Formation « modèles de prévision »

Utilisation des modèles vectoriels autorégressifs (VAR) pour prévoir à un horizon lointain

Intervenants

Morgane Glotain
Alain Quartier-la-Tente

29-31 mars 2017

Introduction

Avec un étalonnage, on prévoit facilement le trimestre coïncident (T) mais comment faire pour la prévision à l'horizon $T+1$ et $T+2$?
Deux solutions :

1. rechercher des étalonnages en mettant les variables retardées
2. faire une modélisation multivariée : toutes les variables sont endogènes, *i.e* : on prévoit également les soldes d'opinion, ils n'ont pas à être prolongés hors du modèle $\rightarrow$ modèles vectoriels autorégressifs (VAR)

Introduction

Les modèles VAR naissent avec la critique de Sims (1980) :
« les modèles macro-économétriques imposent des a priori économiques sans aucune justification statistique. Ainsi, l'exogénéité de certaines variables (par exemple, celles liées à la politique économique) est postulée, mais elle n'est pas formellement testée. De même, le choix des formes fonctionnelles (restrictions, exclusion de certaines variables, structure de retards) relèvent souvent de choix arbitraires. »

Sommaire

Écriture mathématique (1/2)

Postulat : chaque variable est liée au passée de toutes les variables

À la date t, soient N variables $y_{1,t},\dots,y_{N,t}$ et $Y_t=(y_{1,t},\dots,y_{N,t})$. La représentation VAR($p$) de ces variables, avec p le nombre de retards est :

$$Y_t=\mu+\Phi_1 Y_{t-1}+\cdots+\Phi_p Y_{t-p}+\varepsilon_t$$

$$\text{avec } \Phi_i=\begin{pmatrix} a^i_{1,1} & a^i_{1,2} & \cdots & a^i_{1,N} \\ \vdots & \vdots & & \vdots \\ a^i_{N,1} & a^i_{N,2} & \cdots & a^i_{N,N} \end{pmatrix} \text{ et } \varepsilon_t \sim BB(0,\Omega)$$

$$\iff \Phi(L)Y_t=\mu+\varepsilon_t$$

Écriture mathématique (1/2)

Postulat : chaque variable est liée au passée de toutes les variables

À la date t, soient N variables $y_{1,t},\dots,y_{N,t}$ et $Y_t = (y_{1,t},\dots,y_{N,t})$. La représentation VAR($p$) de ces variables, avec p le nombre de retards est :

$$Y_t = \mu + \Phi_1 Y_{t-1} + \dots + \Phi_p Y_{t-p} + \varepsilon_t$$

$$\text{avec } \Phi_i = \begin{pmatrix} a^i_{1,1} & a^i_{1,2} & \cdots & a^i_{1,N} \\ \vdots & \vdots & & \vdots \\ a^i_{N,1} & a^i_{N,2} & \cdots & a^i_{N,N} \end{pmatrix} \text{ et } \varepsilon_t \sim BB(0,\Omega)$$

$$\iff \Phi(L) Y_t = \mu + \varepsilon_t$$

$\rightarrow$ Modèle non contraint : pas d'*a priori* sur les variables

Hypothèse centrale : toutes les variables sont endogènes

Écriture mathématique (2/2)

Le modèle $Y_t = \mu + \Phi_1 Y_{t-1} + \cdots + \Phi_p Y_{t-p} + \varepsilon_t$ est **stable** si :

les racines de $\det(I_n - \Phi_1 z + \cdots + \Phi_p z^p)$ sont de module ≥ 1

$$\iff \text{ les valeurs propres de la matrice } \begin{pmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_p \\ I_N & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots \\ 0 & 0 & I_N & 0 \end{pmatrix}$$

sont de module ≤ 1 (fonction `roots` sous R)

Écriture mathématique (2/2)

Le modèle $Y_t = \mu + \Phi_1 Y_{t-1} + \cdots + \Phi_p Y_{t-p} + \varepsilon_t$ est **stable** si :

les racines de $\det(I_n - \Phi_1 z + \cdots + \Phi_p z^p)$ sont de module ≥ 1

$\iff$ les valeurs propres de la matrice $\begin{pmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_p \\ I_N & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots \\ 0 & 0 & I_N & 0 \end{pmatrix}$

sont de module ≤ 1 (fonction `roots` sous R)

$\rightarrow$ le modèle VAR classique peut encore être amélioré en rajoutant des variables exogènes (*e.g.* : indicatrices)

Estimation du modèle VAR

Si la variable Y_t est stationnaire, le modèle VAR(p) est un modèle de type SUR – *seemingly unrelated regression* – (N équations avec les mêmes régresseurs : cas particulier du théorème de Zellner).

$\Longrightarrow$ chaque équation peut être estimée **séparement** et de manière **efficace** par MCO.

Estimation du modèle VAR

Si la variable Y_t est stationnaire, le modèle VAR(p) est un modèle de type SUR – *seemingly unrelated regression* – (N équations avec les mêmes régresseurs : cas particulier du théorème de Zellner).

$\Longrightarrow$ chaque équation peut être estimée **séparement** et de manière **efficace** par MCO.

De plus, le TCL assure que les coefficients soient *asymptotiquement normaux*

$\rightarrow$ Comment déterminer p et les variables explicatives ?

Choix du nombre de retards et des variables utilisées (1/2)

Quel ordre p choisir ?

Problème :

- si p trop grand : on augmente le RMSE des prévisions
- si p trop petit : risque d'autocorrélation des résidus

Choix du nombre de retards et des variables utilisées (1/2)

Quel ordre p choisir ?

Problème :

- si p trop grand : on augmente le RMSE des prévisions
- si p trop petit : risque d'autocorrélation des résidus

Le choix de p se fait :

- économiquement
- avec les tests de nullité des coefficients
- en fonction de la blancheur des résidus (autocorrélation et hétéroscédasticité)
- des critères d'information (AIC, BIC, HQ)

Choix du nombre de retards et des variables utilisées (2/2)

Comment choisir les variables explicatives ?

Problème : pas d'algorithme de recherche

Choix du nombre de retards et des variables utilisées (2/2)

Comment choisir les variables explicatives ?

Problème : pas d'algorithme de recherche

$\rightarrow$ Proposition : utiliser les variables d'un modèle « simple » ou des modèles obtenus par recherche automatique sur le trimestre coïncident

Choix du nombre de retards et des variables utilisées (2/2)

Comment choisir les variables explicatives ?

Problème : pas d'algorithme de recherche

$\rightarrow$ Proposition : utiliser les variables d'un modèle « simple » ou des modèles obtenus par recherche automatique sur le trimestre coïncident

Pour prévoir l'évolution PIB du trimestre t avec les soldes d'opinion du trimestre t il faut utiliser dans la modélisation VAR les soldes d'opinion avance du trimestre $t+1$

Sommaire

Avantages et inconvénients de la modélisation VAR

✓ pas d'hypothèse sur les « exogènes »
✓ peu d'hypothèses
✓ pas besoin de prolonger les soldes
✗ convergent rapidement (exclut prévision à un horizon supérieur à deux trimestres)
✗ analyse complexe des erreurs de prévision (« boîte noire »)
✗ l'information des soldes d'opinion sur la production future est limitée

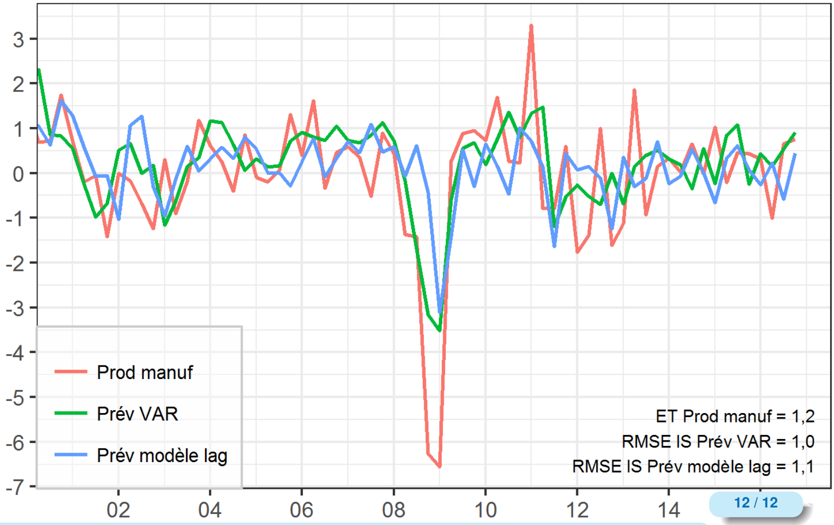
$\Longrightarrow$ modèles VAR permettent essentiellement de détecter une tendance

Application sous R

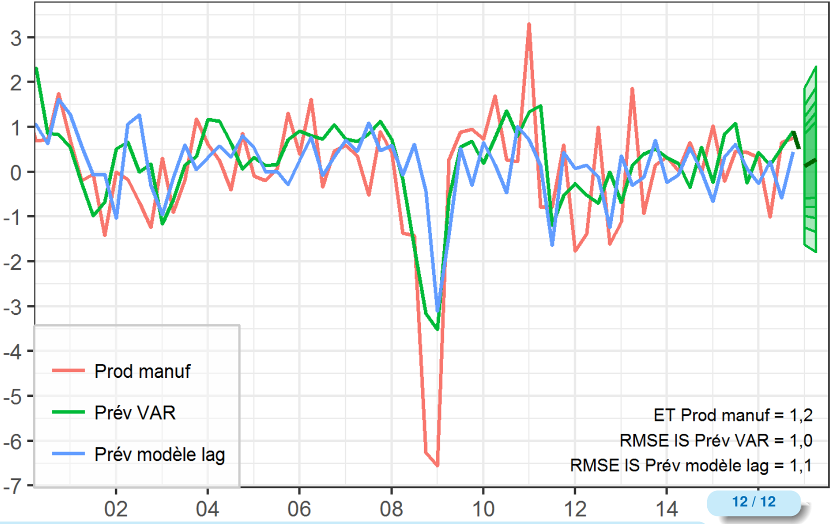
Sous R, le package qui peut être utilisé est le package `vars` :

- `VAR` pour estimer le modèle VAR (option `p` pour le nombre de retards). Ex pour créer un VAR(2) avec une constante sur les donnees Canada (`data(Canada)`) : `VAR(Canada, p = 2, type = "const")`
- `VARselect` pour rechercher le nombre de retards avec un critère d'information
- une fois le modèle estimé (fonction `VAR`), vérification de la spécification :
  - `roots` pour les valeurs propres de la matrice compagnon (doivent être de module inférieur à 1)
  - `serial.test` pour tester l'autocorrélation
  - `arch.test` pour tester l'hétéroscédasticité

Exemple graphique



Exemple graphique



Merci de votre attention !

Morgane Glotain morgane.glotain@insee.fr
Alain Quartier-la-Tente alain.quartier-la-tente@insee.fr