Sims, Stock, and Watson (1990), Toda and Phillips (1993, 1994), Toda and Yamamoto (1995), Dolado and Lütkepohl (1996) and Yamada and Toda (1998)]. To meet the objective of producing simple procedures that can be implemented by least squares methods, we propose to deal with such problems by using an extension to the case

of multiple horizon autoregressions of the lag extension technique suggested by Choi (1993) for inference on univariate autoregressive models and by Toda and Yamamoto (1995) and Dolado and Lütkepohl (1996) for inference on standard VAR models. This extension will allow us to use standard asymptotic theory in order to test non-causality at different horizons without making assumption on the presence of unit roots and cointegrating relations. Finally, to alleviate the problems of finite-sample unreliability of asymptotic approximations in VAR models (on both stationary and nonstationary series), we propose the use of bootstrap methods to implement the proposed test statistics.

In section 2, we describe the model considered and introduce the notion of autoregression at horizon h [or (p, h)-autoregression] which will be the basis of our method. In section 3, we study the estimation of (p, h)-autoregressions and the asymptotic distribution of the relevant estimators for stable VAR processes. In section 4, we study the testing of non-causality at various horizons for stationary processes, while in section 5, we consider the case of processes that may be integrated. In section 6, we illustrate the procedures on a monthly VAR model of the U.S. economy involving a monetary variable (nonborrowed reserves), an interest rate (federal funds rate), prices (GDP deflator) and real GDP, over the period 1965-1996. We conclude in section 7.

2. Multiple horizon autoregressions

In this section, we develop the notion of "autoregression at horizon h" and the relevant notations. Consider a VAR (p) process of the form:

$$W(t) = \mu(t) + \sum_{k=1}^{p} \pi_k W(t-k) + a(t), \quad t = 1, \ldots, T, \tag{2.1}$$

where $W(t) = \left(w_{1t}, w_{2t}, \ldots, w_{mt}\right)'$ is an $m \times 1$ random vector, $\mu(t)$ is a deterministic trend, and

$$\begin{aligned} \mathrm{E}\left[a(s)\, a(t)'\right] &= \Omega, \text{ if } s = t, \\ &= 0, \text{ if } s \neq t, \end{aligned} \tag{2.2}$$

$$\det(\Omega) \neq 0. \tag{2.3}$$

The most common specification for $\mu(t)$ consists in assuming that $\mu(t)$ is a constant vector, *i.e.*

$$\mu(t) = \mu\,, \tag{2.4}$$

although other deterministic trends could also be considered.

The VAR (p) in equation (2.1) is an autoregression at horizon 1. We can then also write for the observation at time $t+h$:

$$W(t+h) = \mu^{(h)}(t) + \sum_{k=1}^{p} \pi_k^{(h)} W(t+1-k) + \sum_{j=0}^{h-1} \psi_j a(t+h-j)\ ,\ t = 0, \ldots, T-h\,,$$

where $\psi_0 = I_m$ and $h < T$. The appropriate formulas for the coefficients $\pi_k^{(h)}$, $\mu^{(h)}(t)$ and ψ_j are given in Dufour and Renault (1998), namely:

$$\pi_k^{(h+1)} = \pi_{k+h} + \sum_{l=1}^{h} \pi_{h-l+1}\pi_k^{(l)} = \pi_{k+1}^{(h)} + \pi_1^{(h)}\pi_k, \tag{2.5}$$

$$\pi_1^{(0)} = I_m,\ \pi_k^{(1)} = \pi_k\ , \tag{2.6}$$

$$\mu^{(h)}(t) = \sum_{k=0}^{h-1} \pi_1^{(k)} \mu(t+h-k)\,, \quad \psi_h = \pi_1^{(h)},\ \forall h \geq 0\,. \tag{2.7}$$

The ψ_h matrices are the impulse response coefficients of the process, which can also be obtained from the formal series:

$$\psi(z) = \pi(z)^{-1} = I_m + \sum_{k=1}^{\infty} \psi_k z^k\,, \quad \pi(z) = I_m - \sum_{k=1}^{\infty} \pi_k z^k\,. \tag{2.8}$$

Equivalently, the above equation for $W(t+h)$ can be written in the following way:

$$\begin{aligned} W(t+h)' &= \mu^{(h)}(t)' + \sum_{k=1}^{p} W(t+1-k)'\,\pi_k^{(h)\prime} + u^{(h)}(t+h)' \\ &= \mu^{(h)}(t)' + W(t,\,p)'\,\pi^{(h)} + u^{(h)}(t+h)'\,, \end{aligned} \tag{2.9}$$

for $t = 0, \ldots, T-h$ and where

$$W(t,\,p)' = \left[W(t)',\, W(t-1)',\, \ldots,\, W(t-p+1)'\right],$$

$$\pi^{(h)} = \left[\pi_1^{(h)}, \ldots, \pi_p^{(h)}\right]',$$

$$u^{(h)}(t+h)' = \left[u_1^{(h)}(t+h), \ldots, u_m^{(h)}(t+h)\right] = \sum_{j=0}^{h-1} a(t+h-j)'\psi_j'.$$

It is straightforward to see that $u^{(h)}(t+h)$ has a non-singular covariance matrix.

We call (2.9) an "autoregression of order p at horizon h" or a "(p, h)-autoregression". In the sequel, we will assume that the deterministic part of each autoregression is a linear function of a finite-dimensional parameter vector, *i.e.*

$$\mu^{(h)}(t) = \gamma(h)D^{(h)}(t) \tag{2.10}$$

where $\gamma(h)$ is a $m \times n$ coefficient vector and $D^{(h)}(t)$ is a $n \times 1$ vector of deterministic regressors. If $\mu(t)$ is a constant vector, *i.e.* $\mu(t) = \mu$, then $\mu^{(h)}(t)$ is simply a constant vector (which may depend on h):

$$\mu^{(h)}(t) = \mu_h. \tag{2.11}$$

To derive inference procedures, it will be convenient to consider a number of alternative formulations of (p, h)-autoregression autoregressions.

a) Matrix (p, h)-autoregression _ First, we can put (2.9) in matrix form, which yields:

$$w_h(h) = \overline{W}_p(h)\Pi^{(h)} + U_h(h), \; h = 1, \ldots, H, \tag{2.12}$$

where $w_h(k)$ and $U_h(k)$ are $(T-k+1) \times m$ matrices and $\overline{W}_p(k)$ is a $(T-k+1) \times (n+mp)$ matrix defined as

$$w_h(k) = \begin{bmatrix} W(0+h)' \\ W(1+h)' \\ \vdots \\ W(T-k+h)' \end{bmatrix} = [w_1(h, k), \ldots, w_m(h, k)], \tag{2.13}$$

$$\overline{W}_p(k) = \begin{bmatrix} W_p(0)' \\ W_p(1)' \\ \vdots \\ W_p(T-k)' \end{bmatrix}, \quad W_p(t) = \begin{bmatrix} D^{(h)}(t)' \\ W(t,p) \end{bmatrix}, \tag{2.14}$$

$$\Pi^{(h)} = \begin{bmatrix} \gamma(h)' \\ \pi^{(h)} \end{bmatrix} = [\beta_1(h), \beta_2(h), \ldots, \beta_m(h)], \tag{2.15}$$

$$U_h(k) = \begin{bmatrix} u^{(h)}(0+h)' \\ u^{(h)}(1+h)' \\ \vdots \\ u^{(h)}(T-k+h)' \end{bmatrix} = [u_1(h,k), \ldots, u_m(h,k)], \tag{2.16}$$

$$u_i(h,k) = \left[u_i^{(h)}(0+h), u_i^{(h)}(1+h), \ldots, u_i^{(h)}(T-k+h)\right]'. \tag{2.17}$$

We shall call the formulation (2.12) a "(p, h)-autoregression in matrix form".

b) Rectangular stacked (p, H)-autoregression _ To get the same regressor matrix on the right-hand side of (2.12), we can also consider:

$$w_h(H) = \overline{W}_p(H)\Pi^{(h)} + U_h(H), \quad h = 1, \ldots, H. \tag{2.18}$$

This, however, involves losing observations. Using (2.18), we can also stack the H systems above as follows:

$$w_H = \overline{W}_p(H)\Pi_H + U_H \tag{2.19}$$

where w_H and U_H are $(T-H+1)\times(mH)$ matrices and $W_p(H)$ is an $(mp)\times(mH)$ matrix such that

$$\begin{aligned} w_H &= [w_1(H), w_2(H), \ldots, w_H(H)], \\ \Pi_H &= [\Pi^{(1)}, \Pi^{(2)}, \ldots, \Pi^{(H)}], \\ U_H &= [U_1(H), U_2(H), \ldots, U_H(H)]. \end{aligned}$$

Since the elements of U_H are linear transformations of the random vectors $a(t)$, $t =$

$1, \ldots, T$, which contain Tm random variables, it is clear that the vector $vec(U_m)$ will have a singular covariance matrix when

$$Tm < (T-H)\,mH = TmH - mH^2,$$

which will be the case when $H \geq 2$ and $Tm > H$.

c) Vec-stacked $(p,\ H)$-autoregression _ We can also write equation (2.9) as

$$\begin{aligned} w(t+h) &= \left[I_m \otimes W_p(t)'\right] \bar{\Pi}^{(h)} + \bar{u}^{(h)}(t+h) \\ &= \overline{W}_p(t)'\, \bar{\Pi}^{(h)} + \bar{u}^{(h)}(t+h),\ t = 0, \ldots, T-h, \end{aligned} \tag{2.20}$$

where

$$\bar{\Pi}^{(h)} = vec\left(\Pi^{(h)}\right) = \begin{bmatrix} \beta_1^{(h)} \\ \beta_2^{(h)} \\ \vdots \\ \beta_m^{(h)} \end{bmatrix},$$

$$\overline{W}_p(t)' = \begin{bmatrix} W_p(t)' & 0 & \cdots & 0 \\ 0 & W_p(t)' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_p(t)' \end{bmatrix},$$

which yields the linear model

$$\bar{w}_h = Z_h \bar{\Pi}^{(h)} + \bar{u}_h \tag{2.21}$$

where

$$\bar{w}_h = \begin{bmatrix} W(0+h) \\ W(1+h) \\ \vdots \\ W(T) \end{bmatrix} = vec\left[w_h(h)'\right],$$

$$Z_h = \begin{bmatrix} \overline{W}_p(0)' \\ \overline{W}_p(1)' \\ \vdots \\ \overline{W}_p(T-h)' \end{bmatrix} = \begin{bmatrix} I_m \otimes W_p(0)' \\ I_m \otimes W_p(1)' \\ \vdots \\ I_m \otimes W_p(T-h)' \end{bmatrix},$$

$$\bar{u}_h = \begin{bmatrix} u^{(h)}(0+h) \\ u^{(h)}(1+h) \\ \vdots \\ u^{(h)}(T) \end{bmatrix} = vec\left[U_h(h)'\right].$$

It is also possible to stack together the models (2.21) for $h = 1, \ldots, H$:

$$\bar{w}(H) = Z(H)\,\bar{\Pi}_H + \bar{u}(H) \tag{2.22}$$

where

$$\bar{w}(H) = \begin{bmatrix} \bar{w}_1 \\ \bar{w}_2 \\ \vdots \\ \bar{w}_H \end{bmatrix}, \; \bar{u}(H) = \begin{bmatrix} \bar{u}_1 \\ \bar{u}_2 \\ \vdots \\ \bar{u}_H \end{bmatrix}, \; Z(H) = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & Z_H \end{bmatrix}.$$

d) Individual $(p,\ H)$-autoregressions _ Consider finally a single dependent variable

$$W_i(t+h) = W_p(t)'\,\beta_i(h) + u_i^{(h)}(t+h)\ ,\ t = 0, \ldots, T-H, \tag{2.23}$$

for $1 \leq h \leq H$, where $1 \leq i \leq m$. We can also write:

$$\widetilde{W}_i(t+H) = \left[I_H \otimes W_p(t)'\right]\bar{\beta}_i(H) + \tilde{u}_i(t+H)\ ,\ t = 0, \ldots, T-H\ , \tag{2.24}$$

where

$$\widetilde{W}_i(t+H) = \begin{bmatrix} W_i(t+1) \\ W_i(t+2) \\ \vdots \\ W_i(t+H) \end{bmatrix}, \; \tilde{u}_i(t+H) = \begin{bmatrix} u_i(t+1) \\ u_i(t+2) \\ \vdots \\ u_i(t+H) \end{bmatrix}, \; \bar{\beta}_i(H) = \begin{bmatrix} \beta_i(1) \\ \beta_i(2) \\ \vdots \\ \beta_i(H) \end{bmatrix},$$

which yields the linear model

$$\widetilde{W}_i(H) = \tilde{Z}_H \bar{\beta}_i(H) + \tilde{u}_H \tag{2.25}$$

where

$$\widetilde{W}_i(H) = \begin{bmatrix} \widetilde{W}_i(0+H) \\ \widetilde{W}_i(1+H) \\ \vdots \\ \widetilde{W}_i(T) \end{bmatrix}, \; \bar{\beta}_i(H) = \begin{bmatrix} \bar{\beta}_i(1) \\ \bar{\beta}_i(2) \\ \vdots \\ \bar{\beta}_i(H) \end{bmatrix},$$

$$\tilde{Z}_H = \begin{bmatrix} I_H \otimes W_p(0)' \\ I_H \otimes W_p(1)' \\ \vdots \\ I_H \otimes W_p(T-H)' \end{bmatrix}, \; \tilde{u}_H = \begin{bmatrix} \tilde{u}_i(0+H) \\ \tilde{u}_i(1+H) \\ \vdots \\ \tilde{u}_i(T) \end{bmatrix}.$$

In the sequel, we shall focus on prediction equations for individual variables and the matrix $(p,\, h)$-autoregressive form of the system in (2.12).

3. Estimation of (p, h) autoregressions

Let us now consider each autoregression of order p at horizon h as given by (2.12):

$$w_h(h) = \overline{W}_p(h)\, \Pi^{(h)} + U_h(h) \;, \; h = 1, \ldots, H\,. \tag{3.1}$$

We can estimate (3.1) by ordinary least squares (OLS), which yields the estimator:

$$\begin{aligned}\hat{\Pi}^{(h)} &= \left[\overline{W}_p(h)'\overline{W}_p(h)\right]^{-1}\overline{W}_p(h)'w_h(h) \\ &= \Pi^{(h)} + \left[\overline{W}_p(h)'\overline{W}_p(h)\right]^{-1}\overline{W}_p(h)'U_h(h)\ ,\end{aligned}$$

hence

$$\sqrt{T}\left[\hat{\Pi}^{(h)} - \Pi^{(h)}\right] = \left[\frac{1}{T}\overline{W}_p(h)'\overline{W}_p(h)\right]^{-1}\frac{1}{\sqrt{T}}\overline{W}_p(h)'U_h(h)$$

where

$$\begin{aligned}\frac{1}{T}\overline{W}_p(h)'\overline{W}_p(h) &= \frac{1}{T}\sum_{t=0}^{T-h} W_p(t)\,W_p(t)'\ , \\ \frac{1}{\sqrt{T}}\overline{W}_p(h)'U_h(h) &= \frac{1}{\sqrt{T}}\sum_{t=0}^{T-h} W_p(t)\,u^{(h)}(t+h)'\ .\end{aligned}$$

Suppose now that

$$\frac{1}{T}\sum_{t=0}^{T-h} W_p(t)\,W_p(t)' \xrightarrow[T\to\infty]{\mathsf{p}} \Gamma_p \quad \text{with} \quad \det(\Gamma_p) \neq 0\,. \tag{3.2}$$

In particular, this will be the case if the process $W(t)$ is second-order stationary, strictly indeterministic and regular, in which case

$$\mathsf{E}\left[W_p(t)\,W_p(t)'\right] = \Gamma_p\,,\ \forall t. \tag{3.3}$$

Cases where the process does not satisfy these conditions are covered in section 5. Further, since

$$u^{(h)}(t+h) = a(t+h) + \sum_{k=1}^{h-1}\psi_k a(t+h-k)$$

(where, by convention, any sum of the form $\sum_{k=1}^{h-1}$ with $h < 2$ is zero), we have:

$$\begin{aligned}\mathsf{E}\left[W_p(t)\,u^{(h)}(t+h)'\right] &= 0\ ,\ \text{for } h = 1,\,2,\,\ldots\,, \\ \mathsf{V}\left\{vec\left[W_p(t)\,u^{(h)}(t+h)'\right]\right\} &= \Delta_p(h)\ .\end{aligned}$$

If the process $W(t)$ is strictly stationary with i.i.d. innovations $a(t)$, we can write:

$$\mathsf{E}[W_p(s)\,u_i^{(h)}(s+h)\,u_j^{(h)}(t+h)\,W_p(t)'] = \Gamma_{ij}(p,\,h,\,t-s) = \Gamma_{ij}(p,\,h,\,s-t) \quad (3.4)$$

where $1 \leq i \leq m$, $1 \leq j \leq m$, with

$$\begin{aligned}\Gamma_{ij}(p,\,h,\,0) &= \mathsf{E}\left[W_p(t)\,u_i^{(h)}(t+h)\,u_j^{(h)}(t+h)\,W_p(t)'\right] \\ &= \sigma_{ij}(h)\,\mathsf{E}\left[W_p(t)\,W_p(t)'\right] = \sigma_{ij}(h)\,\Gamma_p\,, \qquad (3.5)\end{aligned}$$

$$\Gamma_{ij}(p,\,h,\,t-s) = 0\,, \quad \text{if } |t-s| \geq h\,. \qquad (3.6)$$

In this case,

$$\Delta_p(h) = [\sigma_{ij}(h)\,\Gamma_p]_{i,\,j=1,\,\ldots,\,m} = \Sigma(h) \otimes \Gamma_p \qquad (3.7)$$

where $\Sigma(h)$ is nonsingular, and thus $\Delta_p(h)$ is also nonsingular. The nonsingularity of $\Sigma(h)$ follows from the identity

$$u^{(h)}(t+h) = \left[\psi_{h-1},\,\psi_{h-2},\,\ldots,\,\psi_1,\,I_m\right]\left[a(t+1)',\,a(t+2)',\,\ldots,\,a(t+h)'\right]'.$$

Under usual regularity conditions,

$$\frac{1}{\sqrt{T}}\sum_{t=0}^{T-h} vec[W_p(t)\,u^{(h)}(t+h)'] \xrightarrow[T\to\infty]{\mathsf{L}} N\left[0,\,\bar{\Delta}_p(h)\right] \qquad (3.8)$$

where $\bar{\Delta}_p(h)$ is a nonsingular covariance matrix which involves the variance and the autocovariances of $W_p(t)\,u^{(h)}(t+h)'$ [and possibly other parameters, if the process $W(t)$ is not linear]. Then,

$$\begin{aligned}&\sqrt{T}\,vec\left[\hat{\Pi}^{(h)} - \Pi^{(h)}\right] \\ &= \left\{I_m \otimes \left[\frac{1}{T}\overline{W}_p(h)'\,\overline{W}_p(h)\right]^{-1}\right\} vec\left[\frac{1}{\sqrt{T}}\overline{W}_p(h)'\,U_h(h)\right] \\ &= \left\{I_m \otimes \left[\frac{1}{T}\overline{W}_p(h)'\,\overline{W}_p(h)\right]^{-1}\right\} \frac{1}{\sqrt{T}}\sum_{t=0}^{T-h} vec\left[W_p(t)\,u^{(h)}(t+h)'\right] \\ &\xrightarrow[T\to\infty]{\mathsf{L}} N\left[0,\,\left(I_m \otimes \Gamma_p^{-1}\right)\bar{\Delta}_p(h)\left(I_m \otimes \Gamma_p^{-1}\right)\right]. \qquad (3.9)\end{aligned}$$

For convenience, we shall summarize the above observations in the following proposition.

Proposition 3.1 ASYMPTOTIC NORMALITY OF LS IN A (p, h) STATIONARY VAR. *Under the assumptions* (2.1), (3.2), *and* (3.8), *the asymptotic distribution of* $\sqrt{T}\, vec\left[\hat{\Pi}^{(h)} - \Pi^{(h)}\right]$ *is* $N[0, \Sigma(\hat{\Pi}^{(h)})]$, *where* $\Sigma(\hat{\Pi}^{(h)}) = \left(I_m \otimes \Gamma_p^{-1}\right) \bar{\Delta}_p(h) \left(I_m \otimes \Gamma_p^{-1}\right)$.

4. Causality tests based on stationary (p, h)-autoregressions

Consider the i-th equation $(1 \leq i \leq m)$ in system (2.12):

$$\bar{w}_i(h) = \overline{W}_p(h)\, \beta_i(h) + \bar{u}_i(h), \quad 1 \leq i \leq m, \tag{4.1}$$

where $\bar{w}_i(h) = w_i(h, h)$ and $\bar{u}_i(h) = u_i(h, h)$, where $w_i(h, h)$ and $u_i(h, h)$ are defined in (2.13) and (2.16). We wish to test:

$$H_0(h) : R\beta_i(h) = r \tag{4.2}$$

where R is a $q \times (n + mp)$ matrix of rank q. In particular, if we wish to test the hypothesis that w_{jt} does not cause w_{it} at horizon h [*i.e.*, using the notation of Dufour and Renault (1998), $w_j \underset{h}{\not\rightarrow} w_i \mid I_{(j)}$, where $I_{(j)}(t)$ is the Hilbert space generated by the basic information set $I(t)$ and the variables $w_{k\tau}$, $\omega < \tau \leq t$, $k \neq j$, ω being an appropriate starting time $(\omega \leq -p + 1)$], the restriction would take the form:

$$H_{j \not\rightarrow i}^{(h)} : \pi_{ijk}^{(h)} = 0, \quad k = 1, \ldots, p, \tag{4.3}$$

where $\pi_k^{(h)} = \left[\pi_{ijk}^{(h)}\right]_{i,\, j=1,\ldots,m}$, $k = 1, \ldots, p$. In other words, the null hypothesis takes the form of a set of zero restrictions on the coefficients of $\beta_i(h)$ as defined in (2.15). The matrix of restrictions R in this case takes the form $R = R(j)$, where $R(j) \equiv [\delta_1(j), \delta_2(j), \ldots, \delta_p(j)]'$ is a $p \times (n + mp)$ matrix, $\delta_k(j)$ is a $(n + pm) \times 1$ vector

whose elements are all equal to zero except for a unit value at position $n+(k-1)m+j$, i.e. $\delta_k(j) = [\delta(1, n+(k-1)m+j), \ldots, \delta(n+pm, n+(k-1)m+j)]'$, $k = 1, \ldots, p$, with $\delta(i, j) = 1$ if $i = j$, and $\delta(i, j) = 0$ if $i \neq j$. Note also that the conjunction of the hypothesis $H_{j \nrightarrow i}^{(h)}$, $h = 1, \ldots, (m-2)p+1$, is sufficient to obtain noncausality at all horizons [see Dufour and Renault (1998, section 4)]. Non-causality up to horizon H is the conjunction of the hypothesis $H_{j \nrightarrow i}^{(h)}$, $h = 1, \ldots, H$.

We have:

$$\hat{\beta}_i(h) = \beta_i(h) + \left[\overline{W}_p(h)'\overline{W}_p(h)\right]^{-1}\overline{W}_p(h)'\bar{u}_i(h),$$

hence

$$\sqrt{T}\left[\hat{\beta}_i(h) - \beta_i(h)\right] = \left[\frac{1}{T}\overline{W}_p(h)'\overline{W}_p(h)\right]^{-1}\frac{1}{\sqrt{T}}\sum_{t=0}^{T-h} W_p(t)\, u_i^{(h)}(t+h).$$

Under standard regularity conditions [see White (1999, chap. 5-6)],

$$\sqrt{T}\left[\hat{\beta}_i(h) - \beta_i(h)\right] \xrightarrow[T\to\infty]{L} N\left[0, \mathsf{V}(\hat{\beta}_i)\right]$$

with $\det\left[\mathsf{V}(\hat{\beta}_i)\right] \neq 0$, where $\mathsf{V}(\hat{\beta}_i)$ can be consistently estimated:

$$\hat{V}_T(\hat{\beta}_i) \xrightarrow[T\to\infty]{\mathsf{p}} \mathsf{V}(\hat{\beta}_i).$$

More explicit forms for $\hat{V}_T(\hat{\beta}_i)$ will be discussed below. Note also that

$$\Gamma_p = \operatorname*{plim}_{T\to\infty}\frac{1}{T}\overline{W}_p(h)'\overline{W}_p(h), \quad \det(\Gamma_p) \neq 0.$$

Let

$$\begin{aligned} V_{ip}(T) &= \mathsf{Var}\left[\frac{1}{\sqrt{T}}\overline{W}_p(h)'\bar{u}_i(h)\right] = \frac{1}{T}\mathsf{Var}\left[\sum_{t=0}^{T-h} W_p(t)\, u_i^{(h)}(t+h)\right] \\ &= \frac{1}{T}\left\{\sum_{t=0}^{T-h}\mathsf{E}\left[W_p(t)\, u_i^{(h)}(t+h)\, u_i^{(h)}(t+h)\, W_p(t)'\right]\right. \end{aligned}$$

$$+\sum_{\tau=1}^{h-1}\sum_{t=\tau+1}^{T-h}\Big\{\mathsf{E}\big[W_p(t)\,u_i^{(h)}(t+h)\,u_i^{(h)}(t-\tau+h)\,W_p(t-\tau)'\big]$$
$$+\mathsf{E}\big[W_p(t-\tau)\,u_i^{(h)}(t-\tau+h)\,u_i^{(h)}(t+h)\,W_p(t)'\big]\Big\}.$$

Let us assume that

$$V_{ip}(T) \underset{T\to\infty}{\longrightarrow} V_{ip}\,,\ \det(V_{ip}) \neq 0\,, \tag{4.4}$$

where V_{ip} can be estimated by a computable consistent estimator $\hat{V}_{ip}(T)$:

$$\hat{V}_{ip}(T) \underset{T\to\infty}{\xrightarrow{\mathsf{p}}} V_{ip}\,. \tag{4.5}$$

Then,

$$\sqrt{T}\left[\hat{\beta}_i(h) - \beta_i(h)\right] \underset{T\to\infty}{\xrightarrow{\mathsf{p}}} N\left[0,\ \Gamma_p^{-1}V_{ip}\Gamma_p^{-1}\right]$$

so that $\mathsf{V}(\hat{\beta}_i) = \Gamma_p^{-1}V_{ip}\Gamma_p^{-1}$. Further, in this case,

$$\hat{V}_T(\hat{\beta}_i) = \hat{\Gamma}_p^{-1}\hat{V}_{ip}(T)\,\hat{\Gamma}_p^{-1} \underset{T\to\infty}{\xrightarrow{\mathsf{p}}} \mathsf{V}(\hat{\beta}_i)\,,$$
$$\hat{\Gamma}_p = \frac{1}{T}\sum_{t=0}^{T-h} W_p(t)\,W_p(t)' = \frac{1}{T}\overline{W}_p(h)'\,\overline{W}_p(h) \underset{T\to\infty}{\xrightarrow{\mathsf{p}}} \Gamma_p\,.$$

We can thus state the following proposition.

Proposition 4.1 ASYMPTOTIC DISTRIBUTION OF TEST CRITERION FOR NON-CAUSALITY AT HORIZON h IN A STATIONARY VAR. *Suppose the assumptions of Proposition* **3.1** *hold jointly with* (4.4) – (4.5). *Then, under any hypothesis of the form* $H_0(h)$ *in* (4.2), *the asymptotic distribution of*

$$\mathcal{W}[H_0(h)] = T\left[R\hat{\beta}_i(h) - r\right]'\left[R\hat{V}_T(\hat{\beta}_i)R'\right]^{-1}\left[R\hat{\beta}_i(h) - r\right] \tag{4.6}$$

is $\chi^2(q)$. *In particular, under the hypothesis* $H_{j\nrightarrow i}^{(h)}$ *of non-causality at horizon h from* w_{jt} *to* w_{it} $\left(w_j \underset{h}{\nrightarrow} w_i \,|\, I_{(j)}\right)$, *the asymptotic distribution of the corresponding statistic* $\mathcal{W}[H_0(h)]$ *is* $\chi^2(p)$.

The problem now consists in estimating V_{ip}. Let $\widehat{\widehat{u}}_i(h) = \left[\hat{u}_i^{(h)}(t+h) : t = 0, \ldots, T-h\right]'$ be the vector of OLS residuals from the regression

(4.1), $\hat{g}_i^{(h)}(t+h) = W_p(t)\,\hat{u}_i^{(h)}(t+h)$, and set

$$R_i^{(h)}(\tau) = \frac{1}{T-h}\sum_{t=\tau}^{T-h} \hat{g}_i^{(h)}(t+h)\,\hat{g}_i^{(h)}(t+h-\tau)',\ \tau = 0,\ 1,\ 2,\ \ldots\ .$$

If the innovations are i.i.d. or, more generally, if (3.6) holds, a natural estimator of V_{ip}, which would take into account the fact that the prediction errors $u^{(h)}(t+h)$ follow an MA$(h-1)$ process, is given by:

$$\hat{V}_{ip}^{(W)}(T) = R_i^{(h)}(0) + \sum_{\tau=1}^{h-1}\left[R_i^{(h)}(\tau) + R_i^{(h)}(\tau)'\right]\ .$$

Under regularity conditions studied by White (1999, Section 6.3),

$$\hat{V}_{ip}^{(W)}(T) - V_{ip} \xrightarrow[T\to\infty]{\mathrm{p}} 0.$$

A problem with $\hat{V}_{ip}^{(W)}(T)$ is that it is not necessarily positive-definite.

An alternative estimator which is automatically positive-semidefinite is the one suggested by Doan and Litterman (1983), Gallant (1987) and Newey and West (1987a):

$$\hat{V}_{ip}^{(NW)}(T) = R_i^{(h)}(0) + \sum_{\tau=1}^{m(T)-1} \kappa(\tau,\, m(T))\left[R_i^{(h)}(\tau) + R_i^{(h)}(\tau)'\right]\ , \qquad (4.7)$$

where $\kappa(\tau,\, m) = 1 - [\tau/(m+1)]$, $\lim_{T\to\infty} m(T) = \infty$, and $\lim_{T\to\infty}\left[m(T)/T^{1/4}\right] = 0$. Under the regularity conditions given by Newey and West (1987a),

$$\hat{V}_{ip}^{(NW)}(T) - V_{ip} \xrightarrow[T\to\infty]{} 0\,.$$

Other estimators that could be used here includes various heteroskedasticity-autocorrelation-consistent (HAC) estimators; see Andrews (1991), Andrews and Monahan (1992), Cribari-Neto, Ferrari, and Cordeiro (2000), Cushing and McGarvey (1999), Den Haan and Levin (1997), Hansen (1992), Newey and McFadden (1994), Wooldridge (1989).

The cost of having a simple procedure that sidestep all the nonlinearities associated with the non-causality hypothesis is a loss of efficiency. There are two places where we

are not using all information. The constraints on the $\pi_k^{(h)}$'s are giving information on the ψ_j's and we are not using it. We are also estimating the VAR by OLS and correcting the variance-covariance matrix instead of doing a GLS-type estimation. These two sources of inefficiencies could potentially be overcome but it would lead to less user-friendly procedures.

The asymptotic distribution provided by Proposition **4.1**, may not be very reliable in finite samples, especially if we consider a VAR system with a large number of variables and/or lags. Due to autocorrelation, a larger horizon may also affect the size an power of the test. So an alternative to using the asymptotic distribution chi-square of $\mathcal{W}[H_0(h)]$, consists in using Monte Carlo test techniques [see Dufour (2002)] or bootstrap methods [see, for example, Paparoditis (1996), Paparoditis and Streitberg (1991), Kilian (1998a, 1998b)]. In view of the fact that the asymptotic distribution of $\mathcal{W}[H_0(h)]$ is nuisance-parameter-free, such methods yield asymptotically valid tests when applied to $\mathcal{W}[H_0(h)]$ and typically provide a much better control of test level in finite samples. It is also possible that using better estimates would improve size control, although this is not clear, for important size distortions can occur in multivariate regressions even when unbiased efficient estimators are available [see, for example, Dufour and Khalaf (2002)].

5. Causality tests based on nonstationary (p, h)-autoregressions

In this section, we study how the tests described in the previous section can be adjusted in order to allow for non-stationary possibly integrated processes. In particular, let us assume that

$$W(t) = \mu(t) + \eta(t), \tag{5.1}$$

$$\mu(t) = \delta_0 + \delta_1 t + \cdots + \delta_q t^q, \quad \eta(t) = \sum_{k=1}^{p} \pi_k \eta(t-k) + a(t), \tag{5.2}$$

$t = 1, \ldots, T$, where $\delta_0, \delta_1, \ldots, \delta_q$ are $m \times 1$ fixed vectors, and the process $\eta(t)$ is at most $I(d)$ where d is an integer greater than or equal to zero. Typical values for d are 0, 1 or 2.

Under the above assumptions, we can also write:

$$W(t) = \gamma_0 + \gamma_1 t + \cdots + \gamma_q t^q + \sum_{k=1}^{p} \pi_k W(t-k) + a(t) \ , \quad t = 1, \ldots, T, \quad (5.3)$$

where $\gamma_0, \gamma_1, \ldots, \gamma_q$ are $m \times 1$ fixed vectors (which depend on $\delta_0, \delta_1, \ldots, \delta_q$, and $\pi_1, \ldots, \pi_p$); see Toda and Yamamoto (1995). Under the specification (5.3), we have:

$$W(t+h) = \mu^{(h)}(t) + \sum_{k=1}^{p} \pi_k^{(h)} W(t+1-k) + u^{(h)}(t+h) \ , \quad t = 0, \ldots, T-h. \quad (5.4)$$

where $\mu^{(h)}(t) = \gamma_0^{(h)} + \gamma_1^{(h)} t + \cdots + \gamma_q^{(h)} t^q$ and $\gamma_0^{(h)}, \gamma_1^{(h)}, \ldots, \gamma_q^{(h)}$ are $m \times 1$ fixed vectors. For $h = 1$, this equation is identical with (5.3). For $h \geq 2$, the errors $u^{(h)}(t+h)$ follow a MA$(h-1)$ process as opposed to being i.i.d.. For any integer j, we have:

$$\begin{aligned} W(t+h) &= \mu^{(h)}(t) + \sum_{\substack{k=1 \\ k \neq j}}^{p} \pi_k^{(h)} \left[W(t+1-k) - W(t+1-j) \right] \\ &\quad + \left(\sum_{k=1}^{p} \pi_k^{(h)} \right) W(t+1-j) + u^{(h)}(t+h) \ , \end{aligned} \quad (5.5)$$

$$\begin{aligned} W(t+h) - W(t+1-j) &= \mu^{(h)}(t) + \sum_{\substack{k=1 \\ k \neq j}}^{p} \pi_k^{(h)} \left[W(t+1-k) - W(t+1-j) \right] \\ &\quad - \left(I_m - \sum_{k=1}^{p} \pi_k^{(h)} \right) W(t+1-j) + u^{(h)}(t+h) \ , \end{aligned} \quad (5.6)$$

for $t = 0, \ldots, T-h$. The two latter expressions can be viewed as extensions to (p, h)-autoregressions of the representations used by Dolado and Lütkepohl (1996, pp. 372-373) for VAR(p) processes. Further, on taking $j = p+1$ in (5.6), we see that

$$W(t+h) - W(t-p) = \mu^{(h)}(t) + \sum_{k=1}^{p} A_k^{(h)} \Delta W(t+1-k)$$

$$+ B_p^{(h)} W(t-p) + u^{(h)}(t+h) \tag{5.7}$$

where $\Delta W(t) = W(t) - W(t-1)$, $A_k^{(h)} = \sum_{j=1}^{k} \pi_k^{(h)}$,and $B_k^{(h)} = A_k^{(h)} - I_m$. Equation (5.7) may be interpreted as an error-correction form at the horizon h,with base $W(t-p)$.

Let us now consider the extended autoregression

$$\begin{aligned} W(t+h) &= \mu^{(h)}(t) + \sum_{k=1}^{p} \pi_k^{(h)} W(t+1-k) \\ &\quad + \sum_{k=p+1}^{p+d} \pi_k^{(h)} W(t+1-k) + u^{(h)}(t+h) \ , \end{aligned} \tag{5.8}$$

$t = d, \ldots, T-h$. Under model (5.3), the actual values of the coefficient matrices $\pi_{p+1}^{(h)}, \ldots, \pi_{p+d}^{(h)}$ are equal to zero ($\pi_{p+1}^{(h)} = \cdots = \pi_{p+d}^{(h)} = 0$), but we shall estimate the (p, h)-autoregressions without imposing any restriction on $\pi_{p+1}^{(h)}, \ldots, \pi_{p+d}^{(h)}$.

Now, suppose the process $\eta(t)$ is either $I(0)$ or $I(1)$, and we take $d = 1$ in (5.8). Then, on replacing p by $p+1$ and setting $j = p$ in the representation (5.6), we see that

$$\begin{aligned} W(t+h) - W(t-p-1) &= \mu^{(h)}(t) + \sum_{k=1}^{p} \pi_k^{(h)} \left[W(t+1-k) - W(t-p-1) \right] \\ &\quad - B_{p+1}^{(h)} W(t-p-1) + u^{(h)}(t+h) \ , \end{aligned} \tag{5.9}$$

where $B_{p+1}^{(h)} = \left(I_m - \sum_{k=1}^{p+1} \pi_k^{(h)} \right)$. In the latter equation, $\pi_1^{(h)}, \ldots, \pi_p^{(h)}$ all affect trend-stationary variables (in an equation where a trend is included along with the other coefficients). Using arguments similar to those of Sims, Stock, and Watson (1990), Park and Phillips (1989) and Dolado and Lütkepohl (1996), it follows that the estimates of $\pi_1^{(h)}, \ldots, \pi_p^{(h)}$ based on estimating (5.9) by ordinary least squares (without restricting $B_{p+1}^{(h)}$) _ or, equivalently, those obtained from (5.8) without restricting $\pi_{p+1}^{(h)}$ _ are asymptotically normal with the same asymptotic covariance matrix as the one obtained for a stationary process of the type studied in section 4.[6] Consequently, the asymptotic distribution of the statistic $\mathcal{W}[H_{j \nrightarrow i}^{(h)}]$ for testing the null hypothesis $H_{j \nrightarrow i}^{(h)}$ of non-causality at horizon h from w_j to w_i $\left(w_j \underset{h}{\nrightarrow} w_i \,|\, I_{(j)} \right)$, based on estimating (5.8),

[6]For related results, see also Choi (1993), Toda and Yamamoto (1995), Yamamoto (1996), Yamada and Toda (1998), and Kurozumi and Yamamoto (2000).

is $\chi^2(p)$. When computing $H^{(h)}_{j\to i}$ as defined in (4.3), it is important that only the coefficients of $\pi^{(h)}_1, \ldots, \pi^{(h)}_p$ are restricted (but not $\pi^{(h)}_{p+1}$).

If the process $\eta(t)$ is integrated up to order d, where $d \geq 0$, we can proceed similarly and add d extra lags to the VAR process studied. Again, the null hypothesis is tested by considering the restrictions entailed on $\pi^{(h)}_1, \ldots, \pi^{(h)}_p$. Further, in view of the fact the test statistics are asymptotically pivotal under the null hypothesis, it is straightforward to apply bootstrap methods to such statistics. Note finally that the precision of the VAR estimates in such augmented regressions may eventually be improved with respect to the OLS estimates considered here by applying bias corrections such as those proposed by Kurozumi and Yamamoto (2000)]. Adapting and applying such corrections to (p, h)-autoregressions would go beyond the scope of the present paper.

6. Empirical illustration

In this section, we present an application of these causality tests at various horizons to macroeconomic time series. The data set considered is the one used by Bernanke and Mihov (1998) in order to study United States monetary policy. The data set considered consists of monthly observations on nonborrowed reserves (NBR, also denoted w_1), the federal funds rate (r, w_2), the GDP deflator (P, w_3) and real GDP (GDP, w_4). The monthly data on GDP and GDP deflator were constructed by state space methods from quarterly observations [see Bernanke and Mihov (1998) for more details]. The sample goes from January 1965 to December 1996 for a total of 384 observations. In what follows, all the variables were first transformed by a logarithmic transformation.

Before performing the causality tests, we must specify the order of the VAR model. First, in order to get apparently stationary time series, all variables were transformed by taking first differences of their logarithms. In particular, for the federal funds rate, this helped to mitigate the effects of a possible break in the series in the years 1979-1981.[7] Starting with 30 lags, we then tested the hypothesis of K lags versus $K+1$

[7]Bernanke and Mihov (1998) performs tests for arbitrary break points, as in Andrews (1993), and don't find significant evidence of a break point. They use a VAR(13) with two additional variables (total bank reserves and Dow-Jones index of spot commodity prices and they normalize both reserves by a 36-month moving average of total reserves.)

Table 27: Rejection frequencies using the asymptotic distribution and the simulated procedure when the true DGP is an i.i.d. Gaussian sequence

a) i.i.d. Gaussian sequence

$h =$	1	2	3	4	5	6	7	8	9	10	11	12
Asymptotic												
5% level	27.0	27.8	32.4	36.1	35.7	42.6	47.9	48.5	51.0	55.7	59.7	63.6
10% level	37.4	39.4	42.2	46.5	47.8	52.0	58.1	59.3	60.3	66.3	69.2	72.5
Bootstrap												
5% level	5.5	5.7	4.7	6.5	4.0	5.1	5.5	3.9	4.7	6.1	5.2	3.8
10% level	10.0	9.1	10.1	10.9	9.6	10.6	10.2	9.4	9.5	10.9	10.3	8.9

b) VAR(16) without causality up to horizon h

Asymptotic												
5% level	24.1	27.9	35.8	37.5	55.9	44.3	52.3	55.9	54.1	60.1	62.6	72.0
10% level	35.5	38.3	46.6	47.2	65.1	55.0	64.7	64.6	64.8	69.8	72.0	79.0
Bootstrap												
5% level	6.0	5.1	3.8	6.1	4.6	4.7	4.4	4.5	4.3	6.3	4.9	5.8
10% level	9.8	8.8	8.7	10.4	10.3	9.9	8.7	7.4	10.3	11.1	9.3	9.7

lags using the LR test presented in Tiao and Box (1981). This led to a VAR(16) model. Tests of a VAR(16) against a VAR(K) for $K = 17, \ldots, 30$ also failed to reject the VAR(16) specification, and the AIC information criterion [see McQuarrie and Tsai (1998, chapter 5)] is minimized as well by this choice. Calculations were performed using the Ox program (version 3.00) working on Linux [see Doornik (1999)].

Vector autoregressions of order p at horizon h were estimated as described in section 4 and the matrix $\hat{V}_{ip}^{(NW)}$, required to obtain covariance matrices, were computed using formula (4.7) with $m(T) - 1 = h - 1$.

On looking at the values of the test statistics and their corresponding p-values at various horizons it quickly becomes evident that the $\chi^2(q)$ asymptotic approximation of the statistic $\mathcal{W}$ in equation (4.6) is very poor. As a simple Monte Carlo experiment, we replaced the data by a 383×4 matrix of random draw from an $N(0, 1)$, ran the same tests and looked at the rejection frequencies over 1000 replications using the asymptotic critical value. The results are in Table 27a. We see important size distortions even for the tests at horizon 1 where there is no moving average part.

We next illustrate that the quality of the asymptotic approximation is even worse

when we move away from an i.i.d. Gaussian setup to a more realistic case. We now take as the DGP the VAR(16) estimated with our data in first difference but we impose that some coefficients are zero such that the federal funds rate does not cause GDP up to horizon h and then we test the $r \underset{h}{\nrightarrow} GDP$ hypothesis. The constraints of non-causality from j to i up to horizon h that we impose are:

$$\hat{\pi}_{ijl} = 0 \quad \text{for} \quad 1 \leq l \leq p\,, \tag{6.1}$$

$$\hat{\pi}_{ikl} = 0 \quad \text{for} \quad 1 \leq l \leq h,\ 1 \leq k \leq m\,. \tag{6.2}$$

Rejection frequencies for this case are given in Table 27b.

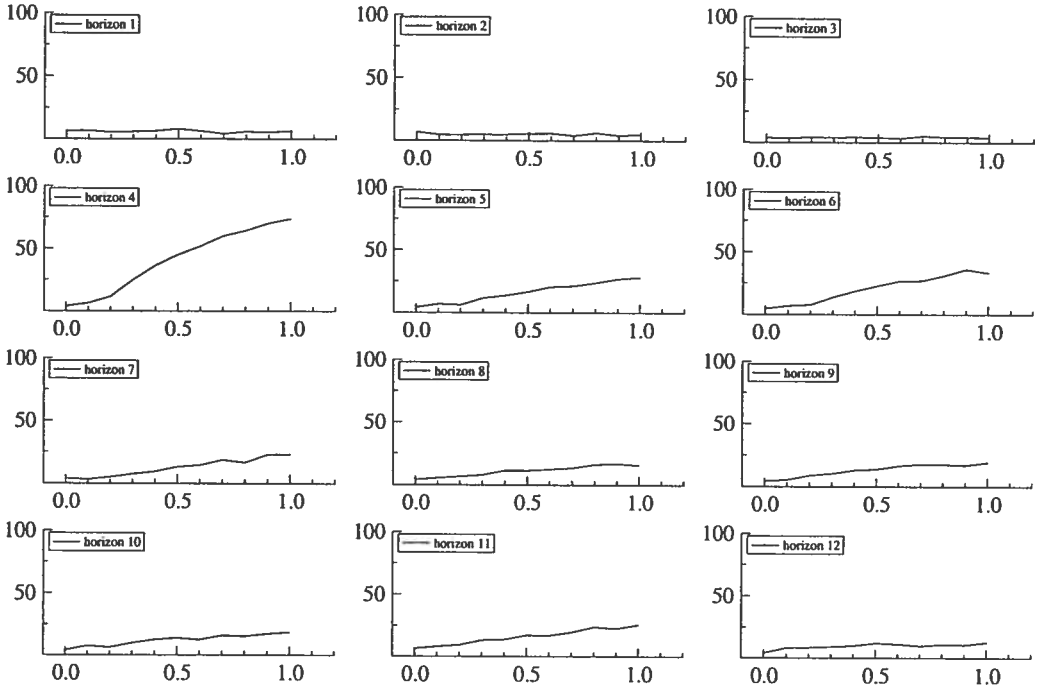
In light of these results we computed the p-values by doing a parametric bootstrap, *i.e.* doing an asymptotic Monte Carlo test based on a consistent point estimate [see Dufour (2002)]. The procedure to test the hypothesis $w_j \underset{h}{\nrightarrow} w_i \mid I_{(j)}$ is the following.

1. An unrestricted VAR(p) model is fitted for the horizon one, yielding the estimates $\hat{\Pi}^{(1)}$ and $\hat{\Omega}$ for $\Pi^{(1)}$ and Ω.
2. An unrestricted (p, h)-autoregression is fitted by least squares, yielding the estimate $\hat{\Pi}^{(h)}$ of $\Pi^{(h)}$.
3. The test statistic $\mathcal{W}$ for testing noncausality at the horizon h from w_j to w_i $\left[H^{(h)}_{j \nrightarrow i} : w_j \underset{h}{\nrightarrow} w_i \mid I_{(j)}\right]$ is computed. We denote by $\mathcal{W}^{(h)}_{j \nrightarrow i}(0)$ the test statistic based on the actual data.
4. N simulated samples from (2.9) are drawn by Monte Carlo methods, using $\Pi^{(h)} = \hat{\Pi}^{(h)}$ and $\Omega = \hat{\Omega}$ [and the hypothesis that $a(t)$ is Gaussian]. We impose the constraints of non-causality, $\hat{\pi}^{(h)}_{ijk} = 0$, $k = 1, \ldots, p$. Estimates of the impulse response coefficients are obtained from $\hat{\Pi}^{(1)}$ through the relations described in equations (2.5) and (2.6). We denote by $\mathcal{W}^{(h)}_{j \nrightarrow i}(n)$ the test statistic for $H^{(h)}_{j \nrightarrow i}$ based on the n-th simulated sample $(1 \leq n \leq N)$.
5. The simulated p-value $\hat{p}_N[\mathcal{W}^{(h)}_{j \nrightarrow i}(0)]$ is obtained, where

$$\hat{p}_N[x] = \left\{1 + \sum_{n=1}^{N} I[\mathcal{W}^{(h)}_{j \nrightarrow i}(n) - x]\right\} / (N+1)\,,$$

$I[z] = 1$ if $z \geq 0$ and $I[z] = 0$ if $z < 0$.

Figure 18: Power of the test at the 5% level for given horizons. The abscissa (x axis) represents the values of θ.



6. The null hypothesis $H_{j \nrightarrow i}^{(h)}$ is rejected at level α if $\hat{p}_N[\mathcal{W}_{j \nrightarrow i}^{(0)}(h)] \leq \alpha$.

From looking at the results in Table 27, we see that we get a much better size control by using this bootstrap procedure. The rejection frequencies over 1000 replications (with $N = 999$) are very close to the nominal size. Although the coefficients ψ_j's are functions of the π_i's we do not constrain them in the bootstrap procedure because there is no direct mapping from $\pi_k^{(h)}$ to π_k and ψ_j. This certainly produces a power loss but the procedure remains valid because the $\hat{\psi}_j$'s are computed with the $\hat{\pi}_k$, which are consistent estimates of the true π_k both under the null and alternative hypothesis. To illustrate that our procedure has power for detecting departure from the null hypothesis of non-causality at a given horizon we ran the following Monte Carlo experiment. We again took a VAR(16) fitted on our data in first differences and we imposed the constraints (6.1) - (6.2) so that there was no causality from r to GDP up to horizon 12 (DGP under the null hypothesis). Next the value of one coefficient previously set to zero was changed to induce causality from r to GDP at horizons 4 and higher:

Table 28: Causality tests and simulated p-values for series in first differences (of logarithm) for the horizons 1 to 12. p-values are reported in parenthesis.

h			1	2	3	4	5	6	7	8	9	10	11	12
NBR	$\nrightarrow$	r	38.5205	26.3851	24.3672	22.5684	24.2294	27.0748	21.4347	17.2164	21.8217	18.4775	18.5379	19.6482
			(0.041)	(0.240)	(0.327)	(0.395)	(0.379)	(0.313)	(0.550)	(0.799)	(0.603)	(0.780)	(0.824)	(0.789)
r	$\nrightarrow$	NBR	22.5390	20.8621	17.2357	17.4222	17.8944	18.6462	25.2059	40.9896	41.6882	38.1656	41.2203	36.1278
			(0.386)	(0.467)	(0.706)	(0.738)	(0.734)	(0.735)	(0.495)	(0.115)	(0.142)	(0.214)	(0.206)	(0.346)
NBR	$\nrightarrow$	P	50.5547	45.2498	49.1408	33.8545	33.8943	35.3923	39.1767	40.0218	36.4089	39.0916	28.4933	26.5439
			(0.004)	(0.014)	(0.009)	(0.121)	(0.162)	(0.157)	(0.099)	(0.141)	(0.234)	(0.202)	(0.485)	(0.609)
P	$\nrightarrow$	NBR	17.0696	18.9504	16.6880	21.2381	28.7264	20.0359	18.2291	23.6704	23.3419	27.8427	30.9171	36.5159
			(0.689)	(0.601)	(0.741)	(0.550)	(0.307)	(0.664)	(0.771)	(0.596)	(0.649)	(0.492)	(0.462)	(0.357)
NBR	$\nrightarrow$	GDP	27.8029	25.0122	25.5123	23.6799	15.0040	17.5748	16.6781	19.6643	29.9020	34.0930	32.3917	34.8183
			(0.184)	(0.302)	(0.294)	(0.393)	(0.830)	(0.757)	(0.790)	(0.746)	(0.368)	(0.300)	(0.370)	(0.357)
GDP	$\nrightarrow$	NBR	17.6338	20.6568	24.5334	18.3220	18.3123	32.8746	33.9979	39.6701	40.7356	26.3424	40.2262	53.3482
			(0.644)	(0.520)	(0.348)	(0.670)	(0.682)	(0.211)	(0.242)	(0.145)	(0.161)	(0.551)	(0.216)	(0.092)
r	$\nrightarrow$	P	32.2481	32.9207	32.0362	25.0124	25.2441	25.8110	29.2553	29.6021	36.0771	43.2116	30.2530	20.8982
			(0.104)	(0.108)	(0.138)	(0.342)	(0.383)	(0.394)	(0.328)	(0.351)	(0.241)	(0.138)	(0.422)	(0.763)
P	$\nrightarrow$	r	22.4385	16.4455	14.3073	14.2932	14.0148	16.7138	11.5599	16.5731	13.0697	14.5759	15.0317	24.4077
			(0.413)	(0.670)	(0.790)	(0.826)	(0.844)	(0.774)	(0.951)	(0.809)	(0.936)	(0.909)	(0.899)	(0.659)
r	$\nrightarrow$	GDP	26.3362	28.6645	35.9768	38.8680	37.6788	39.8145	64.1500	80.2396	89.6998	101.4143	105.9138	110.3551
			(0.262)	(0.159)	(0.073)	(0.059)	(0.082)	(0.079)	(0.006)	(0.002)	(0.002)	(0.002)	(0.001)	(0.003)
GDP	$\nrightarrow$	r	42.5327	43.0078	49.8366	41.1082	41.7945	36.0965	29.5472	25.7898	27.9807	39.6366	39.0477	40.1545
			(0.016)	(0.017)	(0.005)	(0.033)	(0.044)	(0.109)	(0.263)	(0.429)	(0.366)	(0.148)	(0.178)	(0.205)
P	$\nrightarrow$	GDP	20.6903	24.1099	27.4106	23.3585	22.9095	18.5543	20.8172	23.6942	30.5340	28.8286	24.9477	25.7552
			(0.495)	(0.322)	(0.233)	(0.423)	(0.497)	(0.713)	(0.639)	(0.555)	(0.375)	(0.414)	(0.612)	(0.629)
GDP	$\nrightarrow$	P	24.3368	24.4925	24.6125	22.8160	26.7900	40.0825	36.4855	49.6161	46.2574	36.3197	26.8520	24.0113
			(0.329)	(0.365)	(0.380)	(0.470)	(0.348)	(0.081)	(0.157)	(0.058)	(0.072)	(0.262)	(0.540)	(0.666)

Table 29: Causality tests and simulated p-values for series in first differences (of logarithm) for the horizons 13 to 24. p-values are reported in parenthesis.

h			13	14	15	16	17	18	19	20	21	22	23	24
NBR	$\nrightarrow$	r	20.7605	21.1869	18.6062	17.6750	22.2838	34.0098	28.7769	33.7855	66.1538	38.7272	28.8194	34.8015
			(0.783)	(0.771)	(0.852)	(0.905)	(0.811)	(0.532)	(0.696)	(0.632)	(0.143)	(0.558)	(0.803)	(0.708)
r	$\nrightarrow$	NBR	42.2924	43.8196	32.8113	28.1775	31.2326	31.7587	25.6942	34.6680	37.7470	51.4196	33.2897	27.3362
			(0.245)	(0.254)	(0.533)	(0.707)	(0.625)	(0.651)	(0.832)	(0.682)	(0.625)	(0.406)	(0.755)	(0.869)
NBR	$\nrightarrow$	P	38.0451	39.0293	43.2101	37.9049	22.4428	20.0337	38.8919	66.1541	55.0460	64.8568	54.8929	42.2509
			(0.336)	(0.351)	(0.315)	(0.413)	(0.835)	(0.883)	(0.488)	(0.149)	(0.255)	(0.180)	(0.332)	(0.562)
P	$\nrightarrow$	NBR	29.5457	29.4543	27.2425	28.7547	31.4851	44.1254	36.0319	47.1608	50.2743	45.8078	54.1961	56.7865
			(0.565)	(0.667)	(0.688)	(0.703)	(0.627)	(0.400)	(0.581)	(0.427)	(0.379)	(0.486)	(0.401)	(0.399)
NBR	$\nrightarrow$	GDP	30.3906	25.2694	29.7274	23.3347	16.7784	18.7922	26.8942	30.2550	39.0652	43.6661	56.9724	67.7764
			(0.493)	(0.716)	(0.603)	(0.814)	(0.942)	(0.901)	(0.752)	(0.745)	(0.545)	(0.505)	(0.314)	(0.251)
GDP	$\nrightarrow$	NBR	44.2264	41.5795	59.3301	65.7647	53.3579	50.4645	51.5163	43.1132	40.7051	36.3960	38.7162	40.8466
			(0.219)	(0.292)	(0.106)	(0.086)	(0.225)	(0.282)	(0.311)	(0.466)	(0.513)	(0.659)	(0.665)	(0.644)
r	$\nrightarrow$	P	27.3588	37.7170	37.2499	37.1005	27.3052	35.4128	39.2469	53.2675	57.1530	54.6753	69.6377	59.2184
			(0.588)	(0.341)	(0.381)	(0.443)	(0.700)	(0.534)	(0.477)	(0.266)	(0.229)	(0.314)	(0.164)	(0.293)
P	$\nrightarrow$	r	22.3720	26.3588	28.4930	23.0152	26.4211	36.3632	34.3922	17.4269	16.9652	31.9336	28.8377	30.0013
			(0.750)	(0.608)	(0.633)	(0.794)	(0.747)	(0.512)	(0.536)	(0.956)	(0.958)	(0.715)	(0.801)	(0.795)
r	$\nrightarrow$	GDP	123.8280	80.3638	95.0727	83.2773	76.6169	84.1068	76.8979	81.3505	65.5025	71.3334	63.1942	64.8020
			(0.001)	(0.018)	(0.005)	(0.018)	(0.033)	(0.040)	(0.055)	(0.073)	(0.161)	(0.153)	(0.241)	(0.286)
GDP	$\nrightarrow$	r	35.7086	33.0200	40.1713	30.3227	20.5582	19.1144	17.7386	22.6410	38.8175	38.7110	38.9987	25.0577
			(0.319)	(0.420)	(0.273)	(0.545)	(0.838)	(0.898)	(0.930)	(0.856)	(0.515)	(0.565)	(0.549)	(0.860)
P	$\nrightarrow$	GDP	9.4290	9.9870	12.1148	15.1682	14.2883	21.2701	29.0528	47.5841	66.5988	59.2137	71.5165	67.1851
			(0.988)	(0.994)	(0.974)	(0.947)	(0.970)	(0.868)	(0.708)	(0.363)	(0.162)	(0.256)	(0.165)	(0.231)
GDP	$\nrightarrow$	P	29.7642	37.0095	33.4676	42.2190	31.1573	52.3757	51.9567	40.3790	31.6598	54.0684	65.8284	53.0823
			(0.521)	(0.351)	(0.470)	(0.328)	(0.605)	(0.238)	(0.226)	(0.459)	(0.675)	(0.244)	(0.215)	(0.342)

Table 30: Summary of causality relations at various horizons for series in first difference

	h		1	2	3	4	5	6	7	8	9	10	11	12
NBR	$\nrightarrow$	r	**											
r	$\nrightarrow$	NBR												
NBR	$\nrightarrow$	P	**	**	**				*					
P	$\nrightarrow$	NBR												
NBR	$\nrightarrow$	GDP												
GDP	$\nrightarrow$	NBR												*
r	$\nrightarrow$	P												
P	$\nrightarrow$	r												
r	$\nrightarrow$	GDP			*	*	*	*	**	**	**	**	**	**
GDP	$\nrightarrow$	r	**	**	**	**	**							
P	$\nrightarrow$	GDP												
GDP	$\nrightarrow$	P						*		*	*			

	h		13	14	15	16	17	18	19	20	21	22	23	24
NBR	$\nrightarrow$	r												
r	$\nrightarrow$	NBR												
NBR	$\nrightarrow$	P												
P	$\nrightarrow$	NBR												
NBR	$\nrightarrow$	GDP												
GDP	$\nrightarrow$	NBR				*								
r	$\nrightarrow$	P												
P	$\nrightarrow$	r												
r	$\nrightarrow$	GDP	**	**	**	**	**	**	*	*				
GDP	$\nrightarrow$	r												
P	$\nrightarrow$	GDP												
GDP	$\nrightarrow$	P												

Note _ The symbols $\star$ and $\star\star$ indicate rejection of the non-causality hypothesis at the 10% and 5% levels respectively.

$\pi_3(1, 3) = \theta$. As θ increases from zero to one the strength of the causality from r to GDP is higher. Under this setup, we could compute the power of our simulated test procedure to reject the null hypothesis of non-causality at a given horizon. In Figure 18, the power curves are plotted as a function of θ for the various horizons. The level of the tests was controlled through the bootstrap procedure. In this experiment we took again $N = 999$ and we did 1000 simulations. As expected, the power curves are flat at around 5% for horizons one to three since the null is true for these horizons. For horizons four and up we get the expected result that power goes up as θ moves from zero to one, and the power curves gets flatter as we increase the horizon.

Now that we have shown that our procedure does have power we present causality tests at horizon one to 24 for every pair of variables in tables 28 and 29. For every horizon we have twelve causality tests and we group them by pairs. When we say that

a given variable cause or does not cause another, it should be understood that we mean the growth rate of the variables. The p-values are computed by taking $N = 999$. Table 30 summarize the results by presenting the significant results at the 5% and 10% level.

The first thing to notice is that we have significant causality results at short horizons for some pairs of variables while we have it at longer horizons for other pairs. This is an interesting illustration of the concept of causality at horizon h of Dufour and Renault (1998).

The instrument of the central bank, the nonborrowed reserves, cause the federal funds rate at horizon one, the prices at horizon 1, 2, 3 and 9 (10% level). It does not cause the other two variables at any horizon and except the GDP at horizon 12 and 16 (10% level) nothing is causing it. We see that the impact of variations in the nonborrowed reserves is over a very short term. Another variable, the GDP, is also causing the federal funds rates over short horizons (one to five months).

An interesting result is the causality from the federal funds rate to the GDP. Over the first few months the funds rate does not cause GDP, but from horizon 3 (up to 20) we do find significant causality. This result can easily be explained by, e.g. the theory of investment. Notice that we have the following indirect causality. Nonborrowed reserves do not cause GDP directly over any horizon, but they cause the federal funds rate which in turn causes GDP. Concerning the observation that there are very few causality results for long horizons, this may reflect the fact that, for stationary processes, the coefficients of prediction formulas converge to zero as the forecast horizon increases.

Using the results of Proposition 4.5 in Dufour and Renault (1998), we know that for this example the highest horizon that we have to consider is 33 since we have a VAR(16) with four time series. Causality tests for the horizons 25 through 33 were also computed but are not reported. Some p-values smaller or equal to 10% are scattered over horizons 30 to 33 but no discernible pattern emerges.

We next consider extended autoregressions to illustrate the results of section 5. To cover the possibility that the first difference of the logarithm of the four series may not be stationary, we ran extended autoregressions on the series analyzed. Since we used a VAR(16) with non-zero mean for the first difference of the series a VAR(17), *i.e.* $d = 1$, with a non-zero mean was fitted. The Monte Carlo samples with $N = 999$ are drawn

in the same way as before except that the constraints on the VAR parameters at horizon h is $\hat{\pi}_{jik}^{(h)} = 0$ for $k = 1, \ldots, p$ and not $k = 1, \ldots, p + d$.

Results of the extended autoregressions are presented in Table 31 (horizons 1 to 12) and 32 (horizons 13 to 24). Table 33 summarize these results by presenting the significant results at the 5% and 10% level. These results are very similar to the previous ones over all the horizons and variable every pairs. A few causality tests are not significant anymore ($GDP \nrightarrow r$ at horizon 5, $r \nrightarrow GDP$ at horizons 5 and 6) and some causality relations are now significant ($r \nrightarrow P$ at horizon one) but we broadly have the same causality patterns.

7. Conclusion

In this paper, we have proposed a simple linear approach to the problem of testing non-causality hypotheses at various horizons in finite-order vector autoregressive models. The methods described allow for both stationary (or trend-stationary) processes and possibly integrated processes (which may involve unspecified cointegrating relationships), as long as an upper bound is set on the order of integration. Further, we have shown that these can be easily implemented in the context of a four-variable macroeconomic model of the U.S. economy.

Several issues and extensions of interest warrant further study. The methods we have proposed were, on purpose, designed to be relatively simple to implement. This may, of course, involve efficiency losses and leave room for improvement. For example, it seems quite plausible that more efficient tests may be obtained by testing directly the nonlinear causality conditions described in Dufour and Renault (1998) from the parameter estimates of the VAR model. However, such procedures will involve difficult distributional problems and may not be as user-friendly as the procedures described here. Similarly, in nonstationary time series, information about integration order and the cointegrating relationships may yield more powerful procedures, although at the cost of complexity. These issues are the topics of on-going research.

Another limitation comes from the fact we consider VAR models with a known finite order. We should however note that the asymptotic distributional results established

Table 31: Causality tests and simulated p-values for extended autoregressions at the horizons 1 to 12

h			1	2	3	4	5	6	7	8	9	10	11	12
NBR	$\nrightarrow$	r	37.4523	24.0148	24.9365	22.9316	25.4508	26.6763	19.5377	19.3805	21.9278	19.8541	21.2947	19.5569
			(0.051)	(0.337)	(0.309)	(0.410)	(0.333)	(0.333)	(0.660)	(0.688)	(0.609)	(0.716)	(0.710)	(0.795)
r	$\nrightarrow$	NBR	25.9185	17.7977	16.6185	17.3820	19.6425	20.2032	39.8496	44.0172	43.9928	39.4217	36.7802	35.0883
			(0.281)	(0.650)	(0.747)	(0.759)	(0.673)	(0.676)	(0.129)	(0.075)	(0.130)	(0.207)	(0.286)	(0.375)
NBR	$\nrightarrow$	P	50.8648	44.8472	56.5028	33.8112	36.1026	42.1714	38.7204	38.8076	35.3353	33.8049	28.6205	26.9863
			(0.004)	(0.015)	(0.007)	(0.152)	(0.124)	(0.061)	(0.152)	(0.161)	(0.259)	(0.316)	(0.498)	(0.609)
P	$\nrightarrow$	NBR	20.7491	17.2772	16.3891	26.1610	29.7329	18.7583	22.7821	23.1743	27.0076	24.3763	31.1869	38.3349
			(0.506)	(0.704)	(0.764)	(0.339)	(0.276)	(0.730)	(0.605)	(0.629)	(0.508)	(0.655)	(0.484)	(0.336)
NBR	$\nrightarrow$	GDP	27.2102	23.8072	25.9561	24.3384	14.4893	17.8928	15.9036	17.9592	30.7472	33.5333	33.7687	36.3343
			(0.224)	(0.346)	(0.317)	(0.402)	(0.837)	(0.732)	(0.821)	(0.817)	(0.368)	(0.333)	(0.355)	(0.348)
GDP	$\nrightarrow$	NBR	16.1322	19.4471	20.6798	19.1035	29.1229	36.1053	37.1194	40.3578	43.7835	37.7337	48.3004	52.2442
			(0.746)	(0.571)	(0.537)	(0.658)	(0.269)	(0.166)	(0.167)	(0.163)	(0.138)	(0.247)	(0.125)	(0.123)
r	$\nrightarrow$	P	32.5147	31.2909	25.6717	24.4449	21.7640	25.4747	27.6313	31.4581	43.2611	38.2020	30.6394	19.9687
			(0.100)	(0.128)	(0.352)	(0.390)	(0.568)	(0.397)	(0.353)	(0.311)	(0.111)	(0.212)	(0.420)	(0.812)
P	$\nrightarrow$	r	22.7374	15.9453	15.2001	15.1933	16.2334	15.7472	13.4196	15.2506	14.4324	15.8589	21.3637	21.9949
			(0.415)	(0.704)	(0.762)	(0.790)	(0.768)	(0.830)	(0.905)	(0.859)	(0.909)	(0.887)	(0.749)	(0.731)
r	$\nrightarrow$	GDP	27.0435	29.5913	37.5271	35.7130	34.9901	35.1715	79.9402	92.6009	94.9068	107.6638	108.0581	138.0570
			(0.244)	(0.158)	(0.061)	(0.094)	(0.117)	(0.164)	(0.002)	(0.001)	(0.004)	(0.001)	(0.003)	(0.001)
GDP	$\nrightarrow$	r	41.8475	41.9449	44.7597	38.0358	30.4776	28.1840	30.5867	27.0745	26.5876	39.9502	39.8618	33.4855
			(0.032)	(0.019)	(0.019)	(0.060)	(0.209)	(0.286)	(0.250)	(0.369)	(0.431)	(0.157)	(0.181)	(0.347)
P	$\nrightarrow$	GDP	23.7424	26.7148	24.6605	24.6507	23.3233	19.8483	20.3581	28.9318	29.0384	27.2509	26.3318	22.6991
			(0.368)	(0.237)	(0.342)	(0.373)	(0.479)	(0.668)	(0.666)	(0.376)	(0.427)	(0.505)	(0.558)	(0.740)
GDP	$\nrightarrow$	P	25.1264	24.1941	25.5683	19.2127	37.5984	38.0318	37.6254	45.2219	45.2458	36.9912	23.1687	22.9934
			(0.278)	(0.358)	(0.350)	(0.639)	(0.108)	(0.110)	(0.153)	(0.088)	(0.092)	(0.237)	(0.647)	(0.698)

Table 32: Causality tests and simulated p-values for extended autoregressions at the horizons 13 to 24

	h		13	14	15	16	17	18	19	20	21	22	23	24
NBR	$\nrightarrow$	r	19.5845	30.2847	17.4893	24.9745	21.2814	27.4800	36.9567	27.8970	59.3731	34.9153	27.3241	31.9143
			(0.814)	(0.513)	(0.901)	(0.750)	(0.858)	(0.706)	(0.537)	(0.770)	(0.236)	(0.640)	(0.832)	(0.771)
r	$\nrightarrow$	NBR	41.7360	47.8501	30.9613	29.1809	35.7654	22.9966	29.0869	29.1396	38.0675	52.5512	29.6133	25.4872
			(0.282)	(0.212)	(0.592)	(0.692)	(0.540)	(0.847)	(0.757)	(0.766)	(0.665)	(0.401)	(0.838)	(0.917)
NBR	$\nrightarrow$	P	36.7787	37.7357	33.4273	35.3241	20.6330	23.6911	44.5735	55.5420	53.7340	68.0270	52.3332	47.5614
			(0.359)	(0.366)	(0.512)	(0.455)	(0.860)	(0.825)	(0.379)	(0.249)	(0.290)	(0.165)	(0.363)	(0.481)
P	$\nrightarrow$	NBR	29.5049	39.2076	18.0831	30.6486	39.5517	34.8363	39.9608	43.6563	40.6713	44.0254	61.6914	62.9346
			(0.582)	(0.401)	(0.920)	(0.671)	(0.441)	(0.606)	(0.535)	(0.471)	(0.551)	(0.553)	(0.296)	(0.286)
NBR	$\nrightarrow$	GDP	30.9525	22.0737	30.1165	22.7429	17.2546	22.2686	28.6752	31.8817	39.5031	53.6466	58.8413	79.0569
			(0.501)	(0.822)	(0.599)	(0.793)	(0.937)	(0.863)	(0.749)	(0.717)	(0.591)	(0.367)	(0.327)	(0.153)
GDP	$\nrightarrow$	NBR	46.4538	38.0424	64.2269	60.8792	57.5798	57.2237	42.0851	43.1683	43.4379	41.7141	39.8292	41.7123
			(0.186)	(0.347)	(0.071)	(0.125)	(0.169)	(0.205)	(0.453)	(0.491)	(0.501)	(0.568)	(0.631)	(0.632)
r	$\nrightarrow$	P	30.7883	37.3585	30.4331	35.8788	26.9960	39.2961	44.7334	46.6740	56.9680	53.1830	67.3818	61.2522
			(0.503)	(0.364)	(0.566)	(0.448)	(0.712)	(0.429)	(0.358)	(0.352)	(0.223)	(0.310)	(0.182)	(0.241)
P	$\nrightarrow$	r	25.3085	33.1027	28.7362	33.5758	33.8991	39.3031	29.4228	14.3095	17.8588	29.9572	25.0347	29.1903
			(0.654)	(0.450)	(0.624)	(0.546)	(0.525)	(0.451)	(0.707)	(0.984)	(0.957)	(0.764)	(0.883)	(0.825)
r	$\nrightarrow$	GDP	109.5052	84.9556	88.4433	80.9836	79.9549	75.1199	94.8986	65.5929	67.2913	71.6331	75.5123	67.4315
			(0.002)	(0.011)	(0.021)	(0.016)	(0.031)	(0.073)	(0.026)	(0.147)	(0.136)	(0.164)	(0.151)	(0.254)
GDP	$\nrightarrow$	r	34.8185	41.4218	38.2761	28.5326	22.6116	16.7992	20.8097	31.8769	38.7083	34.4663	35.6279	20.5007
			(0.340)	(0.264)	(0.318)	(0.606)	(0.809)	(0.923)	(0.866)	(0.644)	(0.519)	(0.649)	(0.637)	(0.949)
P	$\nrightarrow$	GDP	8.8039	8.9511	17.0182	9.7608	16.4772	19.6942	46.3240	41.7429	64.5928	60.2875	56.0315	72.2823
			(0.995)	(0.995)	(0.933)	(0.995)	(0.923)	(0.916)	(0.349)	(0.484)	(0.207)	(0.250)	(0.332)	(0.215)
GDP	$\nrightarrow$	P	23.8773	39.5163	34.3832	34.5734	36.8350	54.8088	54.8048	36.4102	29.4543	58.0095	58.1341	59.5283
			(0.709)	(0.331)	(0.468)	(0.488)	(0.472)	(0.212)	(0.218)	(0.557)	(0.739)	(0.254)	(0.265)	(0.286)

Table 33: Summary of causality relations at various horizons for series in first difference with extended autoregressions

	h		1	2	3	4	5	6	7	8	9	10	11	12
NBR	$\nrightarrow$	r	$\star$											
r	$\nrightarrow$	NBR								$\star$				
NBR	$\nrightarrow$	P	$\star\star$	$\star\star$	$\star\star$			$\star$						
P	$\nrightarrow$	NBR												
NBR	$\nrightarrow$	GDP												
GDP	$\nrightarrow$	NBR												
r	$\nrightarrow$	P	$\star$											
P	$\nrightarrow$	r												
r	$\nrightarrow$	GDP			$\star$	$\star$			$\star\star$	$\star\star$	$\star\star$	$\star\star$	$\star\star$	$\star\star$
GDP	$\nrightarrow$	r	$\star\star$	$\star\star$	$\star\star$	$\star$								
P	$\nrightarrow$	GDP												
GDP	$\nrightarrow$	P								$\star$	$\star$			

	h		13	14	15	16	17	18	19	20	21	22	23	24
NBR	$\nrightarrow$	r												
r	$\nrightarrow$	NBR												
NBR	$\nrightarrow$	P												
P	$\nrightarrow$	NBR												
NBR	$\nrightarrow$	GDP												
GDP	$\nrightarrow$	NBR			$\star$									
r	$\nrightarrow$	P												
P	$\nrightarrow$	r												
r	$\nrightarrow$	GDP	$\star\star$	$\star\star$	$\star\star$	$\star\star$	$\star\star$	$\star$	$\star\star$					
GDP	$\nrightarrow$	r												
P	$\nrightarrow$	GDP												
GDP	$\nrightarrow$	P												

Note _ The symbols $\star$ and $\star\star$ indicate rejection of the non-causality hypothesis at the 10% and 5% levels respectively.

in this paper continue to hold as long as the order p of the model is selected according to a consistent order selection rule [see Dufour, Ghysels, and Hall (1994), Pötscher (1991)]. So this is not an important restriction. Other problems of interest would consist in deriving similar tests applicable in the context of VARMA or VARIMA models, as well as more general infinite-order vector autoregressive models, using finite-order VAR approximations based on data-dependent truncation rules [such as those used by Lütkepohl and Poskitt (1996c) and Lütkepohl and Saikkonen (1997)]. These problems are also the topics of on-going research.

Chapter 4: Backtesting Value-at-Risk: a duration-based approach[8]

1. Motivation

Financial risk model evaluation or *backtesting* is a key part of the internal model's approach to market risk management as laid out by the Basle Committee on Banking Supervision (1996). However, existing backtesting methods such as those developed in Christoffersen (1998), have relatively small power in realistic small sample settings. Methods suggested in Berkowitz (2001) fare better, but rely on information such as the shape of the left tail of the portfolio return distribution, which is often not available. By far the most common risk measure is Value-at-Risk (VaR), which is defined as a conditional quantile of the return distribution, and it says nothing about the shape of the tail to the left of the quantile.

We will refer to an event where the ex-post portfolio loss exceeds the ex-ante VaR measure as a *violation*. Of particular importance in backtesting is the clustering of violations. An institution's internal risk management team as well as external supervisors explicitly want to be able to detect clustering in violations. Large losses which occur in rapid succession are more likely to lead to disastrous events such as bankruptcy.

In the previous literature, due to the lack of real portfolio data, the evaluation of VaR techniques were largely based on artificial portfolios. Examples in this tradition include Beder (1995), Christoffersen, Hahn, and Inoue (2001), Hendricks (1996), Kupiec (1995), Marshall and Siegel (1997), and Pritsker (1997). But recently, Berkowitz and O'Brien (2002) have reported on the performance of actual VaR forecasts from six large (and anonymous) U.S. commercial banks.[9] Figure 19 reproduces a picture from their paper which shows the VaR exceedences from the six banks reported in standard deviations of the portfolio returns. Even though the banks tend to be conservative–they

[8]This chapter has originally been published undert the title "Backtesting Value-at-Risk: A Duration-Based Approach" in the *Journal of Financial Econometrics*, 2004, volume 2, number 1, pp. 84-108, by permission of Oxford University Press.

[9]Barone-Adesi, Giannopoulos, and Vosper (2002) provides another example using real-life portfolio returns.

have fewer than expected violations–the exceedences are large and appear to be clustered in time and across banks. The majority of violations appear to take place during the August 1998 Russia default and ensuing LTCM debacle. From the perspective of a regulator worried about systemic risk, rejecting a particular bank's risk model due to the clustering of violations is particularly important if the violations also happen to be correlated across banks.

The detection of violation clustering is particularly important because of the widespread reliance on VaRs calculated from the so-called Historical Simulation (HS) technique. In the HS methodology, a sample of historical portfolio returns using current portfolio weights is first constructed. The VaR is then simply calculated as the *unconditional* quantile from the historical sample. The HS method thus largely ignores the last 20 years of academic research on conditional asset return models. Time variability is only captured through the rolling historical sample. In spite of forceful warnings, such as Pritsker (2001), the model-free nature of the HS technique is viewed as a great benefit by many practitioners. The widespread use of HS the technique motivates us to focus attention on backtesting VaRs calculated using this method.

While alternative methods for calculating portfolio measures such as the VaR have been investigated in for example Jorion (2001), and Christoffersen (2003), available methods for backtesting are still relatively few. Our contribution is thus the exploration of a new tool for backtesting based on the duration of days between the violations of the risk metric. The chief insight is that if the one-day-ahead VaR model is correctly specified for coverage rate, p, then, every day, the conditional expected duration until the next violation should be a constant $1/p$ days. We suggest various ways of testing this null hypothesis and we conduct a Monte Carlo analysis which compares the new tests to those currently available. Our results show that in many realistic situations, the duration based tests have better power properties than the previously suggested tests. The size of the tests is easily controlled using the Monte Carlo testing approach of Dufour (2002). This procedure is described in detail below.

We hasten to add that the sort of omnibus backtesting procedures suggested here are meant as complements to–and not substitutes for–the statistical diagnostic tests carried out on various aspects of the risk model in the model estimation stage. The

tests suggested in this paper can be viewed either as a final diagnostic for an internal model builder or alternatively as a feasible diagnostic for an external model evaluator for whom only limited, aggregate portfolio information is available.

Our paper is structured as follows: Section 2 outlines the previous first-order Markov tests, Section 3 suggests the new duration-based tests, and Section 4 discusses details related to the implementation of the tests. Section 5 contains Monte Carlo evidence on the performance of the tests. Section 6 considers backtesting of tail density forecasts, and Section 7 concludes.

2. Extant Procedures for Backtesting Value-at-Risk

Consider a time series of daily ex-post portfolio returns, R_t, and a corresponding time series of ex-ante Value-at-Risk forecasts, $VaR_t(p)$ with promised coverage rate p, such that ideally $\Pr_{t-1}\left(R_t < -VaR_t(p)\right) = p$. The negative sign arises from the convention of reporting the VaR as a positive number.

Define the hit sequence of VaR_t violations as

$$I_t = \begin{cases} 1, \text{ if } R_t < -VaR_t\left(p\right) \\ 0, \text{ else} \end{cases} . \tag{2.1}$$

Notice that the hit sequence appears to discard a large amount of information regarding the size of violations etc. Recall, however, that the VaR forecast does not promise violations of a certain magnitude, but rather only their conditional frequency, i.e. p. This is a major drawback of the VaR risk measure which we will discuss in Section 6.

Christoffersen (1998) tests the null hypothesis that

$$I_t \sim i.i.d.\ Bernoulli(p)$$

against the alternative that

$$I_t \sim i.i.d.\ Bernoulli(\pi)$$

and refers to this as the test of correct unconditional coverage (uc)

$$H_{0,uc} : \pi = p \tag{2.2}$$

which is a test that on average the coverage is correct. The above test implicitly assumes that the hits are independent an assumption which we now test explicitly. In order to test this hypothesis an alternative is defined where the hit sequence follows a first order Markov sequence with switching probability matrix

$$\Pi = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix} \tag{2.3}$$

where π_{ij} is the probability of an i on day $t-1$ being followed by a j on day t. The test of independence (ind) is then

$$H_{0,ind} : \pi_{01} = \pi_{11}. \tag{2.4}$$

Finally one can combine the two tests in a test of conditional coverage (cc)

$$H_{0,cc} : \pi_{01} = \pi_{11} = p \tag{2.5}$$

The idea behind the Markov alternative is that clustered violations represent a signal of risk model misspecification. Violation clustering is important as it implies repeated severe capital losses to the institution which together could result in bankruptcy.

Notice however, that the Markov first-order alternative may have limited power against general forms of clustering. The first point of this paper is to establish more general tests for clustering which nevertheless only rely on information in the hit sequence. Throughout the paper we implicitly assume that the VaR is for a one-day horizon. To apply this backtesting framework to an horizon of more than one day, we would have to use non-overlapping observations.[10]

[10]We implicitly assume that we observe the return process as least as frequently as we compute the VaR.

3. Duration-Based Tests of Independence

The above tests are reasonably good at catching misspecified risk models when the temporal dependence in the hit-sequence is of a simple first-order Markov structure. However we are interested in developing tests which have power against more general forms of dependence but which still rely on estimating only a few parameters.

The intuition behind the duration-based tests suggested below is that the clustering of violations will result in an excessive number of relatively short and relatively long no-hit durations, corresponding to market turbulence and market calm respectively. Motivated by this intuition we consider the duration of time (in days) between two VaR violations (i.e. the no-hit duration) as

$$D_i = t_i - t_{i-1} \tag{3.6}$$

where t_i denotes the day of violation number i.[11]

Under the null hypothesis that the risk model is correctly specified, the no-hit duration should have no memory and a mean duration of $1/p$ days. To verify the no memory property note that under the null hypothesis we have the discrete probability distribution

$$\begin{aligned}
\Pr(D=1) &= p \\
\Pr(D=2) &= (1-p)\,p \\
\Pr(D=2) &= (1-p)^2\,p \\
&\cdots \\
\Pr(D=d) &= (1-p)^{d-1}\,p.
\end{aligned}$$

A duration distribution is often best understood by its hazard function, which has the intuitive definition of the probability of a getting a violation on day D after we have gone $D-1$ days without a violation. The above probability distribution implies a flat

[11] For a general introduction to duration modeling, see Kiefer (1988) and Gouriéroux (2000).

discrete hazard function as the following derivation shows

$$\begin{aligned}\lambda(d) &= \frac{\Pr(D=d)}{1-\sum_{j<d}\Pr(D=j)} \\ &= \frac{(1-p)^{d-1}p}{1-\sum_{j=0}^{d-2}(1-p)^{j}p} \\ &= p.\end{aligned}$$

The only memory free (continuous)[12] random distribution is the exponential, thus we have that under the null the distribution of the no-hit durations should be

$$f_{\exp}(D;p) = p\exp(-pD). \tag{3.7}$$

In order to establish a statistical test for independence we must specify a (parsimonious) alternative which allows for duration dependence. As a very simple case, consider the Weibull distribution where

$$f_W(D;a,b) = a^b bD^{b-1}\exp\left(-(aD)^b\right). \tag{3.8}$$

The Weibull distribution has the advantage that the hazard function has a closed form representation, namely

$$\lambda_W(D) \equiv \frac{f_W(D)}{1-F_W(D)} = a^b bD^{b-1} \tag{3.9}$$

where the exponential distribution appears as a special case with a flat hazard, when $b=1$. The Weibull will have a decreasing hazard function when $b<1$, which corresponds to an excessive number of very short durations (very volatile periods) and an excessive number of very long durations (very tranquil periods). This could be evidence of misspecified volatility dynamics in the risk model.

Due to the bankruptcy threat from VaR violation clustering the null hypothesis of independence is of particular interest. We therefore want to explicitly test the null

[12]Notice that we use a continuous distribution even though we are counting time in days. This discreteness bias will be acounted for in the Monte Carlo tests. The exponential distribution can also be viewed as the continuous time limit of the above discrete time process. See Poirier (1995).

hypothesis

$$H_{0,ind} : b = 1. \tag{3.10}$$

We could also use the Gamma distribution under the alternative hypothesis. The p.d.f. in this case is

$$f_{\Gamma}(D; a, b) = \frac{a^b D^{b-1} \exp(-aD)}{\Gamma(b)} \tag{3.11}$$

which also nests the exponential when $b = 1$. In this case we therefore also have the independence test null hypothesis as

$$H_{0,ind} : b = 1. \tag{3.12}$$

The Gamma distribution does not have a closed-form solution for the hazard function, but the first two moments are $\frac{b}{a}$ and $\frac{b}{a^2}$ respectively, so the notion of excess dispersion which is defined as the variance over the squared expected value is simply $\frac{1}{b}$. Note that the average duration in the exponential distribution is $1/p$, and the variance of durations is $1/p^2$, thus the notion of excess dispersion is 1 in the exponential distribution.

The above duration tests can potentially capture higher order dependence in the hit sequence by simply testing the unconditional distribution of the durations. Dependence in the hit sequence may show up as an excess of relatively long no-hit durations (quiet periods) and an excess of relatively short no-hit durations, corresponding to violation clustering. However, in the above tests, any information in the ordering of the durations is completely lost. The information in the temporal ordering of no-hit durations could be captured using the framework of Engle and Russel (1998)'s Exponential Autoregressive Conditional Duration (EACD) model. In the EACD(1,0) model, the conditional expected duration takes the following form

$$E_{i-1}[D_i] \equiv \psi_i = \omega + \alpha D_{i-1} \tag{3.13}$$

with $\alpha \in [0, 1)$. Assuming an underlying exponential density with mean equal to one,

the conditional distribution of the duration is

$$f_{EACD}\left(D_i|\psi_i\right) = \frac{1}{\psi_i}\exp\left(-\frac{D_i}{\psi_i}\right). \tag{3.14}$$

The null of independent no-hit durations would then correspond to

$$H_{0,ind} : \alpha = 0. \tag{3.15}$$

Excess dispersion in the EACD(1,0) model is defined as

$$V[D_i]/E[D_i]^2 = \frac{1}{1-2\alpha^2} \tag{3.16}$$

so that the ratio of the standard deviation to the mean duration is above one if $\alpha > 0$.

In our test specifications, the information set only contains past durations, but it could be extended to include all the conditioning information used to compute the VaR for example. This would translate into adding variables other than D_{i-1} into the right-hand side of equation (3.13).

4. Test Implementation

We will first discuss the specific implementation of the hit sequence tests suggested above. Later, we will simulate observations from a realistic portfolio return process and calculate risk measures from the popular Historical Simulation risk model, which in turn provides us with hit sequences for testing.

4.1. Implementing the Markov Tests

The likelihood function for a sample of T i.i.d. observations from a Bernoulli variable, I_t, with known probability p is written as

$$L\left(I,p\right) = p^{T_1}\left(1-p\right)^{T-T_1} \tag{4.17}$$

where T_1 is the number of ones in the sample. The likelihood function for an i.i.d. Bernoulli with unknown probability parameter, π_1, to be estimated is

$$L(I, \pi_1) = \pi_1^{T_1} (1 - \pi_1)^{T - T_1}. \tag{4.18}$$

The ML estimate of π_1 is

$$\hat{\pi}_1 = T_1/T \tag{4.19}$$

and we can thus write a likelihood ratio test of unconditional coverage as

$$LR_{uc} = -2 (\ln L(I, \hat{\pi}_1) - \ln L(I, p)). \tag{4.20}$$

For the independence test, the likelihood under the alternative hypothesis is

$$L(I, \pi_{01}, \pi_{11}) = (1 - \pi_{01})^{T_0 - T_{01}} \pi_{01}^{T_{01}} (1 - \pi_{11})^{T_1 - T_{11}} \pi_{11}^{T_{11}} \tag{4.21}$$

where T_{ij} denotes the number of observations with a j following an i. The ML estimates are

$$\hat{\pi}_{01} = T_{01}/T_0 \tag{4.22}$$

$$\hat{\pi}_{11} = T_{11}/T_1 \tag{4.23}$$

and the independence test statistic is

$$LR_{ind} = 2 (\ln L(I, \hat{\pi}_{01}, \hat{\pi}_{11}) - \ln L(I, \hat{\pi}_1)). \tag{4.24}$$

Finally the test of conditional coverage is written as

$$LR_{cc} = 2 (\ln L(I, \hat{\pi}_{01}, \hat{\pi}_{11}) - \ln L(I, p)). \tag{4.25}$$

We note that all the tests are carried out conditioning on the first observation. The tests are asymptotically distributed as χ^2 with degree of freedom one for the uc and ind tests and two for the cc test. But we will rely on finite sample p-values below.

Finally, as a practical matter, if the sample at hand has $T_{11} = 0$, which can easily happen in small samples and with small coverage rates, then we calculate the first-order Markov likelihood as

$$L(I, \pi_{01}, \pi_{11}) = (1 - \pi_{01})^{T_0 - T_{01}} \pi_{01}^{T_{01}} \tag{4.26}$$

and carry out the tests as above.

4.2. Implementing the Weibull and EACD Tests

In order to implement our tests based on the duration between violations we first need to transform the hit sequence into a duration series D_i. While doing this transformation we also create the series C_i to indicate if a duration is censored ($C_i = 1$) or not ($C_i = 0$). Except for the first and last duration the procedure is straightforward, we just count the number of days between each violation and set $C_i = 0$. For the first observation if the hit sequence starts with 0 then D_1 is the number of days until we get the first hit. Accordingly $C_1 = 1$ because the observed duration is left-censored. If instead the hit sequence starts with a 1 then D_1 is simply the number of days until the second hit and $C_1 = 0$.

The procedure is similar for the last duration. If the last observation of the hit sequence is 0 then the last duration, $D_{N(T)}$, is the number of days after the last 1 in the hit sequence and $C_{N(T)} = 1$ because the spell is right-censored. In the same manner if the last observation of the hit sequence is a 1 then $D_{N(T)} = t_{N(T)} - t_{N(T)-1}$ and $C_{N(T)} = 0$.

The contribution to the likelihood of an uncensored observation is its corresponding p.d.f. For a censored observation, we merely know that the process lasted at least D_1 or $D_{N(T)}$ days so the contribution to the likelihood is not the p.d.f. but its survival function $S(D_i) = 1 - F(D_i)$. Combining the censored and uncensored observations, the log-likelihood is

$$\begin{aligned} \ln L(D;\Theta) &= C_1 \ln S(D_1) + (1 - C_1) \ln f(D_1) + \sum_{i=2}^{N(T)-1} \ln(f(D_i)) \\ &\quad + C_{N(T)} \ln S(D_{N(T)}) + (1 - C_{N(T)}) \ln f(D_{N(T)}). \end{aligned} \tag{4.27}$$

Once the durations are computed and the truncations taken care of, then the likelihood ratio tests can be calculated in a straightforward fashion. The only added complication is that the ML estimates are no longer available in closed form, they must be found using numerical optimization.[13] For the unrestricted EACD likelihood this implies maximizing simultaneously over two parameters, α and ω. For the unrestricted Weibull likelihood, we only have to numerically maximize it over one parameter since for a given value of b, the first order condition with respect to a as an explicit solution:[14]

$$\hat{a} = \left(\frac{N(T) - C_1 - C_{N(T)}}{\sum_{i=1}^{N(T)} D_i^b} \right)^{1/b} . \tag{4.28}$$

4.3. Finite Sample Inference

While the large-sample distributions of the likelihood ratio tests we have suggested above are well-known,[15] they may not lead to reliable inference in realistic risk management settings. The nominal sample sizes can be reasonably large, say two to four years of daily data, but the scarcity of violations of for example the 1% VaR renders the effective sample size small. In this section, we therefore introduce the Dufour (2002) Monte Carlo testing technique.

For the case of a continuous test statistic, the procedure is the following. We first generate N independent realizations of the test statistic, LR_i, $i = 1, \ldots, N$. We denote by LR_0 the test computed with the original sample. Under the hypothesis that the risk model is correct we know that the hit sequence is i.i.d. Bernoulli with the mean equal to the coverage rate in our application. We thus benefit from the advantage of not having nuisance parameters under the null hypothesis.

We next rank LR_i, $i = 0, \ldots, N$ in non-decreasing order and obtain the Monte

[13] We have also investigated LM tests which require less numerical optimization than do LR tests. However, in finite sample simulations we found that the power in the LM tests were lower than in the LR tests, thus we only report LR results below.

[14] For numerical stability, we recommend working with a^b instead of a, since b can take values close to zero.

[15] Testing $\alpha = 0$ in the EACD(1,0) model presents a potential difficulty asymptotically in that it is on the boundary of the parameter space. However, the MC method we apply is valid even in this case. See Andrews (2001) for more details.

Carlo p-value $\hat{p}_N(LR_0)$ where

$$\hat{p}_N(LR_0) = \frac{N\hat{G}_N(LR_0)+1}{N+1} \tag{4.29}$$

with

$$\hat{G}_N(LR_0) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(LR_i > LR_0\right) \tag{4.30}$$

where $\mathbf{1}(*)$ takes on the value 1 if $*$ is true and the value 0 otherwise.

When working with binary sequences the test values can only take a countable number of distinct values. Therefore, we need a rule to break ties between the test value obtained from the sample and those obtained from Monte Carlo simulation under the null hypothesis. The tie-breaking procedure is as follows: For each test statistic, LR_i, $i = 0, \ldots, N$, we draw an independent realization of a Uniform distribution on the $[0; 1]$ interval. Denote these draws by U_i, $i = 0, \ldots, N$. The Monte-Carlo p-value is now given by

$$\tilde{p}_N(LR_0) = \frac{N\tilde{G}_N(LR_0)+1}{N+1} \tag{4.31}$$

with

$$\tilde{G}_N(LR_0) = 1-\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(LR_i < LR_0\right)+\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(LR_i = LR_0\right)\mathbf{1}\left(U_i \geq U_0\right). \tag{4.32}$$

There are two additional advantages of using a simulation procedure. The first is that possible systematic biases arising from the use of continuous distributions to study discrete processes are accounted for. They will appear both in LR_0 and LR_i. The second is that Monte-Carlo testing procedures are consistent even if the parameter value is on the boundary of the parameter space. Bootstrap procedures on the other hand could be inconsistent in this case.

5. Backtesting VaRs from Historical Simulation

We now assess the power of the proposed duration tests in the context of a Monte Carlo study. Consider a portfolio where the returns are drawn from a GARCH(1,1)-t(d) model

with an asymmetric leverage effect, that is

$$\begin{aligned} R_{t+1} &= \sigma_{t+1}\sqrt{((d-2)/d)}z_{t+1}, \text{ with} \\ \sigma^2_{t+1} &= \omega + \alpha\sigma^2_t\left(\sqrt{((d-2)/d)}z_t - \theta\right)^2 + \beta\sigma^2_t \end{aligned}$$

where the innovation z_{t+1}s are drawn independently from a Student's t (d) distribution. Notice that the innovations have been rescaled to ensure that the conditional variance of return will be σ^2_{t+1}.

In the simulations below we choose the following parameterization

$$\begin{aligned} \alpha &= 0.1 \\ \theta &= 0.5 \\ \beta &= 0.85 \\ \omega &= 3.9683e-6 \\ d &= 8 \end{aligned}$$

where ω is set to target an annual standard deviation of 0.20. The parameters imply a daily volatility persistence of 0.975, a mean of zero, a conditional skewness of zero, and a conditional (excess) kurtosis of 1.5. This particular DGP is constructed to form a realistic representation of an equity portfolio return distribution.[16]

The risk measurement method under study is the popular Historical Simulation (HS) technique. It takes the VaR on a certain day to be simply the unconditional quantile of the past T_e daily observations. Specifically

$$VaR^p_{t+1} = -Percentile(\{R_\tau\}^t_{\tau=t-T_e+1}, 100p).$$

From the return sample and the above VaR, we are implicitly assuming that \$1 is invested each day. Equivalently, the VaR can be interpreted as being calculated in percent of the portfolio value.

In practice, the sample size is often determined by practical considerations such as

[16]The parameter values are similar to estimates of this GARCH model on daily S&P500 returns (not reported here), and to estimates on daily FX returns published in Bollerslev (1987).

the amount of effort involved in valuing the current portfolio holdings using past prices on the underlying securities. For the purposes of this Monte Carlo experiment, we set $T_e = 250$ or $T_e = 500$ corresponding to roughly one or two years of trading days.

In practice the VaR coverage rate, p, is typically chosen to be either 1% or 5%, and below we assess the power to reject the HS model using either of those rates. Figure 20 shows a return sample path from the above GARCH-t(d) process along with the 1% and 5% VaRs from the HS model (with $T_e = 500$). Notice the peculiar step-shaped VaRs resulting from the HS method. Notice also the infrequent changes in the 1% VaR.[17]

The 1% VaR exceedences from the return sample path are shown in Figure 21 reported in daily standard deviations of returns. The simulated data in Figure 21 can thus be compared with the real-life data in Figure 19, which was taken from Berkowitz and O'Brien (2002). Notice that the simulated data shares the stylized features with the real-life data in Figure 19. [18]

Before calculating actual finite sample power in the suggested tests we want to give a sense of the appropriateness of the duration dependence alternative. To this end we simulate one very long realization (5 million observations) of the GARCH return process and calculate 1% and 5% VaRs from Historical Simulation with a rolling set of 500 in-sample returns. The zero-one hit sequence is then calculated from the ex-post daily returns and the ex-ante VaRs, and the sequence of durations between violations is calculated from the hit sequence. From this duration sequence we fit a Weibull distribution and calculate the hazard function from it. We also estimate nonparametrically the empirical hazard function of the simulated durations via the Kaplan-Meier product-limit estimator of the survival function [see Kiefer (1988)]. These Weibull and empirical hazards are estimated over intervals of 10 days so if there is a probability p of getting a hit at each day then the probability that a given duration will last 10 days or less is

$$\sum_{i=1}^{10} \Pr(D = i) \quad = \quad \sum_{i=1}^{10} (1-p)^{i-1} p$$

[17] When $T_e = 250$ and $p = 1\%$, the VaR is calculated as the simple average between the second and third lowest return.

[18] Note that we have simulated 1,000 observations in Figure 21, while Figure 19 contains between 550 and 750 observations per bank.

$$= 1-(1-p)^{10}.$$

For p equal to 1% and 5% we get a constant hazard of 0.0956 and 0.4013 respectively over a 10-day interval.

We see in Figure 22 that the hazards are distinctly downward sloping which corresponds to positive duration dependence. The relevant flat hazard corresponding to i.i.d. violations is superimposed for comparison. Figure 22 also shows that the GARCH and the Weibull hazards are reasonably close together which suggests that the Weibull distribution offers a useful alternative hypothesis in this type of tests.

Figure 23 shows the duration dependence via simple histograms of the duration between the violations from the Historical Simulation VaRs. The top panel again shows the 1% VaR and the bottom panel shows the 5% VaR.

Data and other resource constraints often force risk managers to backtest their models on relatively short backtesting samples. We therefore conduct our power experiment with samples sizes from 250 to 1,500 days in increments of 250 days. Thus our backtesting samples correspond to approximately one through six years of daily returns.

Below we simulate GARCH returns, calculate HS VaR and the various tests in 5,000 Monte Carlo replications. We present three types of results. We first present the raw power results, which are simply calculated as the frequency of rejections of the null hypothesis in the simulation samples for which we can perform the tests. In order to compute the p-values of the tests we simulate $N = 9999$ hit sequence samples under the null hypothesis that the sequences are distributed i.i.d. Bernoulli(p).

In the simulations, we reject the samples for which we cannot compute the tests. For example, to compute the independence test with the Markov model, we need at least one violation otherwise the LR test is equal to zero when we calculate the likelihood from equation (4.26). Similarly, we need at least one non-censored duration and an additional possibly censored duration to perform the Weibull[19] and EACD independence tests. This of course constitutes a nontrivial sample selection rule for the smallest sample sizes and the 1% VaR coverage rate in particular. We therefore also present the sample selection frequency, i.e. the fraction of simulated samples for which we can compute

[19]The likelihood of the Weibull distribution can be unbounded when we have only one uncensored observation. When this happens we discard the sample.

each test. Finally we report effective power, which corresponds to multiplying the raw power by the sample selection frequency.

The results of the Monte Carlo simulations are presented in Tables 1 through 6. We report the empirical rejection frequencies (power) for the Markov, Weibull and EACD independence tests for various significance test levels, VaR coverage rates, and backtesting sample sizes. Table 1 reports power for a Historical Simulation risk model with $T_e = 500$ observations in the rolling estimation samples. Table 2 gives the sample selection frequencies, that is, the fraction of samples drawn which were possible to to use for calculating the tests. Table 3 reports effective power which is simply the power entries from Table 1 multiplied by the relevant sample selection frequency in Table 2. Tables 4 through 6 shows the results when the rolling samples for VaR calculation contains $T_e = 250$ observations. Notice that we focus solely on the independence tests here because the historical simulation risk models under study are correctly specified unconditionally.

The results are quite striking. The main result in Table 1 is that for inference samples of 750 days and above the Weibull test is always more powerful than the Markov and EACD tests in rejecting the HS risk models. This result holds across inference sample sizes, VaR coverage rates and significance levels chosen. The differences in power are sometimes very large. For example in Table 1 using a 1% significance level, the 5% VaR in a sample of 1,250 observations has a Weibull rejection frequency of 69.2% and a Markov rejection frequency of only 39.5%. The Weibull test clearly appears to pick up dependence in the hit violations which is ignored by the Markov test.

For an inference sample size 500 the ranking of tests depends on the inference sample size, VaR coverage rate and significance level in question. Typically either the Markov or the EACD test performs the best.

For an inference sample size of 250, the power is typically very low in any of the three tests. This is a serious issue as the backtesting guide for market risk capital requirements uses a sample size of one year when assessing model adequacy.[20] The EACD test is often the most powerful in the case of 250 inference observations, which is curious as the performance of the EACD test is quite sporadic for larger sample sizes.

[20]We thank an anonymous referee for pointing out this important issue.

Generally, the EACD appears to do quite well at smaller sample sizes but relatively poorly at larger sample sizes. We suspect that the nonlinear estimate of the α parameter is poorly behaved in this application.

Table 2 shows the sample selection frequencies corresponding to the power calculations in Table 1. As expected the sample rejection issue is the most serious for inference samples of 250 observations. For inference samples of 500 and above virtually no samples are rejected.

Table 3 reports the effective power calculated as the power in Table 1 multiplied by the relevant sample selection frequency in Table 2. Comparing Tables 1 and 3 it is clear that test which has the highest power in any given case in Table 1 also has the highest power in Table 3. But the levels of power are of course lower in Table 3 compared with Table 1 but only dramatically so for inference samples of 250 observations.

Tables 4 shows the power calculations for the case when the VaR is calculated on 250 in-sample observations rather than 500 as was the case in Tables 1 through 3. The overall picture from Table 1 emerges again: The Weibull test is always best for inference samples of 750 observations and above. For samples of 500 the rankings vary case by case and for 250 observations, the power is generally very low.

Table 5 reports the sample selection frequencies corresponding to Table 4. In this case the sample selection frequencies are even higher than in Table 2. For a VaR coverage rate of 5% the rejection frequencies are negligible for all sample sizes.

Table 6 shows the effective power from Table 4. Again we simply multiply the power in Table 4 with the sample selection frequency in Table 5. Notice again that the most powerful test in Table 4 is also the most powerful test in Table 6. Notice also that for most entries the power numbers in Table 6 are very similar to those in Table 4.

Comparing numbers across Tables 1 and 4 and across Tables 3 and 6, we note that the HS VaR with $T_e = 500$ rolling sample observations often has a higher rejection frequency than the HS VaR with $T_e = 250$ rolling sample observations. This result is interesting because practitioners often work very hard to expand their data bases enabling them to increase their rolling estimation sample period. Our results suggest that such efforts may be misguided because lengthening the size of the rolling sample does not necessarily eliminate the distributional problems with Historical Simulation.

6. Backtesting Tail Density Forecasts

The choice of Value-at-Risk as a portfolio risk measure can be criticized on several fronts. Most importantly, the quantile nature of the VaR implies that the shape of the return distribution to the left of the VaR is ignored. Particularly in portfolios with highly nonlinear distributions, such as those including options, this shortcoming can be crucial. Theoreticians have criticized the VaR measure both from a utility-theoretic perspective (Artzner, Delbaen, Eber, and Heath (1999) and from a dynamic trading perspective [Basak and Shapiro (2000). Although some of these criticisms have recently been challenged Cuoco, He, and Issaenko (2001)], it is safe to say that risk managers ought to be interested in knowing the entire distribution of returns, and in particular the left tail. Backtesting distributions rather than VaRs then becomes important.

Consider the standard density forecast evaluation approach[21] of calculating the uniform transform variable

$$U_t = F_t(R_t)$$

where $F_t(*)$ is the a priori density forecast for time t. The null hypothesis that the density forecast is optimal corresponds to

$$U_t \sim i.i.d.\ Uniform(0,1).$$

Berkowitz (2001) argues that the bounded support of the uniform variable renders standard inference difficult. One is forced to rely on nonparametric tests which have notoriously poor small sample properties. He suggests a simple transformation using the inverse normal c.d.f.

$$Z_t = \Phi^{-1}(U_t)$$

after which the hypothesis

$$Z_t \sim i.i.d.\ Normal(0,1)$$

can easily be tested.

[21] See for example Diebold, Gunther, and Tay (1998).

Berkowitz further argues that confining attention to the left tail of the distribution has particular merit in the backtesting of risk models where the left tail contains the largest losses that are most likely to impose bankruptcy risk. He defines the censored variable

$$Z_t^* = \begin{cases} Z_t, \text{ if } R_t < VaR_t \\ \Phi^{-1}\left(VaR_t\right), \text{ else} \end{cases}$$

and tests the null that

$$Z_t^* \sim Censored\ Normal(0, 1, VaR_t).$$

We note first that Berkowitz (2001) only tests the unconditional distribution of Z_t^*. The information in the potential clustering of the VaR exceedences is ignored.

Second, note that the censored variable complication is not needed. If we want to test that the transforms of the $100p$ percent largest losses are themselves uniform, then we can simply multiply the subset of the uniform by $1/p$, apply the transformation and test for standard normality again.[22] That is

$$U_i^{**} = \begin{cases} U_t/p, \text{ if } R_t < VaR_t \\ \text{Else not defined} \end{cases}$$

We then have that

$$Z_i^{**} = \Phi^{-1}\left(U_i^{**}\right) \sim i.i.d.\ Normal(0, 1).$$

Note that due to the censoring there is no notion of time in the sequence Z_i^{**}. We might want to make a joint analysis of both Z_i^{**} and the duration between violations D_i. To do this we would like to write a joint density for these two processes under the alternative. We know that under the null hypothesis that the risk model is correctly specified the Z_i^{**} should be i.i.d. $N(0, 1)$, D_i should be i.i.d. exponential with mean $1/p$, and the processes should be independent. The question is how to write a joint density for these two processes as the alternative hypothesis knowing that, for example,

[22] We are grateful to Nour Meddahi for pointing this out.

the marginal p.d.f. of D_i is a Weibull and some other p.d.f. for Z_i^{**}? Copulas provide a useful tool for doing so.

A (bivariate) copula is a function C from $[0;1] \times [0;1]$ to $[0;1]$ with the following properties:

1. For every u, v in $[0;1]$,

$$C(u,0) = 0 = C(0,v)$$

and

$$C(u,1) = u \quad \text{and} \quad C(1,v) = v.$$

2. For every u_1, u_2, v_1, v_2 in $[0;1]$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$,

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0.$$

In order to explain how copulas can be used we apply Sklar's theorem [Nelsen (1998)], which states: Let H be a joint distribution function with margins F and G. Then there exists a copula C such that for all x, y in $\mathbb{R}$,

$$H(x,y) = C(F(x), G(y)).$$

If F and G are continuous then C is unique. Conversely, if C is a copula and F and G are distribution functions then H is a joint distribution function with marginal densities F and G.

So if we have two densities under the alternative (e.g. $f(D_i)$ and $g(Z_i^{**})$) then we can easily construct a joint density by applying a copula. Suppose the considered bivariate copula $C(u,v;\theta)$ is a function of a unique parameter θ and that we have $C(u,v;\theta_0) = uv$ and $C(u,v;\theta) \neq uv$ for $\theta \neq \theta_0$. This gives us a basis for a test because $C(F(x), G(y); \theta_0) = F(x)G(y)$ means that x and y are independent.

An example of such a copula is the *Ali-Mikhail-Haq* family of copulas where

$$C(u,v;\theta) = \frac{uv}{1-\theta(1-u)(1-v)}; \qquad \theta \in [-1,1]$$

and we have $C(u, v; \theta) = uv$ if $\theta = 0$. A possible alternative hypothesis could be that D_i is i.i.d. Weibull(a, b), Z_i^{**} is i.i.d. $N(\mu, \sigma^2)$ and $C(u, v; \theta)$ is from the *Ali-Mikhail-Haq* family of copulas. We could then test

$$\begin{aligned} H_0 &: \quad a = p,\ b = 1,\ \mu = 0,\ \sigma = 1,\ \theta = 0 \\ H_1 &: \quad \text{at least one of these equalities does not hold} \end{aligned}$$

in a likelihood ratio framework similar to the one considered for the VaR tests above. Another useful approach could be the graphical procedure proposed by Fermanian and Scaillet (2003). We plan to the pursue the implementation of this procedure in future work.

7. Conclusions and Directions for Future Work

We have presented a new set of procedures for backtesting risk models. The chief insight is that if the one-day VaR model is correctly specified for coverage rate, p, then, every day, the conditional expected duration until the next violation should be a constant $1/p$ days. We suggest various ways of testing this null hypothesis and we conduct a Monte Carlo analysis which compares the new tests to those currently available. Our results show that in many of the situations we consider, the duration-based tests have much better power properties than the previously suggested tests. The size of the tests is easily controlled through finite sample p-values, which we calculate using Monte Carlo simulation.

The majority of financial institutions use VaR as a risk measure, and many calculate VaR using the so-called Historical Simulation approach. While the main focus of our paper has thus been backtesting VaRs from Historical Simulation, we also suggest extensions to density and density tail backtesting.

The immediate potential extensions to our Monte Carlo results are several. First, it may be interesting to calculate the power of the tests with different GARCH specifications using for example Engle and Lee (1999) and Hansen (1994). Second, we could consider structural breaks in the underlying return models, such as those investigated by

Andreou and Ghysels (2002). Finally, Hamilton and Jorda (2002) have recently introduced a class of dynamic hazard models. Exploring these for the purpose of backtesting could be interesting.

We could also consider more complicated portfolios including options and other derivatives. Examining the duration patterns from misspecified risk models in this case could suggest other alternative hypotheses than the ones suggested here. We leaves these extensions for future work.

Finally we stress that the current regulator practice of requiring backtesting on samples of only 250 daily observations is likely to prove futile as the power to reject misspecified risk models is very low in this case.

Figure 19: Value-at-Risk exceedences from six major commercial banks [from Berkowitz and O'Brien (2002)].

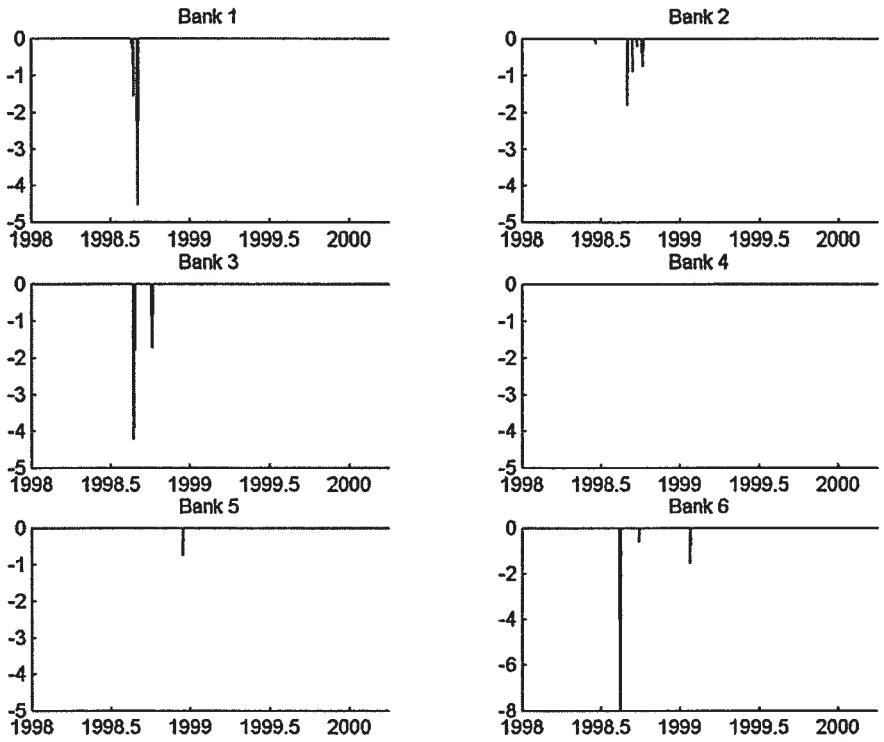


Figure 20: GARCH-t(d) Simulated Portfolio Returns with 1% and 5% Value-at-Risk from Historical Simulation with $T_e = 500$.

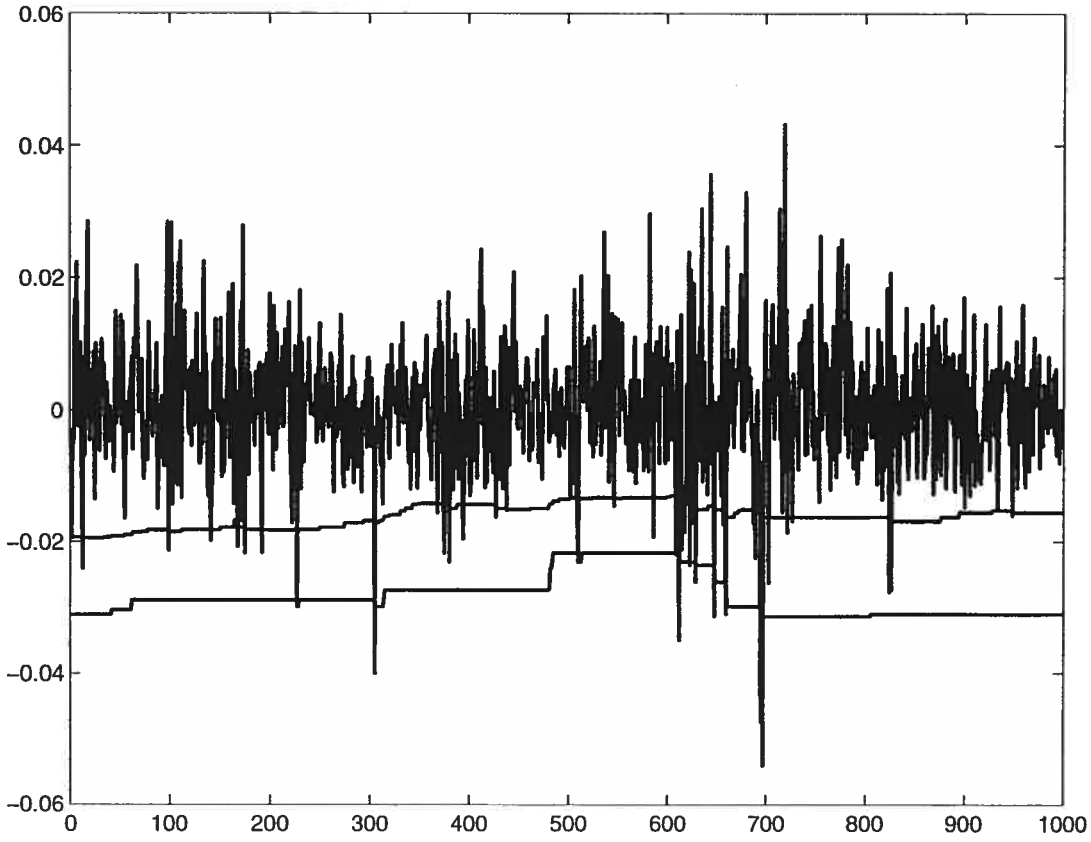


Figure 21: GARCH-t(d) simulated portfolio returns with exeedences of 1% VaRs from Historical Simulation with $T_e = 500$ reported in standard deviations of returns.

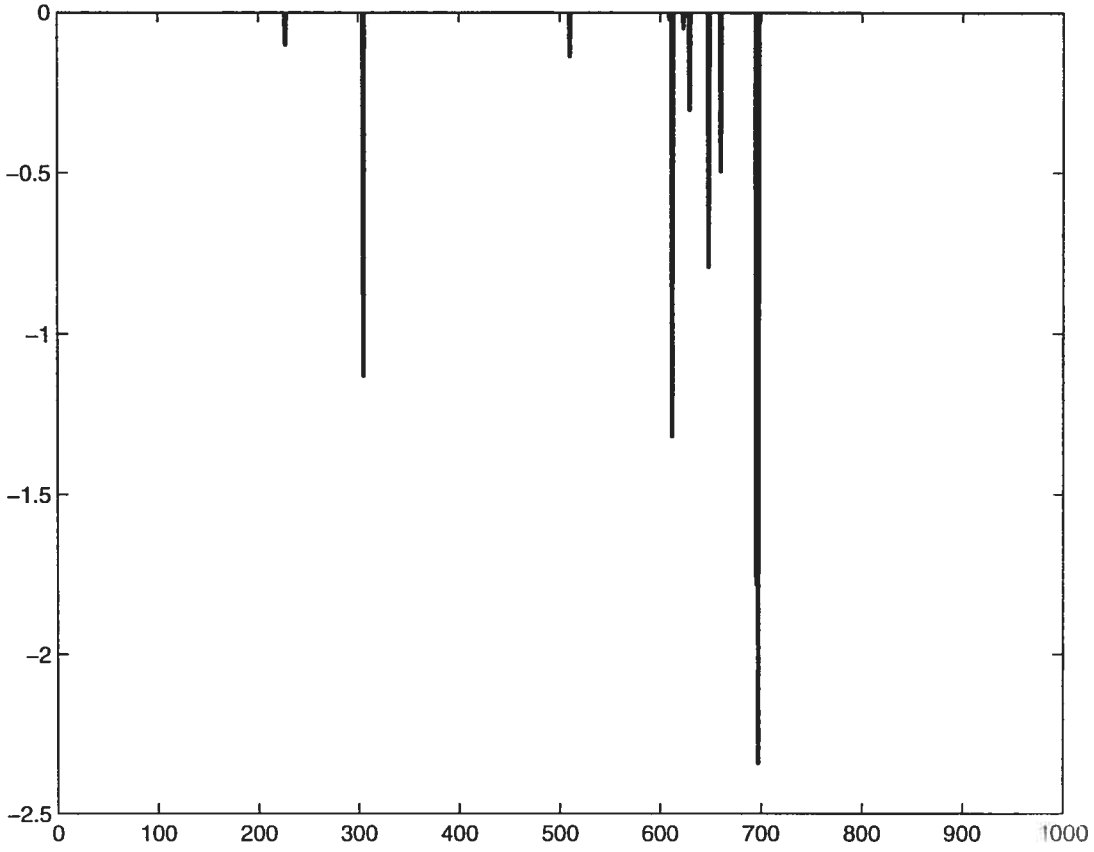


Figure 22: Data-based and Weibull-based hazard functions of durations between VaR violations. Historical Simulation risk model on GARCH-t(d) portfolio returns with $T_e = 500$.

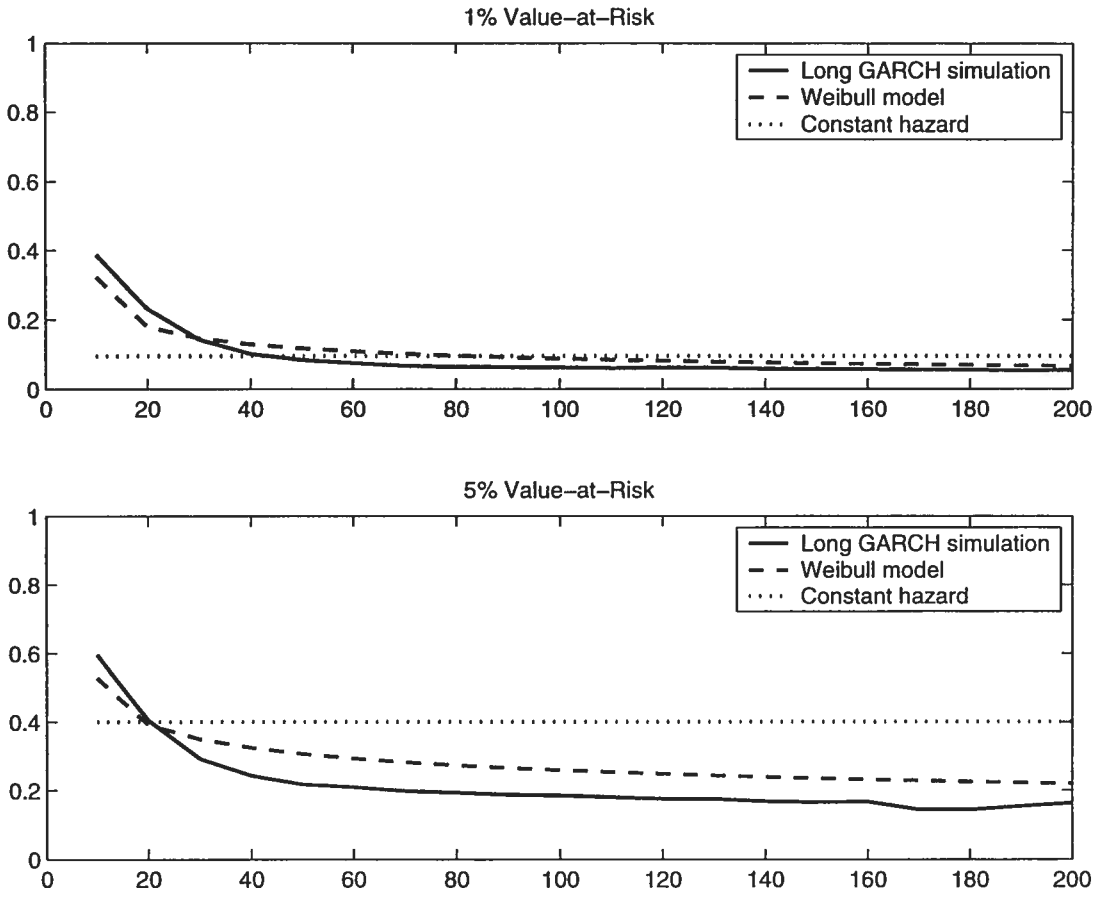


Figure 23: Histograms of duration between VaR violations GARCH-t(d) portfolio returns Historical Simulation risk model with $T_e = 500$.

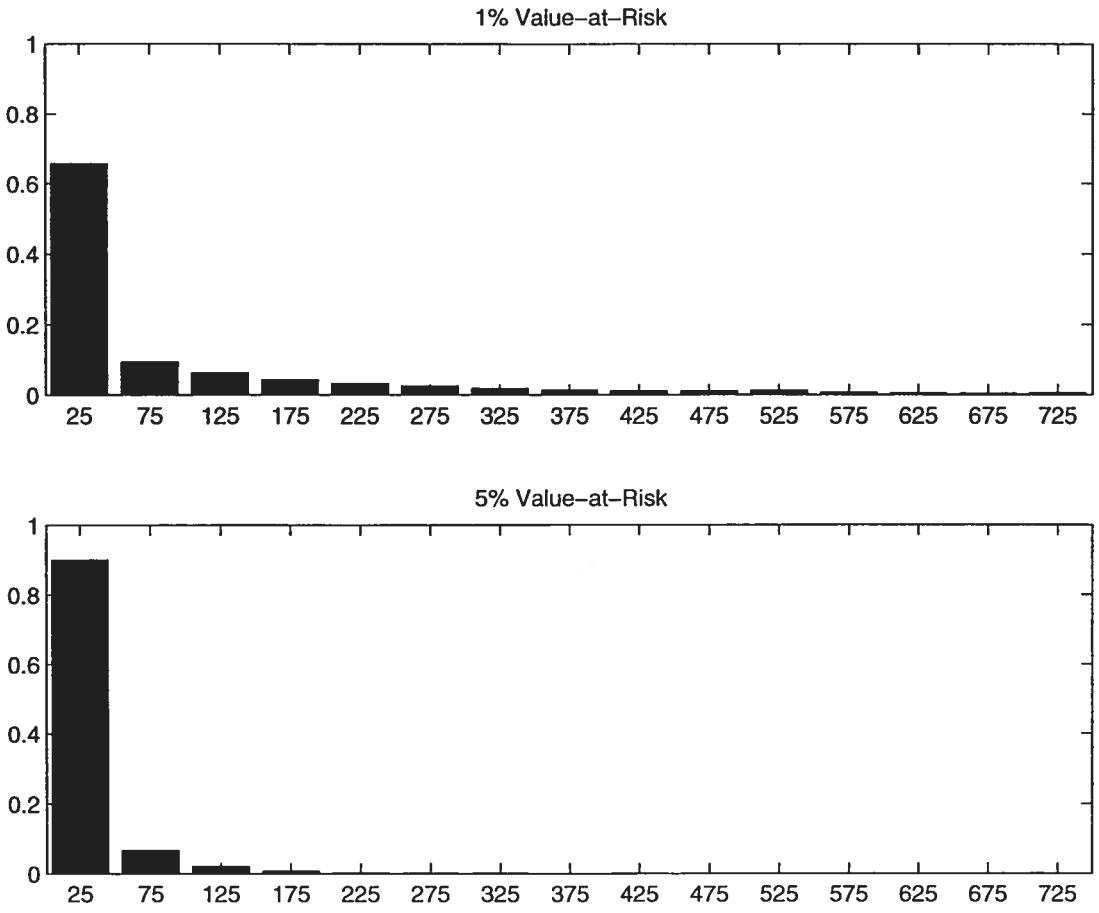


Table 34: Power of independence tests. Historical Simulation VaR calculated on 500 GARCH(1,1)-t(d) returns

Significance level: 1%

	Coverage rate: 1%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.060	0.018	0.150
500	0.105	0.114	0.164
750	0.157	0.236	0.167
1000	0.224	0.378	0.159
1250	0.266	0.484	0.145
1500	0.308	0.596	0.132

Significance level: 5%

	Coverage rate: 1%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.263	0.104	0.234
500	0.307	0.267	0.250
750	0.290	0.415	0.251
1000	0.360	0.546	0.253
1250	0.382	0.674	0.237
1500	0.427	0.752	0.222

Significance level: 10%

	Coverage rate: 1%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.330	0.195	0.278
500	0.370	0.369	0.303
750	0.435	0.536	0.311
1000	0.523	0.648	0.303
1250	0.514	0.758	0.291
1500	0.543	0.820	0.271

	Coverage rate: 5%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.107	0.052	0.159
500	0.215	0.238	0.324
750	0.271	0.413	0.389
1000	0.339	0.546	0.440
1250	0.395	0.692	0.493
1500	0.434	0.750	0.514

	Coverage rate: 5%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.205	0.152	0.273
500	0.296	0.403	0.440
750	0.367	0.607	0.501
1000	0.443	0.734	0.555
1250	0.530	0.833	0.601
1500	0.627	0.882	0.638

	Coverage rate: 5%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.257	0.235	0.342
500	0.351	0.509	0.504
750	0.429	0.706	0.563
1000	0.533	0.810	0.615
1250	0.654	0.895	0.661
1500	0.735	0.927	0.700

Table 35: Sample selection frequency. Historical Simulation VaR calculated on 500 GARCH(1,1)-t(d) returns.

	Coverage rate: 1%				Coverage rate: 5%	
Test:	Markov	Weibull	EACD	Test:	Markov	Weibull	EACD
Sample size				Sample size			
250	0.778	0.589	0.598	250	0.987	0.972	0.974
500	0.956	0.891	0.896	500	1.000	1.000	0.999
750	0.998	0.987	0.986	750	1.000	1.000	1.000
1000	1.000	0.999	0.997	1000	1.000	1.000	1.000
1250	1.000	1.000	1.000	1250	1.000	1.000	1.000
1500	1.000	1.000	1.000	1500	1.000	1.000	1.000

Table 36: Effective power of independence tests. Historical Simulation VaR calculated on 500 GARCH(1,1)-t(d) returns.

Significance level: 1%

	Coverage rate: 1%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.047	0.011	0.090
500	0.100	0.101	0.147
750	0.156	0.233	0.164
1000	0.224	0.378	0.158
1250	0.266	0.484	0.145
1500	0.308	0.596	0.132

Significance level: 5%

	Coverage rate: 1%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.205	0.061	0.140
500	0.294	0.238	0.224
750	0.289	0.410	0.247
1000	0.360	0.545	0.253
1250	0.382	0.674	0.237
1500	0.427	0.752	0.222

Significance level: 10%

	Coverage rate: 1%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.257	0.115	0.167
500	0.353	0.329	0.271
750	0.434	0.529	0.307
1000	0.522	0.647	0.302
1250	0.514	0.758	0.291
1500	0.543	0.820	0.271

	Coverage rate: 5%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.105	0.051	0.154
500	0.215	0.238	0.324
750	0.271	0.413	0.389
1000	0.339	0.546	0.440
1250	0.395	0.692	0.493
1500	0.434	0.750	0.514

	Coverage rate: 5%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.203	0.148	0.266
500	0.296	0.403	0.440
750	0.367	0.607	0.501
1000	0.443	0.734	0.555
1250	0.530	0.833	0.601
1500	0.627	0.882	0.638

	Coverage rate: 5%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.253	0.229	0.333
500	0.351	0.509	0.504
750	0.429	0.706	0.563
1000	0.533	0.810	0.615
1250	0.654	0.895	0.661
1500	0.735	0.927	0.700

Table 37: Power of Independence Tests. Historical Simulation VaR Calculated on 250 GARCH(1,1)-t(d) Returns.

Significance level: 1%

	Coverage rate: 1%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.059	0.005	0.114
500	0.079	0.053	0.098
750	0.108	0.133	0.069
1000	0.153	0.222	0.045
1250	0.203	0.310	0.035
1500	0.230	0.420	0.029

	Coverage rate: 5%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.115	0.068	0.189
500	0.212	0.247	0.288
750	0.244	0.388	0.346
1000	0.299	0.500	0.345
1250	0.344	0.622	0.394
1500	0.385	0.695	0.393

Significance level: 5%

	Coverage rate: 1%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.217	0.072	0.195
500	0.278	0.196	0.183
750	0.254	0.313	0.132
1000	0.290	0.406	0.105
1250	0.305	0.536	0.084
1500	0.321	0.634	0.070

	Coverage rate: 5%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.210	0.169	0.311
500	0.295	0.421	0.408
750	0.346	0.603	0.456
1000	0.413	0.707	0.480
1250	0.499	0.796	0.497
1500	0.582	0.849	0.537

Significance level: 10%

	Coverage rate: 1%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.285	0.166	0.251
500	0.336	0.304	0.236
750	0.401	0.437	0.182
1000	0.467	0.535	0.149
1250	0.463	0.645	0.123
1500	0.459	0.736	0.101

	Coverage rate: 5%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.266	0.250	0.380
500	0.354	0.530	0.475
750	0.419	0.700	0.517
1000	0.507	0.790	0.553
1250	0.631	0.862	0.569
1500	0.688	0.896	0.606

Table 38: Sample selection frequency. Historical Simulation VaR calculated on 250 GARCH(1,1)-t(d) returns.

	Coverage rate: 1%				Coverage rate: 5%	
Test:	Markov	Weibull	EACD	Test:	Markov	Weibull	EACD
Sample size				Sample size			
250	0.877	0.695	0.706	250	0.997	0.993	0.993
500	0.994	0.975	0.976	500	1.000	1.000	1.000
750	1.000	0.999	0.999	750	1.000	1.000	1.000
1000	1.000	1.000	1.000	1000	1.000	1.000	1.000
1250	1.000	1.000	1.000	1250	1.000	1.000	1.000
1500	1.000	1.000	1.000	1500	1.000	1.000	1.000

Table 39: Effective power of independence tests. Historical Simulation VaR calculated on 250 GARCH(1,1)-t(d) returns.

Significance level: 1%

	Coverage rate: 1%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.051	0.004	0.080
500	0.078	0.052	0.095
750	0.108	0.133	0.069
1000	0.153	0.222	0.045
1250	0.203	0.310	0.035
1500	0.230	0.420	0.029

	Coverage rate: 5%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.115	0.068	0.187
500	0.212	0.247	0.288
750	0.244	0.388	0.346
1000	0.299	0.500	0.345
1250	0.344	0.622	0.394
1500	0.385	0.695	0.393

Significance level: 5%

	Coverage rate: 1%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.190	0.050	0.138
500	0.276	0.191	0.179
750	0.254	0.313	0.132
1000	0.290	0.406	0.105
1250	0.305	0.536	0.084
1500	0.321	0.634	0.070

	Coverage rate: 5%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.209	0.167	0.308
500	0.295	0.421	0.408
750	0.346	0.603	0.456
1000	0.413	0.707	0.480
1250	0.499	0.796	0.497
1500	0.582	0.849	0.537

Significance level: 10%

	Coverage rate: 1%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.250	0.116	0.177
500	0.334	0.296	0.230
750	0.401	0.436	0.181
1000	0.467	0.535	0.149
1250	0.463	0.645	0.123
1500	0.459	0.736	0.101

	Coverage rate: 5%		
Test:	Markov	Weibull	EACD
Sample size			
250	0.266	0.248	0.378
500	0.354	0.530	0.475
750	0.419	0.700	0.517
1000	0.507	0.790	0.553
1250	0.631	0.862	0.569
1500	0.688	0.896	0.606

References

AKAIKE, H. (1973): "Information Theory and an Extension of the Maximum Likelihood Principle," in *2nd International Symposium on Information Theory*, ed. by B. N. Petrov, and F. Csaki, pp. 267–281. Akademiai Kiado, Budapest.

AMEMIYA, T., AND R. Y. WU (1972): "The Effect of Aggregation on Prediction in the Autoregressive Model," *Journal of the American statistical Association*, 67(339), 628–632.

ANDERSEN, T., T. BOLLERSLEV, F. X. DIEBOLD, AND P. LABYS (2001): "The Distribution of Realized Exchange Rate Volatility," *Journal of the American Statistical Association*, 96, 42–55.

ANDREOU, E., AND E. GHYSELS (2002): "Quality Control for Value at Risk: Monitoring Disruptions in the Distribution of Risk Exposure," Manuscript, University of North Carolina.

ANDREWS, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

——— (1993): "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821–856.

——— (1999): "Estimation when a parameter is on a boundary," *Econometrica*, 67(6), 1341–1383.

——— (2001): "Testing when a parameter is on the boundary of the maintained hypothesis," *Econometrica*, 69(3), 683–734.

ANDREWS, D. W. K., AND J. C. MONAHAN (1992): "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 60, 953–966.

ARTZNER, P., F. DELBAEN, J.-M. EBER, AND D. HEATH (1999): "Coherent Measures of Risk," *Mathematical Finance*, 9, 203–228.

BAILLIE, R. T., AND T. BOLLERSLEV (1990): "A Multivariate Generalized ARCH Approach to Modeling Risk Premia in Forward Foreign Exchange Rate Markets," *Journal of International Money and Finance*, 9, 309–324.

BARNARD, J., R. MCCULLOCH, AND X.-L. MENG (2000): "Modelling Covariance Matrices in Terms of Standard Deviations and Correlations, with Application to Shrinkage," *Statistica Sinica*, 10(4).

BARONE-ADESI, G. K., K. GIANNOPOULOS, AND L. VOSPER (2002): "Backtesting Derivative Portfolios with Filtered Historical Simulation (FHS)," *European Financial Management*, 8, 31–58.

BARTEL, H., AND LÜTKEPOHL (1998): "Estimating the Kronecker indices of cointegrated echelon-form VARMA models," *Econometrics Journal*, 1, C76–C99.

BASAK, S., AND A. SHAPIRO (2000): "Value at Risk Based Risk Management: Optimal Policies and Asset Prices," *Review of Financial Studies*, 14, 371–405.

BASLE COMMITTEE ON BANKING SUPERVISION (1996): *Amendment to the Capital Accord to Incorporate Market Risks.*Basle.

BEDER, T. (1995): "VaR: Seductive but Dangerous," *Financial Analysts Journal*, 51(5), 12–24.

BERKOWITZ, J. (2001): "Testing density forecasts, with applications to risk management," *JBES*, 19(4), 465–474.

BERKOWITZ, J., AND J. O'BRIEN (2002): "How Accurate are the Value-at-Risk Models at Commercial Banks?," *Journal of Finance*, 57, 1093–1112.

BERNANKE, B. S., AND I. MIHOV (1998): "Measuring Monetary Policy," *The Quarterly Journal of Economics*, 113(3), 869–902.

BOLLERSLEV, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity," *journal of Econometrics*, 31, 307–327.

——— (1987): "A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return," *REStat*, 69, 542–547.

——— (1990): "Modelling the Coherence in Short-Run Nominal Exchange Rates: A Multivariate Generalized ARCH Model," *Review of Economics and Statistics*, 72, 498–505.

BOLLERSLEV, T., AND R. F. ENGLE (1993): "Common Persistence in Conditional Variances," *Econometrica*, 61(1), 167–186.

BOLLERSLEV, T., R. F. ENGLE, AND J. M. WOOLDRIDGE (1988): "A Capital Asset Pricing Model with Time-Varying Covariances," *The Journal of Political Economy*, 96(1), 116–131.

BOSQ, D. (1998): *Nonparametric Statistics for Stochastic Processes - Estimation and Prediction*, no. 110 in Lecture Notes in statistics. Springer-Verlag, Berlin, second edn.

BOUDJELLABA, H., J.-M. DUFOUR, AND R. ROY (1992): "Testing Causality Between Two Vectors in Multivariate ARMA Models," *Journal of the American Statistical Association*, 87, 1082–1090.

CHOI, I. (1993): "Asymptotic Normality of the Least-Squares Estimates for Higher Order Autoregressive Integrated Processes with some Applications," *Econometric Theory*, 9, 263–282.

CHRISTOFFERSEN, P. (1998): "Evaluating Interval Forecasts," *International Economic Review*, 39, 841–862.

——— (2003): *Elements of Financial Risk Management*. Academic Press, San Diego.

CHRISTOFFERSEN, P., J. HAHN, AND A. INOUE (2001): "Testing and Comparing Value-at-Risk Measures," *Journal of Empirical Finance*, 8, 325–342.

COOPER, D. M., AND E. F. WOOD (1982): "Identifying Multivariate Time Series Models," *Journal of Time Series Analysis*, 3, 153–164.

CRIBARI-NETO, F., S. L. P. FERRARI, AND G. M. CORDEIRO (2000): "Improved Heteroskedasticity-Consistent Covariance Matrix Estimators," *Biometrika*, 87(4), 907–918.

CUOCO, D., H. HE, AND S. ISSAENKO (2001): "Optimal Dynamic Trading Strategies with Risk Limits," Manuscript, Yale University.

CUSHING, M. J., AND M. G. MCGARVEY (1999): "Covariance Matrix Estimation," in *Generalized Method of Moments Estimation*, ed. by L. Mátyás, chap. 3, pp. 63–95. Cambridge University Press, Cambridge, U.K.

DAVIDIAN, M., AND R. J. CARROLL (1987): "Variance Function Estimation," *Journal of the American Statistical Association*, 82, 1079–1091.

DAVYDOV, Y. A. (1968): "Convergence of Distributions Generated by Stationary Stochastic Processes," *Theory of Probability and its Applications*, pp. 691–696.

DEISTLER, M., AND E. J. HANNAN (1981): "Some Properties of the Parametrization of ARMA Systems with Unknown Order," *Journal of Multivariate Analysis*, 11, 474–484.

DEMPSTER, A. P., N. M. LAIRD, AND D. B. RUBIN (1977): "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, 39(1), 1–38, With discussion.

DEN HAAN, W. J., AND A. LEVIN (1997): "A Practitioner's Guide to Robust Covariance Matrix Estimation," in *Handbook of Statistics 15: Robust Inference*, ed. by G. S. Maddala, and C. R. Rao, chap. 15. North-Holland, Amsterdam.

DIEBOLD, F., T. GUNTHER, AND A. TAY (1998): "Evaluating Density Forecasts, with Applications to Financial Risk Management," *International Economic Review*, 39, 863–883.

DIEBOLD, F. X., AND M. NERLOVE (1989): "The Dynamics of Exchange Rate Volatility: A Multivariate Latent Factor ARCH Model," *Journal of Applied Econometrics*, 4(1), 1–21.

DING, Z., AND R. F. ENGLE (2001): "Large Scale Conditional Covariance Matrix Modeling, Estimation and Testing," Discussion Paper FIN-01-029, Stern School of Business, New York University.

DOAN, T., AND R. LITTERMAN (1983): *RATS User's Manual.* VAR Econometrics, Minneapolis.

DOLADO, J. J., AND H. LÜTKEPOHL (1996): "Making Wald Tests Work for Cointegrated VAR Systems," *Econometric Reviews*, 15, 369–386.

DOORNIK, J. A. (1999): *Object-Oriented Matrix Programming Using Ox, 3rd ed.* Timberlake Consultants Press and Oxford, Oxford, U.K., www.nuff.ox.ac.uk/Users/Doornik.

DOUKHAN, P. (1995): *Mixing - Properties and Examples*, no. 85 in Lecture Notes in Statistics. Springer-Verlag.

DROST, F. C. (1993): "Temporal Aggregation of Time-Series," in *Econometric Analysis of Financial Markets*, ed. by J. Kaehler, and P. Kugler, pp. 11–21. Physica-Verlag, New York.

DROST, F. C., AND T. E. NIJMAN (1993): "Temporal Aggregation of GARCH Processes," *Econometrica*, 61(4), 909–927.

DUFOUR, J.-M. (2002): "Monte Carlo Tests with Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics in Econometrics," *Journal of Econometrics*, forthcoming.

DUFOUR, J.-M., E. GHYSELS, AND A. HALL (1994): "Generalized Predictive Tests and Structural Change Analysis in Econometrics," *International Economic Review*, 35, 199–229.

DUFOUR, J.-M., AND L. KHALAF (2002): "Simulation Based Finite and Large Sample Tests in Multivariate Regressions," *Journal of Econometrics*, 111(2), 303–322.

DUFOUR, J.-M., AND D. PELLETIER (2003): "Linear estimation of univariate and multivariate volatility models," Unpublished manuscript.

DUFOUR, J.-M., D. PELLETIER, AND É. RENAULT (2002): "Short Run and Long Run Causality in Time Series: Inference," Discussion paper, Centre interuniversitaire de

recherche en analyse des organisations (CIRANO) and Centre interuniversitaire de recherche en économie quantitative (CIREQ), Université de Montréal.

DUFOUR, J.-M., AND E. RENAULT (1998): "Short-Run and Long-Run Causality in Time Series: Theory," *Econometrica*, 66, 1099–1125.

DURBIN, J. (1959): "Efficient Estimation of Parameter in Moving Average Models," *Biometrika*, 46, 306–316.

DURBIN, J. (1960a): "Estimation of parameters in time-series regression models," *Journal of the Royal Statistical Society, Series B*, 22, 139–153.

DURBIN, J. (1960b): "The Fitting of Time Series Models," *Revue de l'Institut International de Statistique*, 28, 233.

ENGLE, R. (2002): "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models," *Journal of Business and Economic Statistics*, 20(3), 339–350.

ENGLE, R., AND G. LEE (1999): "A Permanent and Transitory Component Model of Stock Return Volatility," in *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W.J. Granger*, ed. by R. Engle, and H. White, pp. 475–497. Oxford University Press.

ENGLE, R., AND J. RUSSEL (1998): "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," *Econometrica*, 66, 1127–1162.

ENGLE, R., AND K. SHEPPARD (2001): "Theoretical and Empirical Properties of Dynamic Conditional Correlation Multivariate GARCH," UCSD Discussion Paper 2001-15.

ENGLE, R. F. (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50(4), 987–1008.

ENGLE, R. F., AND K. F. KRONER (1995): "Multivariate Simultaneous Generalized ARCH," *Econometric Theory*, 11, 122–150.

ENGLE, R. F., V. K. NG, AND M. ROTHSCHILD (1990): "Asset Pricing with a Factor-ARCH Covariance Structure, Empirical Estimates for Treasury Bills," *Journal of Econometrics*, 45, 213–237.

ENGLE, R. F., AND R. SUSMEL (1993): "Common Volatility in International Equity Markets," *Journal of Business and Economic Statistics*, 11(2), 167–176.

FERMANIAN, J.-D., AND O. SCAILLET (2003): "Nonparametric Estimation of Copulas for Time Series," *Journal of Risk*, 5, 25–54.

FLORES DE FRUTOS, R., AND G. R. SERRANO (2002): "A Generalized Least Squares Estimation Method for VARMA Models," *Statistics*, 36(4), 303–316.

FRANCQ, C., AND J.-M. ZAKOÏAN (1998): "Estimating Linear Representations of Nonlinear Processes," *Journal of Statistical Planning and Inference*, 68, 145–165.

GALBRAITH, J. W., A. ULLAH, AND V. ZINDE-WALSH (2000): "Estimation of the Vector Moving Average Model by Vector Autoregression," *Econometric Reviews*, 21(2), 205–219.

GALBRAITH, J. W., AND V. ZINDE-WALSH (1994): "A Simple Noniterative Estimator for Moving Average Models," *Biometrika*, 81(1), 143–155.

——— (1997): "On Simple, Autoregression-Based Estimation and Identification Techniques for ARMA Models," *Biometrika*, 84(3), 685–696.

GALLANT, A. R. (1987): *Nonlinear Statistical Models*. John Wiley & Sons, New York.

GEWEKE, J. (1984): "Inference and Causality in Economic Time series," in *Handbook of Econometrics, Volume 2*, ed. by Z. Griliches, and M. D. Intrilligator, pp. 1102–1144. North-Holland, Amsterdam.

GEWEKE, J., AND G. AMISANO (2001): "Compound Markov Mixture Models with Applications in Finance," Unpublished manuscript.

GOURIÉROUX, C. (2000): *Econometrics of Qualitative Dependent Variables*. Cambridge University Press, Translated by Paul B. Klassen.

GOURIÉROUX, C., AND A. MONFORT (1997): *Time Series and Dynamic Models*. Cambridge University Press, Cambridge, U.K.

GRANGER, C. W. J. (1969): "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica*, 37, 424–459.

GRANGER, C. W. J., AND Z. DING (1993): "Some Properties of Absolute Return - An Alternative Measure of Risk," Discussion Paper 93-38, University of California, San Diego.

HAMILTON, J., AND O. JORDA (2002): "A Model of the Federal Funds Rate Target," *Journal of Political Economy*, 110, 1135–1167.

HAMILTON, J. D. (1989): "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57(2), 357–384.

——— (1994): *Time Series Analysis*. Princeton University Press, Princeton, New Jersey.

——— (1996): "Specification Testing in Markov-Switching Time-Series Models," *Journal of Econometrics*, 70, 127–157.

HANNAN, E. J. (1969): "The Identification of Vector Mixed Autoregressive- Moving Average Systems," *Biometrika*, 57, 223–225.

——— (1971): "The Identification Problem for Multiple Equation Sysytem with Moving Average Errors," *Econometrica*, 39, 751–766.

——— (1976): "The Identification and Parameterization of ARMAX and State Space Forms," *Econometrica*, 44(4), 713–723.

HANNAN, E. J., AND M. DEISTLER (1988): *The Statistical Theory of Linear Systems*. John Wiley & Sons, New York.

HANNAN, E. J., W. T. M. DUNSMUIR, AND M. DEISTLER (1980): "Estimation of Vector ARMAX Models," *Journal of Multivariate Analysis*, 10, 275–295.

HANNAN, E. J., AND L. KAVALIERIS (1984a): "A Method for Auregressive-Moving Average Estimation," *Biometrika*, 72(2), 273–280.

——— (1984b): "Multivariate Linear Time Series Models," *Advances in Applied Probability*, 16, 492–561.

——— (1986): "Regression, Autoregression Models," *Journal of Time Series Analysis*, 7(1), 27–49.

HANNAN, E. J., AND B. QUINN (1979): "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society, Series B*, 41, 190–191.

HANNAN, E. J., AND J. RISSANEN (1982): "Recursive Estimation of Mixed Autoregressive-Moving-Average Order," *Biometrika*, 69, 81–94, Errata 70 (1982), 303.

——— (1983): "Errata: "Recursive Estimation of Mixed Autoregressive-Moving Average Order"," *Biometrika*, 70(1), 303.

HANSEN, B. (1994): "Autoregressive Conditional Density Estimation," *International Economic Review*, 35, 705–730.

HANSEN, B. E. (1992): "Consistent Covariance Matrix Estimation for Dependent Heterogeneous Processes," *Econometrica*, 60, 967–972.

HARVEY, C., E. RUIZ, AND N. SHEPHARD (1994): "Multivariate Stochastic Variance Models," *Review of Economic Studies*, 61, 247–264.

HENDRICKS, D. (1996): "Evaluation of Value-at-Risk Models Using Historical Data," *Economic Policy Review*, pp. 39–69, Federal Reserve Bank of New York.

HILLMER, S. C., AND G. C. TIAO (1979): "Likelihood Function of Stationary Multiple Autoregressive Moving Average Models," *Journal of the American Statistical Association*, 74(367), 652–660.

HONG, Y. (1996): "Consistent Testing for Serial Correlation of Unknown Form," *Econometrica*, 64(4), 837–864.

HORN, R. G., AND C. A. JOHNSON (1985): *Matrix Analysis*. Cambridge University Press, Cambridge, U.K.

HUANG, D., AND L. GUO (1990): "Estimation of Nonstationary ARMAX Models Based on the Hannan-Rissanen Method," *The Annals of Statistics*, 18(4), 1729–1756.

HURVICH, C. M., AND C.-L. TSAI (1989): "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297–307.

IBRAGIMOV, I. A. (1962): "Some Limit Theorems for Stationary Processes," *Theory of Probability and its Applications*, 7, 349–382.

JORION, P. (2001): *Value-at-Risk: The New Benchmark for Controlling Financial Risk*. McGraw-Hlill, second edn.

KEATING, J. W. (2002): "Structural Inference with Long-Run Recursive Empirical Models," *Macroeconomic Dynamics*, 6(2), 266–283.

KIEFER, N. M. (1988): "Economic Duration Data and Hazard Functions," *Journal of Economic Literature*, 26, 646–679.

KILIAN, L. (1998a): "Confidence Intervals for Impulse Responses Under Departures From Normality," *Econometric Reviews*, 17(1), 1–29.

KILIAN, L. (1998b): "Small-Sample Confidence Intervals for Impulse Response Functions," *Review of Economics and Statistics*, 80, 218–230.

KIM, C.-J. (1994): "Dynamic linear models with Markov-switching," *Journal of Econometrics*, 60(1-2), 1–22.

KIM, S., N. SHEPHARD, AND S. CHIB (1998): "Stochastic Volatility: Likelihood Inference and Comparison With ARCH Models," *Review of Economic Studies*, 65, 361–393.

KING, M., E. SENTANA, AND S. WADHWANI (1994): "Volatility and Links Between National Stock Markets," *Econometrica*, 62(4), 901–933.

KOREISHA, S. G., AND T. M. PUKKILA (1989): "Fast Linear Estimation Methods for Vector Autoregressive Moving-Average Models," *Journal of Time Series Analysis*, 10(4), 325–339.

——— (1990a): "A Generalized Least-Squares Approach for Estimation of Autoregressive-Moving-Average Models," *Journal of Time Series Analysis*, 11(2), 139–151.

——— (1990b): "Linear Methods for Estimating ARMA and Regression Models with Serial Correlation," *Communications in Statistics, Part B -Simulation and Computation*, 19, 71–102.

——— (1995): "A Comparison Between Different Order-Determination Criteria for Identification of ARIMA Models," *Journal of Business and Economic Statistics*, 13, 127–131.

——— (1999): "The Selection of the Order and Identification of Nonzero Elements in the Polynomial Matrices of Vector Autoregressive Processes," *Journal of Statistical Computation and Simulation*, 62, 207–235.

KUPIEC, P. (1995): "Techniques for Verifying the Accuracy of Risk Measurement Models," *Journal of Derivatives*, 3, 73–84.

KUROZUMI, E., AND T. YAMAMOTO (2000): "Modified Lag Augmented Vector Autoregressions," *Econometric Reviews*, 19, 207–231.

LEDOIT, O., P. SANTA CLARA, AND M. WOLF (2003): "Flexible Multivariate GARCH Modeling With an Application to International Stock Markets," *Review of Economics and Statistics*, 85, forthcoming.

LEWIS, R., AND G. C. REINSEL (1985): "Prediction of Multivariate Time Series by Autoregressive Model Fitting," *Journal of Multivariate Analysis*, 16, 393–411.

LUCEÑO, A. (1994): "A Fast Algorithm For The Exact Likelihood Of Stationary And Partially Nonstationnary Vector Autoregressive-Moving Average Processes," *Biometrika*, 81(3), 555–565.

LÜTKEPOHL, H. (1985): "Comparison of Criteria for Estimating the Order of a Vector Autoregression Process," *Journal of Time Series Analysis*, 6(1), 35–52.

——— (1991): *Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin.

——— (1993a): *Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin, second edn.

——— (1993b): "Testing for Causation Between Two Variables in Higher Dimensional VAR Models," in *Studies in Applied Econometrics*, ed. by H. Schneeweiss, and K. Zimmermann. Springer-Verlag, Heidelberg.

LÜTKEPOHL, H., AND M. M. BURDA (1997): "Modified Wald Tests under Nonregular Conditions," *Journal of Econometrics*, 78, 315–332.

LÜTKEPOHL, H., AND H. CLAESSEN (1997): "Analysis of Cointegrated VARMA Processes," *Journal of Econometrics*, 80, 223–239.

LÜTKEPOHL, H., AND D. S. POSKITT (1996a): "Consistent Estimation of the Number of Cointegration Relations in a Vector Autoregressive Model," Discussion Paper 74, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin.

LÜTKEPOHL, H., AND D. S. POSKITT (1996b): "Specification of Echelon-Form VARMA Models," *Journal of Business and Economic Statistics*, 14, 69–80.

——— (1996c): "Testing for Causation Using Infinite Order Vector Autoregressive Processes," *Econometric Theory*, 12, 61–87.

LÜTKEPOHL, H., AND P. SAIKKONEN (1997): "Impulse Response Analysis in Infinite Order Cointegrated Vector Autoregressive Processes," *Journal of Econometrics*, 81, 127–157.

MALLOWS, C. L. (1973): "Some comments on Cp," *technometrics*, 15, 661–675.

MAURICIO, J. A. (2002): "An Algorithm for the Exact Likelihood of a Stationary Vector Autoregressive-Moving Average Model," *Journal of Time Series Analysis*, 23(4), 473–486.

MCMILLIN, W. D. (2001): "The Effect of Monetary Policy Shocks: Comparing Contemporaneous versus Long-Run Identifying Restrictions," *Southern Economic Journal*, 67(3), 618–636.

MCQUARRIE, A. D. R., AND C.-L. TSAI (1998): *Regression and Time Series Model Selection*. World Scientific, Singapore.

NELSEN, R. (1998): *An Introduction to Copulas*, vol. 139 of *Lectures Notes in Statistics*. Springer Verlag.

NEWBOLD, P. (1982): "Causality Testing in Economics," in *Time Series Analysis: Theory and Practice 1*, ed. by O. D. Anderson. North-Holland, Amsterdam.

NEWEY, W. K., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Volume IV*, ed. by R. F. Engle, and D. L. McFadden, chap. 36, pp. 2111–2245. North-Holland, Amsterdam.

NEWEY, W. K., AND K. D. WEST (1987a): "Hypothesis Testing with Efficient Method of Moments Estimators," *International Economic Review*, 28, 777–787.

——— (1987b): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

NG, V., R. F. ENGLE, AND M. ROTHSCHILD (1992): "A Multi-Dynamic-Factor Model for Stock Returns," *Journal of Econometrics*, 52, 245–266.

NIJMAN, T. H., AND F. PALM (1990): "Parameter Identification in ARMA Processes in the Presence of Regular But Incomplete Sampling," *Journal of Time Series Analysis*, 11(3), 239–248.

NSIRI, S., AND R. ROY (1992): "On the Identification of ARMA Echelon-Form Models," *Canadian Journal of Statistics*, 20(4), 369–386.

NSIRI, S., AND R. ROY (1996): "Identification of Refined ARMA Echelon Form Models for Multivariate Time Series," *Journal of Multivariate Analysis*, 56, 207–231.

PAGAN, A. (1986): "Two Stage and Related Estimators and Their Applications," *The Review of Economic Studies*, 53(4), 517–538.

PALM, F., AND T. H. NIJMAN (1984): "Missing Observations in the Dynamic Regression Model," *Econometrica*, 52(6), 1415–1436.

PAPARODITIS, E. (1996): "Bootstrapping Autoregressive and Moving Average Parameter Estimates of Infinite Order Vector Autoregressive Processes," *Journal of Multivariate Analysis*, 57, 277–296.

PAPARODITIS, E., AND B. STREITBERG (1991): "Order Identification Statistics in Stationary Autoregressive Moving-Average Models: Vector Autocorrelations and the Bootstrap," *Journal of Time Series Analysis*, 13(5), 415–434.

PARK, J. Y., AND P. C. B. PHILLIPS (1989): "Statistical Inference in Regressions with Integrated Processes: Part 2," *Econometric Theory*, 5, 95–131.

PIERCE, D. A., AND L. D. HAUGH (1977): "Causality in Temporal Systems: Characterizations and Survey," *Journal of Econometrics*, 5, 265–293.

POIRIER, D. J. (1995): *Intermediate Statistics and Econometrics*. MIT Press, Cambridge, MA.

POSKITT, D. S. (1987): "A Modified Hannan-Rissanen Strategy for Mixed Autoregressive-Moving Average Oder Determination," *Biometrika*, 74(4), 781–790.

——— (1992): "Identification of Echelon Canonical Forms for Vector Linear Processes Using Least Squares," *The Annals of Statistics*, 20, 195–215.

PÖTSCHER, B. (1991): "Efffects of Model Selection on Inference," *Econometric Theory*, 7, 163–185.

PRITSKER, M. (1997): "Evaluating Value at Risk Methodologies: Accuracy versus Computational Time," *Journal of Financial Services Research*, 12, 201–242.

——— (2001): "The Hidden Dangers of Historical Simulation," Finance and Economics Discussion Series 2001-27, Board of Governors of the Federal Reserve System, Washington.

PUKKILA, T., S. KOREISHA, AND A. KALLINEN (1990): "The Identification of ARMA Models," *Biometrika*, 77(3), 537–548.

RENAULT, E., K. SEKKAT, AND A. SZAFARZ (1998): "Testing for Spurious Causality in Exchange Rates," *Journal of Empirical Finance*, 5, 47–66.

RIVERS, D., AND Q. VUONG (2002): "Model selection tests for nonlinear dynamic models," *Econometrics Journal*, 5, 1–39.

ROSS, S. R. (1993): *Introduction to Probability Models*. Academic Press, fifth edn.

SAIKKONEN, P. (1986): "Asymptotic Properties of some Preliminary Estimators for Autoregressive Moving Average Time Series Models," *Journal of Time Series Analysis*, 7, 133–155.

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

SCHWERT, G. W. (1989): "Why Does Stock Market Volatility Change Over Time," *Journal of Finance*, 44(5), 1115–1153.

SCHWERT, G. W., AND P. J. SEGUIN (1990): "Heteroskedasticity in Stock Returns," *The Journal of Finance*, 45(4), 1129–1155.

SHEA, B. L. (1989): "The Exact Likelihood of a Vector Autoregressive Moving Average Model," *Applied Statistics*, 38(1), 161–184.

SIMS, C. (1980): "Macroeconomics and Reality," *Econometrica*, 48, 1–48.

SIMS, C. A., J. H. STOCK, AND M. W. WATSON (1990): "Inference in Linear Time Series Models with some Unit Roots," *Econometrica*, 58, 113–144.

SPLIID, H. (1983): "A Fast Estimation Method for the Vector Autoregressive Moving Average Model with Exogenous Variables," *Journal of the American Statistical Association*, 78(384), 843–849.

SUGIURA, N. (1978): "Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections," *Communications in Statistics - Theory and Methods*, 7, 13–26.

TAYLOR, S. (1986): *Modelling Financial Time Series*. John Wiley and Sons, New York.

TIAO, G. C., AND G. E. P. BOX (1981): "Modeling Multiple Time Series with Applications," *Journal of the American Statistical Association*, 76, 802–816.

TIAO, G. C., AND R. S. TSAY (1983): "Consistency Properties of Least Squares Estimates of Autoregressive Parameters in ARMA Models," *The Annals of Statistics*, 11, 856–871.

——— (1989): "Model Specification in Multivariate Time Series," *Journal of the Royal Statistical Society, Series B*, 51(2), 157–213.

TODA, H. Y., AND P. C. B. PHILLIPS (1993): "Vector Autoregressions and Causality," *Econometrica*, 61, 1367–1393.

TODA, H. Y., AND P. C. B. PHILLIPS (1994): "Vector Autoregression and Causality: A Theoretical Overview and Simulation Study," *Econometric Reviews*, 13, 259–285.

TODA, H. Y., AND T. YAMAMOTO (1995): "Statistical Inference in Vector Autoregressions with Possibly Integrated Processes," *Journal of Econometrics*, 66, 225–250.

TSAY, R. S. (1989a): "Identifying Multivariate Time Series Models," *Journal of Time Series Analysis*, 10, 357–372.

TSAY, R. S. (1989b): "Parsimonious Parameterization of Vector Autoregressive Moving Average Models," *Journal of Business and Economic Statistics*, 7(3), 327–341.

TSAY, R. S. (1991): "Two Canonical Forms for Vector VARMA Processes," *Statistica Sinica*, 1, 247–269.

TSE, Y. K., AND K. C. TSUI (2002): "A Multivariate Generalized Autoregressive Conditional Heteroscedasticity Model With Time-Varying Correlations," *Journal of Business and Economic Statistics*, 20(3), 351–362.

WAHLBERG, B. (1989): "Estimation of Autoregressive Moving-Average Models Via High-Order Autoregressive Approximations," *Journal of Time Serie Analysis*, 10(3), 283–299.

WALLIS, K. F. (1977): "Multiple time series analysis and the final form of econometric models," *Econometrica*, 45(6), 1481–1497.

WHITE, H. (1999): *Asymptotic Theory for Econometricians*. Academic Press, San Diego, Florida, revised edn.

WIENER, N. (1956): "The Theory of Prediction," in *The Theory of Prediction*, ed. by E. F. Beckenback, chap. 8. McGraw-Hill, New York.

WOOLDRIDGE, J. M. (1989): "A Computationally Simple Heteroskedasticity and Serial Correlation Robust Standard Error for the Linear Regression Model," *Economics Letters*, 31, 239–243.

YAMADA, H., AND H. Y. TODA (1998): "Inference in Possibly Integrated Vector Autoregressive Models: Some Finite Sample Evidence," *Journal of Econometrics*, 86, 55–95.

YAMAMOTO, T. (1996): "A Simple Approach to the Statistical Inference in Linear Time Series Models Which May Have Some Unit Roots," *Hitotsubashi Journal of Economics*, 37, 87–100.

ZELLNER, A., AND F. PALM (1974): "Time series analysis and simultaneous equation econometric model," *Journal of Econometrics*, 2(1), 17–54.

ZHAO-GUO, C. (1985): "The Asymptotic Efficiency of a Linear Procedure of Estimation for ARMA Models," *Journal of Time Series Analysis*, 6(1), 53–62.

Conclusions générales

Dans cette thèse, nous avons étudié plusieurs problèmes d'économétrie des séries chronologiques et de la finance. Les différents sujets abordés ont pour thème commun la malédiction de la dimension qui est intrinsèque de l'étude des séries chronologiques multivariées.

Dans le premier essai, nous étudions la modélisation de séries multivariées à l'aide de modèles VARMA. Notre but est de présenter une méthode visant à simplifier l'utilisation de ces modèles. Notre contribution se fait en deux points. Premièrement, nous introduisons deux nouvelles représentations VARMA identifiées qui ont la propriété d'avoir une partie MA très simple (dans un cas, l'opérateur est diagonal et dans l'autre, il est scalaire), ce qui peut faciliter leur utilisation. Deuxièmement, nous proposons l'utilisation d'une méthode d'estimation qui ne nécessite que trois régressions linéaires. Cette méthode est une généralisation de celle proposée par Hannan and Rissanen (1982) pour les modèles ARMA univariés. Nous démontrons que cette méthode d'estimation est valide sous des hypothèses faibles pour le processus qui gouverne les innovations (non corrélées et fortement mélangeantes) et nous introduisons un critère d'information modifié qui donne une estimation convergente des ordres des opérateurs AR et MA pour nos représentations. Dans une application empirique, nous illustrons les gains qu'on peut réaliser en utilisant des modèles VARMA plutôt que des modèles VAR pour obtenir des coefficients d'impulsion.

Dans le deuxième essai, nous proposons un nouveau modèle de volatilité multivarié, le modèle *Regime Switching Dynamic Correlation*. Nous décomposons les covariances en corrélations et écarts types. La matrice de corrélation suit un modèle à changement de régime : elle est constante à l'intérieur d'un régime, mais différentes d'un régime à un autre. Les transitions entre les régimes suivent une chaîne de Markov. Ce modèle ne souffre pas d'une malédiction de la dimension et il permet le calcul analytique d'espérance conditionnelle à plusieurs horizons pour la matrice de corrélation et la matrice de variance. Nous illustrons également au moyen d'une application empirique que ce modèle peut avoir une performance inter-échantillon supérieure à celle du modèle *Dynamic Conditional Correlation* de Engle (2002).

Dans le troisième essai, nous présentons des méthodes afin de tester des hypothèses de causalité à divers horizons, tel que défini dans Dufour et Renault (1998). Nous étudions en détail le cas des modèles VAR et nous proposons des méthodes basées sur des autorégressions à différents horizons. Bien que les hypothèses de non-causalité considérées soient non linéaires, ces méthodes ne requièrent que des méthodes de régression linéaire et la théorie asymptotique gaussienne habituelle. Nous les appliquons à un modèle VAR de l'économie américaine.

Dans le quatrième essai, nous proposons une méthode de tests statistiques pour l'évaluation des modèles utilisés pour calculer la Valeur-à-Risque d'un portefeuille. Les modèles de volatilité, tels que celui dont il est question dans le deuxième essai, servent notamment à construire la distribution des rendements d'un portefeuille pour un nombre donné de périodes dans l'avenir. Cette distribution est l'ingrédient essentiel aux calculs de la VaR et il est important d'évaluer si elle est bien spécifiée. Les tests que nous proposons sont basés sur les durées, calculées en nombre de jours, entre les violations de la VaR. À l'aide de simulations Monte Carlo, nous montrons qu'en situation réaliste, ces tests ont plus de puissance que les tests avancés précédemment, notamment ceux de Christoffersen (1998).

L'étude des séries chronologiques multivariées et des problèmes dus à la malédiction de la dimension est un sujet de recherche intéressant. Les résultats présentés dans les différents essais de cette thèse ouvrent la voie à d'autres avenues de recherche. Par exemple, nous pouvons revisiter avec la méthodologie VARMA présentée dans le premier essai de nombreux résultats obtenus en macroéconomie à l'aide de modèles VAR, entre autre les tests de causalité à plusieurs horizons dans les modèles VARMA. On pourrait améliorer les résultats présentés dans le deuxième essai, i.e. notre modèle de volatilité multivarié, en permettant un lien entre les corrélations et les écarts types. Le défi est de le faire sans introduire une malédiction de la dimension.