Université de Montréal

PRÉVISIONS ROBUSTES POUR SÉRIES TEMPORELLES MULTIVARIÉES

par

Christian Gagné

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en statistique

juillet 2007

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

PRÉVISIONS ROBUSTES POUR SÉRIES TEMPORELLES MULTIVARIÉES

présenté par

Christian Gagné

a été évalué par un jury composé des personnes suivantes :

Roch Roy

(président-rapporteur)

Pierre Duchesne

(directeur de recherche)

Martin Bilodeau

(membre du jury)

Mémoire accepté le:

_____

SOMMAIRE

Dans ce mémoire, la problématique abordée est celle de la prévision dans les modèles de séries temporelles autorégressifs vectoriels avec variables exogènes (VARX). Les estimateurs des moindres carrés conditionnels (MCC) ne sont pas robustes en présence de valeurs aberrantes. Pour obtenir des estimateurs robustes, la méthode introduite par Duchesne (2005) ainsi que Bou Hamad et Duchesne (2005) est généralisée pour les modèles VARX. La distribution asymptotique des nouveaux estimateurs est étudiée pour obtenir en particulier la matrice de variance asymptotique des nouveaux estimateurs robustes, que nous dénoterons RA-VARX (pour *robust autocovariance estimators* dans la classe des modèles VARX). Les intervalles de prévision usuels sont basés sur les estimateurs non robustes MCC. En présence de valeurs aberrantes, par exemple celles de type additif, ces prévisions peuvent être fortement biaisées. Plus précisément, l'apparition de valeurs aberrantes peut rendre les intervalles de prévision conditionnels usuels invalides. La nouvelle méthodologie robuste est utilisée pour développer des intervalles de prévision conditionnels robustes qui tiennent en considération la variabilité associée à l'estimation des paramètres. Cette source de variation peut être considérable, même dans un échantillon de taille modérée. Dans une étude de simulation, les propriétés en échantillons finis des intervalles de prévision conditionnels robustes sont étudiées sous différents scénarios de contamination pour les valeurs aberrantes. Les nouveaux intervalles conditionnels robustes sont comparés aux intervalles conditionnels basés sur les estimateurs MCC.

Mots clés : Séries temporelles multivariées, intervalles de prévision, robustesse.

SUMMARY

In this article, robust estimation and prediction in multivariate autoregressive models with exogenous variables (VARX) are considered. The conditional least squares estimators (CLS) are known to be non robust when outliers occur. To obtain robust estimators, the method introduced in Duchesne (2005) and Bou Hamad and Duchesne (2005) is generalized for VARX models. The asymptotic distribution of the new estimators is studied and from this is obtained in particular the asymptotic covariance matrix of the robust estimators which we will call RA-VARX (for *robust autocovariance estimators* in the VARX class of models). Classical conditional prediction intervals normally rely on estimators such as the usual non robust CLS estimators. In the presence of outliers, such as additive outliers, these classical predictions can be severely biased. More generally, the occurrence of outliers may invalidate the usual conditional prediction intervals. Consequently, the new robust methodology is used to develop robust conditional prediction intervals which take into account parameter estimation uncertainty. In a simulation study, we investigate the finite sample properties of the robust conditional prediction intervals under several scenarios for the occurrence of the outliers, and the new intervals are compared to non-robust intervals based on classical CLS estimators.

Keywords : Multivariate time series, prediction intervals, robust.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

Mes remerciements vont d'abord et avant tout à mon directeur de recherche qui par son dynamisme et son enthousiasme m'a permi de compléter avec succès mes études graduées. Son soutien moral et financier furent essentiels et je ne peux qu'exprimer ma gratitude la plus sincère envers lui. Il a permi à ce projet d'être soumis dans un journal scientifique reconnu et sa contribution à l'élaboration du projet a permi d'en faire une étude plus complète.

INTRODUCTION

On peut s'intéresser à la prévision de certains aspects de l'évolution d'un système, aspects qui pourraient consister en une seule mesure. Par contre, il est fort possible que de résumer l'évolution d'un système en une seule mesure risque d'être insuffisant. Dans ces conditions, on pourrait croire qu'un ensemble de mesures sur ce système permettrait de mieux suivre son évolution. Prévoir un ensemble de mesures grâce à une structure linéaire est ce que permet les modèles linéaires multivariés. Les propriétés statistiques de ces modèles sont exposées dans Lütkepohl (2005), entre autres. Un processus rencontré dans les applications économiques et physiques est le modèle autorégressif vectoriel avec variables exogènes (VARX).

Dans les applications pratiques, les paramètres d'un modèle VARX peuvent être estimés avec la méthode des moindres carrés conditionnels, par les moindres carrés généralisés ou encore par maximum de vraisemblance. La méthode des moindres carrés conditionnels (MCC) sera exposée dans le prochain chapitre. Voir Lütkepohl (2005) pour une description des autres méthodes.

Par contre, les estimateurs MCC (ainsi que les méthodes similaires) sont très sensibles à la présence de valeurs aberrantes. Par exemple, les estimateurs MCC peuvent être fortement biaisés lorsque les observations ne sont pas tirées du processus VARX que nous dénoterons $\{\mathbf{Y}_t\}$ mais du processus $\{\mathbf{Y}_t + \mathbf{W}_t\}$, disons, où $\{\mathbf{Y}_t\}$ et $\{\mathbf{W}_t\}$ sont indépendants avec $\{\mathbf{W}_t\}$ une séquence indépendante et identiquement distribuée (iid) qui génère les valeurs aberrantes (souvent avec une faible probabilité). Les estimateurs MCC sont très sensibles aux valeurs aberrantes additives, ces dernières étant appelées ainsi puisqu'elles affectent la variable endogène $\mathbf{Y}_t$ de façon additive.

Le but de ce mémoire est de proposer des intervalles de prévision robustes pour les modèles VARX. Même si l'estimation robuste a déjà été étudiée dans la littérature des séries temporelles, la construction d'intervalles de prévision robustes a été peu étudiée, au mieux de notre connaissance. Puisque les estimateurs MCC peuvent être biaisés lors de l'apparition des valeurs aberrantes, on peut s'attendre à ce qu'une prévision ponctuelle basée sur ces mêmes estimateurs soit aussi biaisée. De plus, les intervalles de prévision sont aussi affectés par les valeurs aberrantes puisqu'ils sont calculés avec l'erreur quadratique moyenne (EQM) de prévision, qui est une quantité habituellement estimée avec des méthodes non-robustes. Ainsi, les longueurs de ces intervalles de prévision seront possiblement beaucoup trop larges, ce qui n'est pas souhaitable en pratique. Ce phénomème est connu pour les modèles autorégressifs intégrés moyennes mobiles (dits ARIMA pour *autoregressive integrated moving average*), voir Ledolter (1989) et Chatfield (2001), qui ont présenté l'effet des aberrants additifs sur les intervalles de prévision. Ces considérations pratiques suggèrent le développement d'estimateurs robustes pour les modèles VARX, ainsi que la construction d'intervalles robustes de prévision.

Ainsi, ce mémoire présentera le développement d'estimateurs robustes pour les modèles VARX. La méthode proposée est une généralisation multivariée des méthodes d'estimation robustes proposées par Duchesne (2005) et Bou Hamad et Duchesne (2005) pour les modèles autorégressifs univariés avec variables exogènes (ARX). La distribution asymptotique sera étudiée. En utilisant un argument similaire à celui donné par Bou Hamad et Duchesne (2005), mais adapté au cas multivarié et en utilisant le théorème de Cramér-Wold, les estimateurs proposés convergent en distribution vers une loi normale. La matrice de variance asymptotique est explicitement présentée, ce qui sera une quantité essentielle pour la construction d'intervalles de prévision.

Dans le contexte des séries temporelles, il est d'usage dans les applications avec données réelles d'ignorer la variabilité due à l'estimation des paramètres. En s'inspirant des travaux de Yamamoto (1981), les intervalles construits dans ce mémoire tiendront aussi compte de la variabilité due à l'estimation des paramètres.

En effet, comme il a été démontré par Schmidt (1977), ignorer la variabilité due à l'estimation des paramètres dans la construction d'intervalles de prévision peut donner lieu à des intervalles de prévision beaucoup trop courts, particulièrement pour des échantillons de petites tailles.

Un des aspects étudiés dans ce mémoire est le comportement en échantillons finis des prévisions ponctuelles ainsi que des intervalles de prévision, classiques et robustes, dans divers scénarios de contamination par des valeurs aberrantes. Leurs propriétés asymptotiques sont connues, dans des conditions idéales (par exemple, lorsque le système peut être décrit par un modèle VARX parfaitement observé). L'apparition de valeurs aberrantes peut se faire selon une vaste panoplie de scénarios. Ainsi, il est très pertinent d'un point de vue pratique d'évaluer le comportement des estimateurs, des prévisions et des intervalles de prévision tirés de ces estimateurs lorsque les tailles échantillonnales sont modérées ou petites et lorsqu'il y a présence de valeurs aberrantes. Les choix des stratégies de contamination sont très importants et devraient refléter des problèmes potentiels en pratique. Les simulations démontreront que les estimateurs robustes produisent des prévisions beaucoup plus stables en présence de valeurs aberrantes. De plus, les intervalles de prévision développés à partir de ces estimateurs se verront comparés à ceux obtenus avec les estimateurs MCC pour constater qu'ils produisent des taux de couverture beaucoup plus près des valeurs nominales.

Chapitre 1

PRÉLIMINAIRES

1.1. Série temporelle multivariée

Dans cette section, des notions générales sur les processus linéaires multivariés sont présentées. Plusieurs notions fondamentales introduites dans le cas d'une seule variable, c'est-à-dire dans le cas univarié, trouvent leurs extensions dans le cas de plusieurs variables composant le système, cas que nous appellons multivarié. Pour une exposition des notions sur les processus linéaires univariés, voir par exemple Brockwell et Davis (2002). Parmi les concepts fondamentaux, nous décrirons les propriétés de stationnarité et de stabilité des processus vectoriels.

1.1.1. Processus stationnaire

Nous commençons dans un premier temps par définir ce qu'est une série temporelle.

Définition 1. *Une série temporelle multivariée est une réalisation finie de taille n d'un processus stochastique multivarié $\mathbf{Y} = \{\mathbf{Y}_t, t \in \mathbb{Z}\}$, où le vecteur aléatoire $\mathbf{Y}_t = (Y_t(1), \ldots, Y_t(d))^\top$ est de dimension d et $Y_t(i)$ représente la $i^{\text{ième}}$ composante de $\mathbf{Y}_t$.*

Les propriétés du deuxième ordre d'un processus stochastique sont spécifiées par le vecteur des moyennes:

$$\boldsymbol{\mu}_t = E(\mathbf{Y}_t) = (\mu_t(1), \ldots, \mu_t(d))^\top,$$

et les matrices de covariance,

$$\mathbf{\Gamma}(t, t-h) = E\{(\mathbf{Y}_t - \boldsymbol{\mu}_t)(\mathbf{Y}_{t-h} - \boldsymbol{\mu}_{t-h})^\top\} = [\gamma_{ij}(t, t-h)]_{i,j=1}^d.$$

Les covariances croisées $\gamma_{ij}(t, t-h)$ donnent non seulement une indication de la dépendance entre les observations d'une même série (lorsque $i = j$), mais aussi entre les observations de différentes séries (lorsque $i \neq j$).

Une classe de processus importante possède une propriété qui garantit une certaine uniformité dans le temps. Cette propriété est dite celle de stationnarité.

Définition 2. *Un processus est stationnaire au second ordre si les deux premiers moments restent invariants par rapport à l'indice temporel t. Plus précisément:*

$$E(\mathbf{Y}_t) = \boldsymbol{\mu}, \ \forall t,$$

$$\mathbf{\Gamma}(t, t-h) = E\left[\{\mathbf{Y}_t - E(\mathbf{Y}_t)\}\{\mathbf{Y}_{t-h} - E(\mathbf{Y}_{t-h})\}^\top\right] = \mathbf{\Gamma}(h), \ \forall h.$$

De plus, on présume que $E\{Y_t^2(i)\}$ est finie, $i = 1, \ldots, d$.

Un processus stationnaire fondamental est le *bruit blanc.*

Définition 3. *Le processus $\{\mathbf{a}_t\}$ est dit bruit blanc si:*

$$E(\mathbf{a}_t) = \mathbf{0}, \ E(\mathbf{a}_t\mathbf{a}_t^\top) = \Sigma_{\mathbf{a}}, \ E(\mathbf{a}_t\mathbf{a}_s^\top) = \mathbf{0}, \ \forall t \neq s.$$

En général, il n'est pas nécessaire que les éléments soient indépendants mais seulement non-corrélés. Lorsque les vecteurs aléatoires $\mathbf{a}_t$ sont non-corrélés mais dépendants, nous parlons alors de bruit blanc faible. Si les éléments de la séquence sont aussi indépendants alors on dit que le bruit blanc est fort.

Pour caractériser la dépendance linéaire d'un processus, on utilise la matrice d'autocorrélation qui résume l'intensité de la dépendance linéaire entre les éléments d'un processus.

Définition 4. *La matrice d'autocorrélation au délai k, $\mathbf{R}(k)$, est donnée par:*

$$\mathbf{R}(k) = \mathbf{D}^{-1}\mathbf{\Gamma}(k)\mathbf{D}^{-1},$$

où

$$\mathbf{D}^{-1} = \begin{pmatrix} 1/\sqrt{\gamma_{11}(0)} & & 0 \\ & \ddots & \\ 0 & & 1/\sqrt{\gamma_{dd}(0)} \end{pmatrix},$$

et la matrice $\mathbf{\Gamma}(k)$ *est définie dans la Définition 2.*

De façon plus précise, les éléments de la matrice $\mathbf{R}(h)=[\rho_{ij}(h)]_{i,j=1}^{d}$ sont:

$$\rho_{ij}(h) = \frac{\gamma_{ij}(h)}{\gamma_{ii}(0)\gamma_{jj}(0)},$$

où

$$\gamma_{ij}(h) = \mathrm{Cov}\{Y_t(i), Y_{t-h}(j)\}.$$

Ayant exposé la notion de stationnarité, on peut maintenant introduire une notion plus restrictive qui permet d'assurer la stationnarité lorsque certaines conditions sont satisfaites. Cette notion est appellée la stabilité d'une série temporelle.

1.1.2. Processus VAR(p) stable

Dans la classe des processus stationnaires, il existe une sous-classe importante. Les processus linéaires stables sont stationnaires et la condition qui les caractérisent est plus facilement vérifiable. Un processus peut être non-stable tout en étant stationnaire mais ces cas ne sont pas présentés ici. Nous commençons par définir dans un premier temps les processus autorégressifs vectoriels d'ordre p.

Définition 5. *Un processus multivarié est dit vectoriel autorégressif d'ordre p, noté VAR(p), si:*

$$\mathbf{Y}_t = \boldsymbol{\nu} + \mathbf{\Phi}_1\mathbf{Y}_{t-1} + \mathbf{\Phi}_2\mathbf{Y}_{t-2} + \ldots + \mathbf{\Phi}_p\mathbf{Y}_{t-p} + \mathbf{a}_t, \ \forall t,$$

où les matrices $\mathbf{\Phi}_i$, $i = 1, ..., p$ *sont les paramètres autorégressifs,* $\boldsymbol{\nu}$ *est un vecteur de dimension d,* $\{\mathbf{a}_t\}$ *est un bruit blanc et* $\mathbf{Y}_t$ *est le vecteur observé au temps t.*

1.1.2.1. *Processus VAR(1)*

Un cas particulier que nous examinons de plus près est le processus VAR(1):

$$\mathbf{Y}_t = \boldsymbol{\nu} + \boldsymbol{\Phi}_1 \mathbf{Y}_{t-1} + \mathbf{a}_t. \tag{1.1.1}$$

En procédant à des remplacements successifs, il est possible d'obtenir une expression pour $\mathbf{Y}_t$ en fonction des erreurs:

$$\begin{aligned}
\mathbf{Y}_1 &= \boldsymbol{\nu} + \boldsymbol{\Phi}_1 \mathbf{Y}_0 + \mathbf{a}_1, \\
\mathbf{Y}_2 &= \boldsymbol{\nu} + \boldsymbol{\Phi}_1 \mathbf{Y}_1 + \mathbf{a}_2 = \boldsymbol{\nu} + \boldsymbol{\Phi}_1(\boldsymbol{\nu} + \boldsymbol{\Phi}_1 \mathbf{Y}_0 + \mathbf{a}_1) + \mathbf{a}_2, \\
&= (\mathbf{I}_d + \boldsymbol{\Phi}_1)\boldsymbol{\nu} + \boldsymbol{\Phi}_1 \mathbf{a}_1 + \mathbf{a}_2 + \boldsymbol{\Phi}_1^2 \mathbf{Y}_0.
\end{aligned}$$

En continuant les récursions, on peut représenter $\{\mathbf{Y}_t\}$ de la façon suivante:

$$\mathbf{Y}_t = (\mathbf{I}_d + \boldsymbol{\Phi}_1 + \ldots + \boldsymbol{\Phi}_1^{t-1})\boldsymbol{\nu} + \boldsymbol{\Phi}_1^t \mathbf{Y}_0 + \sum_{i=0}^{t-1} \boldsymbol{\Phi}_1^i \mathbf{a}_{t-i}.$$

À la limite, il est possible de donner un sens précis au processus suivant:

$$\mathbf{Y}_t = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \boldsymbol{\Phi}_1^i \mathbf{a}_{t-i}, \tag{1.1.2}$$

avec

$$\boldsymbol{\mu} = (\mathbf{I}_d - \boldsymbol{\Phi}_1)^{-1} \boldsymbol{\nu}, \tag{1.1.3}$$

où $\boldsymbol{\mu}$ représente l'espérance du processus $\{\mathbf{Y}_t\}$. Cette représentation sous forme de somme infinie permet d'établir des conditions dites de stabilité afin que la somme infinie dans le membre de droite de (1.1.2) converge et afin que l'espérance (1.1.3) existe. Cette forme est aussi la représentation dite moyenne mobile d'un processus VAR(1).

Ayant défini les processus VAR(p), il est particulièrement simple de fournir la condition de stabilité d'un processus VAR(1).

Définition 6. *Un processus VAR(1) est dit stable si les valeurs propres de* $\boldsymbol{\Phi}_1$ *sont toutes inférieures à 1 en module. De manière équivalente, il faut vérifier que:*

$$\det(\mathbf{I}_d - \boldsymbol{\Phi}_1 z) \neq 0, \ |z| < 1.$$

Dans le cas d'un processus stable, (1.1.2) converge car les coefficients $\boldsymbol{\Phi}_1^i$ sont absolument sommables, c'est-à-dire $\sum_{i=0}^{\infty} |\boldsymbol{\Phi}_1^i| < \infty$. De plus, cette condition

garantit l'existence de l'espérance. Notons que la condition de stabilité est équivalente à la notion de causalité rencontrée en économétrie ainsi que dans l'analyse des séries temporelles.

Il est possible d'écrire le processus plus général VAR(p) de façon plus compacte de sorte qu'il soit possible de récupérer les résultats obtenus du processus VAR(1). Ainsi, un processus VAR(p) s'exprime comme:

$$\mathbf{Y}_t^* = \boldsymbol{\nu}^* + \boldsymbol{\Phi}^* \mathbf{Y}_{t-1}^* + \mathbf{a}_t^*, \tag{1.1.4}$$

avec

$$\mathbf{Y}_t^* = \begin{pmatrix} \mathbf{Y}_t \\ \mathbf{Y}_{t-1} \\ \vdots \\ \vdots \\ \mathbf{Y}_{t-p+1} \end{pmatrix}, \quad \mathbf{Y}_{t-1}^* = \begin{pmatrix} \mathbf{Y}_{t-1} \\ \mathbf{Y}_{t-2} \\ \vdots \\ \vdots \\ \mathbf{Y}_{t-p} \end{pmatrix},$$

$$\mathbf{a}_t^* = \begin{pmatrix} \mathbf{a}_t \\ \mathbf{0} \\ \vdots \\ \vdots \\ \mathbf{0} \end{pmatrix}, \qquad \boldsymbol{\nu}^* = \begin{pmatrix} \boldsymbol{\nu} \\ \mathbf{0} \\ \vdots \\ \vdots \\ \mathbf{0} \end{pmatrix},$$

$$\boldsymbol{\Phi}^* = \begin{pmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_2 & \dots & \boldsymbol{\Phi}_{p-1} & \boldsymbol{\Phi}_p \\ \mathbf{I}_d & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_d & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{I}_d & \mathbf{0} \end{pmatrix}.$$

Ceci est la représentation VAR(1) d'un VAR(p). De plus, sous l'hypothèse de stabilité, on peut exprimer $\mathbf{Y}_t^*$ en utilisant la représentation suivante:

$$\mathbf{Y}_t^* = \boldsymbol{\mu}^* + \sum_{i=0}^{\infty} \boldsymbol{\Phi}^{*i} \mathbf{a}_{t-i}. \tag{1.1.5}$$

De la même façon que la condition de stabilité a été établie pour un processus VAR(1), on peut établir cette condition pour un processus VAR(p).

Définition 7. *Un processus VAR(p) est dit stable si les valeurs propres de* $\mathbf{\Phi}^*$ *sont inférieures à* 1 *en module ou de façon équivalente si* $det(\mathbf{I}-\mathbf{\Phi}^* z) \neq 0, |z| < 1$.

Maintenant, si l'on veut une expression pour le vecteur original $\mathbf{Y}_t$, on introduit l'opérateur:

$$\mathbf{E}_i = \mathbf{e}_i \otimes \mathbf{I}_d.$$

où $\mathbf{e}_i$ est un vecteur de 0 sauf en position i qui contient un 1. En particulier, on considère l'opérateur $\mathbf{E}_1^\top = (\mathbf{I}_d : \mathbf{0} : \ldots : \mathbf{0})$.

Cet opérateur extrait les d premières composantes du vecteur ou matrice auquel il est multiplié. On peut donc utiliser les formules précédentes tirées de la représentation compacte VAR(1) d'un processus VAR(p) afin d'obtenir une représentation de $\mathbf{Y}_t$ en fonction des erreurs:

$$\mathbf{Y}_t = \mathbf{E}_1^\top \mathbf{Y}_t^* = \mathbf{E}_1^\top \boldsymbol{\mu}^* + \mathbf{E}_1^\top \sum_{i=0}^{\infty} \mathbf{\Phi}^{*i} \mathbf{a}_{t-i}^*.$$

Il faut noter que la stabilité est une caractéristique qui implique la stationnarité mais l'inverse n'est pas vrai.

1.1.3. La représentation moyenne mobile d'un processus VAR

La section précédente a introduit la représentation VAR(1) d'un VAR(p) donnée par (1.1.4) de laquelle la représentation (1.1.5) a été obtenue. Cette dernière forme (1.1.5) peut être utilisée pour déterminer la fonction d'autocovariance ainsi que la moyenne et l'erreur de prévision.

Théorème 1.1.1. *Un processus VAR(p) peut s'écrire de la façon suivante:*

$$\mathbf{Y}_t = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \boldsymbol{\phi}_i^* \mathbf{a}_{t-i}, \tag{1.1.6}$$

où $\boldsymbol{\phi}_i^* = \mathbf{E}_1^\top \mathbf{\Phi}^{*i} \mathbf{E}_1$ *et* $\mathbf{a}_{t-i} = \mathbf{E}_1^\top \mathbf{a}_{t-i}^*$.

DÉMONSTRATION. Premièrement, on remarque que $\mathbf{a}_t^* = \mathbf{E}_1 \mathbf{E}_1^\top \mathbf{a}_t^*$ car $\mathbf{a}_t^* = \mathbf{e}_1 \otimes \mathbf{a}_t$, $\mathbf{E}_1^\top = (\mathbf{I}_d : \mathbf{0} : \ldots : 0)$ et $\mathbf{E}_1 \mathbf{E}_1^\top = (\mathbf{e}_1 \otimes \mathbf{I}_d)(\mathbf{e}_1^\top \otimes \mathbf{I}_d) = \mathbf{e}_1 \mathbf{e}_1^\top \otimes \mathbf{I}_d$. Ainsi,

$$\mathbf{E}_1 \mathbf{E}_1^\top \mathbf{a}_t^* = (\mathbf{e}_1 \mathbf{e}_1^\top \otimes \mathbf{I}_d)(\mathbf{e}_1 \otimes \mathbf{a}_t) = \mathbf{e}_1 \otimes \mathbf{a}_t = \mathbf{a}_t^*.$$

On peut donc écrire le vecteur des valeurs observées comme:

$$\mathbf{Y}_t = \mathbf{E}_1^\top \mathbf{Y}_t^* = \mathbf{E}_1^\top \boldsymbol{\mu}^* + \sum_{i=0}^{\infty} \mathbf{E}_1^\top \boldsymbol{\Phi}^{*i} \mathbf{E}_1 \mathbf{E}_1^\top \mathbf{a}_{t-i}^*.$$

Maintenant, on pose $\boldsymbol{\phi}_i^* = \mathbf{E}_1^\top \boldsymbol{\Phi}^{*i} \mathbf{E}_1$ et $\mathbf{a}_{t-i} = \mathbf{E}_1^\top \mathbf{a}_{t-i}^*$. Nous obtenons ainsi $\mathbf{Y}_t = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \boldsymbol{\phi}_i^* \mathbf{a}_{t-i}$. □

On peut ainsi tirer plusieurs expressions pour caractériser la fonction d'autocovariance. Cette fonction est fondamentale puisque la stationnarité se décrit par les propriétés du deuxième ordre.

1.1.4. La fonction d'autocovariance d'un processus VAR(p)

Comme présenté plus haut, la première expression pour la fonction d'autocovariance fait appel à la représentation moyenne mobile d'un processus VAR(p).

Théorème 1.1.2. *La fonction d'autocovariance* $\boldsymbol{\Gamma}(h)$ *d'un processus VAR(p) est donnée par*:

$$\boldsymbol{\Gamma}(h) = \sum_{i=0}^{\infty} \boldsymbol{\phi}_{h+i}^* \boldsymbol{\Sigma}_{\mathbf{a}} \boldsymbol{\phi}_i^{*\top}. \tag{1.1.7}$$

DÉMONSTRATION.

$$\begin{aligned} \boldsymbol{\Gamma}(h) &= \mathrm{E}\left\{(\mathbf{Y}_t - \boldsymbol{\mu})(\mathbf{Y}_{t-h} - \boldsymbol{\mu})^\top\right\}, \\ &= \mathrm{E}\left\{\left(\textstyle\sum_{i=0}^{h-1} \boldsymbol{\phi}_i^* \mathbf{a}_{t-i} + \sum_{i=0}^{\infty} \boldsymbol{\phi}_{h+i}^* \mathbf{a}_{t-h-i}\right)\left(\sum_{i=0}^{\infty} \boldsymbol{\phi}_i^* \mathbf{a}_{t-h-i}\right)^\top\right\}, \\ &= \textstyle\sum_{i=0}^{\infty} \boldsymbol{\phi}_{h+i}^* \boldsymbol{\Sigma}_{\mathbf{a}} \boldsymbol{\phi}_i^{*\top}. \end{aligned}$$

□

On peut voir que la condition de stabilité assure que la fonction d'autocovariance existe car la somme infinie devient sommable.

1.1.4.1. *Calcul de la fonction d'autocorrélation*

L'expression (1.1.7) est utile pour voir l'influence de la stabilité sur un processus mais n'est pas pratique lorsque vient le temps de calculer cette fonction compte tenu de la somme infinie. Il est cependant possible de développer des expressions récursives pour cette fonction.

Théorème 1.1.3. *La fonction d'autocovariance d'un processus VAR(1) satisfait la forme récursive suivante:*

$$\boldsymbol{\Gamma}(k) = \boldsymbol{\Phi}_1 \boldsymbol{\Gamma}(k-1), \ \forall\ k \geq 1.$$

DÉMONSTRATION. En utilisant la relation (1.1.3), nous obtenons $\boldsymbol{\nu} = (\mathbf{I}_d - \boldsymbol{\Phi}_1)\boldsymbol{\mu}$. En multipliant (1.1.1) par $\mathbf{Y}_{t-k} - \boldsymbol{\mu}$ nous obtenons:

$$(\mathbf{Y}_t - \boldsymbol{\mu})(\mathbf{Y}_{t-k} - \boldsymbol{\mu})^\top = \boldsymbol{\Phi}_1(\mathbf{Y}_{t-1} - \boldsymbol{\mu})(\mathbf{Y}_{t-k} - \boldsymbol{\mu})^\top + \mathbf{a}_t(\mathbf{Y}_{t-k} - \boldsymbol{\mu})^\top.$$

En prenant l'espérance et en utilisant la représentation moyenne mobile du processus, on obtient le résultat désiré:

$$\begin{aligned} \boldsymbol{\Gamma}(k) &= \boldsymbol{\Phi}_1 \boldsymbol{\Gamma}(k-1) + E\{\mathbf{a}_t(\mathbf{Y}_{t-k} - \boldsymbol{\mu})^\top\}, \\ &= \boldsymbol{\Phi}_1 \boldsymbol{\Gamma}(k-1). \end{aligned}$$

□

Le théorème suivant donne la variance de $\mathbf{Y}_t$.

Théorème 1.1.4. $Var(\mathbf{Y}_t) = \boldsymbol{\Gamma}(0) = \boldsymbol{\Phi}_1 \boldsymbol{\Gamma}(0) \boldsymbol{\Phi}_1^\top + \boldsymbol{\Sigma}_\mathbf{a}$.

DÉMONSTRATION. En effet, en utilisant la définition de la variance de $\mathbf{Y}_t$ et en utilisant encore une fois l'écriture en fonction du bruit blanc:

$$\begin{aligned} E\left\{(\mathbf{Y}_t - \boldsymbol{\mu})(\mathbf{Y}_t - \boldsymbol{\mu})^\top\right\} &= \boldsymbol{\Phi}_1 E\left\{(\mathbf{Y}_{t-1} - \boldsymbol{\mu})(\mathbf{Y}_t - \boldsymbol{\mu})^\top\right\} + E\left\{\mathbf{a}_t(\mathbf{Y}_t - \boldsymbol{\mu})^\top\right\}, \\ &= \boldsymbol{\Phi}_1 E\left\{(\mathbf{Y}_{t-1} - \boldsymbol{\mu})(\boldsymbol{\Phi}_1(\mathbf{Y}_{t-1} - \boldsymbol{\mu}) + \mathbf{a}_t)^\top\right\} \\ &\quad + E\left\{\mathbf{a}_t[\boldsymbol{\Phi}_1(\mathbf{Y}_{t-1} - \boldsymbol{\mu}) + \mathbf{a}_t]^\top\right\}, \\ &= \boldsymbol{\Phi}_1 \boldsymbol{\Gamma}(0) \boldsymbol{\Phi}_1^\top + \boldsymbol{\Sigma}_\mathbf{a}. \end{aligned}$$

□

Nous pouvons dégager la fonction d'autocovariance d'un processus VAR(p). Ceci fait l'objet du théorème suivant.

Théorème 1.1.5. *La fonction d'autocovariance d'un processus VAR(p) s'écrit de façon récursive* $\boldsymbol{\Gamma}(k) = \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{\Gamma}(k-i)$.

DÉMONSTRATION. On multiplie par $(\mathbf{Y}_{t-k} - \boldsymbol{\mu})$ l'équation $\mathbf{Y}_t = \boldsymbol{\nu} + \boldsymbol{\Phi}_1 \mathbf{Y}_{t-1} + \boldsymbol{\Phi}_2 \mathbf{Y}_{t-2} + \ldots + \boldsymbol{\Phi}_p \mathbf{Y}_{t-p} + \mathbf{a}_t$ pour prendre ensuite l'espérance. □

Les récursions dans le Théorème 1.1.3 et la formule pour $\mathbf{\Gamma}(0)$ dans le Théorème 1.1.4 composent ce qui est souvent appelé le système d'équations de Yule-Walker pour un processus VAR(1). Plus précisément, les équations de Yule-Walker d'un processus VAR(p) sont:

$$\mathbf{\Gamma}(0) = \mathbf{\Phi}_1\mathbf{\Gamma}^\top(1) + \cdots + \mathbf{\Phi}_p\mathbf{\Gamma}^\top(p) + \mathbf{\Sigma}_\mathbf{a},$$

et

$$\mathbf{\Gamma}(h) = \mathbf{\Phi}_1\mathbf{\Gamma}(h-1) + \cdots + \mathbf{\Phi}_p\mathbf{\Gamma}(h-p), \ h > 0.$$

Ces formules permettent d'obtenir $\mathbf{\Gamma}(h)$ pour $h \geq p$, si on connaît déjà les coefficients autorégressifs et si $\mathbf{\Gamma}(p-1), \ldots, \mathbf{\Gamma}(0)$ sont déterminées. Afin de trouver les valeurs initiales $\mathbf{\Gamma}(h)$, $h = 0, 1, \ldots, p-1$, Lütkepohl (2005) décrit un algorithme reposant sur la représentation VAR(1) d'un processus VAR(p). Voir Lütkepohl (2005, pp. 28-29) pour plus de détails.

1.1.5. L'opérateur retard

Il est possible de déterminer les matrices coefficients $\boldsymbol{\phi}_i^*$ de façon plus directe, sans faire usage de la représentation VAR(1) d'un processus VAR(p). Pour ce faire, il faut introduire l'opérateur retard B qui est tel que $B\mathbf{Y}_t = \mathbf{Y}_{t-1}$ et $B^k\mathbf{Y}_t = \mathbf{Y}_{t-k}$. Ceci permet d'écrire:

$$\mathbf{\Phi}(B)\mathbf{Y}_t = \boldsymbol{\nu} + \mathbf{a}_t, \tag{1.1.8}$$

où $\mathbf{\Phi}(B) = \mathbf{I}_d - \mathbf{\Phi}_1 B - \mathbf{\Phi}_2 B^2 - \ldots - \mathbf{\Phi}_p B^p$.

Définition 8. *Soit* $\mathbf{\Phi}(B) = \mathbf{I}_d - \mathbf{\Phi}_1 B - \mathbf{\Phi}_2 B^2 - \ldots - \mathbf{\Phi}_p B^p$, *alors l'opérateur inverse de* $\mathbf{\Phi}(B)$ *est* $\mathbf{\Phi}^{-1}(B)$ *et est tel que* $\mathbf{\Phi}^{-1}(B)\mathbf{\Phi}(B) = \mathbf{I}_d$. *De plus,* $\mathbf{\Phi}^{-1}(B)$ *admet la représentation* $\mathbf{\Phi}^{-1}(B) = \sum_{i=0}^{\infty} \boldsymbol{\phi}_i^* B^i$.

Ainsi, en multipliant (1.1.8) par $\mathbf{\Phi}^{-1}(B)$, nous obtenons:

$$\begin{aligned} \mathbf{\Phi}^{-1}(B)\mathbf{\Phi}(B)\mathbf{Y}_t &= \mathbf{\Phi}^{-1}(B)\boldsymbol{\nu} + \mathbf{\Phi}^{-1}(B)\mathbf{a}_t, \\ \mathbf{Y}_t &= \sum_{i=0}^{\infty} \boldsymbol{\phi}_i^* \boldsymbol{\nu} + \sum_{i=0}^{\infty} \boldsymbol{\phi}_i^* \mathbf{a}_{t-i}, \end{aligned}$$

avec $\boldsymbol{\mu} = \sum_{i=0}^{\infty} \boldsymbol{\phi}_i^* \boldsymbol{\nu}$. Bien sûr, l'opérateur $\mathbf{\Phi}^{-1}(B)$ existe sous certaines conditions, que nous donnons dans le théorème suivant.

Théorème 1.1.6. *L'opérateur $\Phi(B)$ est inversible si $det\{\Phi(z)\} \neq 0$, $\forall z$ tel que $|z| < 1$.*

Une preuve du Théorème 1.1.6 se trouve dans Brockwell et Davis (1991, pp. 408-409). Cette condition est aussi celle qui définie la stabilité d'un processus VAR(p).

1.1.5.1. *Identification des poids de la représentation infinie*

Le Théorème 1.1.1 a introduit une représentation sous forme de somme infinie à travers la relation (1.1.6). Il a été dit que cette représentation existe sous des conditions de stabilité du processus. Les coefficients de cette représentation peuvent être obtenus facilement grâce à la représentation VAR(1) d'un processus VAR(p). Par contre, il pourrait être souhaitable de pouvoir obtenir les coefficients de la représentaion infinie directement en fonction de ceux de la représentation récursive. Les coefficients matriciels que nous avons notées $\boldsymbol{\phi}_i^*$ peuvent être calculés grâce à l'identification des coefficients des puissances de B dans le système suivant:

$$\begin{aligned} \mathbf{I}_d &= \boldsymbol{\phi}_0^*, \\ \mathbf{0} &= \boldsymbol{\phi}_1^* - \boldsymbol{\phi}_0^* \boldsymbol{\Phi}_1, \\ &\vdots \\ \mathbf{0} &= \boldsymbol{\phi}_i^* - \textstyle\sum_{j=1}^{i} \boldsymbol{\phi}_{i-j}^* \boldsymbol{\Phi}_j. \end{aligned}$$

Ainsi,

$$\begin{aligned} \boldsymbol{\phi}_0^* &= \mathbf{I}_d, \\ \boldsymbol{\phi}_i^* &= \textstyle\sum_{j=1}^{i} \boldsymbol{\phi}_{i-j}^* \boldsymbol{\Phi}_j. \end{aligned}$$

On pose $\boldsymbol{\Phi}_j = \mathbf{0}$ lorsque $j > p$.

En particulier, pour un modèle VAR(1), $\mathbf{Y}_t = \boldsymbol{\Phi}_1 \mathbf{Y}_{t-1} + \mathbf{a}_t$ et les équations précédentes entraînent que les coefficients $\boldsymbol{\phi}_i^* = \boldsymbol{\Phi}_1^i$. C'est le résultat qui avait été obtenu auparavant lors du développement en somme infinie d'un processus VAR(p) sous sa représentation VAR(1).

1.1.6. Processus VARX(p,s)

Le processus VARX est une extension des processus VAR auquel on permet d'ajouter des variables supplémentaires, disons $\mathbf{X}_t$, non-corrélées avec le processus d'erreur qui compose le modèle. De plus, on suppose que le processus $\{\mathbf{X}_t\}$ est stationnaire et sans perte de généralité, on peut présumer que le processus $\{\mathbf{X}_t\}$ est de moyenne $\mathbf{0}$. Ces variables sont dites exogènes et permettent d'expliquer la variable d'intérêt, dite endogène. Dans cette section, un processus multivarié autorégressif de dimension d et de délai p, avec m variables exogènes de délai s, est introduit. Plus précisément, le processus s'écrit de la manière suivante:

$$\mathbf{Y}_t = \boldsymbol{\nu} + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \mathbf{Y}_{t-i} + \sum_{i=0}^{s} \mathbf{V}_i \mathbf{X}_{t-i} + \mathbf{a}_t,$$

avec $\dim(\mathbf{Y}_t)=d \times 1$, $\dim(\mathbf{X}_t)=m \times 1$ et $\dim(\mathbf{a}_t)=d \times 1$. De plus, $\dim(\boldsymbol{\Phi}_i)=d \times d$, $\dim(\mathbf{V}_i)=d \times m$ et $\dim(\boldsymbol{\nu})=d \times 1$, où $\dim(\cdot)$ représente la dimension d'une matrice.

On peut écrire ce modèle sous une forme VARX(1,0):

$$\mathbf{Y}_t^* = \boldsymbol{\nu}^* + \boldsymbol{\Phi}^* \mathbf{Y}_{t-1}^* + \mathbf{V}^* \mathbf{X}_t^* + \mathbf{a}_t^*,$$

avec

$$\mathbf{V}^* = \begin{pmatrix} \mathbf{V}_0 & \mathbf{V}_2 & \dots & \mathbf{V}_{s-1} & \mathbf{V}_s \\ \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{I} & \mathbf{0} \end{pmatrix},$$

$$\mathbf{X}_t^* = \begin{pmatrix} \mathbf{X}_t \\ \mathbf{X}_{t-1} \\ \vdots \\ \vdots \\ \mathbf{X}_{t-s} \end{pmatrix}, \quad \boldsymbol{\nu}^* = \begin{pmatrix} \boldsymbol{\nu} \\ \vdots \\ \boldsymbol{\nu} \end{pmatrix}.$$

Ayant maintenant décrit les modèles considérés et décrit certaines de leurs propriétés, il est de mise de présenter la méthode utilisée pour estimer leurs

paramètres. Nous décrirons dans la prochaine section la méthode des moindres carrés conditionnels (MCC).

1.1.7. Estimation

Nous commençons par décrire l'estimation d'un processus VAR(p).

1.1.7.1. *Estimation VAR(p)*

L'estimation des paramètres peut se faire grâce à la méthode des moindres carrés. Le modèle usuel est réécrit de sorte que l'on obtient une forme semblable à un modèle linéaire de régression. Ainsi:

$$\mathbf{Y}_t = \boldsymbol{\nu} + \boldsymbol{\Phi}_1 \mathbf{Y}_{t-1} + \boldsymbol{\Phi}_2 \mathbf{Y}_{t-2} + \ldots + \boldsymbol{\Phi}_p \mathbf{Y}_{t-p} + \mathbf{a}_t,$$

peut s'écrire de façon plus compacte comme:

$$\mathbf{W} = \mathbf{BZ} + \mathbf{A},$$

avec

$$\begin{aligned}
\mathbf{W} &= (\mathbf{Y}_1, \ldots, \mathbf{Y}_n), \\
\mathbf{B} &= (\boldsymbol{\nu}, \boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p), \\
\mathbf{Z}_t &= (1, \mathbf{Y}_t^\top, \ldots, \mathbf{Y}_{n-p+1}^\top)^\top, \\
\mathbf{Z} &= (\mathbf{Z}_0, \ldots, \mathbf{Z}_{n-1}), \\
\mathbf{A} &= (\mathbf{a}_1, \ldots, \mathbf{a}_n).
\end{aligned}$$

En empilant les colonnes, on retrouve la forme standard avec laquelle il est particulièrement commode d'utiliser les méthodes habituelles pour obtenir les estimateurs. Ainsi, en vectorisant et en utilisant la propriété (A.2.2) dans l'appendice, on trouve:

$$\mathbf{w} = (\mathbf{Z}^\top \otimes \mathbf{I}_n)\boldsymbol{\beta} + \mathbf{a},$$

où

$$\begin{aligned} \boldsymbol{\beta} &= \text{vec}(\mathbf{B}), \\ \mathbf{w} &= \text{vec}(\mathbf{W}), \\ \mathbf{a} &= \text{vec}(\mathbf{A}). \end{aligned}$$

On obtient les estimateurs MCC en optimisant le critère:

$$S = \mathbf{a}^\top (\mathbf{I}_n \otimes \boldsymbol{\Sigma}_\mathbf{a})^{-1} \mathbf{a}, \tag{1.1.9}$$

$$= \text{tr}(\mathbf{A}^\top \boldsymbol{\Sigma}_\mathbf{a}^{-1} \mathbf{A}), \tag{1.1.10}$$

où $\mathbf{I}_n \otimes \boldsymbol{\Sigma}_\mathbf{a}$ est la matrice de variance du vecteur $\mathbf{a}$ et $\Sigma_\mathbf{a} = E(\mathbf{a}_t \mathbf{a}_t^\top)$. De plus, $\text{tr}(\cdot)$ représente la trace d'une matrice.

Le processus $\{\mathbf{a}_t\}$ étant un bruit blanc, ses composantes sont non-corrélées dans le temps. Cet aspect explique la forme diagonale de la matrice $\mathbf{I}_n \otimes \boldsymbol{\Sigma}_\mathbf{a}$. Les estimateurs s'expriment sous la forme:

$$\hat{\boldsymbol{\beta}} = \{(\mathbf{Z}\mathbf{Z}^\top)^{-1} \mathbf{Z} \otimes \mathbf{I}_n\} \mathbf{w}.$$

Les estimateurs des moindres carrés et les estimateurs des moindres carrés généralisés sont les mêmes pour ce modèle. En fait, il est à noter que le critère (1.1.9) correspond à un critère de type moindres carrés généralisés. Cependant, si on adopte le critère usuel des moindres carrés :

$$\tilde{S}(\boldsymbol{\beta}) = \mathbf{a}^\top \mathbf{a},$$

alors l'estimateur des moindres carrés de $\boldsymbol{\beta}$ obtenu en optimisant $\tilde{S}$ ou l'estimateur des moindres carrés généralisés optimisant S coïncident. Pour de plus amples détails, voir Lütkepohl (2005, pp. 70-71).

L'idée entourant la reformulation d'un modèle VAR sous la forme d'un modèle usuel de régression linéaire multiple se généralise facilement pour la construction d'estimateurs de type moindres carrés dans le cadre des modèles VARX. La prochaine section introduit les estimateurs MCC pour les modèles VARX(p, s).

1.1.7.2. *Estimation VARX(p,s)*

Pour obtenir les estimateurs MCC dans un modèle VARX(p,s), il suffit de réécrire les équations précédentes en tenant compte de la présence de la variable exogène :

$$\begin{aligned}\mathbf{Y}_t &= \boldsymbol{\nu} + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \mathbf{Y}_{t-i} + \sum_{j=0}^{s} \mathbf{V}_j \mathbf{X}_{t-j} + \mathbf{a}_t, \\ \mathbf{W} &= \mathbf{B}\mathbf{Z} + \mathbf{A},\end{aligned}$$

avec

$$\begin{aligned}\mathbf{W} &= (\mathbf{Y}_1, \ldots, \mathbf{Y}_n), \\ \mathbf{B} &= (\boldsymbol{\nu}, \boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p, \mathbf{V}_0, \ldots, \mathbf{V}_s), \\ \mathbf{Z}_t &= (1, \mathbf{Y}_t^\top, \ldots, \mathbf{Y}_{n-p+1}^\top, \mathbf{X}_t^\top, \ldots, \mathbf{X}_{n-s}^\top)^\top, \\ \mathbf{Z} &= (\mathbf{Z}_0, \ldots, \mathbf{Z}_{n-1}), \\ \mathbf{A} &= (\mathbf{a}_1, \ldots, \mathbf{a}_n).\end{aligned}$$

Maintenant, il suffit d'empiler les colonnes de ces matrices pour obtenir une forme vectorielle contenant l'ensemble des observations. Ainsi, toujours en vectorisant et en utilisant la propriété (A.2.2) dans l'appendice, on trouve:

$$\mathbf{w} = (\mathbf{Z}^\top \otimes \mathbf{I}_n)\boldsymbol{\beta} + \mathbf{a},$$

où

$$\begin{aligned}\boldsymbol{\beta} &= \text{vec}(\mathbf{B}), \\ \mathbf{w} &= \text{vec}(\mathbf{W}), \\ \mathbf{a} &= \text{vec}(\mathbf{A}).\end{aligned}$$

Ainsi, les estimateurs des moindres carrés pour un modèle VARX prennent la forme:

$$\hat{\boldsymbol{\beta}} = \{(\mathbf{Z}\mathbf{Z}^\top)^{-1}\mathbf{Z} \otimes \mathbf{I}_n\}\mathbf{w}.$$

Ce sont ces estimateurs qui seront utilisés dans l'article subséquent lorsque nous procéderons à l'estimation des paramètres d'un modèle VARX particulier, lorsque la technique des moindres carrés sera adoptée.

1.2. PRÉVISION

Malgré certaines similitudes entre les séries temporelles et l'analyse de régression, certaines différences fondamentales existent entre les deux champs d'étude. La méthode d'estimation des paramètres par les moindres carrés nous amène à formuler le modèle de série temporelle comme un modèle de régression, ce qui peut sembler une approche naturelle lorsque l'on parle d'un modèle autorégressif.

Cependant, une différence survient lorsque l'on considère le calcul des prévisions, au moins d'un point de vue pratique. Pour le calcul des prévisions selon un certain modèle de régression, il est naturel de tenir compte de la variabilité due à l'estimation des paramètres alors que dans un modèle de série temporelle, il est d'usage d'ignorer cette variabilité dans le calcul de l'erreur de prévision, qui s'effectue en présumant les paramètres connus avec certitude. Cette habitude est possiblement due au fait que la nature récursive des séries temporelles rend l'inclusion de cette variabilité beaucoup plus difficile. Les travaux de Yamamoto (1981) ont permis d'obtenir un terme dans l'expression de l'erreur qui permet de corriger cette dernière pour la variabilité due à l'estimation des paramètres. En utilisant $\mathbf{E}_1^\top = (\mathbf{I}_d : \mathbf{0} : \ldots : \mathbf{0})$, il est possible d'utiliser la représentation sous forme de somme infinie. Ainsi, une observation au temps t prend la forme $\mathbf{Y}_t = \mathbf{E}_1^\top \mathbf{Y}_t^* = \mathbf{E}_1^\top \boldsymbol{\mu}^* + \mathbf{E}_1^\top \sum_{i=0}^{\infty} \boldsymbol{\Phi}^{*i} \mathbf{a}_{t-i}^*$ et une observation au temps $t+h$ s'écrit $\mathbf{Y}_{t+h} = \mathbf{E}_1^\top \mathbf{Y}_{t+h}^* = \mathbf{E}_1^\top \boldsymbol{\mu}^* + \mathbf{E}_1^\top \sum_{i=0}^{\infty} \boldsymbol{\Phi}^{*i} \mathbf{a}_{t+h-i}^*$.

Une prévision s'obtient en prenant l'espérance conditionnelle de $\mathbf{Y}_{t+h}$ sachant $\mathbf{Y}_t^*$. Le théorème suivant explicite l'espérance conditionnelle dans un modèle VAR(p).

Théorème 1.2.1. *Soit un processus VAR(p). L'espérance conditionnelle à l'horizon h, sachant* $\mathbf{Y}_t^*$, *s'écrit:* $E(\mathbf{Y}_{t+h}|\mathbf{Y}_t^*) = E(\mathbf{E}_1^\top \mathbf{Y}_{t+h}^*|\mathbf{Y}_t^*) = \mathbf{E}_1^\top \boldsymbol{\mu}^* + \mathbf{E}_1^\top \boldsymbol{\Phi}^{*h}(\mathbf{Y}_t^* - \boldsymbol{\mu}^*)$.

DÉMONSTRATION. Remarquons que $\mathbf{Y}_{t+h} = \mathbf{E}_1^\top \mathbf{Y}^*_{t+h}$. Conséquemment:

$$\begin{aligned}
\mathbf{Y}_{t+h} &= \mathbf{E}_1^\top \mathbf{Y}^*_{t+h} = \mathbf{E}_1^\top \boldsymbol{\mu}^* + \mathbf{E}_1^\top \sum_{i=0}^{h-1} \boldsymbol{\Phi}^{*i} \mathbf{a}^*_{t+h-i} + \mathbf{E}_1^\top \sum_{i=0}^{\infty} \boldsymbol{\Phi}^{*h+i} \mathbf{a}^*_{t-i}, \\
&= \mathbf{E}_1^\top \boldsymbol{\mu}^* + \mathbf{E}_1^\top \sum_{i=0}^{h-1} \boldsymbol{\Phi}^{*i} \mathbf{a}^*_{t+h-i} + \mathbf{E}_1^\top \boldsymbol{\Phi}^{*h} \sum_{i=0}^{\infty} \boldsymbol{\Phi}^{*i} \mathbf{a}^*_{t-i}, \\
&= \mathbf{E}_1^\top \boldsymbol{\mu}^* + \mathbf{E}_1^\top \sum_{i=0}^{h-1} \boldsymbol{\Phi}^{*i} \mathbf{a}^*_{t+h-i} + \mathbf{E}_1^\top \boldsymbol{\Phi}^{*h} (\mathbf{Y}^*_t - \boldsymbol{\mu}^*).
\end{aligned}$$

Ainsi, en prenant l'espérance conditionnelle:

$$\begin{aligned}
E(\mathbf{Y}_{t+h}|\mathbf{Y}^*_t) &= E(\mathbf{E}_1^\top \mathbf{Y}^*_{t+h}|\mathbf{Y}^*_t) = \mathbf{E}_1^\top \boldsymbol{\mu}^* + \mathbf{E}_1^\top \sum_{i=0}^{h-1} \boldsymbol{\Phi}^{*i} E(\mathbf{a}^*_{t+h-i}|\mathbf{Y}^*_t) \\
&\quad + \mathbf{E}_1^\top \boldsymbol{\Phi}^{*h} E\{(\mathbf{Y}^*_t - \boldsymbol{\mu}^*)|\mathbf{Y}^*_t\}, \\
&= \mathbf{E}_1^\top \boldsymbol{\mu}^* + \mathbf{E}_1^\top \boldsymbol{\Phi}^{*h} (\mathbf{Y}^*_t - \boldsymbol{\mu}^*).
\end{aligned}$$

□

L'espérance conditionnelle représente le meilleur prédicteur selon le critère de l'erreur quadratique moyenne. Par contre, il demeure une incertitude lors du calcul des prévisions qui doit être quantifiée. Ainsi, l'erreur de prévision s'écrit:

$$\mathbf{e}_{t+h} = \mathbf{Y}_{t+h} - E(\mathbf{Y}_{t+h}|\mathbf{Y}^*_t) = \mathbf{E}_1^\top \sum_{i=0}^{h-1} \boldsymbol{\Phi}^{*i} \mathbf{a}^*_{t+h-i}.$$

Ces formules sont à la base des méthodes de prévision pour ces modèles mais ne tiennent pas compte d'un aspect important en pratique: les paramètres ne sont pas connus et il faut les estimer. Le prédicteur utilisé en pratique est l'espérance conditionnelle dans laquelle on remplace les paramètres inconnus par des estimateurs.

Définition 9. *Soit $\mathbf{Y}_t(h)$ le prédicteur au temps t pour l'horizon de prévision h. Ce prédicteur prend la forme:*

$$\mathbf{Y}_t(h) = \mathbf{E}_1^\top \hat{\boldsymbol{\mu}}^* + \mathbf{E}_1^\top \hat{\boldsymbol{\Phi}}^{*h} (\mathbf{Y}^*_t - \hat{\boldsymbol{\mu}}^*).$$

Il est cependant pratique de représenter le prédicteur sous forme vectorielle.

Théorème 1.2.2. *Le prédicteur au temps t pour l'horizon de prévision h s'écrit: $\mathbf{Y}_t(h) = \hat{\boldsymbol{\alpha}} \mathbf{W}_h$ avec $\mathbf{W}_h = (\mathbf{Y}^{*\top}_t, 1)^\top$ et $\hat{\boldsymbol{\alpha}} = (\mathbf{E}_1^\top \hat{\boldsymbol{\Phi}}^{*h}, \mathbf{E}_1^\top (\mathbf{I} - \hat{\boldsymbol{\Phi}}^{*h}) \hat{\boldsymbol{\mu}}^*)$.*

DÉMONSTRATION.

$$\begin{aligned}\mathbf{Y}_t(h) &= \mathbf{E}_1^\top \hat{\boldsymbol{\mu}}^* + \mathbf{E}_1^\top \hat{\boldsymbol{\Phi}}^{*h}(\mathbf{Y}_t^* - \hat{\boldsymbol{\mu}}^*),\\ &= \mathbf{E}_1^\top (\mathbf{I} - \hat{\boldsymbol{\Phi}}^{*h})\hat{\boldsymbol{\mu}}^* + \mathbf{E}_1^\top \hat{\boldsymbol{\Phi}}^{*h}\mathbf{Y}_t^*,\\ &= \hat{\boldsymbol{\alpha}}\mathbf{W}_h.\end{aligned}$$

□

Le prochain théorème permet d'écire le prédicteur dans une forme alternative.

Théorème 1.2.3. *Le prédicteur* $\mathbf{Y}_t(h)$ *s'écrit:*

$$\mathbf{Y}_t(h) = \mathbf{W}_h^* \hat{\boldsymbol{\alpha}}^*,$$

avec $\mathbf{W}_h^* = \mathbf{W}_h^\top \otimes \mathbf{I}_d$ *et* $\hat{\boldsymbol{\alpha}}^* = \text{vec}(\hat{\boldsymbol{\alpha}})$.

DÉMONSTRATION. D'abord, on remarque que: $\mathbf{Y}_t(h) = \text{vec}\{\mathbf{Y}_t(h)\} = \text{vec}(\hat{\boldsymbol{\alpha}}\mathbf{W}_h)$. Maintenant, en utilisant l'identité $\text{vec}(\mathbf{AB}) = (\mathbf{B}^\top \otimes \mathbf{I})\text{vec}(\mathbf{A})$, (voir aussi (A.2.3) dans l'appendice), on obtient:

$$\mathbf{Y}_t(h) = \{\mathbf{W}_h^\top \otimes \mathbf{I}_d\}\{\text{vec}(\hat{\boldsymbol{\alpha}})\}.$$

Ainsi, on pose $\mathbf{W}_h^* = \mathbf{W}_h^\top \otimes \mathbf{I}_d$ et $\hat{\boldsymbol{\alpha}}^* = \text{vec}(\hat{\boldsymbol{\alpha}})$ pour obtenir le résultat. □

Théorème 1.2.4. *L'erreur de prévision au temps* t *pour l'horizon* h *prend la forme:*

$$\mathbf{e}_{t+h} = \mathbf{E}_1^\top \sum_{i=0}^{h-1} \boldsymbol{\Phi}^{*i}\mathbf{a}_{t+h-i}^* + \mathbf{W}_h^*(\boldsymbol{\alpha}^* - \hat{\boldsymbol{\alpha}}^*).$$

DÉMONSTRATION.

$$\begin{aligned}\mathbf{e}_{t+h} &= \mathbf{Y}_{t+h} - \mathbf{Y}_t(h) = \mathbf{E}_1^\top \sum_{i=0}^{h-1} \boldsymbol{\Phi}^{*i}\mathbf{a}_{t+h-i}^* + \mathbf{E}_1^\top \boldsymbol{\Phi}^{*h}(\mathbf{Y}_t^* - \boldsymbol{\mu}^*) - \mathbf{E}_1^\top \hat{\boldsymbol{\Phi}}^{*h}(\mathbf{Y}_t^* - \hat{\boldsymbol{\mu}}^*),\\ &= \mathbf{E}_1^\top \sum_{i=0}^{h-1} \boldsymbol{\Phi}^{*i}\mathbf{a}_{t+h-i}^* + (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})\mathbf{W}_h.\end{aligned}$$

Maintenant, on remarque que pour un vecteur $\mathbf{a}$ on a $\text{vec}(\mathbf{a}) = \mathbf{a}$. Ce qui permet d'obtenir: $(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})\mathbf{W}_h = \text{vec}\{(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})\mathbf{W}_h\} = \mathbf{W}_h^*(\boldsymbol{\alpha}^* - \hat{\boldsymbol{\alpha}}^*)$. Ainsi,

$$\mathbf{e}_{t+h} = \mathbf{E}_1^\top \sum_{i=0}^{h-1} \boldsymbol{\Phi}^{*i}\mathbf{a}_{t+h-i}^* + \mathbf{W}_h^*(\boldsymbol{\alpha}^* - \hat{\boldsymbol{\alpha}}^*).$$

□

Le fait d'avoir des paramètres estimés plutôt que de présumer qu'ils sont connus avec certitude entraine que l'erreur de prévision tient maintenant compte de cette estimation. La mesure de l'incertitude entourant le calcul des intervalles de prévision peut s'effectuer en adoptant une approche conditionnelle ou inconditionnelle. Ansley et Kohn (1986) fournissent plusieurs arguments convaincants qui penchent en faveur d'une approche conditionnelle. L'erreur quadratique moyenne (EQM) de prévision conditionnelle fait l'objet du prochain théorème.

Théorème 1.2.5. *L'EQM de prévision conditionnelle à* $\mathbf{Y_t}^*$ *s'écrit:*

$$\begin{aligned} E(\mathbf{e}_{t+h}\mathbf{e}_{t+h}^{\top}|\mathbf{Y_t}^*) &= \sum_{j=0}^{h-1}\mathbf{E}_1^{\top}\boldsymbol{\Phi}^{*j}E(\mathbf{a}_{t+h-j}^*\mathbf{a}_{t+h-j}^{*\top})\boldsymbol{\Phi}^{*j\top}\mathbf{E}_1 \\ &\quad +\mathbf{W}_h^*\frac{\partial\boldsymbol{\alpha}^*}{\partial\boldsymbol{\lambda}^{\top}}\mathrm{Var}(\hat{\boldsymbol{\lambda}})\frac{\partial\boldsymbol{\alpha}^{*\top}}{\partial\boldsymbol{\lambda}}\mathbf{W}_h^{\top}+\mathbf{o}_p(n^{-1}), \end{aligned}$$

avec $\boldsymbol{\lambda}^{\top}=(\mathrm{vec}(\mathbf{E}_1^{\top}\boldsymbol{\Phi}^*)^{\top},\boldsymbol{\mu}^{\top})$.

DÉMONSTRATION. On remarque en premier lieu que:

$$\begin{aligned} \mathbf{e}_{t+h} &= \mathrm{vec}\{\mathbf{e}_{t+h}\} \\ &= \sum_{j=0}^{n-1}\mathrm{vec}\{\mathbf{E}_1^{\top}\boldsymbol{\Phi}^{*j}\mathbf{a}_{t+h-j}\}+\mathbf{W}_h^*(\boldsymbol{\alpha}^*-\hat{\boldsymbol{\alpha}}^*). \end{aligned}$$

On procède maintenant à une expansion en série de Taylor du premier ordre. On remarque que:

$$\hat{\boldsymbol{\alpha}}^*=\boldsymbol{\alpha}^*+\frac{\partial\hat{\boldsymbol{\alpha}}^*}{\partial\boldsymbol{\lambda}^{\top}}(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda})+\mathbf{o}_p(n^{-1/2}),$$

où

$$\frac{\partial\hat{\boldsymbol{\alpha}}^*}{\partial\boldsymbol{\lambda}^{\top}}=\left(\frac{\partial\hat{\boldsymbol{\alpha}}^*}{\partial\mathrm{vec}(\mathbf{E}_1^{\top}\boldsymbol{\Phi}^*)^{\top}},\frac{\partial\hat{\boldsymbol{\alpha}}^*}{\partial\boldsymbol{\mu}^{\top}}\right).$$

Ainsi,

$$\mathbf{W}_h^*(\hat{\boldsymbol{\alpha}}^*-\boldsymbol{\alpha}^*)=\mathbf{W}_h^*\frac{\partial\hat{\boldsymbol{\alpha}}^*}{\partial\boldsymbol{\lambda}^{\top}}(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda})+\mathbf{o}_p(n^{-1/2}).$$

De plus,

$$\mathbf{e}_{t+h}=\sum_{j=0}^{h-1}\mathrm{vec}(\mathbf{E}_1\boldsymbol{\Phi}^{*j}\mathbf{a}_{t+h-j})-\mathbf{W}_h^*\frac{\partial\hat{\boldsymbol{\alpha}}^*}{\partial\boldsymbol{\lambda}^{\top}}(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda})+\mathbf{o}_p(n^{-1/2}).$$

Conséquemment, l'espérance conditionnelle devient:

$$E(\mathbf{e}_{t+h}\mathbf{e}_{t+h}^{\top}|\mathbf{Y_t}^*) = \sum_{j=0}^{h-1} \mathbf{E}_1^{\top}\boldsymbol{\Phi}^{*j}E(\mathbf{a}^*_{t+h-j}\mathbf{a}^{*\top}_{t+h-j})\boldsymbol{\Phi}^{*j\top}\mathbf{E}_1 + \mathbf{W}_h^*\frac{\partial\boldsymbol{\alpha}^*}{\partial\boldsymbol{\lambda}^{\top}}\mathrm{Var}(\hat{\boldsymbol{\lambda}})\frac{\partial\boldsymbol{\alpha}^{*\top}}{\partial\boldsymbol{\lambda}}\mathbf{W}_h^{*\top} + \mathbf{o}_p(n^{-1}).$$

□

1.3. Valeurs aberrantes

Dans la pratique, il arrive que le système $\{\mathbf{Y}_t\}$ ne soit pas parfaitement bien observé. Ainsi, certaines erreurs pourraient survenir, et la source de ces erreurs n'est pas toujours claire pour l'analyste. Formellement, on pourrait trouver que les données proviennent majoritairement d'un certain modèle, et c'est ce dernier que nous voudrions décrire aussi précisément que possible. Autrement formulé, les données pourraient provenir majoritairement d'un premier modèle, et occasionellement d'un autre ; ces deux modèles pourraient être fort différents, et exprimer simplement un seul et unique modèle qui les réconcilie pourrait s'avérer compliqué. On désire alors modéliser la portion des données provenant du phénomène majoritaire et les autres données sont considérées aberrantes.

Les valeurs aberrantes peuvent apparaître de façon additive sur la valeur observée. Par exemple, si l'on considère le processus $\{\mathbf{Y}_t\}$, une contamination additive voudrait dire que le processus observé sera $\{\mathbf{Y}_t + \mathbf{W}_t\}$, où $\{\mathbf{W}_t\}$ représente le processus de contamination et est indépendant de $\{\mathbf{Y}_t\}$. C'est ce méchanisme de contamination dit additif qui sera considéré par la suite.

Un autre méchanisme possible est un processus de contamination des erreurs $\{\mathbf{U}_t\}$ du processus linéaire $\boldsymbol{\phi}(B)\mathbf{Y}_t = \mathbf{U}_t$. Ainsi, le processus contaminé sera plutôt $\boldsymbol{\phi}(B)\mathbf{Y}_t = \mathbf{U}_t + \mathbf{W}_t$. Ce méchanisme de contamination a pour effet d'affecter non seulement l'observation au temps t mais aussi les observations subséquantes. Ce méchanisme ne sera pas considéré plus en détail dans ce mémoire. De toute façon, il est habituellement reconnu que les estimateurs sont moins affectés par ce genre de contamination. Voir par exemple, Rousseeuw et Leroy (1987, pp. 275-277) ou Bustos et Yohai (1986).

1.3.1. Mesure d'éloignement

La distance de Mahalanobis mesure l'éloignement d'une observation par rapport à sa distribution. Elle est définie comme suit:

$$M = \{\mathbf{x} - E(\mathbf{x})\}^\top \Sigma_{\mathbf{x}}^{-1} \{\mathbf{x} - E(\mathbf{x})\}.$$

où $\Sigma_{\mathbf{x}} = \text{Var}(\mathbf{x})$. Ainsi, on peut détecter qu'une observation est éloignée en utilisant sa distance de Mahalanobis associée. Par exemple, si $\mathbf{x} \sim N_p(\boldsymbol{\mu}_{\mathbf{x}}, \Sigma_{\mathbf{x}})$ alors $M \sim \chi^2_p$. Dans le cas d'une observation univariée, cela se résume à considérer $\{x - E(x)\}^2/\sigma_x^2 \sim \chi^2_1$.

1.3.2. Estimation robuste

Les estimateurs robustes, par construction, devraient afficher un comportement plus stable lorsque les données sont contaminées. Autrement dit, les estimateurs robustes devraient fournir des valeurs similaires à ce qui serait obtenu s'il n'y avait pas eu d'erreurs dans les données. Une façon d'arriver à cet objectif est d'attribuer des poids aux données et de réduire les poids des observations douteuses.

Les estimateurs RA-ARX consistent essentiellement à modifier les équations du premier ordre qui définissent les estimateurs MCC. Cette modification se résume à remplacer les mesures de dépendance (les autocorrélations et autocorrélations croisées) par des versions robustes. Les estimateurs qui sont solution du nouveau système d'équations devraient être plus robustes sous certaines conditions.

1.3.2.1. *Estimation univariée RA-ARX*

L'estimation dans un contexte de valeurs aberrantes pour un modèle linéaire univarié contenant des variables exogènes, dit modèle ARX, a déjà été traité par Duchesne (2005). Le développement de ces estimateurs permet d'avoir un aperçu de la méthode utilisée pour obtenir des estimateurs équivalents dans un contexte multivarié.

Soit le modèle univarié ARX(p,s):

$$Y_t = \nu + \sum_{i=1}^{p} \phi_{t-i} Y_{t-i} + \sum_{i=0}^{s} \theta_i X_{t-i} + a_t,$$

que l'on peut également écrire de façon plus compacte:

$$\phi(B)(Y_t - \mu) = \theta(B) X_t + a_t,$$

et où comme d'habitude $\{a_t\}$ est un bruit blanc et $\{X_t\}$ est stationnaire.

Théorème 1.3.1. *Les estimateurs des moindres carrés satisfont le système d'équations*:

$$\begin{aligned}
\sum_{k=0}^{s} \theta_k \sum_{j=0}^{n-k-i-1} \pi_j r_{ax}(k+j+i) & \\
+\sigma_a^2 \sum_{j=0}^{n-p-i-1} \pi_j r_{aa}(j+i;\eta) &= 0,\ i = 1\ldots, p, \\
r_{ax}(i) &= 0,\ i = 0, 1, \ldots, s, \\
\sum_{t=1}^{n} a_t &= 0,
\end{aligned}$$

avec

$$r_{xy}(k) = n^{-1} \frac{\sum_{i=k+1}^{n} x_i y_{i-k}}{\sigma_x \sigma_y} \tag{1.3.1}$$

et $\phi^{-1}(B) = \sum_{i=0}^{\infty} \pi_i B^i$. *De plus,* σ_x^2 *et* σ_y^2 *dénotent les variances de* $\{X_t\}$ *et* $\{Y_t\}$ *respectivement.*

DÉMONSTRATION. On obtient les estimateurs MCC en optimisant:

$$S = \sum_{t=1}^{n} a_t^2.$$

La condition de premier ordre pour un optimum est:

$$\frac{\partial S}{\partial \boldsymbol{\beta}^\top} = 2 \sum_{t=1}^{n} a_t \frac{\partial a_t}{\partial \boldsymbol{\beta}^\top} = 0,$$

où $\boldsymbol{\beta} = (\phi_1, \ldots, \phi_p, \theta_0, \theta_1, \ldots, \theta_s, \mu)^\top$ et:

$$\begin{aligned}
\frac{\partial a_t}{\partial \phi_i} &= -(Y_{t-i} - \mu), \quad i = 1, \ldots, p, \\
\frac{\partial a_t}{\partial \theta_i} &= X_{t-i}, \quad i = 0, 1, \ldots, s, \\
\frac{\partial a_t}{\partial \mu} &= -(1 - \sum_{i=1}^{p} \phi_i).
\end{aligned}$$

En injectant $\phi(B)(Y_t - \mu) = \theta(B)X_t + a_t$ dans les équations plus haut, on a:

$$\begin{aligned}
\frac{\partial S}{\partial \phi_i} &= \sum_{t=i+1}^{n} a_t \{\phi(B)^{-1}\theta(B)X_{t-i} + \phi(B)^{-1}a_{t-i}\} = 0, \quad i = 1, \ldots, p, \\
\frac{\partial S}{\partial \theta_i} &= \sum_{t=i+1}^{n} a_t X_{t-i} = 0, \quad i = 0, 1, \ldots, s, \\
\frac{\partial S}{\partial \mu} &= \sum_{t=1}^{n} a_t = 0.
\end{aligned}$$

En rappelant que $\phi^{-1}(B) = \sum_{i=0}^{\infty} \pi_i B^i$, on réécrit le système:

$$\begin{aligned}
\frac{\partial S}{\partial \phi_i} &= \sum_{k=0}^{s} \theta_k \sum_{j=0}^{n-k-i-1} \pi_j \sum_{t=i+1}^{n} a_t X_{t-i-k-j} \\
&\quad + \sum_{j=0}^{n-p-i-1} \pi_j \sum_{t=i+1}^{n} a_t a_{t-i-j} = 0, \text{ pour } i = 1, \ldots, p, \\
\frac{\partial S}{\partial \theta_i} &= \sum_{t=i+1}^{n} a_t X_{t-i} = 0, \text{pour } i = 0, 1, \ldots, s, \\
\frac{\partial S}{\partial \mu} &= \sum_{t=1}^{n} a_t = 0.
\end{aligned}$$

Maintenant en utilisant (1.3.1) , on obtient:

$$\begin{aligned}
&\sum_{k=0}^{s} \theta_k \sum_{j=0}^{n-k-i-1} \pi_j r_{ax}(j+i+k) \\
&\quad + \sigma_a^2 \sum_{j=0}^{n-p-i-1} \pi_j r_{aa}(j+i) = 0, \; i = 1 \ldots, p, \qquad (1.3.2) \\
&\qquad r_{ax}(i) = 0, \; i = 0, 1, \ldots, s \qquad (1.3.3) \\
&\qquad \sum_{t=1}^{n} a_t = 0.
\end{aligned}$$

□

À partir de l'expression des estimateurs des moindres carrés à travers les équations (1.3.2) et (1.3.3), on voit que ces estimateurs ne dépendent des données qu'à travers les mesures de dépendance r_{aa} et r_{ax}. L'idée qui contribue à accentuer la robustesse des estimateurs RA-ARX est d'utiliser des mesures de dépendance robustes (*Robust Autocovariance*, d'où le RA dans RA-ARX).

Définition 10. *Les estimateurs robustes RA-ARX satisfont le système d'équations suivant:*

$$\begin{aligned}
\sum_{k=0}^{s}\theta_k \sum_{j=0}^{n-i-k-1}\pi_j\gamma_{ax}(j+i+k;\psi)+\sigma_a\sum_{j=0}^{n-p-i-1}\pi_j\gamma_{aa}(j+i;\eta) &= 0,\ i=1,\ldots,p,\\
\gamma_{ax}(i;\psi) &= 0,\ i=0,1,\ldots,s\\
\sum_{t=1}^{n}\psi(a_t) &= 0.
\end{aligned}$$

Les quantités γ_{aa} et γ_{ax} sont des mesures de dépendance et sont définies comme suit:

$$\begin{aligned}
\gamma_{ax}(i;\psi) &= n^{-1}\sum_{t=i+1}^{n}\psi(a_t/\sigma_a)X_{t-i},\\
\gamma_{aa}(i;\eta) &= n^{-1}\sum_{t=i+1}^{n}\eta(a_t/\sigma_a, a_{t-i}/\sigma_a).
\end{aligned}$$

Ces mesures dépendent des fonctions $\eta(\cdot,\cdot)$ et $\psi(\cdot)$. On note que si $\eta(u,v)=uv$ alors on retrouve le système original ayant les estimateurs MCC comme solution. De manière générale, dans un contexte d'estimateurs robustes, on présume que les fonctions $\eta(\cdot,\cdot)$ et $\psi(\cdot)$ sont continues. La fonction $\psi(\cdot)$ est supposée impaire et la choisir bornée permet de limiter l'influence des observations potentiellement suspectes. Des choix robustes peuvent consister à choisir $\eta(u,v)=\psi(u)v$ où

$$\psi(u;c)=\text{sign}(u)\min(|u|,c),$$

et c est une constante de robustesse à spécifier. D'autres suggestions sont discutées dans Bou Hamad et Duchesne (2005) et dans l'article qui se trouve au prochain chapitre.

1.3.3. Récapitulatif

Dans ce chapitre, il a été question d'introduire des notions fondamentales sur les processus multivariés. Les notions de stationnarité et de stabilité d'un processus ont été présentées ainsi que les propriétés de la matrice d'autocovariance. La classe de modèles VAR ainsi que la classe plus générale VARX ont été décrites avec les estimateurs des moindres carrés conditionnels pour ces deux classes respectives.

La prévision dans la classe des modèles VAR a été présenté. Il a été question d'erreur de prévision qui considère la variabilité due à l'estimation des paramètres. Cet aspect sera repris dans l'article qui suit pour la classe plus générale des modèles VARX lorsqu'il sera question de construire des intervalles de prévision.

Enfin, l'estimation robuste en présence de valeurs aberrantes a été discutée. Plus particulièrement, le développement des estimateurs robustes RA-ARX a été exposé, ce qui illustre l'essentiel de la démarche utilisée dans le cas multivarié pour obtenir les estimateurs RA-VARX qui seront développés dans l'article suivant.

Chapitre 2

L'ARTICLE

Le présent chapitre est une copie intégrale de l'article intitulé "On robust forecasting in dynamic vector time series models" soumis à une revue scientifique avec un comité de lecture. La première version de l'article a été soumise en 2006. Il a été révisé et raccourci en mars 2007. Puisque la première version contient plus de détails théoriques et empiriques, c'est cette dernière que nous incluons ici. Le premier auteur de l'article est également l'auteur du présent mémoire.

On robust forecasting in dynamic vector time series models*

August 15, 2006

Christian GAGNÉ
Université de Montréal
and
Pierre DUCHESNE
Université de Montréal

ABSTRACT

In this article, robust estimation and prediction in multivariate autoregressive models with exogenous variables (VARX) are considered. The conditional least squares estimators (CLS) are known to be non robust when outliers occur. To obtain robust estimators, the method introduced in Duchesne (2005) and Bou Hamad and Duchesne (2005) is generalized for VARX models. The asymptotic distribution of the new estimators is studied and from this is obtained in particular the asymptotic covariance matrix of the robust estimators. Classical conditional prediction intervals normally rely on estimators such as the usual non robust CLS estimators. In the presence of outliers, such as additive outliers, these classical predictions can be severely biased. More generally, the occurrence of outliers may invalidate the usual conditional prediction intervals. Consequently, the new robust methodology is used to develop robust conditional prediction intervals which take into account parameter estimation uncertainty. In a simulation study, we investigate the finite sample properties of the robust prediction intervals under several scenarios for the occurrence of the outliers, and the new intervals are compared to non-robust intervals based on classical CLS estimators.

Key words and phrases : Multivariate time series; Exogenous variables; Prediction; Robust estimators; Additive outliers.
Mathematics subject classification codes (2000) : primary 62M20; secondary 62M10.
JEL classification : C13, C32, C53.

*Abbreviated title : "On robust forecasting in vector time series models".

Corresponding author : Pierre Duchesne, Université de Montréal; Département de mathématiques et de statistique; C.P. 6128 Succursale Centre-Ville; Montréal, Québec H3C 3J7; Canada. tel : (514) 343-7267; fax : (514) 343-5700; e-mail : [redacted]

1. INTRODUCTION

Dynamic simultaneous equation models are frequently used for forecasting purposes. The statistical properties of these linear systems are discussed in Judge *et al.* (1985), Hannan and Deistler (1988) and Lütkepohl (1993), amongst others. A possible process encountered in economic and physical applications is the vector autoregressive model with stationary exogenous variables (VARX). Let the d-vectors $\mathbf{Y}_t = (Y_t(1), \ldots, Y_t(d))^\top$ and the m-vectors $\mathbf{X}_t = (X_t(1), \ldots, X_t(m))^\top$, $t = 1, \ldots, n$, be observations corresponding to a stationary VARX process of orders p and s, abbreviated VARX(p,s). Then, there exist $d \times d$ matrices $\mathbf{\Phi}_i$, $i = 1, \ldots, p$ and $d \times m$ matrices $\mathbf{V}_i$, $i = 0, 1, \ldots, s$, with $\mathbf{\Phi}_p \neq \mathbf{0}$ and $\mathbf{V}_s \neq \mathbf{0}$, such that :

$$\mathbf{\Phi}(B)\mathbf{Y}_t = \boldsymbol{\theta} + \mathbf{V}(B)\mathbf{X}_t + \mathbf{a}_t, \tag{2.0.1}$$

where $\boldsymbol{\theta}$ denotes a $d \times 1$ constant vector, $\mathbf{\Phi}(B) = \mathbf{I}_d - \sum_{i=1}^{p} \mathbf{\Phi}_i B^i$, $\mathbf{I}_d$ being the $d \times d$ identity matrix, $\mathbf{V}(B) = \sum_{i=0}^{s} \mathbf{V}_i B^i$, B representing the usual backward shift operator. In representation (2.0.1), the random vector $\mathbf{a}_t = (a_t(1), \ldots, a_t(d))^\top$ corresponds to the error term, $t = 1, \ldots, n$. Let $\mathbf{X} = \{\mathbf{X}_t, t \in \mathbb{Z}\}$ and $\mathbf{a} = \{\mathbf{a}_t, t \in \mathbb{Z}\}$ be the exogenous and error stochastic processes, respectively. Process $\mathbf{a}$ is assumed to be a strong white noise, that is $\mathbf{a}_t$, $t \in \mathbb{Z}$, are identically and independently distributed (iid) random vectors with mean zero and regular covariance matrix $\mathbf{\Sigma}$. Without loss of generality the mean of process $\mathbf{X}$ is zero. Furthermore, we suppose that the autocovariance function of $\mathbf{X}$ defined by $\mathbf{\Gamma}_{\mathbf{X}}(j) = E(\mathbf{X}_t \mathbf{X}_{t-j}^\top)$, $j \in \mathbb{Z}$, is absolutely summable, that is $\sum_j ||\mathbf{\Gamma}_{\mathbf{X}}(j)|| < \infty$, where $||\mathbf{A}||$ denotes the Euclidian norm of matrix $\mathbf{A}$. In the following, $\mathbf{X}$ and $\mathbf{a}$ are statistically independent stochastic processes. The VARX process $\mathbf{Y} = \{\mathbf{Y}_t, t \in \mathbb{Z}\}$ is supposed stationary and consequently all the roots of $\det\{\mathbf{\Lambda}(z)\}$, $z \in \mathbb{C}$, lie outside the unit disk, where $\det(\mathbf{A})$ stands for the determinant of matrix $\mathbf{A}$.

In practical applications, the parameters of the VARX model can be estimated by conditional least squares estimators (CLS) or by alternative methods such as generalized least squares, estimated generalized least squares or maximum likelihood estimators, amongst others. See Lütkepohl (1993) for a description of these methods. However, the sensitivity of the CLS estimation procedure (or related techniques) to outliers is well-known. For example, CLS estimators may be strongly biased when the observations are not observed from the process $\{\mathbf{Y}_t\}$ satisfying (2.0.1), but for example from $\{\mathbf{Y}_t + \mathbf{W}_t\}$, where $\{\mathbf{Y}_t\}$ and $\{\mathbf{W}_t\}$ are independent, with $\{\mathbf{W}_t\}$ as an iid sequence that generates the outliers (typically the occurrence of outliers is relatively small, for example with probability less than 5%). The CLS method appears to be very sensitive to the kind of outliers which fall in this so-called additive category, since they affect the endogenous variable $\mathbf{Y}_t$ additively. Robust estimation techniques and robust diagnostics for time series models are introduced in Hampel *et al.* (1986) and Li (2004), amongst others.

The goal of this paper is to propose robust prediction intervals in VARX models. Whereas robust estimation has been widely studied in the time series literature, the construction of robust prediction intervals in dynamic models has received little attention, to the best of our knowledge. From a practical point of view, the empirical evaluation of existing prediction intervals to various kind of outliers seems an important issue. Since outliers can cause bias problems for the CLS estimators of the parameters of the model, it is expected that the usual point predictions based on such estimation technique will also be biased. Furthermore, and maybe more importantly, the classical prediction confidence limits are calculated using the usual mean squared errors of prediction. Since these quantities are estimated in practice using non robust estimation methods, they display also sensitivity to outliers. Thus, the usual prediction intervals are expected to be distorted by outliers, an undesirable outcome in the forecasting practice. In general, outliers may invalidate the predictions and the prediction confidence limits for practical purposes. These phenomena are well-known to occur in ARIMA models, see Ledolter (1989) and Chatfield (2001), who have discussed the effects of additive outliers on predictions intervals. These practical considerations call for the development of robust estimators in VARX models and for the construction of robust prediction intervals.

Since the CLS estimation method displays sensitivity to outliers, one of this paper's primary objectives consists in developing robust estimators in VARX models. The proposed method provides a multivariate generalization of the robust estimation technique elaborated in Duchesne (2005) and Bou Hamad and Duchesne (2005) for univariate autoregressive models with exogenous variables (ARX). The asymptotic distribution of the proposed estimators is studied. Using an argumentation similar to the one given in Bou Hamad and Duchesne (2005) but adapted to the multivariate case using the Cramér-Wold device, the proposed estimators are seen to converge in distribution towards the normal distribution. The asymptotic covariance matrix of the robust estimators is explicitly obtained, which is useful for constructing prediction intervals. Other estimation procedures for vector time series include the robust estimators used by Ben, Martinez and Yohai (1999) in vector autoregressive moving average time series (VARMA) models. See also Li and Hui (1989) who developed alternative robust estimators in VARMA models.

Our new robust estimators allow us to develop new robust conditional prediction intervals in VARX models. The proposed robust prediction intervals are conditional because the coverage properties are conditional on the lagged endogenous and exogenous processes. In time series practice, it is rather common to overlook parameter estimation uncertainty in constructing prediction intervals. However, in small and moderate samples, accounting for parameter estimation

uncertainty can greatly improve the coverage properties of the prediction intervals. Consequently, the robust conditional prediction intervals being proposed here offer accuracy up to $\mathbf{O}_P(n^{-1})$, and they take into account parameter estimation. In a seminal work on prediction in dynamic models, Schmidt (1977) considered the asymptotic distribution of dynamic simulation forecasts and provided some small sample evidence, concluding that a disregard for parameter estimation uncertainty may lead to substantial underestimation of forecast variance in small samples. For vector autoregressive (VAR) processes and VARMA processes, Baillie (1979), Reinsel (1980) and Yamamoto (1981) derived the asymptotic mean squared error of multi-step predictions. The asymptotic distribution of predictions for the more general simultaneous equation model with lagged endogenous variables and VAR errors has been studied in Baillie (1981). The optimal prediction scheme for multi-period predictions of a simultaneous equation autoregressive model with exogenous variables, with VAR errors or vector moving average errors, has been investigated in Yamamoto (1980). Ansley and Kohn (1986) have considered prediction mean squared error for state space models with estimated parameters as well as applications to ARMA models, while Reinsel and Lewis (1987) have investigated forecasting methods for non-stationary multivariate time series. Resampling techniques represent an alternative approach for constructing prediction intervals. Masarotto (1990) and Grigoletto (1998) have proposed bootstrap prediction intervals for autoregressive processes of unknown order p, when p can be estimated consistently. Cao (1999) has reviewed some resampling techniques for predicting time series. Alonso, Peña and Romo (2002) have introduced the so-called sieve bootstrap for estimating prediction intervals. Their study concluded that empirical coverage comes closer to the nominal confidence level when the prediction intervals constructed incorporate the uncertainty due to parameter estimation, specially when the sample size is small. Prediction using estimated parameters in misspecified univariate models has been considered in Kunitomo and Yamamoto (1985), Stine (1987), de Luna (2000) and Kabaila and He (2004). Schorfheide (2005) has recently investigated forecasting under misspecification in VAR models.

The general objective of this paper consists in proposing a robust methodology for use in VARX models, in order to derive reliable forecasts when outliers are suspected to occur in the observed time series. Valid conditional prediction intervals are provided under a correctly specified model, which should offer robustness in the presence of additive outliers. The rest of the paper is organized in the following way. In Section 2, robust estimation is developed in VARX models, generalizing the robust method of Duchesne (2005) and Bou Hamad and Duchesne (2005). The asymptotic distribution of the robust estimators is stated, and a practical algorithm for computing the robust estimators is discussed. In Section 3, new robust prediction intervals are developed, which take into account

parameter estimation uncertainty. In Section 4, we examine the finite sample behavior of point predictions and prediction intervals, when additive outliers are present and absent. Several scenarios generate the additive outliers, including outliers at fixed and random positions, and patches of outliers. Robust prediction intervals based on the robust estimators proposed, and classical prediction intervals relying on the CLS method, are compared with respect to : coverage properties of prediction intervals; empirical biases of the predictions; and empirical prediction mean squared errors. We conclude the paper with a discussion. Some technical details are given in the Appendices.

2. ROBUST ESTIMATORS IN VARX MODELS

2.1 Development of RA-VARX estimators

In Duchesne (2005), robust estimators in ARX models are introduced, generalizing a robust method proposed by Basawa *et al.* (1985) for dynamic regression models with fixed regressors. These estimators are obtained by adapting to dynamic models the robust methodology Bustos and Yohai (1986) initially elaborated for ARMA models. Ben, Martinez and Yohai (1999) proposed robust estimators in VARMA models, extending the RA estimators of Bustos and Yohai (1986) to multivariate time series. Another class of estimators in VARMA models is due to Li and Hui (1989), but the approach taken by Ben, Martinez and Yohai (1999) provides a more natural multivariate generalization, since the resulting estimators enjoy some desirable properties (they are for example affine equivariant).

In ARX models, the system of equations which has the CLS estimators as a solution involves the non robust sample autocovariances of the residuals and cross-covariances between the residuals and the exogenous variables. This suggests that robustness can be achieved by replacing the usual auto- and cross-covariances by robustified versions. Solving the resulting system leads to the so-called RA-ARX estimators. Duchesne (2005) has studied the consistency of the RA-ARX estimators when a perfect ARX model is observed, and Bou Hamad and Duchesne (2005) have derived the asymptotic distribution of the RA-ARX estimators, which is shown to be normal. In particular, the asymptotic covariance matrix of the RA-ARX estimators is obtained; applications of these results include the elaboration of test statistics for diagnostic checking, amongst others.

The process (2.0.1) can be written equivalently as :

$$\mathbf{\Phi}(B)(\mathbf{Y}_t - \boldsymbol{\mu}) = \mathbf{V}(B)\mathbf{X}_t + \mathbf{a}_t, \tag{2.0.2}$$

where $\boldsymbol{\mu} = \mathbf{\Phi}^{-1}(1)\boldsymbol{\theta}$, and we will often resort to the alternative parametrization (2.0.2). Let $\mathbf{\Phi} = (\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_p)$, $\mathbf{V} = (\mathbf{V}_0, \mathbf{V}_1, \ldots, \mathbf{V}_s)$ be $d \times dp$ and $d \times (s+1)m$ matrices corresponding to the autoregressive and exogenous parts, respectively; and $\boldsymbol{\lambda} = (\mathrm{vec}^\top(\mathbf{\Phi}), \mathrm{vec}^\top(\mathbf{V}), \boldsymbol{\mu}^\top)^\top$ be a $d^2p + dm(s+1) + d$ vector, which include all the model parameters. For a matrix $\mathbf{A}$, $\mathrm{vec}(\mathbf{A})$ represents the

vector obtained by stacking the columns of $\mathbf{A}$ (see Harville, 1997, Chapter 16.3). We denote $\mathbf{\Phi}_0$, $\mathbf{V}_0$, $\boldsymbol{\mu}_0$ and $\boldsymbol{\lambda}_0$ as the true values of the parameters. The CLS estimators of $\boldsymbol{\lambda}_0$ are obtained by minimizing the following objective function :

$$S(\boldsymbol{\lambda}) = \frac{1}{2} \sum_{t=(p\vee s)+1}^{n} \mathbf{r}_t^\top(\boldsymbol{\lambda}) \mathbf{\Sigma}^{-1} \mathbf{r}_t(\boldsymbol{\lambda}),$$

where $p \vee s$ denotes the maximum between p and s. For a given value of $\boldsymbol{\lambda}$, the residuals, noted $\mathbf{r}_t(\boldsymbol{\lambda})$, are defined recursively as $\mathbf{r}_t(\boldsymbol{\lambda}) = \mathbf{0}$, for $t \leq 0$, and, for $t = 1, \ldots, n$, by :

$$\begin{aligned} \mathbf{r}_t(\boldsymbol{\lambda}) &= \mathbf{\Phi}(B)(\mathbf{Y}_t - \boldsymbol{\mu}) - \mathbf{V}(B)\mathbf{X}_t, \\ &= \mathbf{\Phi}(B)\dot{\mathbf{Y}}_t - \mathbf{V}(B)\mathbf{X}_t, \end{aligned}$$

where $\dot{\mathbf{Y}}_t = \mathbf{Y}_t - \boldsymbol{\mu}$; and we set $\mathbf{Y}_t = \boldsymbol{\mu}$ and $\mathbf{X}_t = \mathbf{0}$ for $t \leq 0$. We will write simply $\mathbf{r}_t \equiv \mathbf{r}_t(\boldsymbol{\lambda})$ for the residuals when no confusion is possible.

The CLS equations are obtained by equating to zero the derivatives of $S(\boldsymbol{\lambda})$ with respect to the matrices $\mathbf{\Phi}_i$, $i = 1, \ldots, p$; $\mathbf{V}_i$, $i = 0, 1, \ldots, s$; and with respect to the vector $\boldsymbol{\mu}$. To obtain these derivatives, some intermediate results appear useful. Vectorizing $\mathbf{r}_t(\boldsymbol{\lambda})$ and using the well-known identity $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})\mathrm{vec}(\mathbf{B})$, we deduce that :

$$\mathbf{r}_t(\boldsymbol{\lambda}) = \dot{\mathbf{Y}}_t - \sum_{j=1}^{p} (\dot{\mathbf{Y}}_{t-j}^\top \otimes \mathbf{I}_d)\mathrm{vec}(\mathbf{\Phi}_j) - \sum_{j=0}^{s} (\mathbf{X}_{t-j}^\top \otimes \mathbf{I}_d)\mathrm{vec}(\mathbf{V}_j).$$

Consequently, it is not difficult to establish that :

$$\begin{aligned} \frac{\partial \mathbf{r}_t(\boldsymbol{\lambda})}{\partial \mathrm{vec}^\top(\mathbf{\Phi}_j)} &= -(\dot{\mathbf{Y}}_{t-j}^\top \otimes \mathbf{I}_d), \quad j = 1, \ldots, p, \\ \frac{\partial \mathbf{r}_t(\boldsymbol{\lambda})}{\partial \mathrm{vec}^\top(\mathbf{V}_j)} &= -(\mathbf{X}_{t-j}^\top \otimes \mathbf{I}_d), \quad j = 0, 1, \ldots, s, \\ \frac{\partial \mathbf{r}_t(\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}^\top} &= -\mathbf{I}_d + \sum_{i=1}^{p} \mathbf{\Phi}_i = -\mathbf{\Phi}(1). \end{aligned} \tag{2.0.3}$$

Computing the derivatives of the objective function $S(\boldsymbol{\lambda})$ with respect to the model parameters, the system to solve then becomes :

$$\begin{aligned} \frac{\partial S(\boldsymbol{\lambda})}{\partial \mathrm{vec}^\top(\mathbf{\Phi}_j)} &= \sum_{t=1}^{n} \mathbf{r}_t^\top(\boldsymbol{\lambda}) \mathbf{\Sigma}^{-1} \frac{\partial \mathbf{r}_t(\boldsymbol{\lambda})}{\partial \mathrm{vec}^\top(\mathbf{\Phi}_j)} = \mathbf{0}, \quad j = 1, \ldots, p, \\ \frac{\partial S(\boldsymbol{\lambda})}{\partial \mathrm{vec}^\top(\mathbf{V}_j)} &= \sum_{t=1}^{n} \mathbf{r}_t^\top(\boldsymbol{\lambda}) \mathbf{\Sigma}^{-1} \frac{\partial \mathbf{r}_t(\boldsymbol{\lambda})}{\partial \mathrm{vec}^\top(\mathbf{V}_j)} = \mathbf{0}, \quad j = 0, 1, \ldots, s, \\ \frac{\partial S(\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}^\top} &= \sum_{t=1}^{n} \mathbf{r}_t^\top(\boldsymbol{\lambda}) \mathbf{\Sigma}^{-1} \frac{\partial \mathbf{r}_t(\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}^\top} = \mathbf{0}. \end{aligned}$$

Using the preceding intermediate results and transposing, the system of equations can be written alternatively as :

$$\begin{aligned}
\left\{\frac{\partial S(\boldsymbol{\lambda})}{\partial \mathrm{vec}^\top(\boldsymbol{\Phi}_j)}\right\}^\top &= -\sum_{t=1}^{n}(\dot{\mathbf{Y}}_{t-j}\otimes \mathbf{I}_d)\mathrm{vec}\{\boldsymbol{\Sigma}^{-1}\mathbf{r}_t(\boldsymbol{\lambda})\} = \mathbf{0},\ \ j=1,\ldots,p,\\
\left\{\frac{\partial S(\boldsymbol{\lambda})}{\partial \mathrm{vec}^\top(\mathbf{V}_j)}\right\}^\top &= -\sum_{t=1}^{n}(\mathbf{X}_{t-j}\otimes \mathbf{I}_d)\mathrm{vec}\{\boldsymbol{\Sigma}^{-1}\mathbf{r}_t(\boldsymbol{\lambda})\} = \mathbf{0},\ \ j=0,1,\ldots,s,\\
\left\{\frac{\partial S(\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}^\top}\right\}^\top &= -\boldsymbol{\Phi}^\top(1)\sum_{t=1}^{n}\mathrm{vec}\{\boldsymbol{\Sigma}^{-1}\mathbf{r}_t(\boldsymbol{\lambda})\} = \mathbf{0}.
\end{aligned}$$

The last equation implies that $\sum_{t=1}^{n}\mathbf{r}_t(\boldsymbol{\lambda}) = \mathbf{0}$. Consequently, the CLS estimators of $\boldsymbol{\lambda}_0$ also satisfy the following system of equations :

$$\begin{aligned}
\sum_{t=1}^{n}\mathrm{vec}(\boldsymbol{\Sigma}^{-1}\mathbf{r}_t\dot{\mathbf{Y}}_{t-j}^\top) &= \mathbf{0},\ \ j=1,\ldots,p,\\
\sum_{t=1}^{n}\mathrm{vec}(\boldsymbol{\Sigma}^{-1}\mathbf{r}_t\mathbf{X}_{t-j}^\top) &= \mathbf{0},\ \ j=0,1,\ldots,s,\\
\sum_{t=1}^{n}\mathbf{r}_t(\boldsymbol{\lambda}) &= \mathbf{0}.
\end{aligned} \tag{2.0.4}$$

Let $\boldsymbol{\Phi}^{-1}(B) = \sum_{h=0}^{\infty}\mathbf{U}_h(p,\boldsymbol{\Phi})B^h$. A useful expression for process (2.0.2) is given by the infinite representation :

$$\dot{\mathbf{Y}}_t = \sum_{h\geq 0}\sum_{i=0}^{s}\mathbf{U}_h\mathbf{V}_i\mathbf{X}_{t-i-h} + \sum_{h\geq 0}\mathbf{U}_h\mathbf{r}_{t-h},$$

where to simplify the notation we write $\mathbf{U}_h(p,\boldsymbol{\Phi}) \equiv \mathbf{U}_h$, $h \geq 0$. We demonstrate that system (2.0.4) can be expressed in terms of the matricial auto- and cross-covariances. Let

$$\mathbf{C}_{\mathbf{rr}}(i) = n^{-1}\sum_{t=p+1}^{n-i}\mathbf{r}_{t+i}\mathbf{r}_t^\top, \tag{2.0.5}$$

$$\mathbf{C}_{\mathbf{rX}}(i) = n^{-1}\sum_{t=1}^{n-i}\mathbf{r}_{t+i}\mathbf{X}_t^\top, \tag{2.0.6}$$

be the lag-i autocovariance of the residuals, and the lag-i cross-covariance between $\mathbf{r}_t$ and $\mathbf{X}_t$, respectively. Using the initial values for $\mathbf{Y}_t$ and $\mathbf{X}_t$ for $t \leq 0$, and performing a change of variables similar to the one done in Duchesne (2005, pp. 54-55), we see that the CLS estimators satisfying (2.0.4) are also a solution

of the following system, expressed in function of (2.0.5) and (2.0.6) :

$$\begin{aligned}
\operatorname{vec}\Bigg\{\sum_{i=0}^{s}\sum_{h=0}^{n-j}\Sigma^{-1}\mathbf{C}_{\mathbf{rX}}(h+i+j)\mathbf{V}_i^{\top}\mathbf{U}_h^{\top} \\
+\sum_{h=0}^{n-p-j-1}\Sigma^{-1}\mathbf{C}_{\mathbf{rr}}(h+j)\mathbf{U}_h^{\top}\Bigg\} &= \mathbf{0},\ j=1,\ldots,p, \qquad (2.0.7)\\
\operatorname{vec}\left\{\Sigma^{-1}\mathbf{C}_{\mathbf{rX}}(j)\right\} &= \mathbf{0}, j=0,1,\ldots,s, \qquad (2.0.8)\\
\sum_{t=1}^{n}\mathbf{r}_t(\boldsymbol{\lambda}) &= \mathbf{0}.
\end{aligned}$$

Consequently, the system of equations (2.0.8) also leads to the CLS estimators. However, the preceding argumentation suggests that the system relies on matrices (2.0.5) and (2.0.6), which are based on the residuals $\mathbf{r}_t$. The usual residuals computed with the CLS estimators display sensitivity to outliers. In order to limit their influence and to obtain more robust estimators, we proceed as in Ben, Martinez and Yohai (1999), by replacing the residuals $\mathbf{r}_t$ in the auto- and cross-covariances (2.0.5) and (2.0.6) by the modified residuals $\tilde{\mathbf{r}}_t$ as follows. Let $\psi(\cdot)$ be an odd and bounded function and define the weight function :

$$\omega(x) = \frac{\psi(x)}{x}. \qquad (2.0.9)$$

To measure outlying residuals, we introduce $d_t(\boldsymbol{\lambda}, \Sigma)$, the Mahalanobis distance (a metric with respect to the covariance matrix Σ) of the residual $\mathbf{r}_t$:

$$d_t^2(\boldsymbol{\lambda}, \Sigma) = \mathbf{r}_t^{\top}(\boldsymbol{\lambda})\Sigma^{-1}\mathbf{r}_t(\boldsymbol{\lambda}).$$

Residuals $\mathbf{r}_t$ for which $d_t^2(\boldsymbol{\lambda}, \Sigma)$ is large represent potential outliers, far from the majority of all the $\mathbf{r}_t$'s. Then, the modified residuals are defined as :

$$\tilde{\mathbf{r}}_t(\boldsymbol{\lambda}, \Sigma) = \mathbf{r}_t(\boldsymbol{\lambda})\omega\{d_t(\boldsymbol{\lambda}, \Sigma)\}. \qquad (2.0.10)$$

The idea behind modification (2.0.10) consists in reducing gradually the influence of residuals suspected to be outliers. The resulting estimators based on these modified residuals should be robust when additive outliers occur in the endogenous variables. The residuals (2.0.10) rely on the ω function (2.0.9) and a function ψ. Several possibilities exist for choosing these functions. A possible example is offered by the Huber function, which is defined as :

$$\psi_H(u; \vartheta) = \operatorname{sign}(u)\min(|u|, \vartheta),$$

where sign(u) denotes the sign of the real number u. Another choice is given by the redescending bisquare function :

$$\psi_B(u;\vartheta) = \begin{cases} u(1-u^2/\vartheta^2)^2, & \text{if } 0 \leq |u| \leq \vartheta, \\ 0, & \text{elsewhere.} \end{cases}$$

The functions $\psi_H(u;\vartheta)$ and $\psi_B(u;\vartheta)$ are based on a constant ϑ, which is called the tuning parameter. This additional constant controls the degree of robustness and the efficiency : a large ϑ is associated with efficient procedures with respect to CLS estimators under perfectly observed VARX models and, on the other hand, smaller values of ϑ provide more robustness, downweighting more heavily larger residuals. Ben, Martinez and Yohai (1999) provide a table for these parameters in VARMA models.

The robust estimators, that we denote RA-VARX, are now defined as a solution of a robustification of system (2.0.8), using the modified residuals (2.0.10) :

$$\begin{aligned} \text{vec}\Bigg\{ \sum_{i=0}^{s}\sum_{h=0}^{n-j} \boldsymbol{\Sigma}^{-1}\mathbf{C}_{\tilde{\mathbf{r}}\mathbf{X}}(h+i+j)\mathbf{V}_i^\top\mathbf{U}_h^\top \cdot \\ + \sum_{h=0}^{n-p-j-1} \boldsymbol{\Sigma}^{-1}\mathbf{C}_{\tilde{\mathbf{r}}\tilde{\mathbf{r}}}(h+j)\mathbf{U}_h^\top \Bigg\} &= \mathbf{0}, \quad j=1,\ldots,p, \qquad (2.0.11) \\ \text{vec}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{C}_{\tilde{\mathbf{r}}\mathbf{X}}(j)\right\} &= \mathbf{0}, j=0,1,\ldots,s, \qquad (2.0.12) \\ \sum_{t=1}^{n}\tilde{\mathbf{r}}_t(\boldsymbol{\lambda}) &= \mathbf{0}. \end{aligned}$$

The variance of the error term $\mathbf{a}_t$ is estimated robustly as :

$$\tilde{\boldsymbol{\Sigma}} = \frac{c}{n}\sum_{t=1}^{n}\tilde{\mathbf{r}}_t(\boldsymbol{\lambda})\tilde{\mathbf{r}}_t^\top(\boldsymbol{\lambda}), \qquad (2.0.13)$$

where

$$c = \frac{d}{E\{\psi^2(V^{1/2})\}}. \qquad (2.0.14)$$

The random variable V in (2.0.14) follows a χ^2 distribution with d degrees of freedom. This choice of the constant c ensures that $\tilde{\boldsymbol{\Sigma}}$ provides a consistent estimator of $\boldsymbol{\Sigma}$ when $\mathbf{a}_t$ follows a normal distribution. In the next section, the asymptotic distribution of the RA-VARX estimators is investigated.

2.2 Asymptotic distribution of RA-VARX estimators

Let $\{\mathbf{Y}_t\}$ be a stable VARX(p,s) process with stationary exogenous process $\{\mathbf{X}_t\}$ given by (2.0.1).

We introduce the system $\mathbf{L}(\boldsymbol{\lambda}, \Sigma) = (\mathbf{L}_1^\top(\boldsymbol{\lambda}, \Sigma), \ldots, \mathbf{L}_{p+s+2}^\top(\boldsymbol{\lambda}, \Sigma))^\top$ and $\mathbf{L}_j(\boldsymbol{\lambda}, \Sigma) = \sum_{t=1}^n \boldsymbol{\delta}_{jt}(\boldsymbol{\lambda}, \Sigma)$, $j = 1, \ldots, p+s+2$, with :

$$\begin{aligned}
\boldsymbol{\delta}_{jt}(\boldsymbol{\lambda}, \Sigma) &= \sum_{i=0}^{s} \boldsymbol{\Delta}_{\tilde{\mathbf{r}}\mathbf{X},(i+j)t}(\boldsymbol{\lambda}, \Sigma)\mathrm{vec}(\mathbf{V}_i^\top) + \boldsymbol{\delta}_{\tilde{\mathbf{r}}\tilde{\mathbf{r}},jt}(\boldsymbol{\lambda}, \Sigma),\ j = 1, \ldots, p, \\
\boldsymbol{\delta}_{jt}(\boldsymbol{\lambda}, \Sigma) &= \mathrm{vec}(\Sigma^{-1}\tilde{\mathbf{r}}_t\mathbf{X}_{t-j+p+1}^\top),\ j = p+1, \ldots, p+s+1, \qquad (2.0.15) \\
\boldsymbol{\delta}_{p+s+2,t}(\boldsymbol{\lambda}, \Sigma) &= \tilde{\mathbf{r}}_t,
\end{aligned}$$

where in the preceding system we let $\boldsymbol{\Delta}_{\tilde{\mathbf{r}}\mathbf{X},kt}(\boldsymbol{\lambda}, \Sigma) = \sum_{h\geq 0} \mathbf{U}_h \otimes (\Sigma^{-1}\tilde{\mathbf{r}}_t\mathbf{X}_{t-k-h}^\top)$, $k = 1, \ldots, p+s$, and $\boldsymbol{\delta}_{\tilde{\mathbf{r}}\tilde{\mathbf{r}},jt}(\boldsymbol{\lambda}, \Sigma) = \sum_{h\geq 0} \mathrm{vec}(\Sigma^{-1}\tilde{\mathbf{r}}_t\tilde{\mathbf{r}}_{t-j-h}^\top\mathbf{U}_h^\top)$, $j = 1, \ldots, p$. Using these definitions, the RA-VARX estimators are defined as a solution of the system :

$$\mathbf{L}(\tilde{\boldsymbol{\lambda}}, \tilde{\Sigma}) = \sum_{t=1}^{n} \boldsymbol{\delta}_t(\tilde{\boldsymbol{\lambda}}, \tilde{\Sigma}) = \mathbf{0},$$

and (2.0.13). Under suitable regularity conditions on ψ and using a Taylor series expansion as in Bou Hamad and Duchesne (2005), we obtain that :

$$n^{1/2}(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) = -\{n^{-1}D\mathbf{L}(\boldsymbol{\lambda}_0, \tilde{\Sigma})\}^{-1}\{n^{-1/2}\mathbf{L}(\boldsymbol{\lambda}_0, \tilde{\Sigma})\} + \mathbf{o}_P(1),$$

where $D\mathbf{L}$ corresponds to the differential matrix of $\mathbf{L}$ with respect to $\boldsymbol{\lambda}$. According to (2.0.13), $\tilde{\Sigma} \xrightarrow{p} \Sigma_0 = cE\{\mathbf{r}_t\mathbf{r}_t^\top\omega^2(d_t)\}$. Using results similar to those obtained by Ben, Martinez and Yohai (1999, pp. 387-388), it may be proved that under suitable regularity conditions on the function ψ that :

$$\begin{aligned}
n^{-1}D\mathbf{L}(\boldsymbol{\lambda}_0, \tilde{\Sigma}) &= n^{-1}D\mathbf{L}(\boldsymbol{\lambda}_0, \Sigma_0) + \mathbf{o}_P(1), \\
n^{-1/2}\mathbf{L}(\boldsymbol{\lambda}_0, \tilde{\Sigma}) &= n^{-1/2}\mathbf{L}(\boldsymbol{\lambda}_0, \Sigma_0) + \mathbf{o}_P(1).
\end{aligned}$$

Consequently,

$$n^{1/2}(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) = -\{n^{-1}D\mathbf{L}(\boldsymbol{\lambda}_0, \Sigma_0)\}^{-1}\{n^{-1/2}\mathbf{L}(\boldsymbol{\lambda}_0, \Sigma_0)\} + \mathbf{o}_P(1). \qquad (2.0.16)$$

Adopting arguments similar to Bou Hamad and Duchesne (2005), it follows that :

$$\begin{aligned}
n^{-1/2}\mathbf{L}(\boldsymbol{\lambda}_0, \Sigma_0) &\xrightarrow{d} N(\mathbf{0}, \mathbf{A}), \qquad (2.0.17) \\
n^{-1}D\mathbf{L}(\boldsymbol{\lambda}_0, \Sigma_0) &\xrightarrow{p} \mathbf{B}, \qquad (2.0.18)
\end{aligned}$$

where $\mathbf{A}$ and $\mathbf{B}$ are certain matrices. It is assumed that $\mathbf{B}$ is non-singular. Consequently, since the matrix $n^{-1}D\mathbf{L}(\boldsymbol{\lambda}_0, \Sigma_0)$ converges in probability toward a non-singular matrix, it follows that $n^{-1}D\mathbf{L}(\boldsymbol{\lambda}_0, \Sigma_0)$ is invertible for large n. The proofs of (2.0.17) and (2.0.18) involve the central limit theorem for stationary sequences and the ergodic theorem (see Durrett (1995)), and the Cramér-Wold device (see Serfling (1980)). Explicit expressions for $\mathbf{A}$ and $\mathbf{B}$ are given in Appendix 1. Using (2.0.17) and (2.0.18) yield :

$$n^{1/2}(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1^\top}). \qquad (2.0.19)$$

Robust estimators are usually associated with computationally intensive methods. In the next section, the algorithm provided relies on an iterative least squares (ILS) scheme, which is relatively simple to implement. This algorithm should be useful in practical applications.

2.3 Computation of the RA-VARX estimators

The system of equations (2.0.12) which define RA-VARX estimators could be solved using general algorithms. For example, Bou Hamad and Duchesne (2005) have studied the computation of RA-ARX estimators in univariate models using Newton-Raphson type algorithms and this approach could be adapted in the multivariate case. However, this kind of algorithm relies on an appropriate choice of the initial values of the parameters, which often seems a delicate matter. This calls for fast and simpler procedures. In the present situation, a simplified ILS scheme offers a workable and rather easy possibility, given the availability of software for VAR models. In general, computation time is expected to be greatly reduced with ILS procedures. The algorithm that we describe in this section seems relatively easy to implement and is likely to be the most useful in practice.

First, we remark that the RA-ARX estimators for VARX models may be interpreted as CLS estimators applied to a transformed time series. Let us define the modified process :

$$\tilde{\mathbf{Y}}_t(\boldsymbol{\lambda}, \boldsymbol{\Sigma}) = \boldsymbol{\mu} + \boldsymbol{\Phi}^{-1}(B)\mathbf{V}(B)\mathbf{X}_t + \boldsymbol{\Phi}^{-1}(B)\tilde{\mathbf{r}}_t,$$

where the initial conditions $\tilde{\mathbf{Y}}_t = \boldsymbol{\mu}$ and $\mathbf{X}_t = \mathbf{0}$ if $t \leq 0$ and $\tilde{\mathbf{r}}_t \equiv \tilde{\mathbf{r}}_t(\boldsymbol{\lambda}, \boldsymbol{\Sigma})$ are defined in (2.0.10). It can easily be shown that computing the residuals of the modified stochastic process $\{\tilde{\mathbf{Y}}_t\}$ and solving the CLS system of equations (2.0.4), will yield the same result as solving system (2.0.12) based on the modified residuals, $\tilde{\mathbf{r}}_t$, of the original process $\{\mathbf{Y}_t\}$. In other words, the RA-VARX estimators of the original process coincide exactly with the CLS estimators of the modified process. This property suggests an iterative computational algorithm that we describe in the following.

Step 1. Let $\hat{\boldsymbol{\lambda}}^{(0)}$ and $\hat{\boldsymbol{\Sigma}}^{(0)}$ be initial estimators. These estimators could be the usual CLS estimators.

Step 2. Given $\hat{\boldsymbol{\lambda}}^{(i)}$ and $\hat{\boldsymbol{\Sigma}}^{(i)}$, the estimators corresponding to the ith iteration, compute the modified residuals and the Mahalanobis distances :

$$\begin{aligned} \tilde{\mathbf{r}}_t^{(i)} &= \mathbf{r}_t(\hat{\boldsymbol{\lambda}}^{(i)})\omega(d_t^{(i)}), \\ d_t^{2(i)} &= \mathbf{r}_t^{\top}(\hat{\boldsymbol{\lambda}}^{(i)})\hat{\boldsymbol{\Sigma}}^{-1(i)}\mathbf{r}_t(\hat{\boldsymbol{\lambda}}^{(i)}), \end{aligned}$$

and the modified process :

$$\tilde{\mathbf{Y}}^{(i)} = \hat{\boldsymbol{\mu}}^{(i)} + \hat{\boldsymbol{\Phi}}^{-1(i)}(B)\hat{\mathbf{V}}^{(i)}(B)\mathbf{X}_t + \hat{\boldsymbol{\Phi}}^{-1(i)}(B)\tilde{\mathbf{r}}_t^{(i)}.$$

Step 3. Compute the CLS estimators of the modified stochastic process $\{\tilde{\mathbf{Y}}_t\}$ with exogenous process $\{\mathbf{X}_t\}$.

Steps 2 and 3 are repeated until convergence is reached.

In our simulation studies, convergence always occurred with the preceding algorithm when ψ was chosen in the Huber family. For a small sample size (less than $n = 100$), we did encounter some problems of convergence when ψ was taken from the bisquare family. This is not surprising, since it is well-known that calculating robust estimators with redescending ψ functions, such as the bisquare ψ function, often leads to systems of equations with several solutions (see, e.g., Martin (1980)). In VARMA models, Ben, Martinez and Yohai (1999) suggest computing robust estimators with a ψ function in the Huber family and then using these estimators as initial values for algorithms relying on redescending ψ functions. However, even this strategy does not guarantee convergence of the algorithm. To avoid convergence problems when ψ was taken from the bisquare family, we computed $\hat{\Sigma}$ at the first iteration of the algorithm and it has been kept fixed in the ILS scheme. At the final iteration $\hat{\Sigma}$ was re-computed and the algorithm is completed. While it is desirable to estimate all the parameters simultaneously (see, e.g., Bustos and Yohai, 1986), convergence problems were completely avoided using this modification with a sample size as small as $n = 100$. In unreported experiments, estimating simultaneously $\boldsymbol{\lambda}$ and Σ or keeping $\hat{\Sigma}$ fixed according to the preceding algorithm seemed to give very similar results for sample sizes $n = 100$ and $n = 200$, at least in our experiments (note that as the sample size becomes larger the probability of convergence of the algorithm increases, as expected). Consequently, when $\psi(\cdot) = \psi_B(\cdot;\vartheta)$, the first step of the algorithm could be modified as follows :

Step 1'. Let $\hat{\boldsymbol{\lambda}}^{(0)}$ and $\hat{\Sigma}^{(0)}$ be initial estimators calculated using the preceding algorithm with $\psi(\cdot) = \psi_H(\cdot;\vartheta)$. At the first iteration, compute an estimator of Σ and then keep this matrix fixed.

The other steps are then exactly the same as before. At the final iteration, the matrix Σ is re-estimated.

These algorithms have been implemented in R and have been used in the simulation experiments presented in Section 5; the computer code is available upon request by communicating with the authors.

3. ROBUST PREDICTION INTERVALS IN VARX MODELS

In this section we derive a conditional asymptotic mean squared error of multi-step prediction for VARX(p,s) processes. Unconditional mean squared errors in dynamic models have been considered in Baillie (1980, 1981) and Yamamoto (1980), amongst others. As argued in Ansley and Kohn (1986), the unconditional mean squared error may not be an appropriate measure of the actual forecast error, and a measure which is conditioned on the observed data may be preferable.

However, as noted by Ledolter (1989) and Chatfield (2001), classical prediction intervals can be strongly affected by additive outliers. Here, we propose robustified conditional intervals which should be less sensitive to these kinds of outliers.

It is convenient to collapse the VARX(p,s) stochastic process to a VARX(1,0) process by means of the companion form representation :

$$\begin{aligned}\dot{\mathbf{Y}}_t^* &= \begin{pmatrix} \mathbf{\Phi}_1 & \mathbf{\Phi}_2 & \cdots & \mathbf{\Phi}_{p-1} & \mathbf{\Phi}_p \\ \mathbf{I}_d & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_d & \mathbf{0} \end{pmatrix} \dot{\mathbf{Y}}_{t-1}^* + \begin{pmatrix} \mathbf{V}_0 & \mathbf{V}_1 & \cdots & \mathbf{V}_s \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix} \mathbf{X}_t^* + \mathbf{a}_t^*, \\ &= \mathbf{\Phi}^* \dot{\mathbf{Y}}_{t-1}^* + \mathbf{V}_0^* \mathbf{X}_t^* + \mathbf{a}_t^*, \end{aligned} \tag{2.0.20}$$

where $\dot{\mathbf{Y}}_t^* = \mathbf{Y}_t^* - \boldsymbol{\mu}_t^*$, $\boldsymbol{\mu}_t^* = \mathbf{1}_d \otimes \boldsymbol{\mu}$, $\mathbf{1}_d = (1, \ldots, 1)^\top$ is a d-vector such that each component is one; and $\mathbf{Y}_t^* = (\mathbf{Y}_t^\top, \ldots, \mathbf{Y}_{t-p+1}^\top)^\top$, $\mathbf{X}_t^* = (\mathbf{X}_t^\top, \ldots, \mathbf{X}_{t-s}^\top)^\top$, $\mathbf{a}_t^* = (\mathbf{a}_t^\top, \mathbf{0}^\top, \ldots, \mathbf{0}^\top)^\top$ denote vectors of dimension $dp \times 1$, $(s+1)m \times 1$ and $dp \times 1$, respectively; and $\mathbf{\Phi}^*$ and $\mathbf{V}_0^*$ are matrices of dimension $dp \times dp$ and $dp \times m(s+1)$, respectively. The above representation is sometimes called the state variable representation (see Yamamoto (1980)).

At time $n+l$, the model (2.0.2) can be expressed as :

$$\mathbf{Y}_{n+l}^* = \sum_{j=0}^{l-1} \mathbf{\Phi}^{*j} \boldsymbol{\theta}^* + \mathbf{\Phi}^{*l} \mathbf{Y}_n^* + \sum_{j=0}^{l-1} \mathbf{\Phi}^{*j} \mathbf{V}_0^* \mathbf{X}_{n+l-j}^* + \sum_{j=0}^{l-1} \mathbf{\Phi}^{*j} \mathbf{a}_{n+l-j}^*, \tag{2.0.21}$$

where $\boldsymbol{\theta}^* = (\mathbf{I}_{dp} - \mathbf{\Phi}^*)\boldsymbol{\mu}^*$. In order to obtain $\mathbf{Y}_{n+l}$, the random variable that we want to predict, we introduce $\mathbf{E}_i = \mathbf{e}_i \otimes \mathbf{I}_d$, where $\mathbf{e}_i$ corresponds to a d-vector with one in position i, and zero elsewhere. Pre-multiplying the relation (2.0.21) by $\mathbf{E}_1^\top$, we obtain :

$$\begin{aligned}\mathbf{Y}_{n+l} &= \mathbf{E}_1^\top \mathbf{Y}_{n+l}^* = \sum_{j=0}^{l-1} \mathbf{E}_1^\top \mathbf{\Phi}^{*j} \boldsymbol{\theta}^* + \mathbf{E}_1^\top \mathbf{\Phi}^{*l} \mathbf{Y}_n^* + \sum_{j=0}^{l-1} \mathbf{E}_1^\top \mathbf{\Phi}^{*j} \mathbf{V}_0^* \mathbf{X}_{n+l-j}^* \\ &\quad + \sum_{j=0}^{l-1} \mathbf{E}_1^\top \mathbf{\Phi}^{*j} \mathbf{a}_{n+l-j}^*. \end{aligned} \tag{2.0.22}$$

Consequently, the minimum mean squared error predictor made at time n, l periods ahead, is given by the following expression :

$$\mathbf{Y}_n(l) = \sum_{j=0}^{l-1} \mathbf{E}_1^\top \mathbf{\Phi}^{*j} \boldsymbol{\theta}^* + \mathbf{E}_1^\top \mathbf{\Phi}^{*l} \mathbf{Y}_n^* + \sum_{j=0}^{l-1} \mathbf{E}_1^\top \mathbf{\Phi}^{*j} \mathbf{V}_0^* \mathbf{X}_{n+l-j}^*. \tag{2.0.23}$$

Once the model has been estimated, the optimal predictor can be modified to form a practical predictor, by estimating appropriately the matrices $\mathbf{\Phi}_i$, $i = 1, \ldots, p$, and $\mathbf{V}_i$, $i = 0, 1, \ldots, s$ in the block matrices $\mathbf{\Phi}^*$ and $\mathbf{V}_0^*$ defined by (2.0.20).

However, the CLS estimators may be strongly affected in the presence of outliers; the same can be said for a calculable predictor relying on such a non robust method. In order to obtain a robustified predictor, we propose to employ the RA-VARX estimators defined in Section 2 to estimate the model parameters. Let $\tilde{\mathbf{\Phi}}_i$, $i = 1, \ldots, p$; $\tilde{\mathbf{V}}_i$, $i = 0, 1, \ldots, s$; $\tilde{\boldsymbol{\theta}}^* = (\mathbf{I}_{dp} - \tilde{\mathbf{\Phi}}^*)\tilde{\boldsymbol{\mu}}^*$, where $\tilde{\boldsymbol{\mu}}^* = \mathbf{1}_d \otimes \tilde{\boldsymbol{\mu}}$ be the RA-VARX estimators of $\mathbf{\Phi}_i$, $i = 1, \ldots, p$; $\mathbf{V}_i$, $i = 0, 1, \ldots, s$; $\boldsymbol{\theta}^*$ and $\boldsymbol{\mu}^*$, respectively. We introduce the following robust predictor :

$$\tilde{\mathbf{Y}}_n(l) = \sum_{j=0}^{l-1} \mathbf{E}_1^\top \tilde{\mathbf{\Phi}}^{*j} \tilde{\boldsymbol{\theta}}^* + \mathbf{E}_1^\top \tilde{\mathbf{\Phi}}^{*l} \mathbf{Y}_n^* + \sum_{j=0}^{l-1} \mathbf{E}_1^\top \tilde{\mathbf{\Phi}}^{*j} \tilde{\mathbf{V}}_0^* \mathbf{X}_{n+l-j}^*. \tag{2.0.24}$$

As in Baillie (1980) and Yamamoto (1980), we assume that the future values of the exogenous variables are known exactly. Using a first-order Taylor series expansion, we can show that the predictor error at horizon l, noted $\mathbf{e}_{n+l}$, is given by :

$$\begin{aligned} \mathbf{e}_{n+l} &= \mathbf{Y}_{n+l} - \tilde{\mathbf{Y}}_n(l), \\ &= \sum_{j=0}^{l-1} \mathbf{E}_1^\top \mathbf{\Phi}^{*j} \mathbf{a}_{n+l-j}^* + \mathbf{W}_{n,l}^* (\boldsymbol{\alpha}^* - \tilde{\boldsymbol{\alpha}}^*), \end{aligned}$$

where $\mathbf{W}_{n,l}^* = (\mathbf{Y}_n^{*\top}, \mathbf{X}_{n+1}^{*\top}, \ldots, \mathbf{X}_{n+l}^{*\top}, \mathbf{1}_l^\top) \otimes \mathbf{I}_d$ and the vector of parameters in the transformed model satisfies :

$$\begin{aligned} \boldsymbol{\alpha}^* &= (\mathrm{vec}(\mathbf{E}_1^\top \mathbf{\Phi}^{*l})^\top, \mathrm{vec}(\mathbf{E}_1^\top \mathbf{\Phi}^{*l-1} \mathbf{V}_0^*)^\top, \ldots, \\ &\quad \mathrm{vec}(\mathbf{E}_1^\top \mathbf{V}_0^*)^\top, \mathrm{vec}(\mathbf{E}_1^\top \mathbf{\Phi}^{*l-1} \boldsymbol{\theta}^*)^\top, \ldots, \mathrm{vec}(\mathbf{E}_1^\top \boldsymbol{\theta}^*)^\top)^\top. \end{aligned} \tag{2.0.25}$$

Noting that $\mathbf{E}_1^\top \mathbf{\Phi}^* = \mathbf{\Phi}$, $\mathbf{E}_1^\top \mathbf{V}_0^* = \mathbf{V}$, a first order Taylor expansion gives :

$$\tilde{\boldsymbol{\alpha}}^* = \boldsymbol{\alpha}^* + \mathbf{H}_l^* (\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) + \mathbf{o}_P(n^{-1/2}),$$

where $\tilde{\boldsymbol{\lambda}}$ denotes the RA-VARX estimators and $\mathbf{H}_l^* = \partial \boldsymbol{\alpha}^* / \partial \boldsymbol{\lambda}^\top$. This implies that the prediction error can be written as :

$$\mathbf{e}_{n+l} = \sum_{j=0}^{l-1} \mathbf{E}_1^\top \mathbf{\Phi}^{*j} \mathbf{a}_{n+l-j}^* - \mathbf{W}_{n,l}^* \mathbf{H}_l^* (\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) + \mathbf{o}_P(n^{-1/2}).$$

Expressions for the matrix of derivatives $\mathbf{H}_l^*$ are given in Appendix 2. Let the conditional mathematical expectation $E_n^*(\cdot) = E(\cdot | \mathbf{Y}_n^*, \mathbf{X}_{n+1}^*, \ldots, \mathbf{X}_{n+l}^*)$. Assuming the asymptotic independence between $\mathbf{Y}_n^*, \mathbf{X}_{n+1}^*, \ldots, \mathbf{X}_{n+l}^*$ and $\tilde{\boldsymbol{\lambda}}$, it follows that the conditional mean squared prediction error is given by :

$$E_n^*(\mathbf{e}_{n+l} \mathbf{e}_{n+l}^\top) = \sum_{j=0}^{l-1} \mathbf{E}_1^\top \mathbf{\Phi}^{*j} \mathbf{E}_1 \mathbf{\Sigma} \mathbf{E}_1^\top \mathbf{\Phi}^{*j\top} \mathbf{E}_1 + \mathbf{W}_{n,l}^* \mathbf{H}_l^* \mathrm{var}(\tilde{\boldsymbol{\lambda}}) \mathbf{H}_l^{*\top} \mathbf{W}_{n,l}^{*\top} + \mathbf{o}_P(n^{-1}). \tag{2.0.26}$$

Since $\boldsymbol{\lambda}$ is estimated using the RA-VARX estimators, the covariance matrix $\text{var}(\tilde{\boldsymbol{\lambda}})$ in expression (2.0.26) represents the asymptotic covariance matrix of the RA-VARX estimators appearing in formula (2.0.19). Expression (2.0.26) is similar to a result obtained by Baillie (1981) in a zero mean dynamic simultaneous model with lagged endogenous variables and autoregressive errors, where the parameters are estimated by maximum likelihood.

To derive the conditional mean squared prediction error (2.0.26), the asymptotic independence between $\mathbf{Y}_n^*$, $\mathbf{X}_{n+1}^*, \ldots, \mathbf{X}_{n+l}^*$ and the RA-VARX estimator of $\boldsymbol{\lambda}$ is assumed. This is a key assumption which has been verified in many important situations and for many non robust methods, such as maximum likelihood estimators. In fact, if the vector of parameters $\boldsymbol{\lambda}$ is estimated from an independent sample, as in Yamamoto (1976) (see also Ansley and Kohn (1986)), then this hypothesis trivially hold also in our robust framework. However, in the more realistic situation in which the actual observations are used for estimation and forecasting, the asymptotic independence can be easily shown under a Gaussian VARX model for the robust RA-VARX estimators, in the sense that the asymptotic covariance between $\mathbf{Y}_n^*$, $\mathbf{X}_{n+1}^*, \ldots, \mathbf{X}_{n+l}^*$ and the RA-VARX estimator $\tilde{\boldsymbol{\lambda}}$ is precisely zero. This property easily follows since RA-VARX estimators satisfy the asymptotic expansion (2.0.16). For a similar argument in multivariate time series, see Reinsel (1980, pp. 330-331).

Consequently, an estimator of the conditional mean squared prediction error, accurate up to order $\mathbf{O}_P(n^{-1})$, is given by :

$$\hat{E}_n^*(\mathbf{e}_{n+l}\mathbf{e}_{n+l}^\top) = \sum_{j=0}^{l-1} \mathbf{E}_1^\top \tilde{\boldsymbol{\Phi}}^{*j} \mathbf{E}_1 \tilde{\boldsymbol{\Sigma}} \mathbf{E}_1^\top \tilde{\boldsymbol{\Phi}}^{*j\top} \mathbf{E}_1 + \mathbf{W}_{n,l}^* \tilde{\mathbf{H}}_l^* \widehat{\text{var}}(\tilde{\boldsymbol{\lambda}}) \tilde{\mathbf{H}}_l^{*\top} \mathbf{W}_{n,l}^{*\top}, \quad (2.0.27)$$

where each unknown parameter in $\boldsymbol{\Phi}^*$, $\boldsymbol{\Sigma}$, $\mathbf{H}_l$ and $\text{var}(\tilde{\boldsymbol{\lambda}})$ is naturally replaced by the corresponding RA-VARX estimator. Since $\text{var}(\tilde{\boldsymbol{\lambda}}) = \mathbf{O}(n^{-1})$, the second term of expression (2.0.27) is called the $\mathbf{O}_P(n^{-1})$ correction factor; this term is due to parameter estimation uncertainty. For large samples this factor can be omitted with limited loss of accuracy. However, for small and moderate sample sizes, the $\mathbf{O}_P(n^{-1})$ correction factor is expected to improve the finite sample behavior of the prediction intervals. For example, the exact confidence level of the prediction interval should be closer to the nominal confidence level; the proposed intervals are justified by an asymptotic theory. This issue is investigated in the simulation experiments of the next section.

4. SIMULATION RESULTS

In the previous sections, we introduced a new class of robust estimators and we derived robust conditional prediction mean squared errors, accurate up to order $\mathbf{O}_P(n^{-1})$, of multi-step predictions in VARX models. From a practical point of view, it seems natural to inquire about their finite sample properties. In particular, it may be informative for the analyst to investigate empirically the effects of various kinds of outliers on classical point predictions and prediction intervals, based on the usual CLS estimators (with and without the $\mathbf{O}_P(n^{-1})$ correction factor).

Here, we report the simulation results of a small Monte Carlo experiment conducted in order to study the exact prediction biases, prediction mean squared errors (MSE) and the coverage properties of predictions based on CLS estimators and of predictions constructed using the proposed RA-VARX estimators. The following data generating process (DGP) is used :

$$\text{DGP} : (\mathbf{I}_2 - \mathbf{\Phi}_1 B)(\mathbf{Y}_t - \boldsymbol{\mu}) = \mathbf{V}_0 X_t + \mathbf{a}_t, \tag{2.0.28}$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} 1.0 \\ -1.0 \end{pmatrix}, \ \mathbf{\Phi}_1 = \begin{pmatrix} 0.4 & 0.3 \\ 0.3 & 0.4 \end{pmatrix}, \ \mathbf{V}_0 = \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix}.$$

The exogenous process $X = \{X_t, t \in \mathbb{Z}\}$ is assumed to be a sequence of iid uniform $U[-\sqrt{3}, \sqrt{3}]$ random variables, where $U[a, b]$ denotes the uniform distribution on the interval $[a, b]$. The error process $\{\mathbf{a}_t, t \in \mathbb{Z}\}$ represents a Gaussian white noise, with covariance matrix :

$$\mathbf{\Sigma} = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}.$$

The resulting DGP corresponds to a bivariate VARX(1,0) process with a scalar exogenous variable. We consider the sample size $n = 100$. To obtain a realization, we set the initial value $\mathbf{Y}_{-n} = \mathbf{0}$ and we generate $2n + 1 + l_{max}$ observations $\mathbf{Y}_{-n}, \mathbf{Y}_{-n+1}, \ldots, \mathbf{Y}_{-1}, \mathbf{Y}_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_n, \mathbf{Y}_{n+1}, \ldots, \mathbf{Y}_{n+l_{max}}$ according to the DGP given by (2.0.28), where $l_{max} = 12$. We have discarded the first $n + 1$ observations to reduce the impact of the initial value, and we have used the last l_{max} observations for evaluating the predictions; the model (2.0.28) is estimated with observations $(\mathbf{Y}_t, X_t)$, $t = 1, \ldots, n$.

We investigate six strategies for the occurrence of outliers :

Scenario 1 : No contamination.

Scenario 2 : Outliers in the endogenous variables at fixed positions. The outliers are obtained by replacing $\mathbf{Y}_t = (Y_t(1), Y_t(2))^\top$, $t \in \{19, 39, 59, 79, 99, 119, 139, 159, 179, 199\}$, by $\mu(i) - (-1)^i \times 10$, $i = 1, 2$, where $\boldsymbol{\mu} = (\mu(1), \mu(2))^\top$.

Scenario 3 : Outliers in the endogenous variables at fixed positions as in Scenario 2. However, the signs of the contamination are independently and randomly attributed for each variable $Y_t(1)$ and $Y_t(2)$.

Scenario 4 : Outliers in the endogenous variables at random positions, such that an outlier is created with probability 5%. The magnitude of the contamination is 10. For each variable, the position and the signs of the contamination are randomly and independently attributed.

Scenario 5 : Outliers in the endogenous variable at random positions as in Scenario 4. However, for an outlier generated randomly, $Y_t(i)$ is replaced by $\mu(i) - (-1)^i \times 10$, $i = 1, 2$.

Scenario 6 : Patch of outliers. For each variable, two blocks of five successive outliers are created. The position of the blocks is random and created independently for each variable. The sign of the contamination, of a magnitude of 10, is positive for the first variable $Y_t(1)$, and negative for the second variable $Y_t(2)$.

We examined the coverage properties of predictions (at two different confidence levels : 90% and 95% percent), the empirical prediction biases, and the empirical prediction mean squared errors for predicting $Y_t(1)$, $Y_t(2)$, and the linear combination $\boldsymbol{\gamma}^\top \mathbf{Y}_t$, where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^\top$ represents the eigenvector corresponding to the largest eigenvalue obtained maximizing $\text{var}(\boldsymbol{\gamma}^\top \mathbf{Y}_t)$ subject to $\boldsymbol{\gamma}^\top \boldsymbol{\gamma} = 1$. The weights of the linear combination are given by $\boldsymbol{\gamma} = (0.6844, 0.7291)^\top$ and for the purpose of our empirical investigation, the vector $\boldsymbol{\gamma}$ is assumed to be known and fixed in all the simulation experiments. For each scenario, 1000 independent realizations have been generated.

We compare the robust RA-VARX estimators with the usual CLS estimators used for obtaining predictions. The algorithm described in Section 2.3 is used to compute the robust estimators. Prediction limits with and without the $\mathbf{O}_P(n^{-1})$ correction factor are considered. For a sample size of $n = 100$, it appears useful to investigate the potential benefit of increasing the precision of the prediction interval. We construct multi-step predictions for various horizons, that is $l = 1, 6, 12$.

For the robust tests, we include in our experiments Huber and bisquare ψ functions with tuning parameters $\vartheta = 1.49$ and $\vartheta = 5.1$, respectively. For the scale factor (2.0.14) in estimator (2.0.13) defining $\tilde{\boldsymbol{\Sigma}}$, we employed 1.50 and 1.79, respectively. These values are taken from Ben, Martinez and Yohai (1999). If the DGP was a bivariate VAR(1), these values of the robustness constants would deliver an efficiency higher than 90% for the RA estimators comparatively to the CLS estimators under a perfectly generated VAR(1) Gaussian process. When the RA-VARX estimators are computed with the bisquare ψ function, we use the modified algorithm, that is the RA-VARX estimators calculated with the Huber ψ function are taken as the starting values, and $\boldsymbol{\Sigma}$ is estimated at the beginning and the end of the algorithm.

TAB. 2.1. Empirical coverage rates (in percentage) of prediction intervals (P. I.) with and without the $\mathbf{O}_P(n^{-1})$ correction factor (C. F.), prediction biases, prediction mean squared errors, for the prediction of $Y_{n+l}(i)$, $i = 1, 2$ and $\gamma^\top \mathbf{Y}_{n+l}$, for horizons $l = 1, 6, 12$, using CLS estimators and RA-VARX estimators based on Huber and bisquare ψ functions, for the VARX(1,0) model given by (2.0.28), in the case of Scenario 1 (no contamination).

		$Y_{n+l}(1)$			$Y_{n+l}(2)$			$\gamma^\top \mathbf{Y}_{n+l}$		
		$l=1$	$l=6$	$l=12$	$l=1$	$l=6$	$l=12$	$l=1$	$l=6$	$l=12$
		Nominal confidence level : 90%								
CLS	P. I. Coverage	0.886	0.882	0.875	0.890	0.883	0.857	0.891	0.876	0.861
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.913	0.900	0.892	0.913	0.899	0.888	0.911	0.897	0.891
RA-VARX (ψ_H)	P. I. Coverage	0.889	0.876	0.872	0.891	0.887	0.859	0.892	0.876	0.858
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.916	0.905	0.893	0.917	0.904	0.886	0.918	0.901	0.886
RA-VARX (ψ_B)	P. I. Coverage	0.885	0.873	0.868	0.885	0.880	0.857	0.889	0.876	0.859
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.913	0.906	0.891	0.917	0.901	0.889	0.916	0.900	0.886
		Nominal confidence level : 95%								
CLS	P. I. Coverage	0.940	0.937	0.923	0.938	0.931	0.926	0.946	0.934	0.922
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.967	0.954	0.946	0.957	0.949	0.941	0.965	0.949	0.944
RA-VARX (ψ_H)	P. I. Coverage	0.944	0.938	0.928	0.941	0.931	0.920	0.950	0.931	0.924
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.967	0.957	0.949	0.958	0.946	0.943	0.966	0.948	0.944
RA-VARX (ψ_B)	P. I. Coverage	0.943	0.936	0.923	0.940	0.929	0.918	0.944	0.929	0.919
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.965	0.957	0.947	0.957	0.947	0.943	0.966	0.946	0.943
		Empirical prediction bias								
CLS		0.0280	-0.0151	-0.0190	0.0647	0.00278	0.0214	0.0663	-0.0083	0.0026
RA-VARX (ψ_H)		0.0235	-0.0160	-0.0241	0.0638	0.0036	0.0197	0.0626	-0.0083	-0.0022
RA-VARX (ψ_B)		0.0222	-0.0165	-0.0241	0.0630	0.0031	0.0196	0.0612	-0.0090	-0.0022
		Empirical prediction mean squared error								
CLS		0.9802	1.8299	1.8850	1.0340	1.7379	1.9029	1.4523	3.0751	3.2398
RA-VARX (ψ_H)		0.9783	1.8423	1.8873	1.0295	1.7450	1.9043	1.4421	3.0909	3.2409
RA-VARX (ψ_B)		0.9797	1.8424	1.8850	1.0287	1.7466	1.9027	1.4419	3.0923	3.2375

Notes :

(i) The simulation experiments are based on 1000 realizations.

(ii) The linear combination $\gamma^\top \mathbf{Y}_{n+l}$ is such that $\gamma = (0.6844, 0.7291)^\top$.

TAB. 2.2. Empirical coverage rates (in percentage) of prediction intervals (P. I.) with and without the $\mathbf{O}_P(n^{-1})$ correction factor (C. F.), prediction biases, prediction mean squared errors, for the prediction of $Y_{n+l}(i)$, $i = 1, 2$ and $\gamma^\top \mathbf{Y}_{n+l}$, for horizons $l = 1, 6, 12$, using CLS estimators and RA-VARX estimators based on Huber and bisquare ψ functions, for the VARX(1,0) model given by (2.0.28), in the case of Scenario 2 (additive outliers at fixed positions, systematic signs).

		$Y_{n+l}(1)$			$Y_{n+l}(2)$			$\gamma^\top \mathbf{Y}_{n+l}$		
		$l=1$	$l=6$	$l=12$	$l=1$	$l=6$	$l=12$	$l=1$	$l=6$	$l=12$
		Nominal confidence level : 90%								
CLS	P. I. Coverage	1.000	0.997	0.998	0.997	0.997	0.991	0.903	0.878	0.861
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	1.000	1.000	0.998	0.998	0.997	0.992	0.908	0.889	0.879
RA-VARX (ψ_H)	P. I. Coverage	0.902	0.888	0.882	0.905	0.889	0.873	0.895	0.880	0.875
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.911	0.901	0.893	0.913	0.902	0.890	0.905	0.890	0.886
RA-VARX (ψ_B)	P. I. Coverage	0.874	0.857	0.861	0.880	0.872	0.843	0.882	0.863	0.857
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.910	0.902	0.888	0.915	0.903	0.887	0.918	0.893	0.884
		Nominal confidence level : 95%								
CLS	P. I. Coverage	1.000	1.000	0.999	1.000	0.999	0.998	0.951	0.931	0.922
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	1.000	1.000	0.999	1.000	0.999	0.999	0.957	0.944	0.931
RA-VARX (ψ_H)	P. I. Coverage	0.958	0.946	0.938	0.952	0.938	0.931	0.955	0.932	0.925
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.963	0.953	0.949	0.955	0.949	0.940	0.960	0.942	0.939
RA-VARX (ψ_B)	P. I. Coverage	0.933	0.932	0.921	0.935	0.918	0.913	0.939	0.923	0.912
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.968	0.959	0.950	0.957	0.942	0.945	0.965	0.941	0.943
		Empirical prediction bias								
CLS		-0.5568	-0.5293	-0.5127	0.6515	0.5106	0.5109	0.0939	0.0100	0.0216
RA-VARX (ψ_H)		-0.0307	-0.0658	-0.0710	0.1195	0.0570	0.0725	0.0662	-0.0035	0.0042
RA-VARX (ψ_B)		0.0305	-0.0021	-0.0130	0.0522	-0.0057	0.0118	0.0589	-0.0056	-0.0003
		Empirical prediction mean squared error								
CLS		1.3998	2.1412	2.1800	1.6146	2.0622	2.2243	1.4615	3.0512	3.2230
RA-VARX (ψ_H)		0.9794	1.8674	1.9314	1.0409	1.7847	1.9574	1.4355	3.1423	3.3232
RA-VARX (ψ_B)		0.9931	1.8444	1.9237	1.0249	1.7537	1.9448	1.4514	3.0983	3.3148

Notes :

(i) The simulation experiments are based on 1000 realizations.

(ii) The [illegible] $\gamma^\top = (0.6844, 0.7291)^\top$

TAB. 2.3. Empirical coverage rates (in percentage) of prediction intervals (P. I.) with and without the $\mathbf{O}_P(n^{-1})$ correction factor (C. F.), prediction biases, prediction mean squared errors, for the prediction of $Y_{n+l}(i)$, $i = 1, 2$ and $\boldsymbol{\gamma}^\top \mathbf{Y}_{n+l}$, for horizons $l = 1, 6, 12$, using CLS estimators and RA-VARX estimators based on Huber and bisquare ψ functions, for the VARX(1,0) model given by (2.0.28), in the case of Scenario 3 (additive outliers at fixed positions, random signs).

		$Y_{n+l}(1)$			$Y_{n+l}(2)$			$\boldsymbol{\gamma}^\top \mathbf{Y}_{n+l}$		
		$l=1$	$l=6$	$l=12$	$l=1$	$l=6$	$l=12$	$l=1$	$l=6$	$l=12$
		Nominal confidence level : 90%								
CLS	P. I. Coverage	0.999	0.993	0.998	0.997	0.993	0.995	0.988	0.976	0.980
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	1.000	0.994	0.999	0.997	0.994	0.996	0.989	0.976	0.983
RA-VARX (ψ_H)	P. I. Coverage	0.908	0.875	0.892	0.906	0.880	0.899	0.909	0.882	0.891
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.918	0.892	0.906	0.920	0.895	0.918	0.917	0.908	0.910
RA-VARX (ψ_B)	P. I. Coverage	0.863	0.849	0.863	0.855	0.853	0.866	0.855	0.843	0.856
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.904	0.882	0.896	0.905	0.884	0.899	0.894	0.881	0.900
		Nominal confidence level : 95%								
CLS	P. I. Coverage	1.000	0.997	1.000	1.000	0.999	1.000	0.995	0.987	0.996
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	1.000	0.997	1.000	1.000	1.000	1.000	0.996	0.988	0.996
RA-VARX (ψ_H)	P. I. Coverage	0.955	0.937	0.945	0.954	0.935	0.945	0.952	0.936	0.943
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.962	0.954	0.955	0.964	0.948	0.958	0.964	0.945	0.954
RA-VARX (ψ_B)	P. I. Coverage	0.916	0.905	0.912	0.925	0.908	0.918	0.911	0.909	0.917
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.956	0.943	0.949	0.950	0.932	0.951	0.940	0.934	0.944
		Empirical prediction bias								
CLS		-0.0592	-0.0369	-0.0306	-0.0118	-0.0380	-0.0261	-0.0491	-0.0529	-0.0399
RA-VARX (ψ_H)		-0.0250	-0.0409	-0.0341	-0.0009	-0.0478	-0.0278	-0.0178	-0.0629	-0.0436
RA-VARX (ψ_B)		-0.0182	-0.0420	-0.0381	0.0028	-0.0465	-0.0285	-0.0104	-0.0626	-0.0468
		Empirical prediction mean squared error								
CLS		1.5345	2.1533	1.9836	1.6699	2.1525	2.0800	2.6021	3.6769	3.3921
RA-VARX (ψ_H)		1.0784	1.9324	1.8159	1.1088	1.8995	1.8937	1.6942	3.3159	3.1351
RA-VARX (ψ_B)		1.0788	1.9122	1.8088	1.0974	1.8872	1.8990	1.6777	3.2858	3.1331

Notes :

(i) The simulation experiments are based on 1000 realizations.

Table 1 reports the empirical results under Scenario 1. In this scenario, no outliers are created. As expected, point predictions based on CLS estimators for $Y_{n+l}(1)$, $Y_{n+l}(2)$ and $\boldsymbol{\gamma}^\top \mathbf{Y}_{n+l}$, $l = 1, 6, 12$, exhibit small biases and small prediction MSEs. Interestingly, based on 1000 realizations, the prediction biases and MSEs of the robust and non robust methods appear very close, at least in our experiments. The prediction intervals offer reasonable coverage properties at the 90% and 95% confidence levels. Generally, the $\mathbf{O}_P(n^{-1})$ correction term improves the coverage properties, particularly for $\mathbf{Y}_{n+l}(2)$ and for the linear combination $\boldsymbol{\gamma}^\top \mathbf{Y}_{n+l}$ when the horizons l are large.

Empirical results when outliers are located at fixed positions are presented in Table 2. This scenario illustrates that CLS estimators can lead to serious prediction biases and large prediction MSEs. For predicting $Y_{n+l}(1)$ and $Y_{n+l}(2)$, the predictions calculated with CLS estimators can lead to prediction MSEs substantially larger than in Scenario 1 when no outliers are present. The empirical biases and MSEs appeared smaller for the robust methods. Interestingly, the prediction MSEs of the robust estimators are comparable under both Scenarios 1 and 2. Since the outliers are located at fixed positions, of the same magnitude, and such that $\mu(1) + 10$ for the first variable and $\mu(2) - 10$ for the second variable, small prediction bias has been observed for predicting the linear combination. This is not surprising given the definition of the vector $\boldsymbol{\gamma}$. From this experiment, the RA-VARX estimators based on the bisquare ψ function appears to offer the smallest bias in general but the prediction MSE for the Huber and bisquare functions seem rather similar. The prediction intervals based on CLS estimators were strongly affected by the outliers : here, empirical evidence gives severe distortion to the empirical coverage rates for the non robust method. Since the coverage properties of the robust methods are more than reasonable in this situation, this illustrates the merits of using robust methods for constructing prediction intervals in the presence of outliers. Some improvements have been observed using the $\mathbf{O}_P(n^{-1})$ correction factor, particularly for predicting $Y_{n+l}(2)$ and the linear combination in this part of the experiment.

In Table 3, empirical results are presented when outliers are located at fixed positions, but the signs of the contamination are randomly attributed. In this scenario, the prediction biases for predicting $Y_{n+l}(1)$ and the linear combination appeared similar for the robust and non-robust methods. For predicting $Y_{n+l}(2)$, the robust methods offered smaller empirical biases. However, the prediction MSEs based on the CLS estimators were much larger than those based on the RA-VARX estimators. Furthermore, these outliers create a serious inflation of the MSEs for predicting $Y_{n+l}(i)$, $i = 1, 2$ and the linear combination, since the prediction MSEs are much larger in this scenario than the prediction MSEs obtained in Scenario 1. The robust methods offered slightly larger MSEs than the ones corresponding to the outlier-free situation, except for horizon $l = 12$, where the prediction MSEs

TAB. 2.4. Empirical coverage rates (in percentage) of prediction intervals (P. I.) with and without the $\mathbf{O}_P(n^{-1})$ correction factor (C. F.), prediction biases, prediction mean squared errors, for the prediction of $Y_{n+l}(i)$, $i = 1, 2$ and $\boldsymbol{\gamma}^\top \mathbf{Y}_{n+l}$, for horizons $l = 1, 6, 12$, using CLS estimators and RA-VARX estimators based on Huber and bisquare ψ functions, for the VARX(1,0) model given by (2.0.28), in the case of Scenario 4 (additive outliers at random positions, random signs).

		$Y_{n+l}(1)$			$Y_{n+l}(2)$			$\boldsymbol{\gamma}^\top \mathbf{Y}_{n+l}$		
		$l=1$	$l=6$	$l=12$	$l=1$	$l=6$	$l=12$	$l=1$	$l=6$	$l=12$
		Nominal confidence level : 90%								
CLS	P. I. Coverage	0.997	0.992	0.990	0.999	0.993	0.994	0.996	0.978	0.983
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.997	0.992	0.991	0.999	0.995	0.995	0.997	0.981	0.987
RA-VARX (ψ_H)	P. I. Coverage	0.935	0.907	0.904	0.925	0.899	0.898	0.934	0.898	0.899
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.945	0.922	0.925	0.938	0.910	0.916	0.940	0.909	0.911
RA-VARX (ψ_B)	P. I. Coverage	0.860	0.854	0.867	0.876	0.857	0.855	0.877	0.857	0.855
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.914	0.907	0.909	0.924	0.896	0.904	0.925	0.894	0.900
		Nominal confidence level : 95%								
CLS	P. I. Coverage	1.000	0.997	0.998	1.000	1.000	1.000	1.000	0.990	0.998
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	1.000	0.998	0.998	1.000	1.000	1.000	1.000	0.994	0.998
RA-VARX (ψ_H)	P. I. Coverage	0.975	0.957	0.955	0.960	0.937	0.957	0.967	0.940	0.949
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.980	0.960	0.961	0.967	0.950	0.964	0.971	0.950	0.955
RA-VARX (ψ_B)	P. I. Coverage	0.925	0.920	0.917	0.925	0.907	0.918	0.931	0.908	0.919
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.962	0.949	0.955	0.956	0.942	0.952	0.965	0.944	0.954
		Empirical prediction bias								
CLS		0.0140	0.0316	0.0562	-0.0544	0.0457	0.0475	-0.0301	0.0549	0.0731
RA-VARX (ψ_H)		0.0239	0.0416	0.0585	-0.0494	0.0569	0.0591	-0.0197	0.0700	0.0831
RA-VARX (ψ_B)		0.0224	0.0369	0.0599	-0.0486	0.0613	0.0685	-0.0202	0.0700	0.0909
		Empirical prediction mean squared error								
CLS		1.4379	2.1650	1.9471	1.5022	2.2012	2.0446	2.3063	3.7179	3.3764
RA-VARX (ψ_H)		1.0677	1.9083	1.8195	1.0778	1.9615	1.8959	1.6165	3.3247	3.1806
RA-VARX (ψ_B)		1.0626	1.8918	1.8385	1.0434	1.9447	1.9287	1.5746	3.2930	3.2288

Notes :

(i) The simulation experiments are based on 1000 realizations.

(ii) The linear combinations $\boldsymbol{\gamma}^\top \mathbf{Y}_{n+l}$ are such that $\boldsymbol{\gamma} = (0.6644, 0.7391)^\top$

TAB. 2.5. Empirical coverage rates (in percentage) of prediction intervals (P. I.) with and without the $\mathbf{O}_P(n^{-1})$ correction factor (C. F.), prediction biases, prediction mean squared errors, for the prediction of $Y_{n+l}(i)$, $i = 1, 2$ and $\boldsymbol{\gamma}^\top \mathbf{Y}_{n+l}$, for horizons $l = 1, 6, 12$, using CLS estimators and RA-VARX estimators based on Huber and bisquare ψ functions, for the VARX(1,0) model given by (2.0.28), in the case of Scenario 5 (additive outliers at random positions, systematic signs).

		$Y_{n+l}(1)$			$Y_{n+l}(2)$			$\boldsymbol{\gamma}^\top \mathbf{Y}_{n+l}$		
		$l=1$	$l=6$	$l=12$	$l=1$	$l=6$	$l=12$	$l=1$	$l=6$	$l=12$
		Nominal confidence level : 90%								
CLS	P. I. Coverage	0.995	0.990	0.991	0.992	0.985	0.993	0.989	0.978	0.983
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.996	0.991	0.993	0.992	0.986	0.994	0.989	0.982	0.988
RA-VARX (ψ_H)	P. I. Coverage	0.931	0.904	0.911	0.920	0.897	0.902	0.931	0.893	0.904
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.941	0.918	0.923	0.928	0.903	0.923	0.934	0.903	0.917
RA-VARX (ψ_B)	P. I. Coverage	0.856	0.869	0.870	0.870	0.853	0.857	0.874	0.862	0.864
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.914	0.915	0.910	0.924	0.902	0.904	0.919	0.901	0.911
		Nominal confidence level : 95%								
CLS	P. I. Coverage	0.999	0.995	0.995	0.998	0.993	1.000	0.998	0.995	0.997
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.999	0.998	0.995	0.998	0.994	1.000	0.998	0.997	0.998
RA-VARX (ψ_H)	P. I. Coverage	0.966	0.952	0.954	0.958	0.942	0.955	0.963	0.939	0.955
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.969	0.963	0.958	0.965	0.953	0.970	0.965	0.950	0.962
RA-VARX (ψ_B)	P. I. Coverage	0.924	0.918	0.919	0.929	0.914	0.920	0.931	0.912	0.923
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.955	0.949	0.960	0.958	0.949	0.954	0.958	0.947	0.954
		Empirical prediction bias								
CLS		-0.4843	-0.4778	-0.4452	0.4378	0.5491	0.5391	-0.0123	0.0734	0.0884
RA-VARX (ψ_H)		-0.0622	-0.0611	-0.0428	0.0459	0.1500	0.1503	-0.0091	0.0676	0.0803
RA-VARX (ψ_B)		0.0355	0.0519	0.0710	-0.0536	0.0518	0.0593	-0.0147	0.0733	0.0918
		Empirical prediction mean squared error								
CLS		1.7259	2.3457	2.1789	1.7807	2.4958	2.3273	2.3957	3.6386	3.3819
RA-VARX (ψ_H)		1.0911	1.9242	1.8269	1.1053	1.9895	1.8924	1.6525	3.3385	3.1550
RA-VARX (ψ_B)		1.0779	1.9194	1.8293	1.0557	1.9745	1.8837	1.5977	3.3488	3.1742

Notes :

(i) The simulation experiments are based on 1000 realizations.

(ii) The linear combination $\boldsymbol{\gamma}^\top \mathbf{Y}_{n+l}$ is such that $\boldsymbol{\gamma} = (0.6044, 0.7961)^\top$.

TAB. 2.6. Empirical coverage rates (in percentage) of prediction intervals (P. I.) with and without the $\mathbf{O}_P(n^{-1})$ correction factor (C. F.), prediction biases, prediction mean squared errors, for the prediction of $Y_{n+l}(i)$, $i = 1, 2$ and $\boldsymbol{\gamma}^\top \mathbf{Y}_{n+l}$, for horizons $l = 1, 6, 12$, using CLS estimators and RA-VARX estimators based on Huber and bisquare ψ functions, for the VARX(1,0) model given by (2.0.28), in the case of Scenario 6 (patches of outliers).

		$Y_{n+l}(1)$			$Y_{n+l}(2)$			$\boldsymbol{\gamma}^\top \mathbf{Y}_{n+l}$		
		$l=1$	$l=6$	$l=12$	$l=1$	$l=6$	$l=12$	$l=1$	$l=6$	$l=12$
		Nominal confidence level : 90%								
CLS	P. I. Coverage	0.989	0.994	0.996	0.991	0.997	0.991	0.982	0.985	0.983
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.991	0.994	0.996	0.993	0.997	0.993	0.983	0.988	0.985
RA-VARX (ψ_H)	P. I. Coverage	0.911	0.918	0.916	0.918	0.919	0.894	0.904	0.899	0.891
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.921	0.930	0.923	0.926	0.927	0.911	0.908	0.916	0.905
RA-VARX (ψ_B)	P. I. Coverage	0.878	0.864	0.864	0.871	0.859	0.858	0.874	0.866	0.854
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.918	0.904	0.898	0.918	0.906	0.892	0.914	0.904	0.893
		Nominal confidence level : 95%								
CLS	P. I. Coverage	0.997	0.996	0.998	0.999	1.000	0.994	0.999	0.998	0.993
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.997	1.000	0.999	0.999	1.000	0.997	0.999	0.999	0.994
RA-VARX (ψ_H)	P. I. Coverage	0.957	0.958	0.957	0.960	0.960	0.946	0.952	0.944	0.944
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.965	0.968	0.964	0.969	0.967	0.962	0.960	0.951	0.952
RA-VARX (ψ_B)	P. I. Coverage	0.932	0.920	0.916	0.927	0.923	0.908	0.924	0.920	0.914
	P. I. Coverage, $\mathbf{O}_P(n^{-1})$ C. F.	0.965	0.950	0.944	0.969	0.949	0.935	0.960	0.947	0.944
		Empirical prediction bias								
CLS		-0.2211	-0.4416	-0.5699	0.1151	0.3727	0.5010	-0.0674	-0.0305	-0.0247
RA-VARX (ψ_H)		-0.1803	-0.1971	-0.2639	0.0610	0.1138	0.1867	-0.0789	-0.0519	-0.0445
RA-VARX (ψ_B)		-0.0747	0.0086	-0.0466	-0.0399	-0.0781	-0.0187	-0.0802	-0.0511	-0.0455
		Empirical prediction mean squared error								
CLS		1.1568	1.9634	2.1822	1.0919	1.9313	2.2064	1.4486	2.9329	3.1718
RA-VARX (ψ_H)		1.0544	1.7232	1.8755	0.9994	1.7563	1.8993	1.4357	2.8826	3.1154
RA-VARX (ψ_B)		0.9893	1.6498	1.7906	0.9764	1.7436	1.8487	1.4564	2.8749	3.1151

Notes :

(i) The simulation experiments are based on 1000 realizations.

(ii) The linear combination $\boldsymbol{\gamma}^\top \mathbf{Y}_{n+l}$ is such that $\boldsymbol{\gamma} = (0.6844, 0.7291)^\top$.

appeared smaller. Note that, since the sign of the contamination was randomly selected, the prediction MSE of the linear combination, unlike in the previous scenario, was also larger than the one for the robust method. As in the previous scenario, the coverage properties of the robust methods generally display better behaviors than those of the classical method. This scenario illustrates that prediction biases may be small in the presence of outliers for predictions based on the CLS method. However, outliers may lead to very large prediction intervals when a non robust estimation method is used. In this scenario, empirical coverage rates of the prediction intervals based on CLS estimators were not close to the nominal confidence levels, and serious over-coverage has been observed.

The outliers are located at random positions in Scenario 4. The corresponding empirical results are presented in Table 4. In this scenario, the prediction biases appeared small for predicting $Y_{n+l}(i)$, $i = 1, 2$ and the linear combination. However, the distortion of the estimated prediction MSE based on the CLS method appears rather substantial : the classical method offers much higher prediction MSEs than those of the robust methods. Compared to Scenario 1, all the prediction MSEs based on the CLS estimators are larger in this particular experiment. For the robust methods, the MSEs are generally comparable to the empirical results obtained under Scenario 1, but some overestimation of the MSEs appeared at horizon $l = 1$ for predicting $Y_{n+l}(i)$, $i = 1, 2$ and the linear combination. Given the inflation in MSEs of the non-robust methods, the classical prediction intervals exhibit over-coverage as in Scenario 3, while the behavior of the robust prediction intervals were in most cases reasonable. In this scenario, the robust methods based on the Huber and bisquare functions displayed a different behavior. When ψ belongs to the Huber family, the prediction interval gives some over-coverage, particularly at horizon $l = 1$, but when $\psi(\cdot) = \psi_B(\cdot; \vartheta)$, some under-coverage has been observed. The $\mathbf{O}_P(n^{-1})$ correction factor improved the coverage properties in this part of the experiment.

In Scenario 5, presented in Table 5, the outliers are located at random positions, but they are systematically positive for $Y_t(1)$, and systematically negative for $Y_t(2)$. As in Scenario 2, this created large empirical biases for predicting $Y_{n+l}(1)$ and $Y_{n+l}(2)$, but rather negligible ones for predicting the linear combination. However, all the prediction MSEs of the predictions based on the CLS estimators offer a substantial positive bias, compared to the results obtained in the first scenario. As in Scenario 4, over-coverage (under-coverage) has been observed when ψ was in the Huber (bisquare) family.

Finally, Table 6 presents the results of Scenario 6, when the outliers occur in patches of five observations. In this case, important biases of the non-robust methods are observed for $Y_{n+l}(1)$ and $Y_{n+l}(2)$. The empirical biases of the robust method based on the Huber ψ function appear more important than the ones based on the bisquare ψ function, the latter performing the best. For predicting

the linear combination, all the methods performed similarly with respect to biases and prediction MSEs. However, the coverage properties of the robust prediction intervals offered coverage rates much closer to the nominal confidence level. The prediction confidence intervals appeared very satisfactory when ψ was chosen in the Huber family, and generally reasonable when ψ was selected in the bisquare family. The improvement of the correction factor seemed small when ψ was in the Huber family : prediction intervals with and without the correction factors were rather satisfactory in general. However, and rather interestingly, the correction factor improved the coverage properties when ψ was in the bisquare family.

Summarizing :

(1) In general, point predictions and prediction intervals appear unreliable when CLS estimators are used in the presence of outliers. Severe prediction biases and large over-coverage of the prediction intervals can occur.

(2) More precisely, the nature of outliers can affect point predictions very differently. This problem is amplified in multivariate time series. Small prediction biases have been observed for the non robust method when outliers were at random positions with random signs. However, large biases occurred in other experiments ; for example, when outliers with systematic signs and of the same magnitude were created. In a large majority of cases, the scenarios generating outliers have lead to serious over-coverage of prediction intervals based on CLS estimators, at least in our experiments.

(3) In general, predictions based on the robust RA-VARX methods were very reasonable and, in most cases, coverage properties were satisfactory.

(4) When the outliers were located at random positions with random or systematic signs, the predictions, calculated using the robust estimators based on the Huber and bisquare ψ function behaved slightly differently, particularly for small horizons. Robust estimators based on the Huber ψ function offered some over-coverage, while the bisquare ψ function exhibited under-coverage.

(5) In the scenario where the outliers occurred in patches, it appeared preferable to adopt the redescending ψ function in the bisquare family ; the predictions offered small biases and the coverage rates using the $\mathbf{O}_P(n^{-1})$ correction factor were in most cases reasonable, at least in our experiments.

(6) In general, the $\mathbf{O}_P(n^{-1})$ correction factor improved the coverage properties of the robust methods for the moderate sample size $n = 100$, particularly when ψ was chosen in the bisquare family.

5. CONCLUSION

In this article, we considered robust estimation and prediction in multivariate autoregressive models with exogenous variables (VARX). Since the conditional least squares and related procedures are not robust in the presence of outliers, the class of estimators introduced in Duchesne (2005) and Bou Hamad and Duchesne (2005) has been generalized for multivariate time series. The asymptotic distribution of the new estimators has been studied, and from this we obtained in particular the asymptotic covariance matrix of the robust estimators. Based on these results we developed new robust conditional prediction intervals which take into account the variability of parameter estimators. These prediction intervals include a $\mathbf{O}_P(n^{-1})$ correction factor to account for parameter estimation uncertainty, which may, on several important occasions, improve the coverage properties.

In a simulation study, we investigated the finite sample properties of the robust point predictions and of the robust prediction intervals obtained when the parameters were estimated by RA-VARX estimators, these were then compared to the classical point predictions and prediction intervals based on the traditional least squares methodology. Several scenarios for the occurrence of outliers have been considered, including outliers at random and fixed positions, and patches of outliers. We found that the classical predictions were generally unreliable when outliers were present in the time series. Strong prediction biases may be observed and important distortion of the prediction intervals may occur. In our experiments, we found severe over-coverage of the classical intervals. The proposed robust point predictions were much less biased than those based on the least squared method, and the robust conditional prediction intervals offered reasonable coverage properties in most cases. Overall, the proposed robust estimators and robust conditional prediction intervals may be recommended when outliers are suspected in VARX models. If it is true as believed that outliers occur in patches, the RA-VARX estimators based on a redescending ψ function should be appropriate in practical applications.

ACKNOWLEDGEMENTS

Discussions and comments from Christian Léger, W. K. Li and participants in the 34th meeting of the Statistical Society of Canada and of the 2006 Joint Statistical Meeting are gratefully acknowledged. This research is supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Fonds québécois de la recherche sur la nature et les technologies du Québec (P. of Québec).

APPENDIX 1

In this Appendix explicit expressions for the matrices $\mathbf{A}$ and $\mathbf{B}$ are given, where

$\mathbf{A}$ and $\mathbf{B}$ appear in the asymptotic distribution of $\tilde{\boldsymbol{\lambda}}$ stated in (2.0.19). Like Ben, Martinez and Yohai (1999) we assume in the sequel that $E\{\mathbf{a}_t\omega(d_t)\} = \mathbf{0}$. This assumption is satisfied for a large class of distributions : for example, when the distribution of the random vector $\mathbf{a}_t$ is symmetrical.

We first give the asymptotic covariance matrix $\mathbf{A}$, satisfying $\mathbf{A} = (\mathbf{A}_{ij})_{i,j=1,\ldots,p+s+2}$, where the block elements $\mathbf{A}_{ij}$ are given by :

$$\mathbf{A}_{ij} = E\{\boldsymbol{\delta}_{it}(\boldsymbol{\lambda}_0, \boldsymbol{\Sigma}_0)\boldsymbol{\delta}_{jt}^{\top}(\boldsymbol{\lambda}_0, \boldsymbol{\Sigma})\}.$$

To obtain explicit expressions for the $\mathbf{A}_{ij}$'s, the following two lemmas are needed.

Lemma 1. *Consider the VARX(p,s) process $\{\mathbf{Y}_t\}$ with zero mean exogenous process $\{\mathbf{X}_t\}$ defined by (2.0.1). Consider the system (2.0.15) defining the RA-VARX estimators. Then*

(i) $\boldsymbol{\Delta}_{\tilde{\mathbf{r}}\mathbf{X},kt}\mathrm{vec}(\mathbf{V}_i^{\top}) = \sum_{h\geq 0}\mathbf{U}_h\mathbf{V}_i\mathbf{X}_{t-k-h}\otimes\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{r}}_t = \sum_{h\geq 0}(\mathbf{U}_h\mathbf{V}_i\mathbf{X}_{t-k-h}\otimes\boldsymbol{\Sigma}^{-1})\tilde{\mathbf{r}}_t$, $k = 1,\ldots,p+s$;

(ii) $\boldsymbol{\delta}_{\tilde{\mathbf{r}}\tilde{\mathbf{r}},jt} = \sum_{h\geq 0}\mathbf{U}_h\tilde{\mathbf{r}}_{t-j-h}\otimes\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{r}}_t$, $j = 1,\ldots,p$;

(iii) $\boldsymbol{\delta}_{jt} = \mathbf{X}_{t-j+p+1}\otimes\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{r}}_t$, $j = p+1,\ldots,p+s+1$;

(iv) $\boldsymbol{\delta}_{jt} = \left(\sum_{i=0}^{s}\sum_{h\geq 0}\mathbf{U}_h\mathbf{V}_i\mathbf{X}_{t-i-j-h} + \sum_{h\geq 0}\mathbf{U}_h\tilde{\mathbf{r}}_{t-j-h}\right)\otimes\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{r}}_t$, $j = 1,\ldots,p$.

Lemma 2. *Consider $\boldsymbol{\Sigma}_0 = cE(\mathbf{r}_t\mathbf{r}_t^{\top}\omega^2(d_t))$, $d_t^2 = \mathbf{r}_t^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{r}_t$ and $c = d/E\{\psi^2(V^{1/2})\}$, where V follows a χ^2 distribution with d degrees of freedom. Let $\boldsymbol{\Gamma}_{\mathbf{X}}(j) = E(\mathbf{X}_t\mathbf{X}_{t-j}^{\top})$ by the lag-j autocovariance of the zero mean exogenous process $\{\mathbf{X}_t\}$. Then under the same conditions that Lemma 1,*

(i) $E\left\{\boldsymbol{\Delta}_{\tilde{\mathbf{r}}\mathbf{X},(k_1+i)t}\mathrm{vec}(\mathbf{V}_{k_1}^{\top})\mathrm{vec}^{\top}(\mathbf{V}_{k_2}^{\top})\boldsymbol{\Delta}_{\tilde{\mathbf{r}}\mathbf{X},(k_2+j)t}^{\top}\right\} = c^{-1}\sum_{h_1\geq 0}\sum_{h_2\geq 0}\mathbf{U}_{h_1}\mathbf{V}_{k_1}\boldsymbol{\Gamma}_{\mathbf{X}}(k_2-k_1+h_2-h_1+j-1)\mathbf{V}_{k_2}^{\top}\mathbf{U}_{h_2}^{\top}\otimes\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1}$, $i,j\in\{1,\ldots,p\}$;

(ii) $E(\boldsymbol{\delta}_{\tilde{\mathbf{r}}\tilde{\mathbf{r}},it}\boldsymbol{\delta}_{\tilde{\mathbf{r}}\tilde{\mathbf{r}},jt}^{\top}) = c^{-2}\sum_{h\geq 0}\mathbf{U}_h\boldsymbol{\Sigma}_0\mathbf{U}_{h+i-j}^{\top}\otimes\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1}$, $i,j\in\{1,\ldots,p\}$;

(iii) $E\left\{\boldsymbol{\Delta}_{\tilde{\mathbf{r}}\mathbf{X},(k+i)t}\mathrm{vec}(\mathbf{V}_k^{\top})\boldsymbol{\delta}_{jt}^{\top}\right\} = c^{-1}\sum_{h\geq 0}\mathbf{U}_h\mathbf{V}_i\boldsymbol{\Gamma}_{\mathbf{X}}(j-i-h-k-p-1)\otimes\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1}$, $i\in\{1,\ldots,p\}$ *and* $j\in\{p+1,\ldots,p+s+1\}$.

Proofs of Lemma 1 and Lemma 2 :

Concerning the proof of Lemma 1, (i), (ii) and (iii) follows directly using the properties $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^{\top}\otimes\mathbf{A})\mathrm{vec}(\mathbf{B})$, $\mathrm{vec}(\mathbf{ab}^{\top}) = \mathbf{b}\otimes\mathbf{a}$, where $\mathbf{a}$ and $\mathbf{b}$ are d-vectors, and noting also that $(\mathbf{a}\otimes\mathbf{B})\mathbf{C} = (\mathbf{a}\otimes\mathbf{BC})$ if $\mathbf{B}$ and $\mathbf{C}$ are compatible and $\mathbf{a}$ is a vector. The proof of (iv) is a consequence of (i) and (ii). The proof of Lemma 2 involves straightforward algebraic manipulations and is therefore omitted. □

To compute the asymptotic matrix $\mathbf{A}$, it suffices to derive the asymptotic limits of the covariances $\mathrm{cov}(n^{-1/2}\mathbf{L}_i(\boldsymbol{\lambda}_0,\boldsymbol{\Sigma}_0), n^{-1/2}\mathbf{L}_j(\boldsymbol{\lambda}_0,\boldsymbol{\Sigma}_0))$, $i,j\in\{1,\ldots,p+s+2\}$. More precisely, using the stated Lemmas 1 and 2, it follows that the block matrices $\mathbf{A}_{ij}$, $i,j\in\{1,\ldots,p+s+2\}$ composing the blocks of the matrix $\mathbf{A}$ are

given by :

$$\begin{aligned}
\mathbf{A}_{ij} &= c^{-1}\sum_{k_1=0}^{s}\sum_{k_2=0}^{s}\sum_{h_1\geq 0}\sum_{h_2\geq 0}\mathbf{U}_{h_1}\mathbf{V}_{k_1}\boldsymbol{\Gamma}_{\mathbf{X}}(k_2-k_1+h_2-h_1+j-i)\mathbf{V}_{k_2}^{\top}\mathbf{U}_{h_2}^{\top}\otimes\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1} \\
&\quad +c^{-2}\sum_{h\geq 0}\mathbf{U}_h\boldsymbol{\Sigma}_0\mathbf{U}_{h+i-j}^{\top}\otimes\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1},\ i,j\in\{1,\ldots,p\}; \\
\mathbf{A}_{ij} &= c^{-1}\sum_{k=0}^{s}\sum_{h\geq 0}\mathbf{U}_h\mathbf{V}_k\boldsymbol{\Gamma}_{\mathbf{X}}(j-i-h-k-p-1)\otimes\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1},\ i\in\{1,\ldots,p\}, \\
&\quad j\in\{p+1,\ldots,p+s+1\}; \\
\mathbf{A}_{ij} &= \mathbf{0},\ i\in\{1,\ldots,p\},\ j=p+s+2; \\
\mathbf{A}_{ij} &= c^{-1}\boldsymbol{\Gamma}_{\mathbf{X}}(j-i)\otimes\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1},\ i,j\in\{p+1,\ldots,p+s+1\}; \\
\mathbf{A}_{ij} &= \mathbf{0},\ i\in\{p+1,\ldots,p+s+1\},\ j=p+s+2; \\
\mathbf{A}_{ij} &= c^{-1}\boldsymbol{\Sigma}_0, i=j=p+s+2.
\end{aligned}$$

To obtain the asymptotic matrix of variances and covariances of the RA-VARX estimators, we also need explicit expressions for matrix $\mathbf{B}$. The following lemma is needed at various places in the derivation of the asymptotic matrix $\mathbf{B}$.

Lemma 3. *Let* $\tilde{\mathbf{r}}_t = \mathbf{r}_t\omega(d_t)$, $d_t^2 = \mathbf{r}_t^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{r}_t$, $\omega(x) = \psi(x)/x$, $\omega^*(x) = \{\psi'(x)x - \psi(x)\}/x^3$. *For an arbitrary vector* $\boldsymbol{\beta}$, *it follows that :*

$$\frac{\partial\tilde{\mathbf{r}}_t}{\partial\boldsymbol{\beta}^{\top}} = \{\omega(d_t)\mathbf{I}_d + \omega^*(d_t)\mathbf{r}_t\mathbf{r}_t^{\top}\boldsymbol{\Sigma}^{-1}\}\frac{\partial\mathbf{r}_t}{\partial\boldsymbol{\beta}^{\top}}.$$

Proof of Lemma 3 :
Noting that $\partial\mathbf{d}_t/\partial\boldsymbol{\beta}^{\top} = \{\mathbf{r}_t^{\top}\boldsymbol{\Sigma}^{-1}\partial\mathbf{r}_t/\partial\boldsymbol{\beta}^{\top}\}/d_t$ and consequently $\partial\omega(d_t)/\partial\boldsymbol{\beta}^{\top} = \omega^*(d_t)\mathbf{r}_t^{\top}\boldsymbol{\Sigma}^{-1}\partial\mathbf{r}_t/\partial\boldsymbol{\beta}^{\top}$ yield the stated result. □

Let $v = E\{\omega(d_t)\}$, $\mathbf{Q} = E\{\omega(d_t)\mathbf{a}_t\mathbf{a}_t^{\top}\}$, $\mathbf{Q}^* = E\{\omega^*(d_t)\mathbf{a}_t\mathbf{a}_t^{\top}\}$. By definition, the matrix $\mathbf{B} = (\mathbf{B}_{ij})_{i,j=1,\ldots,p+s+2}$ is composed of the following blocks :

$$\mathbf{B} = \left(\frac{\partial\boldsymbol{\delta}_t(\boldsymbol{\lambda}_0,\boldsymbol{\Sigma}_0)}{\partial\mathrm{vec}^{\top}(\boldsymbol{\Phi}_1)}\cdots\frac{\partial\boldsymbol{\delta}_t(\boldsymbol{\lambda}_0,\boldsymbol{\Sigma}_0)}{\partial\mathrm{vec}^{\top}(\boldsymbol{\Phi}_p)}\ \frac{\partial\boldsymbol{\delta}_t(\boldsymbol{\lambda}_0,\boldsymbol{\Sigma}_0)}{\partial\mathrm{vec}^{\top}(\mathbf{V}_0)}\ \frac{\partial\boldsymbol{\delta}_t(\boldsymbol{\lambda}_0,\boldsymbol{\Sigma}_0)}{\partial\mathrm{vec}^{\top}(\mathbf{V}_1)}\cdots\frac{\partial\boldsymbol{\delta}_t(\boldsymbol{\lambda}_0,\boldsymbol{\Sigma}_0)}{\partial\mathrm{vec}^{\top}(\mathbf{V}_s)}\ \frac{\partial\boldsymbol{\delta}_t(\boldsymbol{\lambda}_0,\boldsymbol{\Sigma}_0)}{\partial\boldsymbol{\mu}^{\top}}\right).$$

Using Lemmas 1-3, standard results on vector and matrix differentiation (see, e.g., Lütkepohl, 1993, Appendix A.13), and the derivatives (2.0.3), the blocks $\mathbf{B}_{ij}$, $i,j = 1,\ldots,p+s+2$ can be explicitly given. The following formulas give the

precise expressions for the blocks composing the matrix $\mathbf{B}$:

$$
\begin{aligned}
\mathbf{B}_{ij} &= -\sum_{k_1=0}^{s}\sum_{k_2=0}^{s}\sum_{h_1\geq 0}\sum_{h_2\geq 0}\mathbf{U}_{h_1}\mathbf{V}_{k_1}\boldsymbol{\Gamma}_{\mathbf{X}}(k_2+h_2-k_1-h_1+j-i)\mathbf{V}_{k_2}^{\top}\mathbf{U}_{h_2}^{\top} \\
&\quad \otimes(v\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-1}\mathbf{Q}^{*}\boldsymbol{\Sigma}^{-1}) \\
&\quad - \sum_{h\geq 0}\mathbf{U}_h\mathbf{Q}\mathbf{U}_{h+i-j}^{\top}\otimes(v\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-1}\mathbf{Q}^{*}\boldsymbol{\Sigma}^{-1}),\quad i,j\in\{1,\ldots,p\}; \\
\mathbf{B}_{ij} &= -\sum_{k=0}^{s}\sum_{h\geq 0}\mathbf{U}_h\mathbf{V}_k\boldsymbol{\Gamma}_{\mathbf{X}}(j-i-k-h-p-1)\otimes(v\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-1}\mathbf{Q}^{*}\boldsymbol{\Sigma}^{-1}), \\
&\quad i\in\{1,\ldots,p\}, j\in\{p+1,\ldots,p+s+1\}; \\
\mathbf{B}_{ij} &= \mathbf{0},\quad i\in 1,\ldots,p,\ j=p+s+2; \\
\mathbf{B}_{ij} &= -\sum_{k=0}^{s}\sum_{h\geq 0}\boldsymbol{\Gamma}_{\mathbf{X}}(k+h+j-i+p+1)\mathbf{V}_k^{\top}\mathbf{U}_h^{\top}\otimes(v\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-1}\mathbf{Q}^{*}\boldsymbol{\Sigma}^{-1}), \\
&\quad i\in\{p+1,\ldots,p+s+1\}, j\in\{1,\ldots,p\}; \\
\mathbf{B}_{ij} &= -\{\boldsymbol{\Gamma}_{\mathbf{X}}(j-i)\otimes(v\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-1}\mathbf{Q}^{*}\boldsymbol{\Sigma}^{-1})\},\quad i,j\in\{p+1,\ldots,p+s+1\}; \\
\mathbf{B}_{ij} &= \mathbf{0},\quad i\in\{p+1,\ldots,p+s+1\}, j=p+s+2; \\
\mathbf{B}_{ij} &= \mathbf{0},\quad i=p+s+2, j\in\{1,\ldots,p\}; \\
\mathbf{B}_{ij} &= \mathbf{0},\quad i=p+s+2, j\in\{p+1,\ldots,p+s+1\}; \\
\mathbf{B}_{ij} &= -(v\mathbf{I}+\mathbf{Q}^{*}\boldsymbol{\Sigma}^{-1})\boldsymbol{\Phi}(1),\quad i=j=p+s+2.
\end{aligned}
$$

Note that when no exogenous variables are present in model (2.0.1), we retrieve the asymptotic matrices obtained by Ben, Martinez and Yohai (1999) for VAR(p) processes. Matrices $\mathbf{A}$ and $\mathbf{B}$ also represent multivariate generalizations of univariate results derived in Bou Hamad and Duchesne (2005) for ARX models.

APPENDIX 2

In practice, the computation of the robust conditional prediction limits involve the computation of the matrix of derivatives $\partial\boldsymbol{\alpha}^{*}/\partial\boldsymbol{\lambda}^{\top}$, where $\boldsymbol{\alpha}^{*}$ is defined by (2.0.25). The purpose of this appendix is to give explicit expressions for this matrix.

It is useful at this point to note that :

$$
\boldsymbol{\lambda}=(\mathrm{vec}^{\top}(\boldsymbol{\Phi}),\mathrm{vec}^{\top}(\mathbf{V}),\boldsymbol{\mu}^{\top})^{\top}=(\mathrm{vec}^{\top}(\mathbf{E}_1^{\top}\boldsymbol{\Phi}^{*}),\mathrm{vec}^{\top}(\mathbf{E}_1^{\top}\mathbf{V}_0^{*}),\boldsymbol{\mu}^{\top})^{\top}.
$$

The following lemma is used at various places in the proof.

Lemma 4. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two matrices of dimension* $m \times n$ *and* $p \times n$*, respectively. Then*

$$\partial \mathrm{vec} \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} / \partial \mathrm{vec}^\top(\mathbf{A}) = \mathbf{I}_n \otimes \begin{pmatrix} \mathbf{I}_m \\ \mathbf{0}_{p\times m} \end{pmatrix},$$

where $\mathbf{0}_{p\times m}$ *is the null matrix of dimension* $p \times m$.
Proof of Lemma 4 :
Since

$$\mathrm{vec} \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} = \mathrm{vec} \begin{pmatrix} \mathbf{A} \\ \mathbf{0}_{p\times n} \end{pmatrix} + \mathrm{vec} \begin{pmatrix} \mathbf{0}_{m\times n} \\ \mathbf{B} \end{pmatrix},$$

this implies that :

$$\partial \mathrm{vec} \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} / \partial \mathrm{vec}^\top(\mathbf{A}) = \partial \mathrm{vec} \begin{pmatrix} \mathbf{A} \\ \mathbf{0}_{p\times n} \end{pmatrix} / \partial \mathrm{vec}^\top(\mathbf{A}).$$

Then the conclusion of Lemma 4 follows, noting that $\begin{pmatrix} \mathbf{A} \\ \mathbf{0}_{p\times n} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_m \\ \mathbf{0}_{p\times m} \end{pmatrix} \mathbf{A}$ and using the well-known rule : $\partial \mathrm{vec}(\mathbf{A}_1\mathbf{A}_2)/\partial \boldsymbol{\lambda}^\top = (\mathbf{I}_q \otimes \mathbf{A}_1)\partial \mathrm{vec}(\mathbf{A}_2)/\partial \boldsymbol{\lambda}^\top + (\mathbf{A}_2^\top \otimes \mathbf{I}_n)\partial \mathrm{vec}(\mathbf{A}_1)/\partial \boldsymbol{\lambda}^\top$, $\mathbf{A}_1$ and $\mathbf{A}_2$ being $n \times p$ and $p \times q$ matrices. □

As useful applications of Lemma 4, note that the following relations hold :

$$\frac{\partial \mathrm{vec}(\boldsymbol{\Phi}^*)}{\partial \mathrm{vec}^\top(\mathbf{E}_1^\top \boldsymbol{\Phi}^*)} = \mathbf{I}_{dp} \otimes \mathbf{E}_1, \tag{2.0.29}$$

$$\frac{\partial \mathrm{vec}(\mathbf{V}_0^*)}{\partial \mathrm{vec}^\top(\mathbf{E}_1^\top \mathbf{V}_0^*)} = \mathbf{I}_{m(s+1)} \otimes \mathbf{E}_1. \tag{2.0.30}$$

The next lemma contains the derivatives of $\boldsymbol{\theta}^*$ with respect to $\boldsymbol{\Phi}$ and $\boldsymbol{\mu}$.

Lemma 5. *Let* $\boldsymbol{\Phi}^*$ *and* $\boldsymbol{\theta}^* = (\mathbf{I} - \boldsymbol{\Phi}^*)\boldsymbol{\mu}^*$ *be defined as in relations (2.0.20) and (2.0.21). Then*

$$\frac{\partial \boldsymbol{\theta}^*}{\partial \mathrm{vec}^\top(\mathbf{E}_1^\top \boldsymbol{\Phi}^*)} = -(\boldsymbol{\mu}^{*\top} \otimes \mathbf{E}_1),$$

$$\frac{\partial \boldsymbol{\theta}^*}{\partial \boldsymbol{\mu}^\top} = (\mathbf{I} - \boldsymbol{\Phi}^*)(1_p \otimes \mathbf{I}_d).$$

Proof of Lemma 5 :
The proof follows using standard differentiation rules and relation (2.0.29). □

Using Lemma 5, important intermediate results can be established, which are contained in the next Lemma.

Lemma 6. *Let* $\mathbf{\Phi}^*$, $\mathbf{V}_0^*$, *and* $\boldsymbol{\theta}^* = (\mathbf{I} - \mathbf{\Phi}^*)\boldsymbol{\mu}^*$ *be defined as in relations (2.0.20) and (2.0.21). Then*

$$
\begin{aligned}
(i)\quad & \frac{\partial \mathrm{vec}(\mathbf{E}_1^\top \mathbf{\Phi}^{*l})}{\partial \mathrm{vec}^\top(\mathbf{E}_1^\top \mathbf{\Phi}^*)} &&= \sum_{j=0}^{l-1} (\mathbf{\Phi}^{*\top})^{l-j-1} \otimes \mathbf{E}_1^\top \mathbf{\Phi}^{*j} \mathbf{E}_1, \\
(ii)\quad & \frac{\partial \mathrm{vec}(\mathbf{E}_1^\top \mathbf{\Phi}^{*i} \mathbf{V}_0^*)}{\partial \mathrm{vec}^\top(\mathbf{E}_1^\top \mathbf{\Phi}^*)} &&= \sum_{j=0}^{i-1} \mathbf{V}_0^{*\top} (\mathbf{\Phi}^{*\top})^{i-j-1} \otimes \mathbf{E}_1^\top \mathbf{\Phi}^{*j} \mathbf{E}_1, \\
(iii)\quad & \frac{\partial \mathrm{vec}(\mathbf{E}_1^\top \mathbf{\Phi}^{*i} \mathbf{V}_0^*)}{\partial \mathrm{vec}^\top(\mathbf{E}_1^\top \mathbf{V}_0^*)} &&= \mathbf{I}_{m(s+1)} \otimes \mathbf{E}_1^\top \mathbf{\Phi}^{*i} \mathbf{E}_1, \\
(iv)\quad & \frac{\partial \mathrm{vec}(\mathbf{E}_1^\top \mathbf{\Phi}^{*i} \boldsymbol{\theta}^*)}{\partial \mathrm{vec}^\top(\mathbf{E}_1^\top \mathbf{\Phi}^*)} &&= -(\boldsymbol{\mu}^{*\top} \otimes \mathbf{E}_1^\top \mathbf{\Phi}^{*i} \mathbf{E}_1) + \sum_{j=0}^{i-1} \boldsymbol{\theta}^{*\top} (\mathbf{\Phi}^{*\top})^{i-j-1} \otimes \mathbf{E}_1^\top \mathbf{\Phi}^{*j} \mathbf{E}_1, \\
(v)\quad & \frac{\partial \mathrm{vec}(\mathbf{E}_1^\top \boldsymbol{\theta}^*)}{\partial \mathrm{vec}^\top(\mathbf{E}_1^\top \mathbf{\Phi}^*)} &&= -(\boldsymbol{\mu}^* \otimes \mathbf{E}_1^\top \mathbf{E}_1), \\
(vi)\quad & \frac{\partial \mathrm{vec}(\mathbf{E}_1^\top \mathbf{\Phi}^{*i} \boldsymbol{\theta}^*)}{\partial \boldsymbol{\mu}^\top} &&= \mathbf{E}_1^\top \mathbf{\Phi}^{*i} (\mathbf{I} - \mathbf{\Phi}^*)(1_p \otimes \mathbf{I}_d), \\
(vii)\quad & \frac{\partial \mathrm{vec}(\mathbf{E}_1^\top \boldsymbol{\theta}^*)}{\partial \boldsymbol{\mu}^\top} &&= \mathbf{E}_1^\top (\mathbf{I} - \mathbf{\Phi}^*)(1_p \otimes \mathbf{I}_d).
\end{aligned}
$$

Proof of Lemma 6 :
The proof of the lemma follows using standard differentiation results. In particular relation (2.0.29) and the following rule appears useful at various places for proving (i), (ii), (iv) :

$$
\frac{\partial \mathrm{vec}(\mathbf{A}^h)}{\partial \boldsymbol{\beta}^\top} = \left\{ \sum_{i=0}^{h-1} (\mathbf{A}^\top)^{h-1-i} \otimes \mathbf{A}^i \right\} \frac{\partial \mathrm{vec}(\mathbf{A})}{\partial \boldsymbol{\beta}^\top},
$$

where $\mathbf{A}$ is a square matrix and $\boldsymbol{\beta}$ is a vector, see for example Lütkepohl (1993, p. 471). The relation (iii) is obtained using relation (2.0.30). Finally, formulas (v), (vi) and (vii) are applications of Lemma 5. This concludes the proof of the Lemma. □

We now give an explicit expression for the matrix :

$$
\frac{\partial \boldsymbol{\alpha}^*}{\partial \boldsymbol{\lambda}^\top} = \left(\frac{\partial \boldsymbol{\alpha}^*}{\partial \mathrm{vec}^\top(\mathbf{E}_1^\top \mathbf{\Phi}^*)} \quad \frac{\partial \boldsymbol{\alpha}^*}{\partial \mathrm{vec}^\top(\mathbf{E}_1^\top \mathbf{V}_0^*)} \quad \frac{\partial \boldsymbol{\alpha}^*}{\partial \boldsymbol{\mu}^\top} \right),
$$

where it is useful to recall the definition of the vector $\boldsymbol{\alpha}^*$:

$$
\boldsymbol{\alpha}^* = (\mathrm{vec}(\mathbf{E}_1^\top \mathbf{\Phi}^{*l})^\top, \mathrm{vec}(\mathbf{E}_1^\top \mathbf{\Phi}^{*l-1} \mathbf{V}_0^*)^\top, \ldots, \mathrm{vec}(\mathbf{E}_1^\top \mathbf{V}_0^*)^\top, \mathrm{vec}(\mathbf{E}_1^\top \mathbf{\Phi}^{*l-1} \boldsymbol{\theta}^*)^\top, \ldots, \mathrm{vec}(\mathbf{E}_1^\top \boldsymbol{\theta}^*)^\top
$$

which is a vector composed of $2l+1$ sub-vectors, and $\boldsymbol{\lambda} = (\mathrm{vec}^\top(\mathbf{E}_1^\top \mathbf{\Phi}^*), \mathrm{vec}^\top(\mathbf{E}_1^\top \mathbf{V}_0^*), \boldsymbol{\mu}^\top)^\top$, which includes three sub-vectors. It is convenient to partition this matrix using

blocks : $\partial\boldsymbol{\alpha}^*/\partial\boldsymbol{\lambda}^\top = \mathbf{H}_l^* = (\mathbf{H}_{ij.l}^*)_{i=1,\ldots,2l+1,j=1.2.3}$. As an immediate consequence of Lemma 6, we obtain that the matrix $\mathbf{H}_l^*$ is given by :

$$
\begin{aligned}
\mathbf{H}_{11.l}^* &= \sum_{j=0}^{l-1}(\boldsymbol{\Phi}^{*\top})^{l-j-1}\otimes\mathbf{E}_1^\top\boldsymbol{\Phi}^{*j}\mathbf{E}_1;\\
\mathbf{H}_{i1.l}^* &= \sum_{j=0}^{l-i}\mathbf{V}_0^{*\top}(\boldsymbol{\Phi}^{*\top})^{l-i-j}\otimes\mathbf{E}_1^\top\boldsymbol{\Phi}^{*j}\mathbf{E}_1,\quad i=2,\ldots,l;\\
\mathbf{H}_{(l+1)1.l}^* &= \mathbf{0};\\
\mathbf{H}_{i1.l}^* &= -(\boldsymbol{\mu}^{*\top}\otimes\mathbf{E}_1^\top\boldsymbol{\Phi}^{*2l+1-i}\mathbf{E}_1)+\sum_{j=0}^{2l-i}\boldsymbol{\theta}^{*\top}(\boldsymbol{\Phi}^{*\top})^{2l-i-j}\otimes\mathbf{E}_1^\top\boldsymbol{\Phi}^{*j}\mathbf{E}_1,\\
&\quad i=l+2,\ldots,2l;\\
\mathbf{H}_{(2l+1)1.l}^* &= -(\boldsymbol{\mu}^{*\top}\otimes\mathbf{E}_1^\top\mathbf{E}_1);\\
\mathbf{H}_{12.l}^* &= \mathbf{0};\\
\mathbf{H}_{i2.l}^* &= \mathbf{I}_{m(s+1)}\otimes\mathbf{E}_1^\top\boldsymbol{\Phi}^{*l-i+1}\mathbf{E}_1,\quad i=2,\ldots,l;\\
\mathbf{H}_{(l+1)2.l}^* &= \mathbf{I}_{dm(s+1)};\\
\mathbf{H}_{i2.l}^* &= \mathbf{0};\quad i=l+2,\ldots,2l+1;\\
\mathbf{H}_{i3.l}^* &= \mathbf{0};i=1,\ldots,l+1;\\
\mathbf{H}_{i3.l}^* &= \mathbf{E}_1^\top\boldsymbol{\Phi}^{*2l+1-i}(\mathbf{I}-\boldsymbol{\Phi}^*)(\mathbf{1}_p\otimes\mathbf{I}_d),\quad i=l+2,\ldots,2l;\\
\mathbf{H}_{(2l+1)3.l}^* &= \mathbf{E}_1^\top(\mathbf{I}-\boldsymbol{\Phi}^*)(\mathbf{1}_p\otimes\mathbf{I}_d).
\end{aligned}
$$

REFERENCES

Alonso, A. M., Peña, D. and Romo, J. (2002), 'Forecasting time series with sieve bootstrap', *Journal of Statistical Planning and Inference* **100**, 1–11.

Ansley, C. F. and Kohn, R. (1986), 'Prediction mean squared error for state space models with estimated parameters', *Biometrika* **73**, 467–473.

Baillie, R. T. (1981), 'Prediction from the dynamic simultaneous equation model with vector autoregressive errors', *Econometrica* **49**, 1331–1337.

Baillie, R. T. (1980), 'Prediction from ARMAX models', *Journal of Econometrics* **12**, 365–374.

Baillie, R. T. (1979), 'Asymptotic prediction mean squared error for vector autoregressive models', *Biometrika* **66**, 675–678.

Basawa, I. V., Huggins, R. M. and Staudte, R. G. (1985), 'Robust tests for time series with an application to first-order autoregressive processes', *Biometrika* **72**, 559–571.

Ben, M. G., Martinez, E. J. and Yohai, V. (1999), 'Robust estimation in vector autoregressive moving-average models', *Journal of Time Series Analysis* **20**, 381–399.

Bou Hamad, I. and Duchesne, P. (2005), 'On robust diagnostics at individual lags using RA-ARX estimators', in *Statistical Modeling and Analysis for Complex Data Problems*, edited by P. Duchesne and B. Rémillard, Springer : New York.

Cao, R. (1999), 'An overview of bootstrap methods for estimating and predicting in time series', *Test* **8**, 95–116.

Chatfield, C. (2001), 'Prediction intervals for time-series forecasting', in *Principles of Forecasting, A Handbook for Researchers and Practitioners*, edited by J. Scott Armstrong, Kluwer Academic Publishers : Norwell.

de Luna, X. (2000), 'Prediction intervals based on autoregression forecasts', *The Statistician* **49**, 87–93.

Duchesne, P. (2005), 'Robust and powerful serial correlation tests with new robust estimates in ARX models', *Journal of Time Series Analysis* **26**, 49–81.

Durrett, R. (1995), *Probability, Theory and Examples*, Second Edition, Duxbury Press : Belmont.

Fuller, W. A. and Hasza, D. P. (1981), 'Properties of predictors for autoregressive time series', *Journal of the American Statistical Association* **76**, 155–161.

Grigoletto, M. (1998), 'Bootstrap prediction intervals for autoregressions : some alternatives', *International Journal of Forecasting* **14**, 447–456.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986), *Robust Statistics : The Approach Based on Influence Functions*, Wiley : New York.

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

Harville, D. A. (1997), *Matrix Algebra from a Statistician's Perspective*, Springer-Verlag : Berlin.

Judge, G. G., Hill, R. C., Grittiths, W. E., Lütkepohl, H. and Lee, T.-C. (1985), *The Theory and Practice of Econometrics*, Second Edition, Wiley : New York.

Kabaila, P. and He, Z. (2004), 'The adjustment of prediction intervals to account for errors in parameter estimation', *Journal of Time Series Analysis* **25**, 351–358.

Kunitomo, N. and Yamamoto, T. (1985), 'Properties of predictors in misspecified autoregressive time series models', *Journal of the American Statistical Association* **80**, 941–950.

Ledolter, J. (1989), 'The effect of additive outliers on the forecasts from ARIMA models', *International Journal of Forecasting* **5**, 231–240.

Li, W. K. (2004), *Diagnostic Checks in Time Series*, Chapman & Hall/CRC : New York.

Li, W. K. and Hui, Y. V. (1989), 'Robust multiple time series modelling', *Biometrika* **76**, 309–315.

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

Martin, R. D. (1980), 'Robust estimation of autoregressive models', in *Directions in Time Series*, edited by D.C. Brillinger and G.C. Tiao, Institute of Mathematical Statistics, Hagwood, CA.

Masarotto, G. (1990), 'Bootstrap prediction intervals for autoregressions', *International Journal of Forecasting* **6**, 229–239.

Reinsel, G. (1980), 'Asymptotic properties of prediction errors for the multivariate autoregressive model with estimated parameters', *Journal of the Royal Statistical Society, Series B* **42**, 328–333.

Reinsel, G. and Lewis, R. A. (1987), 'Prediction mean square error for non-stationary multivariate time series using estimated parameters', *Economics Letters* **24**, 57–61.

Schmidt, P. (1977), 'Some small sample evidence on the distribution of dynamic simulation forecasts', *Econometrica* **45**, 997–1006.

Schorfheide, F. (2005), 'VAR forecasting under misspecification', *Journal of Econometrics* **128**, 99–136.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley : New York.

Stine, R. A. (1987), 'Estimating properties of autoregressive forecasts', *Journal of the American Statistical Association* **82**, 1072–1078.

Yamamoto, T. (1981), 'Prediction of multivariate autoregressive-moving average models', *Biometrika* **68**, 485–492.

Yamamoto, T. (1980), 'On the treatment of autocorrelated errors in the multiperiod prediction of dynamic simultaneous equation models', *International Economic Review* **21**, 735–748.

Yamamoto, T. (1976), 'Asymptotic mean square prediction error for an autoregressive model with estimated coefficients', *Applied Statistics* **25**, 123–127.

CONCLUSION

Le but de ce mémoire était de présenter des intervalles de prévision qui peuvent être utilisés avec des estimateurs robustes dans les modèles VARX. Une autre problématique importante reposait sur l'absence de considération pour la variabilité due à l'estimation des paramètres dans les formules existantes pour ces modèles. Les méthodes utilisées pour développer un terme de correction pour la variabilité due à l'estimation des paramètres dans les modèles VAR ont été appliquées pour obtenir un terme de correction dans la classe plus générale des modèles VARX.

Les estimateurs robustes RA-VARX ont été définis pour cette classe de modèles en s'inspirant des estimateurs robustes RA-ARX existants pour les modèles univariés ARX. Les résultats portant sur les estimateurs robustes RA-VAR existants pour les modèles VAR ont aussi été utiles. Leurs propriétés asymptotiques ont été présentées. Notamment, des expression matricielles précises pour la variance asymptotique ont été obtenues.

Des expériences de simulation ont permis d'évaluer la qualité des taux de couverture pour les intervalles de prévision avec différentes fonctions robustes soit la fonction dite de Huber et celle dite Bi-carré dans différents contextes de contamination par des valeurs aberrantes. Ces taux de couverture basés sur les estimateurs robustes RA-VARX ont été comparés aux taux de couverture obtenus avec les estimateurs des moindres carrés sous différent scénarios de contamination par des valeurs aberrantes. L'influence du terme de correction pour la variabilité due à l'estimation des paramètres a été mis en évidence lors de l'usage de la fonction robuste Bi-carré. De façon générale, les taux de couverture pour les prévisions utilisant les estimateurs robustes se sont avérés plus juste en présence de contamination que ceux utilisant les estimateurs MCC.

BIBLIOGRAPHIE

[1] ANSLEY, C. F. ET KOHN, R. (1986), 'Prediction mean squared error for state space models with estimated parameters', *Biometrika* **73**, 467–473.

[2] BAILLIE, R. T. (1979), 'Asymptotic prediction mean squared error for vector autoregressive models', *Biometrika* **66**, 675–678.

[3] BOU HAMAD, I. ET DUCHESNE, P. (2005), 'On robust diagnostics at individual lags using RA-ARX estimators', *Statistical Modeling and Analysis for Complex Data Problems*, édité par P. Duchesne et B. Rémillard, Springer : New York.

[4] BROCKWELL, P. J. ET DAVIS, R. A. (1991), *Time Series : Theory and Methods*, 2^e Édition, Springer-Verlag : New York.

[5] BROCKWELL, P. J. ET DAVIS, R. A. (2002), *Introduction to Time Series and Forecasting*, 2^e Édition, Springer : New York.

[6] BUSTOS, O. ET YOHAI, V. J. (1986), 'Robust estimates for ARMA models', *Journal of the American Statistical Association* **81**, 155–168.

[7] CHATFIELD, C. (2001), 'Prediction intervals for time-series forecasting', dans *Principles of Forecasting, A Handbook for Researchers and Practitioners*, edité par J. Scott Armstrong, Kluwer Academic Publishers : Norwell.

[8] DUCHESNE, P. (2005), 'Robust and powerful serial correlation tests with new robust estimates in ARX models', *Journal of Time Series Analysis* **26**, 49–81.

[9] HARVILLE, D. A. (1997), *Matrix Algebra from a Statistician's Perspective*, Springer-Verlag : Berlin.

[10] LEDOLTER, J. (1989), 'The effect of additive outliers on the forecasts from ARIMA models', *International Journal of Forecasting* **5**, 231–240.

[11] LÜTKEPOHL, H. (2005), *New Introduction to Multiple Time Series Analysis*, Springer : Berlin.

[12] ROUSSEEUW, P. J. ET LEROY A. M. (1987), *Robust Regression And Outlier Detection*. John Wiley & Sons : New York.

[13] SCHMIDT, P. (1977), 'Some small sample evidence on the distribution of dynamic simulation forecasts', *Econometrica* **45**, 997–1006.

[14] YAMAMOTO, T. (1981), 'Prediction of multivariate autoregressive-moving average models', *Biometrika* **68**, 485–492.

Annexe A

PRODUIT DE KRONECKER ET OPÉRATEUR VEC

L'objectif de cette section est de fournir une liste des résultats les plus utiles concernant le produit de Kronecker et l'opérateur vec(·).

Soient deux matrices $\mathbf{A} = (a_{ij})$ et $\mathbf{B} = (b_{ij})$. Le produit de Kronecker entre la matrice $\mathbf{A}$ et la matrice $\mathbf{B}$ est défini comme suit:

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{pmatrix}.$$

A.1. Quelques identités sur les produits de Kronecker

Ce produit obéit à certaines identiés qui sont enumérées ci-dessous:

(A.1.1) $\mathbf{A} \otimes \mathbf{B} \neq \mathbf{B} \otimes \mathbf{A}$ en général,

(A.1.2) $(\mathbf{A} \otimes \mathbf{B})^\top = \mathbf{A}^\top \otimes \mathbf{B}^\top$,

(A.1.3) $\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C}$,

(A.1.4) $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$,

(A.1.5) $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$,

(A.1.6) $\det(\mathbf{AB}) = \det\mathbf{A})^n \det(\mathbf{B})^m$,

(A.1.7) $\mathrm{tr}(\mathbf{A} \otimes \mathbf{B}) = \mathrm{tr}(\mathbf{B})\mathrm{tr}(\mathbf{B})$.

A.2. Quelques identités sur l'opérateur vec

L'opérateur vec$(\cdot)$ permet d'empiler les colonnes d'une matrice pour former un vecteur. Ses propriétés sont enumérées de façon non-exhaustive ci-dessous:

(A.2.1) $\text{vec}(\mathbf{A}+\mathbf{B}) = \text{vec}(\mathbf{A}) + \text{vec}(\mathbf{B})$,

(A.2.2) $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})\text{vec}(\mathbf{B})$,

(A.2.3) $\text{vec}(\mathbf{AB}) = (\mathbf{I} \otimes \mathbf{A})\text{vec}(\mathbf{B}) = (\mathbf{B}^\top \otimes \mathbf{I})\text{vec}(\mathbf{A})$,

(A.2.4) $\text{vec}(\mathbf{ABC}) = (\mathbf{I} \otimes \mathbf{AB})\text{vec}(\mathbf{C}) = (\mathbf{C}^\top\mathbf{B}^\top)\text{vec}(\mathbf{A})$,

(A.2.5) $\text{vec}(\mathbf{B}^\top)^\top\text{vec}(\mathbf{A}) = \text{tr}(\mathbf{BA}) = \text{tr}(\mathbf{AB}) = \text{vec}(\mathbf{A}^\top)^\top\text{vec}(\mathbf{B})$.

D'autres propriétés se trouvent dans Harville (1997).