Un regard bayésien sur les modèles dynamiques de la macroéconomie

Stéphane Adjemian[(*)]

Florian Pelgrin[(**)]

Au cours de la dernière décennie on a pu observer un intérêt croissant du monde institutionnel pour des modèles qui jusqu'alors n'avait pas d'audience en dehors du milieu académique ; les modèles DSGE – pour Dynamic Stochastic General Equilibrium – stipulent que l'évolution observée des variables économiques résulte à tout instant des réponses optimales d'individus face aux chocs qui affectent l'économie. Les travaux de Smets et Wouters montrent qu'un modèle de cycle incorporant suffisamment de rigidités réelles et nominales peut être estimé et surtout proposer une description pertinente de l'économie. Le résultat le plus notable est que la qualité d'ajustement du modèle considéré par Smets et Wouters (une extension du modèle de Christiano, Eichenbaum et Evans) est comparable, voire meilleure, à celle d'un modèle VAR –- un modèle statistique dont les performances en termes de prévisions sont reconnues. En particulier, le modèle DSGE peut fournir des prévisions, in sample ou out of sample, qui sont équivalentes ou qui surpassent les prévisions d'un modèle VAR.

Conjointement aux succès récents des modèles DSGE on a observé une diffusion des méthodes statistiques bayésiennes. En fait, un modèle DSGE, suffisamment riche pour avoir un intérêt pratique, ne peut être estimé autrement qu'en recourant à une approche bayésienne. Les données ne sont généralement pas assez informatives pour identifier avec précision la totalité des paramètres structurels d'un modèle DSGE. L'approche bayésienne fournit un protocole objectif pour compléter l'information apportée par l'échantillon avec une information a priori sur les paramètres structurels. En pratique, l'apport de cette information a priori sur les paramètres revient à déformer, dans certaines directions, la fonction de vraisemblance associée au modèle.

Dans une première section nous présentons brièvement l'approche bayésienne en l'appliquant aux modèles auto-régressifs vectoriels (VAR). En supposant qu'il est possible de caractériser nos croyances sur les paramètres d'un modèle à l'aide d'une distribution de probabilité jointe sur ces paramètres, nous montrons comment obtenir la distribution jointe postérieure en complétant la vraisemblance par la densité a priori. Dans le cas du modèle VAR il est possible d'obtenir une expression analytique pour la distribution postérieure. Nous considérons alors deux situations polaires : (i) le cas où l'économètre ne dispose pas d'information a priori sur les paramètres et (ii) le cas où l'économètre dispose d'une information a priori sur les paramètres. Dans le premier cas, conformément à ce que suggère l'intuition, la distribution postérieure est centrée sur l'estimateur du maximum de vraisemblance (MV). Dans le second cas, la distribution postérieure est centrée sur un mélange (convexe) de l'estimateur du MV et de l'espérance a priori. L'espérance a posteriori est d'autant plus proche de l'estimateur du MV que l'échantillon est informatif relativement à nos croyances a priori. L'intérêt de l'estimation

(*) Université du Maine, Gains et Cepremap.
E-mail: stephane.adjemian@ens.fr
(**) Université de Lausanne -HEC, IEMS et Cirano.
E-mail: florian.pelgrin@unil.ch

bayésienne d'un modèle VAR est qu'en augmentant l'information (c'est-à-dire implicitement la taille de l'échantillon) avec des croyances a priori, on accroît la précision de l'estimation qui est généralement faible dans les modèles VAR à cause du nombre important de paramètres à estimer. Si de nombreux indices nous laissent penser que les séries macro-économiques sont non stationnaires et cointégrées, on peut utiliser cette information afin d'améliorer la précision de l'estimation (et donc des prévisions ou des IRFs obtenues à partir du modèle VAR estimé).

Dans la deuxième section nous montrons en quoi l'approche bayésienne des DSGE est plus complexe que celle des modèles VAR. Il n'est généralement pas possible d'obtenir une expression analytique pour la distribution postérieure des paramètres d'un modèle DSGE. Deux raisons expliquent cette impossibilité : (i) un modèle fondé sur des comportements individuels est généralement non linéaire dans les paramètres structurels que nous cherchons à estimer et (ii) nous n'observons qu'un sous ensemble des variables endogènes du modèle et nous faisons donc appel à un filtre de Kalman pour obtenir des estimations des variables latentes et évaluer la vraisemblance. Puisque nous ne disposons pas d'une expression analytique de la vraisemblance, il n'est pas possible d'obtenir une expression analytique de la distribution a posteriori. Il est alors nécessaire de recourir à des simulations pour caractériser les croyances a posteriori. Nous décrivons les algorithmes de MCMC (Monte Carlo Markov Chain) et plus spécialement l'algorithme du Metropolis-Hastings qui est généralement considéré dans cette littérature.

*Enfin dans la troisième section nous revenons sur la comparaison, voire l'opposition, entre les modèles DSGE et VAR. Une modélisation VAR, relativement à la modélisation DSGE, a l'avantage de n'imposer qu'un nombre limité de contraintes (la liste des variables dans le modèle, le nombre de retards ou la spécification de la partie déterministe). Le coût de cet avantage est la relative faible précision des estimations. Avec la modélisation DSGE, en imposant plus de structure sur les données on introduit implicitement de l'information, qui si celle-ci est "*légitime*", peut éventuellement améliorer les performances relatives du DSGE (par exemple, les RMSE des prévisions). Le problème est la nature* déterministe *de cette information ; si les restrictions du DSGE sont sans rapport avec la forme du processus générateur des données, elles dégraderont les performances relatives du DSGE. Dans cette dernière section, nous montrons qu'il est possible d'utiliser l'information structurelle (le DSGE) pour définir l'information a priori sur le modèle VAR. Le poids de l'information structurelle, c'est-à-dire de la croyance a priori, est endogène et choisi de façon à maximiser les performances en prévision in sample du modèle BVAR (pour Bayesian VAR). Cela revient à poser des restrictions* stochastiques *sur le modèle empirique qui ne seront effectivement utilisées que dans la mesure où elles sont pertinentes. Les VAR et DSGE sont complémentaires, l'approche bayésienne offre la possibilité de les "mélanger" en ne gardant que le meilleur de chaque modélisation.*

Ces dernières années, l'analyse des fluctuations économiques s'est développée autour des Modèles d'Équilibre Général Intertemporels Stochastiques (DSGE). Pour autant, jusqu'à très récemment, l'engouement pour l'approche DSGE comme outil d'analyse de la politique économique est demeuré relativement faible, et l'approche des modèles Vectoriels Autorégressif (VAR) a été (est) souvent privilégiée. Plusieurs raisons expliquent cette préférence. D'une part, la modélisation VAR de la dynamique des variables macroéconomiques impose un nombre très restreint de contraintes et offre une qualité d'ajustement aux données (et des prévisions) relativement bonne. Au contraire, en augmentant le nombre de contraintes sur les données, encourant ainsi le risque d'une mauvaise spécification, les modèles DSGE de la première génération (les modèles de la théorie des cycles réels) se sont traduits par des performances d'ajustement et de prévisions très pauvres. D'autre part, l'émergence d'une approche plus structurelle des modèles VAR (par rapport à l'approche a-théorique, Sims, 1980)– autorisant des procédures d'identification des chocs à partir de restrictions contemporaines, de court terme (Sims, 1986 ; Bernanke, 1986) ou de long terme (Blanchard et Quah, 1986) – ont conduit à exiger que tout modèle théorique puisse reproduire les fonctions de réponse des variables macroéconomiques à des chocs structurels identifiés dans les modèles VAR (Rotemberg et Woodford, 1997 ; Christiano, Eichenbaum et Charles, 2003). Finalement, l'absence d'un traitement économétrique convaincant n'a fait que renforcer la recommandation de Kydland et Prescott (1982) – l'étalonnage est préférable.

Cependant, on a constaté un regain d'intérêt des modèles DSGE et cela essentiellement pour deux raisons : (*i*) les avancées théoriques et notamment la prise en compte de fondements microéconomiques des rigidités nominales et/ou réelles (*ii*) les progrès dans l'estimation et l'évaluation des modèles sur la base de méthodes statistiques formelles[1]. Dans cette perspective, l'idée suivant laquelle de tels modèles sont utiles pour la prévision et l'analyse de la politique économique s'est répandue dans le milieu académique ainsi qu'auprès des institutions internationales et des banques centrales. Parmi toutes ces approches économétriques, la littérature privilégie, pour de *bonnes* et *mauvaises* raisons, la statistique bayésienne. Parmi les *bonnes* raisons, nous pourrions souligner le fait que la fonction de vraisemblance d'un modèle de dimension élevée (de nombreux paramètres à estimer) est souvent *plate* dans certaines directions. En d'autres termes, les données peuvent ne pas être suffisamment informatives pour identifier (avec précision) les paramètres structurels. En déformant la fonction de vraisemblance à l'aide d'informations *a priori* sur les paramètres, c'est-à-dire en privilégiant une approche bayésienne, l'identification devient possible. Il est néanmoins trop souvent ignoré que la mise en œuvre et l'interprétation des résultats de l'estimation bayésienne requièrent un certain nombre d'hypothèses et de conditions de validité, ou que nombre de problèmes rencontrés en économétrie classique ont leur contrepartie en économétrie bayésienne. Toujours est-il que l'approche bayésienne a considérablement favorisé le développement des modèles DSGE comme outil d'analyse et de prévision de la politique monétaire. Dans le même temps, il n'en demeure pas moins que les modèles DSGE et VAR continuent à être opposés et que nombre de papiers cherchent généralement à légitimer leurs résultats en comparant, par exemple, les prévisions (ou tout autre statistique ou quantité d'intérêt) de leur(s) modèle(s) avec ceux d'un VAR (Smets et Wouters, 2002).

L'objet de ce papier est de présenter l'approche bayésienne des modèles VAR et DSGE en mettant en avant les principaux concepts, leur mise en œuvre pratique et les limites sous-jacentes. Nous montrons en quoi les modèles DSGE et VAR sont des outils complémentaires que l'on ne doit pas nécessairement chercher à opposer. Nous n'abordons pas ici certains problèmes importants, comme l'estimation non linéaire des modèles DSGE[2].

L'article est organisé comme suit. Dans une première partie, nous présentons les principaux concepts de l'analyse bayésienne et montrons comment les appliquer dans le cadre des modèles VAR. Une attention particulière est attachée à la nature (informative, non informative, empirique) des croyances *a priori*. Dans une deuxième partie, nous abordons les spécificités de l'approche bayésienne des modèles DSGE. Contrairement aux modèles VAR, il n'est plus possible d'obtenir une expression analytique de la distribution *a posteriori*. Pour remédier à cette difficulté, il est nécessaire de recourir à des méthodes de Monte-Carlo et notamment à la théorie des chaînes de Markov. Dans cette perspective, après avoir dérivé de manière générale la densité *a posteriori* d'un modèle DSGE, nous expliquons les principaux algorithmes d'estimation (algorithme de Metropolis-Hasting, par fonction d'importance). Dans une troisième partie, nous illustrons comment peuvent se combiner les approches VAR et DSGE.

L'approche Bayésienne

Généralités

L'approche bayésienne propose un cadre rigoureux pour (*i*) formaliser nos croyances[3] *a priori* et (*ii*) déterminer comment celles-ci doivent être mises à jour une fois que les données sont observées. Les croyances, *a priori* ou *a posteriori*, sont représentées à l'aide d'une densité de probabilité jointe sur les paramètres d'un modèle. Cette densité jointe caractérise l'incertitude quant au processus générateur des données (DGP, pour *Data Generating Process*), en décrivant une famille (un continuum) de modèles.

Imaginons que nous souhaitions caractériser nos croyances sur le paramètre de Calvo d'une courbe de Phillips. Ce paramètre, ξ_p, est la probabilité pour une firme, en concurrence monopolistique, de ne pas pouvoir ajuster son prix de façon optimale à une date quelconque. Ainsi, nous savons déjà que ce paramètre doit appartenir à l'intervalle [0,1]. Nous pourrions donc utiliser une distribution bêta[4] définie sur cet intervalle. À partir de la probabilité ξ_p, nous pouvons définir le temps moyen pendant lequel une firme ne pourra pas ajuster son prix de façon optimale : $\zeta_p \equiv \frac{1}{1-\xi_p}$. Si par ailleurs, à l'aide d'enquêtes microéconomiques, nous savons que le temps moyen durant lequel une firme ne réajuste pas son prix de façon *optimale* est de 4 trimestres, nous pouvons déduire qu'une valeur pertinente de la probabilité ξ_p est 3/4. L'économiste bayésien pourra donc formaliser son *a priori* sur le paramètre ξ_p en sélectionnant une distribution bêta ayant pour mode 3/4 et en spécifiant une variance mesurant son incertitude sur le paramètre d'intérêt. Il choisira une variance d'autant plus grande qu'il est incertain des évaluations microéconomiques dont il dispose[5]. Notons qu'il pourrait directement poser son *a priori* sur le délai moyen d'attente ζ_p ; ceci conduirait à une distribution différente pour le paramètre ξ_p. Si ξ_p est le seul paramètre du modèle pour lequel nous sommes incertain, *i.e.* si les autres paramètres ont des variances *a priori* nulles, la densité *a priori* sur ce paramètre décrit une famille de DGP indexée par ξ_p : chaque valeur possible de ξ_p correspond à un DGP.

Plus généralement, nous noterons l'*a priori* sur un vecteur de paramètres θ_M associé à un modèle paramétrique $M, \theta_M \equiv (\theta_1^M, \ldots, \theta_{q_M}^M)$, de la façon suivante :

(1) $p_0(\theta_M \mid M)$

Cette densité jointe définit notre incertitude quant aux paramètres θ_M avant que nous ayons porté attention aux données. Il convient de noter que nous raisonnons conditionnellement à un modèle. En toute généralité l'incertitude pourrait aussi porter sur la forme du modèle paramétrique *M*. Plus loin nous omettrons généralement le conditionnement (ainsi que l'indexation) par le modèle pour simplifier les notations.

Nous observons un échantillon $Y_T^* = \{y_t^*\}_{t=1}^T$ où y_t^* est un vecteur de *m* variables. Nous nous limiterons au cas où l'indice *t* représente le temps. La vraisemblance est la densité de l'échantillon conditionnellement au modèle et ses paramètres ; on notera :

(2) $L(\theta_M ; Y_T^*, M) \equiv p(Y_T^* \mid \theta_M, M)$

L'estimateur du maximum de vraisemblance (MV) des paramètres θ_M d'un modèle *M* est la valeur des paramètres qui rend le plus *probable* l'occurrence de l'échantillon à notre disposition. Autrement dit, l'estimateur du MV sélectionne le paramètre θ_M définissant le DGP qui a le plus probablement généré les données. La démarche statistique, classique ou bayésienne, est une démarche d'inversion– il s'agit de remonter des observations aux paramètres du DGP. Un modèle définit la densité d'un ensemble de variables conditionnellement à des paramètres inconnus. L'observation de l'échantillon donne en retour de l'information sur les paramètres. La notation définie par l'équation (2) résume le principe de l'inférence ; la vraisemblance est la densité de l'échantillon Y_T^* sachant les paramètres θ, mais nous écrivons habituellement la vraisemblance comme une fonction des paramètres, *i.e.* formellement nous échangeons les rôles de Y_T^* et θ.

Nous disposons des densités $p_0(\theta_M \mid M)$, qui caractérise l'information postulée *a priori*, et $p(Y_T^* \mid \theta_M, M)$, qui caractérise l'information apportée par les données. On croise ces deux sources d'informations orthogonales, en utilisant le théorème de Bayes, pour obtenir la densité de θ_M connaissant les données Y_T^*, *i.e.* la densité postérieure :

(3) $p_1(\theta_M \mid Y_T^*, M) = \frac{p_0(\theta_M \mid M)\, p(Y_T^* \mid \theta_M, M)}{p(Y_T^* \mid M)}$

avec

(4) $p(Y_T^* \mid M) = \int_{\Theta_M} p_0(\theta_M \mid M) \times p(Y_T^* \mid \theta_M, M)\, d\theta_M$

la densité marginale. Ainsi, la densité postérieure est proportionnelle à la densité *a priori* multipliée par la vraisemblance :

$$p_1(\theta_M \mid Y_T^*, M) \propto p_0(\theta_M \mid M)\, p(Y_T^* \mid \theta_M, M)$$

$$\equiv K(\theta_M \mid Y_T^*, M)$$

Puisque le dénominateur dans (3), la densité marginale, ne dépend pas de θ_M, l'inférence sur les paramètres, par exemple l'évaluation de l'espérance postérieure, peut être mise en œuvre à l'aide du seul noyau postérieur, $K(\theta_M \mid Y_T^*, M)$. On représente nos croyances *a posteriori* en exhibant les propriétés de la distribution *a posteriori*. Nous pouvons représenter graphiquement la densité postérieure marginale de chaque paramètre θ, construire des intervalles contenant α% de la distribution postérieure, ou encore calculer des moments *a posteriori*. Par exemple, la comparaison des variances *a priori* et *a posteriori* peut nous renseigner sur l'information apportée par les données, relativement à celle contenue dans nos croyances *a priori*. Les variances de chaque paramètre sont définies à partir des éléments diagonaux des matrices suivantes :

$$V_0[\theta] = \int_\Theta \theta\theta'\; p_0(\theta)\, d\theta - \left(\int_\Theta \theta p_0(\theta)\, d\theta\right) \left(\int_\Theta \theta p_0(\theta)\, d\theta\right)'$$

et

$$V_1[\theta] = \int_\Theta \theta\theta'\; p_1(\theta \mid Y_T^*)\, d\theta - \left(\int_\Theta \theta p_1(\theta \mid Y_T^*)\, d\theta\right) \left(\int_\Theta \theta p_1(\theta \mid Y_T^*)\, d\theta\right)'$$

Si la variance postérieure d'un paramètre est plus faible que sa variance *a priori,* cela signifie que les données apportent une information supplémentaire sur ce paramètre, relativement à l'information *a priori*. Dans certains cas, il est possible d'obtenir analytiquement la densité postérieure et ses moments[6] ; nous verrons un exemple dans la section suivante. Plus généralement, il est nécessaire de recourir à des algorithmes numériques, pour caractériser la distribution postérieure, *i.e.* pour évaluer les intégrales nécessaires au calcul des moments.

Pour communiquer nos croyances *a posteriori* nous désirons souvent recourir à un média plus synthétique en résumant, à l'image de l'approche classique, la distribution postérieure par un point. On parle alors d'estimation ponctuelle. Réduire la distribution postérieure à un point s'apparente à un choix en univers incertain. Il est donc naturel de construire une estimation ponctuelle en minimisant l'espérance postérieure d'une fonction de perte :

$$(5)\quad \hat{\theta} = \arg\min_a \int_\Theta p_1(\theta \mid Y_T^*, M)\, L(a, \theta)\, d\theta$$

où $L(a,\theta)$ est une fonction associant une perte au choix a si la vraie valeur du paramètre est θ. Si, par exemple, la fonction de perte est quadratique[7] :

$$L(a,\theta) = (a-\theta)^2$$

alors on montre (Zellner, 1971, page 24) que l'estimation ponctuelle doit être l'espérance postérieure de θ. D'autres fonctions de perte aboutiront à d'autres estimations ponctuelles. La médiane postérieure peut être justifiée par la fonction de perte $L(a,\theta) = |a-\theta|$.

Tant que l'inférence porte sur les paramètres d'un modèle, nous pouvons mettre de côté la constante d'intégration, $p(Y_T^* \mid M)$. Cependant, la densité marginale contient une information pertinente si nous désirons comparer différents modèles. L'interprétation de sa définition (4) est directe : la densité marginale est une moyenne des vraisemblances, obtenues pour différentes valeurs des paramètres, pondérées par nos croyances *a priori* sur les valeurs de ces paramètres. Comme cette quantité ne dépend pas des paramètres, elle autorise facilement la comparaison de modèles non emboîtés[8]. Par exemple, si nous disposons de deux modèles alternatifs, *A* et *B*, pour expliquer les données et si nous obtenons $p(Y_T^* \mid A) > p(Y_T^* \mid B)$, alors cela signifie que les données à notre disposition ont plus probablement été générées par le modèle *A* que par le modèle *B*. Cette approche ne fournit pas un test formel ; pour prendre une décision sur le modèle, à l'image de l'estimation ponctuelle, il faut spécifier un critère de perte sur les modèles. Nous supposons implicitement que nous n'avons pas de *préférence a priori* pour un des deux modèles[9]. Plus généralement nous pouvons définir des probabilité *a priori* pour les modèles $I = A, B$. Par exemple, nous pourrions supposer que $p_0(A) > p_0(B)$, c'est-à-dire que le modèle A est a priori plus probable que le modèle B. Par le théorème de Bayes, nous obtenons la probabilité *a posteriori* du modèle I (=*A* ou *B*) :

$$p_1(I \mid Y_T^*) = \frac{p_0(I)\, p(Y_T^* \mid I)}{\Sigma_I\; p_0(I)\, p(Y_T^* \mid I)}$$

L'interprétation est directe, il s'agit d'une version discrète de l'équation (3). Si nous devons sélectionner un modèle, notre préférence ira au modèle qui maximise la densité postérieure. L'élicitation d'une densité de probabilité *a priori* sur la collection de modèles n'est pas une chose aisée ; on ne peut associer une probabilité à un modèle de la même façon que l'on pose une densité *a priori* sur le

paramètre de Calvo d'une courbe de Phillips. C'est pourquoi une densité de probabilité uniforme est souvent utilisée dans la littérature concernée par l'estimation des modèles DSGE. Pour une revue des enjeux de la comparaison de modèles, on peut lire Robert (2006, chapitre 7) ou Sims (2003). Enfin, notons que la comparaison de différents modèles, par l'intermédiaire de leurs densités marginales, ne doit pas nécessairement aboutir à un choix. Dans certaines situations, il peut être souhaitable de combiner plusieurs modèles, en les pondérant par leurs densités marginales respectives(10).

Après l'estimation, le modèle peut être utilisé pour construire des prévisions et des fonctions de réponse. À l'image de l'estimation des paramètres, le paradigme bayésien ne fournit pas des prévisions ponctuelles mais des densités prédictives. Supposons que nous souhaitions établir des prédictions sur y^*_{T+1} un vecteur d'endogènes non encore observées, le but est de construire la densité (prédictive) de ce vecteur.

Cette densité peut être obtenue en intégrant par rapport à θ la densité postérieure jointe de y^*_{T+1} et θ.

$$p(y^*_{T+1} \mid Y^*_T) = \int_\Theta p(y^*_{T+1}, \theta \mid Y^*_T)\, d\theta$$

où la densité sous l'intégrale est définie par :

$$p(y^*_{T+1}, \theta \mid Y^*_T) = p(y^*_{T+1} \mid \theta, Y^*_T)\, p_1(\theta \mid Y^*_T)$$

par le théorème de Bayes. La densité jointe de y^*_{T+1} et θ est le produit de la densité postérieure de θ et de la densité de y^*_{T+1} conditionnelle à θ. Cette dernière densité est directement obtenue à partir de la définition du modèle.

Donnons un exemple dans le cas scalaire. Si le modèle est un processus autorégressif d'ordre un : $y^*_t = \theta y^*_{t-1} + \varepsilon_t$ avec $t = 1,...,T$, $\varepsilon_t \underset{iid}{\sim} N(0,\sigma^2)$ et σ^2, la variance de l'innovation, connue. La distribution de y^*_{T+1} conditionnellement à θ et Y^*_T(10) est gaussienne : $Y_{T+1} \mid Y^*_T, \theta \sim N(\theta y^*_T, \sigma^2)$. La densité prédictive s'écrit finalement :

$$(6)\quad p(y^*_{T+1} \mid Y^*_T) = \int_\Theta p(y^*_{T+1} \mid \theta, Y^*_T)\, p_1(\theta \mid Y^*_T)\, d\theta$$

et s'interprète comme une moyenne des densités conditionnelles y^*_{T+1} sachant θ, pondérées par la densité postérieure de θ.

À partir de cette densité prédictive, on peut construire une prédiction ponctuelle des variables en se donnant une fonction de perte, calculer un intervalle contenant α% de la distribution de y^*_{T+1}, etc. En confrontant la densité prédictive aux réalisations effectives des variables, on peut alors évaluer dans quelle mesure notre modèle tend à sur-estimer ou sous-estimer, par exemple, le taux de croissance à un trimestre du PIB par tête.

Cette comparaison peut fournir un critère d'évaluation du modèle. Si on se rend compte que les réalisations effectives d'une variable se situent systématiquement dans les queues de la densité prédictive, alors on peut conclure que le modèle est mal spécifié vis-à-vis de cette variable.

Le choix des croyances *a priori*

On comprend déjà que le choix des croyances *a priori* est essentiel, dans la mesure où il détermine partiellement les résultats (surtout pour un échantillon de taille réduite comme nous le verrons par la suite). La subjectivité de l'économètre ne peut intervenir que dans la première étape d'élicitation de l'*a priori*, les étapes suivantes (l'évaluation de la vraisemblance,...) sont automatiques et nécessairement objectives sauf quant au choix éventuel d'un critère de perte pour l'estimation ponctuelle. La question du choix des croyances *a priori* est donc crucial. Il est donc important de bien comprendre le rôle de la densité *a priori* dans les résultats, par exemple en menant des exercices de sensibilité aux croyances *a priori*. Ces expériences, en donnant une idée du rôle des *priors*, dévoilent implicitement la forme de la vraisemblance. L'expérience la plus extrême(11) est de considérer un *a priori* non informatif, c'est-à-dire le cas où nous n'avons aucune croyance *a priori* sur les paramètres du modèle. De façon assez surprenante, les statisticiens bayésiens ne parviennent pas à s'accorder sur une chose aussi essentielle que la caractérisation de ce non savoir.

Dans la section précedente nous avons examiné le cas d'un *a priori* informatif sur le paramètre de Calvo définissant le degré de rigidité de l'inflation. Dans ce cas notre connaissance *a priori* provient de l'observation de données microéconomiques, différentes de celles utilisées pour l'estimation du modèle. Lorsque l'information *a priori* est basée sur des données, celles-ci doivent être différentes des données utilisées pour identifier le modèle. Dans le cas contraire la démarcation entre vraisemblance et densité *a priori* devient ambiguë, ce qui paraît inacceptable à de nombreux statisticiens. Notons néanmoins que de non moins nombreux statisticiens utilisent l'échantillon pour définir les croyances *a priori*. Par exemple quand il s'agit de spécifier la densité *a priori* de façon à optimiser les capacités prédictives d'un modèle (comme nous le verrons plus loin. Les croyances *a priori* peuvent aussi être basées sur des considérations purement théoriques. Dans la littérature concernée par l'estimation des modèles DSGE (et aussi des VAR), les croyances *a priori*, indépendamment de l'origine de ces croyances, sont généralement représentées par des densité paramétrées (distribution gaussienne,

gamma,...). Dans certains cas, on parle alors d'*a priori* conjugués, elles sont choisies de façon à ce que la densité *a posteriori* soit de la même famille paramétrique (voir l'exemple du modèle VAR plus loin). La motivation, en facilitant les calculs, est un héritage du passé et essentiellement technique. Aujourd'hui, les progrès de l'informatique nous permettent d'adopter une formulation non paramétrique plus générale. Par exemple nous pourrions caractériser nos croyances *a priori* sur chaque paramètre en spécifiant les quantiles de chaque distribution. Il est vrai que nos croyances sont rarement aussi précises.

Dans certaines situations nos connaissances *a priori* sont faibles. Malheureusement la caractérisation de l'ignorance est toujours sujet à débat. Un exemple frappant est donné par Sims et Uhlig (1991) puis Phillips (1991b), Phillips (1991a) et Sims (1991), qui débatent de la caractérisation de l'ignorance dans un modèle autorégressif d'ordre un et des conséquences sur la détection de racines unitaires.

Une première approche est de considérer un *prior* plat. Pour un paramètre μ qui peut prendre des valeurs entre $-\infty$ et ∞, Jeffrey (1961) propose d'adopter une distribution uniforme entre $-\infty$ et ∞ :

$$p_0(\mu) \propto 1$$

Évidemment cette densité est impropre dans le sens où $\int p_0(\mu)\,d\mu$ est indéfini. Mais c'est précisément cette propriété qui, pour Jeffrey, rend ce *prior* non informatif. En effet, pour tout $a < b < c < d$ réels on ne peut pas dire que $\mu \in [a,b]$ soit *a priori* plus probable que $\mu \in [c,d]$, puisque les probabilités de ces événements sont nulles. Pour un paramètre σ, par exemple un écart type qui peut prendre des valeurs entre 0 et ∞, Jeffrey propose d'adopter une distribution uniforme pour le logarithme de σ entre $-\infty$ et ∞ :

$$p_0(\log\sigma) \propto 1$$

$$\Leftrightarrow\ p_0(\sigma) \propto \frac{1}{\sigma}$$

Comme dans le cas précédent l'intégrale de cette densité est impropre. En particulier, on ne peut définir $\int_0^c p_0(\sigma)\,d\sigma$ et $\int_c^\infty p_0(\sigma)\,d\sigma$, on ne peut dire s'il est plus probable que σ soit supérieur ou inférieur à c [(12)].

On note en passant que cette densité a l'heureuse propriété d'être invariante à une transformation puissance[(13)] : si le *prior* est non informatif pour l'écart type, il en va de même pour la variance (σ^2).

Plus tard, Jeffrey généralisa ce résultat d'invariance et proposa un *prior* non informatif (le plus souvent impropre) basé sur la matrice d'information de Fisher :

$$p_0(\theta) \propto |I(\theta)|^{\frac{1}{2}}$$

avec

$$I(\theta) = E\left[\left(\frac{\partial p(Y_T^*|\theta)}{\partial\theta}\right)\left(\frac{\partial p(Y_T^*|\theta)}{\partial\theta}\right)'\right]$$

La matrice d'information de Fisher quantifie l'information amenée par le modèle et les données sur le paramètre θ. En favorisant les valeurs de θ pour lesquelles l'information de Fisher est plus grande, on diminue l'influence de la loi *a priori* puisque l'information véhiculée par celle-ci est peu différente de l'information provenant de la vraisemblance. La définition de la densité *a priori* est donc liée à la courbure de la vraisemblance. Cette densité *a priori* est invariante à toute reparamétrisation (continue) du modèle (voir Zellner, 1971, annexe du chapitre 2)) pour une description plus détaillée des propriétés d'invariance).

L'utilisation d'un *prior* plat ou d'un *prior* dérivé de la matrice d'information de Fisher pour caractériser l'absence d'information affecte généralement l'inférence. Par exemple, dans un modèle AR(1), voir Phillips (1991b), un *prior* basé sur l'information de Fisher n'est pas équivalent à un prior uniforme (plat). En effet, dans un modèle dynamique, la quantité d'information véhiculée par les données (*i.e.* la vraisemblance) dépend de la valeur du paramètre autorégressif (ρ). Si le paramètre est proche de l'unité, voire égal ou supérieur à un, les données sont plus informatives. Ainsi, pour Phillips, l'utilisation d'un *prior* plat, à l'instar de Zellner (1971) ou Sims et Uhlig (1991), biaise la distribution postérieure de ρ en faveur de la stationnarité. En donnant autant de poids aux valeurs explosives de ρ qu'aux valeurs stationnaires, le *prior* plat ne prend pas en compte le fait que des données générées par un modèle à racine unitaire ou explosif sont plus informatives. Il existe d'autres approches pour caractériser l'ignorance, on peut lire le chapitre 3 de Robert (2006) et plus spécialement la section 5.

Le choix d'une densité *a priori* et ses conséquences sur l'inférence sont l'objet de toutes les critiques de la part des statisticiens ou économètres classiques. Il ne faudrait pourtant pas oublier que le paradigme classique n'est pas plus exempt de choix, aux conséquences non négligeables sur l'inférence. Par exemple, le choix d'une métrique (minimiser la somme des carrés des résidus ou la somme des valeurs absolues des résidus), le choix des variables instrumentales, de modèles auxiliaires ou des conditions de moments, sont rarement discutés même s'ils déterminent les résultats. Dans une

certaine mesure, nous n'avons même plus conscience des choix effectués. L'approche bayésienne est de ce point de vue bien plus transparente.

Comportement asymptotique et approximations

Même si l'approche bayésienne ne repose pas sur des arguments asymptotiques, comme généralement l'approche classique, il est utile de s'interroger sur ses propriétés asymptotiques. Le résultat rassurant est que si les conditions de normalité asymptotique de l'estimateur du maximum de vraisemblance sont réunies[14], alors la distribution postérieure tend vers une gaussienne multivariée. Asymptotiquement, la distribution postérieure est centrée sur l'estimateur du maximum de vraisemblance. Ce résultat, avancé par Laplace, est intuitif puisque lorsque la taille de l'échantillon tend vers l'infini, le poids de l'information *a priori* relativement à l'information contenue dans l'échantillon devient négligeable.

Plus formellement, si on note θ^* l'unique mode de la distribution postérieure obtenu en maximisant le noyau postérieur $K(\theta) \equiv K(\theta_A | Y_T^*, A)$, et s'il est possible d'écrire une approximation de Taylor à l'ordre deux du noyau postérieur autour de θ^*, alors nous avons :

$$\log K(\theta) = \log K(\theta^*) + (\theta - \theta^*)' \left. \frac{\partial \log K(\theta)}{\partial \theta} \right|_{\theta=\theta^*} + \frac{1}{2}(\theta - \theta^*)' \left. \frac{\partial^2 \log K(\theta)}{\partial \theta \, \partial \theta'} \right|_{\theta=\theta^*} (\theta - \theta^*) + O(||\theta - \theta^*||^3)$$

Puisque les dérivées premières sont, par définition, nulles en θ^*, nous avons de façon équivalente :

$$\log K(\theta) = \log K(\theta^*) - \frac{1}{2}(\theta - \theta^*)' [H(\theta^*)]^{-1} (\theta - \theta^*) + O(||\theta - \theta^*||^3)$$

où $H(\theta^*)$ est l'opposé de l'inverse de la matrice hessienne évaluée au mode. Ainsi, en ne considérant que le terme quadratique, le noyau postérieur peut être approximé par :

$$K(\theta) \doteq K(\theta^*) e^{-\frac{1}{2}(\theta - \theta^*)'[H(\theta^*)]^{-1}(\theta - \theta^*)}$$

On reconnaît, à une constante d'intégration près[15], la densité d'une loi normale multivariée. En complétant par la constante d'intégration, nous obtenons finalement une approximation de la densité postérieure $p_1(\theta) \equiv p_1(\theta_A | Y_T^*, A)$:

(7)

$$p_1(\theta) \doteq (2\pi)^{-\frac{q}{2}} |H(\theta^*)|^{-\frac{1}{2}} e^{-\frac{1}{2}(\theta - \theta^*)'[H(\theta^*)]^{-1}(\theta - \theta^*)}$$

Généralement, la matrice hessienne est d'ordre $O(T)$: lorsque la taille de l'échantillon augmente la distribution postérieure se concentre autour du mode. À partir de cette approximation asymptotique on peut alors très facilement calculer, par exemple, des moments postérieurs ou les densités prédictives. Par exemple, l'espérance postérieure de $\varphi(\theta)$ est définie par :

$$E[\varphi(\theta)] = \frac{\int_\Theta \varphi(\theta) \, p(Y_T^* | \theta) \, p_0(\theta) \, d(\theta)}{\int_\Theta p(Y_T^* | \theta) \, p_0(\theta) \, d(\theta)}$$

Tierney et Kadane (1986) montrent que si l'on approxime à l'ordre 2 le numérateur autour du mode de $\varphi(\theta) \, p(Y_T^* | \theta) \, p_0(\theta)$ et le dénominateur autour du mode de $p(Y_T^* | \theta) \, p_0(\theta)$, alors l'erreur d'approximation de l'espérance est d'ordre $O(T^{-2})$. Les erreurs d'approximation du numérateur et du dénominateur, qui sont d'ordre $O(T^{-1})$, se compensent favorablement. L'approche de Tierney et Kadane ne va pas sans poser certains problèmes. Si on cherche à calculer $E[\varphi(\theta)]$ pour différentes fonctions φ, alors il est nécessaire de recourir à une nouvelle maximisation pour chaque paramètre et chaque fonction φ. Par exemple, si on désire calculer les espérances et écart types *a posteriori* pour chacun des k paramètres, il faut recourir à $2k+1$ maximisations, auxquelles il faut rajouter le calcul des matrices hessiennes. Il est alors évident que si k est élevé, une telle approximation peut devenir coûteuse en temps de calculs. Tierney, Kass, et Kadane (1989) propose différentes méthodes pour pallier cette difficulté[16]. Notons néanmoins qu'une approche basée sur des simulations (voir plus bas) devient aussi plus coûteuse lorsque le nombre de paramètres augmente.

Un modèle linéaire : le modèle VAR

Dans cette section, nous considérons un exemple où les résultats peuvent être obtenus analytiquement. Le modèle VAR gaussien, se prête à cet exercice et a l'avantage d'être un outil couramment utilisé en macroéconomie (voir, par exemple, la contribution de Fabrice Collard et Patrick Fève à ce numéro).

Nous considérons un modèle VAR(p) pour caractériser le vecteur $1 \times m$ de variables endogènes y_t^* observées :

$$y_t^* = \sum_{i=1}^{p} y_{t-i}^* \mathbf{A}_i + \varepsilon_t$$

où $\{\mathbf{A}_i\}$ est une suite de matrice $m \times m$ et ε_t est un bruit blanc gaussien, de dimension $1 \times m$ d'espérance nulle et de variance $V[e_t] = \Sigma$. Nous pourrions compléter le modèle avec des variables exogènes, une constante par exemple, mais nous allons à l'essentiel en omettant cette possibilité.

On note $Y_T^* \equiv \{y_t^*\}_{t=-p+1}^{T}$ les données à notre disposition et on note z_t la concaténation horizontale des vecteurs lignes $y_{t-1}^*, y_{t-2}^*, ..., y_{t-p}^*$. En concaténant verticalement les vecteurs lignes y_t^*, z_t et ε_t, pour $t = 1,...,T$, on obtient la représentation matricielle suivante du modèle VAR(p) :

$$Y = ZA + E$$

où Y et E sont des matrices $T \times m$, Z est une matrice $T \times (mp)$ et $A = (\mathbf{A}_1', \mathbf{A}_2', ..., \mathbf{A}_p')'$ la matrice $k \times m$ (avec $k = mp$) regroupant les coefficients auto-régressifs. La vraisemblance associée à ce modèle linéaire gaussien est donnée par :

$$L(A,\Sigma;Y_T^*) = (2\pi)^{-\frac{mT}{2}} |\Sigma|^{-\frac{T}{2}} \times e^{-\frac{1}{2} tr\{(Y-ZA)\Sigma^{-1}(Y-ZA)'\}}$$

L'estimateur du maximum de vraisemblance (MCO) est défini par :

$$\hat{A} = (Z'Z)^{-1} Z'Y$$

et

$$\hat{\Sigma} = T^{-1} (Y - Z\hat{A})' (Y - Z\hat{A})$$

Nous verrons plus loin qu'il est profitable de réécrire la vraisemblance en faisant apparaître l'estimateur des MCO :

$$L(A,\Sigma;Y_T^*) = (2\pi)^{-\frac{mT}{2}} \times |\Sigma|^{-\frac{k}{2}} e^{-\frac{1}{2} tr\{\Sigma^{-1}(A-\hat{A})'Z'Z(A-\hat{A})\}}$$

$$\times |\Sigma|^{-\frac{T-k}{2}} e^{-\frac{1}{2} tr\{\Sigma^{-1}(Y-Z\hat{A})'(Y-Z\hat{A})\}}$$

Aux constantes d'intégration près on reconnaît ici les fonctions de densité de probabilité d'une gaussienne matricielle et d'une inverse Wishart (voir l'annexe A). La vraisemblance se réécrit donc sous la forme suivante :

$$L(A,\Sigma;Y_T^*) = (2\pi)^{-\frac{mT}{2}} \times (2\pi)^{\frac{km}{2}} |Z'Z|^{-\frac{m}{2}}$$

$$\times f_{MN_{k,m}}(A;\hat{A},(Z'Z)^{-1},\Sigma)$$

$$\times \frac{2^{\frac{vm}{2}} \pi^{\frac{m(m-1)}{4}} \Pi_{i=1}^{m} \Gamma\left(\frac{v+1-i}{2}\right)}{|\hat{S}|^{\frac{v}{2}}}$$

$$\times f_{iW_m}(\Sigma;\hat{S},v)$$

Avec $v = T-k-m-1$ les degrés de liberté et $\hat{S} = T\hat{\Sigma}$. Cette écriture nous apprend que la vraisemblance du VAR(p) est proportionnelle au produit de la densité d'une normale matricielle et d'une loi inverse Wishart :

$$(8)\quad L(A,\Sigma;Y_T^*) \propto f_{MN_{k,m}}(A;\hat{A},(Z'Z)^{-1},\Sigma)$$

$$\times f_{iW_m}(\Sigma;\hat{S},v)$$

Cette propriété va nous aider à poser une forme de la densité *a priori* telle que nous puissions obtenir une expression analytique de la densité postérieure.

A priori non informatif

Dans cette section nous supposons que nos croyances sont non informatives en adoptant un *a priori* plat à la Jeffrey :

$$(9)\quad p_0(A,\Sigma) = |\Sigma|^{-\frac{m+1}{2}}$$

On note que dans le cas scalaire, $m = 1$, on retrouve le *prior* suggéré par Jeffrey $(1/\sigma^2)$ décrit plus haut. La densité *a posteriori* satisfait donc :

$$p(A,\Sigma|Y_T^*) \propto |\Sigma|^{-\frac{m+1}{2}} \times L(A,\Sigma;Y_T^*)$$

La densité jointe postérieure est donc proportionnelle au produit d'une loi normale multivariée et d'une loi inverse Wishart :

$$(10)\quad p(A,\Sigma;Y_T^*) \propto f_{MN_{k,m}}(A;\hat{A},(Z'Z)^{-1},\Sigma)$$

$$\times f_{iW_m}(\Sigma;\hat{S},\tilde{v})$$

avec $\tilde{v} = T-k$. Ainsi, la densité postérieure s'écrit sous la forme suivante :

$$(11)\quad A|\Sigma,Y_T^* \sim MN_{k,m}(\hat{A},\Sigma,(Z'Z)^{-1})$$

$$\Sigma|Y_T^* \sim iW_m(\hat{S},\tilde{v})$$

Il n'est pas surprenant de constater que la distribution postérieure de A (conditionnelle à la matrice de variance covariance) est centrée sur l'estimateur du maximum de vraisemblance, puisque notre *a priori* est non informatif. Nous pourrions montrer, en intégrant par rapport à Σ, que la distribution postérieure (marginale) de A est une version matricielle de la loi de Student (voir (Zellner, 1971, chapitre 8)). L'*a priori* de Jeffrey n'affecte que le nombre de degré de liberté de la distribution postérieure de A. On obtient la densité marginale postérieure de Y_T^* en intégrant le noyau postérieur successivement par rapport à Σ et A :

$$(12)\quad p(Y_T^*) = (2\pi)^{-\frac{mT}{2}} \times (2\pi)^{\frac{km}{2}} |Z'Z|^{-\frac{m}{2}} |\hat{S}|^{-\frac{\tilde{v}}{2}}$$

$$\times 2^{\frac{\tilde{v}m}{2}} \pi^{\frac{m(m-1)}{4}} \prod_{i=1}^{m} \Gamma\left(\frac{\tilde{v}+1-i}{2}\right)$$

Cette quantité nous renseigne sur la qualité d'ajustement du modèle VAR(p). On note que la densité marginale de Y_T^* est une fonction décroissante de la taille des erreurs ($|\hat{S}|$). Dans cet exemple, nous pouvons caractériser la distribution postérieure analytiquement. Notons néanmoins que même si nous connaissons l'expression analytique de la distribution de A et Σ, la construction des densités prédictives nécessite une approche par simulations[17], puisque les prévisions sont des fonctions non linéaires des matrices auto-régressives (dont nous connaissons la distribution postérieure). L'intérêt pratique de l'approche bayésienne peut paraître faible dans ce cas, dans la mesure où l'espérance postérieure n'est pas différente de l'estimateur du maximum de vraisemblance.

Un exemple d'a priori informatif

Nous considérons maintenant un *prior* plus informatif qui va écarter l'espérance de la distribution *a posteriori* de l'estimateur du maximum de vraisemblance ; dans un modèle linéaire gaussien, l'espérance *a posteriori* est une combinaison convexe de l'estimateur du maximum de vraisemblance et de l'espérance *a priori*. Afin d'aller à l'essentiel[18], nous adoptons une densité *a priori* dégénérée pour la matrice de variance-covariance des erreurs, en supposant que la matrice Σ est connue (on posera $\Sigma = \hat{\Sigma}$). Enfin nous spécifions le *prior* sur A de la façon suivante :

$$(13)\quad p_0(vec\,A) \sim N(a_0, \Omega_0)$$

où Ω_0 est une matrice symétrique définie positive de dimension $mp \times mp$. En multipliant la vraisemblance par l'expression (13), on établit facilement que le noyau postérieur est :

$$(14a)\ K(A|Y_T^*) = \exp\left\{-\frac{1}{2}(a-a_1)'\Omega_1^{-1}(a-a_1)\right\}$$

$$\times\exp\left\{-\frac{1}{2}[a_0'\Omega_0^{-1}a_0 + \hat{a}'(\Sigma^{-1}\otimes Z'Z)\hat{a} - a_1'\Omega_1^{-1}a_1]\right\}$$

$$\times(2\pi)^{-\frac{km}{2}}|\Omega_0|^{-\frac{1}{2}}(2\pi)^{-\frac{mT}{2}}|\Sigma|^{-\frac{T}{2}}e^{-\frac{1}{2}tr\Sigma^{-1}\hat{S}}$$

$$(14b)\quad \Omega_1 = (\Omega_0^{-1} + \Sigma^{-1}\otimes Z'Z)^{-1}$$

$$(14c)\quad a_1 = \Omega_1[\Omega_0^{-1}a_0 + (\Sigma^{-1}\otimes Z'Z)\,vec\,\hat{A}]$$

La distribution postérieure de A est donc gaussienne : $N(a_1, \Omega_1)$, et son interprétation est immédiate. L'inverse de la variance postérieure (Ω_1^{-1}, que l'on peut interpréter comme une quantification de l'information a posteriori) est égale à la somme de l'inverse de la variance *a priori* (Ω_0^{-1}, l'information *a priori*) et de l'inverse de la variance de l'estimateur du maximum de vraisemblance de A ($\Sigma^{-1}\otimes Z'Z$), l'information apportée par les données). *Ceteris paribus*, quand l'information *a priori* est importante, la matrice de variance-covariance Ω_0 est petite, la variance *a posteriori* est faible. L'espérance postérieure est une combinaison linéaire convexe de l'espérance *a priori*, a_0, et de l'estimateur du maximum de vraisemblance, $vec\,\hat{A}$. Les pondérations respectives sont définies par le contenu informatif des croyances *a priori* et de l'échantillon. Lorsque l'information *a priori* tend vers l'infini, *i.e.* $\Omega_0 \to 0$, l'espérance postérieure tend vers l'espérance *a priori*. Lorsque l'information amenée par les données tend vers l'infini, *i.e.* $\Sigma^{-1}\otimes Z'Z \to 0$, l'espérance *a posteriori* tend vers l'estimateur du maximum de vraisemblance. On peut donc interpréter le paradigme bayésien comme un pont entre la calibration et l'estimation par le maximum de vraisemblance. En notant que, si le modèle est stationnaire[19], $Z'Z$ est généralement d'ordre $O(T)$, on retrouve un résultat conforme aux considérations asymptotiques introduites plus haut : l'espérance postérieure tend vers l'estimateur du maximum de vraisemblance lorsque T tend vers l'infini.

À nouveau, en intégrant le noyau postérieur (14) par rapport aux paramètres auto-régressifs, on obtient une expression analytique de la densité marginale postérieure :

$$(15)\quad p(Y_T^*) = (2\pi)^{\frac{km}{2}}|\Omega_1|^{\frac{1}{2}}$$

$$\times\exp\left\{-\frac{1}{2}[a_0'\Omega_0^{-1}a_0 + \hat{a}'(\Sigma^{-1}\otimes Z'Z)\hat{a} - a_1'\Omega_1^{-1}a_1]\right\}$$

$$\times(2\pi)^{-\frac{km}{2}}|\Omega_0|^{-\frac{1}{2}}(2\pi)^{-\frac{mT}{2}}|\Sigma|^{-\frac{T}{2}}e^{-\frac{1}{2}tr\Sigma^{-1}\hat{S}}$$

qui mesure la qualité d'ajustement du modèle et permet de comparer le VAR à d'autres modèles estimés à l'aide du même échantillon.

La pratique

L'intérêt pratique de l'estimation bayésienne des modèles VAR s'explique par l'équation (14b). Celle-ci établit que la variance postérieure de A est inférieure à la variance de l'estimateur du maximum de vraisemblance, $\hat{A}$, dès lors que l'on apporte de l'information *a priori*. L'estimation des modèles VAR sur des données macroéconomiques pose souvent des problèmes de précision. En effet, un modèle avec cinq variables et quatre retards demande l'estimation de vingt paramètres alors que les échantillons sont habituellement de l'ordre de la centaine d'observations. En incorporant de l'information à l'aide d'une densité *a priori* tout se passe comme si nous augmentions le nombre de degrés de liberté. Ce gain en variance sur les paramètres du modèle, permettra d'obtenir des

prévisions ou des fonctions de réponses plus précises.

On peut faire l'analogie avec l'incorporation de contraintes sur les paramètres d'un modèle estimé dans le paradigme classique. Par exemple, si nous pensons que A doit satisfaire les contraintes linéaires définies par $R \times vecA = b$ (où R est une matrice $r \times mp$, b est un vecteur $r \times 1$ et r le nombre de restrictions linéaires), l'incorporation de ces contraintes lors de l'estimation, *i.e.* l'utilisation de moindres carrés contraints, permet de réduire la variance des estimateurs et aussi l'erreur quadratique moyenne (dans la mesure où la contrainte n'est pas en contradiction avec le processus générateur des données)[20]. Le paradigme bayésien est plus souple, dans le sens où il ne pose pas des contraintes déterministes. Dans certains cas[21], la définition de croyances *a priori* revient à poser une contrainte probabiliste de la forme $R \times vec\,A - b = \varepsilon$, où ε est une variable aléatoire gaussienne. Plus la variance de ε est importante, moins la contrainte sur $vecA$ est forte (plus l'information *a priori* est floue).

La formalisation de l'information *a priori* ne se limite pas au choix de la forme d'une distribution. Dans le cas du modèle BVAR de la section précédente, nous devons aussi choisir les paramètres a_0 et Ω_0. Dans cette perspective, un *prior* qui s'est montré particulièrement efficace quand on cherche à modéliser des séries macroéconomiques est le *prior* de Minnesota[22]. Celui-ci correspond à la croyance *a priori* que les séries observées sont des marches aléatoires indépendantes. L'espérance *a priori* de vec$A$ est alors telle que $E[\mathbf{A}_1] = I_m$ et $E[\mathbf{A}_i] = 0_m$ pour $i = 2, ..., p$. La variance *a priori* de vec$A$ est supposée diagonale. En notant $\omega_{i,j,k}\ (i,j=1,...,m, k=1,...,p)$ la variance associée au paramètre correspondant à la variable j dans l'équation i au retard k, la variance *a priori* est définie par :

$$\omega_{i,i,k} = \frac{\pi_1}{k^{\pi 3}} \quad i=1,...,m \quad \text{et} \quad k=1,...,p$$

$$\omega_{i,j,k} = \frac{\pi_2}{k^{\pi 3}} \frac{\sigma_i}{\sigma_j} \; i=1,...,m \quad j \neq i \; \text{et} \quad k=1,...,p$$

où les hyperparamètres π_h pour h =1,2,3 sont positifs, $\{\sigma_i^2\}$ est l'estimateur de la variance des résidus dans l'estimation d'un AR(p) pour la variable i. Le ratio des écarts types permet de prendre en compte les différences d'échelles entre les différentes variables composant le vecteur des observables. La variance *a priori* décroît lorsque le retard k augmente, ce qui traduit l'idée que plus le retard est important plus nous croyons que la matrice $\mathbf{A}_k$ est nulle. L'hyperparamètre π_3 indique à quelle vitesse la variance *a priori* tend vers zéro. Des valeurs fréquemment utilisées pour π_1 et π_2 sont respectivement 0,05 et 0,005. Cela revient à dire, dans la mesure où σ_i et σ_j sont proches, que nos croyances *a priori* sont plus fortes sur la nullité des termes hors des diagnonales de $\mathbf{A}_k\ (k=1,...,p)$, c'est-à-dire sur l'absence de causalité[23]. Notons que l'*a priori* de Minnesota suppose l'absence de relations de cointégration entre les variables : il y a, *a priori*, autant de racines unitaires que de variables. Néanmoins rien n'empêche l'apparition de relations de long terme dans la distribution postérieure.

Il nous reste à choisir les valeurs des hyperparamètres du *prior* de Minnesota. Cette étape est importante car l'expérience montre que l'inférence postérieure, en particulier l'évaluation de la densité marginale qui nous permet d'évaluer le modèle, est très sensible à ce choix. Si, comme souvent dans la littérature (voir par exemple Smets et Wouters (2002) ou Fernández-Villaverde et Rubio-Ramírez (2001)), l'estimation d'un BVAR ne sert qu'à titre de comparaison afin d'évaluer la qualité d'ajustement d'un modèle DSGE, le choix des hyperparamètres devient crucial. Ce point n'est malheureusement jamais abordé dans la littérature. Le contenu économique d'un modèle VAR étant faible, il paraît difficile de recourir à la théorie pour spécifier la densité *a priori*. Un critère objectif à notre disposition est de choisir les hyperparamètres $(\pi_1, \pi_2 \text{ et } \pi_3)$ qui maximisent les performances en prévisions du modèle BVAR. En spécifiant ainsi les *priors* de notre BVAR, nous savons au moins que nous ne comparons par notre DSGE avec un BVAR aux performances prédictives médiocres. Dans cet esprit Phillips (1996) propose le critère PIC (*Posterior Information Criterion*) que l'on peut minimiser par rapport aux hyperparamètres. Ce critère peut être vu comme une généralisation, au cas non stationnaire, du bien connu critère BIC[24]. Dans le cas du modèle considéré dans la section précédente on choisit les hyperparamètres de la façon suivante :

$$(16)\quad (\pi_1^*, \pi_2^*, \pi_3^*) = \arg \min_{\pi_1,\pi_2,\pi_3} \log|\widetilde{\Sigma}| + \frac{1}{T} \log \frac{|\Omega_0^{-1} + \widetilde{\Sigma}^{-1} \otimes Z'Z|}{|\Omega_0^{-1} + \widetilde{\Sigma}_{T_0}^{-1} \otimes Z_{T_0}' Z_{T_0}|}$$

où $\widetilde{\Sigma}$ est la matrice de variance covariance des innovations au mode postérieur, les matrices indicées par T_0 sont obtenues à partir du sous échantillon 1,...,T_0 (où T_0 est supérieur au nombre de paramètres estimés). Ici nous avons considéré le nombre de retards comme une donnée, mais nous pourrions aussi optimiser par rapport à p le critère PIC (voir Phillips, 1996). À notre connaissance, l'utilisation d'*a priori* objectif pour les modèles BVAR, tel que l'optimisation du critère PIC proposé par Phillips (1996), demeure inappliqué dans la littérature. On peut donc légitimement douter de la pertinence des comparaisons entre BVAR et DSGE effectuées jusqu'à présent[25].

Modèles DSGE

Dans cette partie nous présentons de façon générale les modèles DSGE, puis soulignons les problèmes que peut poser leur estimation. En particulier, nous expliquons pourquoi, à la différence des BVAR, il n'est pas possible d'obtenir une expression analytique de la distribution postérieure. Nous terminons en présentant les méthodes de simulation de Monte Carlo utilisées pour calculer les croyances postérieures.

Résolution et vraisemblance

Nous limitons notre attention aux modèles DSGE que nous pouvons écrire sous la forme suivante :

$$(17)\ \mathrm{E}_t\left[F_\theta(y_{t+1}, y_t, y_{t-1}, \varepsilon_t)\right] = 0$$

avec $\varepsilon_t \sim iid(0,\Sigma)$, un vecteur aléatoire dans R^r, les innovations structurelles, $y_t \in \Lambda \subseteq \mathrm{R}^n$ un vecteur regroupant les variables endogènes, $F : \Lambda^3 \times \mathrm{R}^n \to \Lambda$ une fonction réelle dans C^2 paramétrée par un vecteur réel $\theta \in \Theta \subseteq \mathrm{R}^q$ regroupant l'ensemble des paramètres structurels du modèle. La fonction F est simplement l'ensemble des équations qui définissent un modèle ; on a autant d'équations que de variables endogènes. Le vecteur des variables endogènes, y_t, inclut des variables d'état (endogènes ou exogènes), des variables de choix et des variables statiques [(26)]. On supposera qu'il est possible d'exhiber une unique solution stable et invariante du modèle décrit par l'équation (17) :

$$(18)\ y_t = H_\theta(y_{t-1}, \varepsilon_t)$$

qui exprime les variables endogènes en fonction du passé et des chocs structurels contemporains. La fonction paramétrée H_θ regroupe les *policy rules* et les équations de transition (voir la contribution de Michel Juillard et Tarik Ocaktan à ce numéro). La solution (18), en décrivant une récurrence stochastique non linéaire, définit la distribution jointe d'un ensemble de variables.

Pour estimer les paramètres θ du modèle, ou d'un sous ensemble des paramètres, nous devons évaluer la vraisemblance associée au modèle (17) ou à sa forme réduite (18). Même si la fonction H_θ est linéaire en y_{t-1} et ε_t, cette évaluation ne peut être directe comme dans le cas du modèle VAR examiné plus haut. En effet, l'équation (18) décrit la distribution jointe d'un ensemble de variables qui ne sont pas toutes observées. Afin d'amener le modèle aux données on peut l'écrire sous une forme état-mesure :

$$(19a)\ y_t^* = Z y_t + \eta_t$$

$$(19b)\ y_t = H_\theta(y_{t-1}, \varepsilon_t)$$

où y_t^* est un vecteur $m \times 1$, avec $r \leq m < n$, regroupant les variables observées et Z est une matrice de sélection $m \times n$. On peut éventuellement augmenter l'équation de mesure d'un bruit blanc multivarié, η_t, représentant l'inadéquation des variables théoriques avec les variables observées, ou plus simplement une erreur de mesure. On note $Y_T^* = \{y_t^*\}_{t=1}^T$ l'échantillon à notre disposition et $\psi \in \Psi \subseteq \mathrm{R}^{q+\frac{n(n+1)}{2}+\frac{r(r+1)}{2}}$ le vecteur des paramètres du modèle état-mesure (θ, Σ et éventuellement la matrice de variance-covariance de η_t). La vraisemblance est la densité de l'échantillon, conditionnellement aux paramètres ψ et au modèle défini par (19) :

$$(20)\ L(\psi; Y_T^*) = p(Y_T^* | \psi) = \prod_{t=1}^{T} p(y_t^* | Y_{t-1}^*, \psi)$$

L'évaluation de la densité de y_t^* conditionnellement à Y_{t-1}^* n'est généralement pas directe, dans la mesure où y_t^* dépend de variables endogènes inobservables. Nous pouvons néanmoins utiliser la relation suivante :

$$(21)\ p(y_t^* | Y_{t-1}^*, \psi) = \int_\Lambda p(y_t^* | y_t, \psi)\, p(y_t | Y_{t-1}^*, \psi)\, dy_t$$

La densité de y_t^* conditionnellement à Y_{t-1}^* est obtenue comme la moyenne de la densité de y_t^* sachant y_t, pondérée par la densité de y_t sachant Y_{t-1}^*. La première densité sous l'intégrale est specifiée par l'équation de mesure (19a). L'évaluation de la densité de la prévision des variables latentes, conditionnellement à l'information disponible en t-1, est moins directe, et on doit utiliser un filtre de Kalman. Il s'agit d'une procédure récursive. À chaque date (entre 1 et T) on forme une prévision des variables latentes (y_t sachant Y_{t-1}^*), en utilisant l'équation d'état (y_t sachant y_{t-1}, équation 19b) et une estimation initiale des variables latentes (y_{t-1} sachant Y_{t-1}^*), puis on corrige cette prévision quand une nouvelle observation (y_t^*) augmente l'ensemble d'information. On peut interpréter cette démarche comme une estimation récursive bayésienne du vecteur des variables latentes. En initialisant les variables latentes avec la densité associée à la distribution ergodique des variables latentes définie par l'équation d'état (19b),

$p(y_0 | Y_0^*, \psi) = p(y_0 | \psi) = p(y_\infty | \psi)$, la récursion s'écrit de la façon suivante :

(22a) $p(y_t | Y_{t-1}^*, \psi) = \int_\Lambda p(y_t | y_{t-1}, \psi)$

$$p(y_{t-1} | Y_{t-1}^*, \psi)\, dy_{t-1}$$

(22b)

$$p(y_t | Y_t^*, \psi) = \frac{p(y_t^* | y_t, \psi)\, p(y_t | Y_{t-1}^*, \psi)}{\int_\Lambda p(y_t^* | y_t, \psi)\, p(y_t | Y_{t-1}^*, \psi)\, dy_t}$$

L'interprétation de l'équation de prédiction (22a) est directe : la densité de la prédiction des variables latentes en *t* est la moyenne de la densité de y_t sachant y_{t-1}, définie par l'équation d'état (19b), pondérée par la densité de y_{t-1} sachant Y_{t-1}^* . Cette dernière densité est définie par l'équation de mise à jour (22b) ou la condition initiale. L'équation de mise à jour (22b) est, à l'instar de l'équation (3), une application directe du théorème de Bayes. Le premier terme au numérateur, $p(y_t | Y_{t-1}^*)$, est la densité *a priori* du vecteur des variables latentes. Le second terme, $p(y_t^* | y_t)$, la densité de l'observation sachant l'état obtenu *via* l'équation de mesure (19a), est la vraisemblance. Le dénominateur est la densité marginale de la nouvelle observation.

Puisque nous pouvons, au moins théoriquement, évaluer la vraisemblance associée au modèle DSGE, nous devrions être capable d'estimer ses paramètres. Malheureusement, les équations (21) et (22) nécessitent l'évaluation d'intégrales[27] dans l'espace des variables d'état. Quand le nombre de variables latentes agmente il devient très coûteux d'évaluer ces intégrales (on parle de *curse of dimensionality*). La dérivation de la forme réduite du modèle (18) nécessite également l'évaluation d'intégrales. En pratique, même pour des modèles de dimensions modestes, l'évaluation de la vraisemblance est difficile. Nous devons donc approximer celle-ci. Dans le cas où le modèle état-mesure (19) est linéaire et gaussien, l'évaluation des intégrales devient très simple car les variables latentes et observées sont normalement distribuées à chaque date. Ainsi la dynamique de la distribution des variables latentes est complétement caractérisée par la dynamique de l'espérance et de la variance des variables latentes. On peut trouver une présentation du filtre de Kalman dans ce cas simple dans Gouriéroux et Monfort 1989, chapitre 13), Harvey (1989, chapitre 3) ou encore dans la contribution de Fabrice Collard et Patrick Fève à ce même numéro. Ceci explique pourquoi les modèles DSGE estimés sont généralement (log-) linéarisés autour de l'état stationnaire.

Malgré l'approximation (log-) linéaire de la forme réduite du modèle (H_θ) l'évaluation de la vraisemblance est numérique. Nous ne disposons pas d'une expression analytique, comme dans le cas du modèle VAR, et ne pouvons donc écrire formellement la densité postérieure ou les moments postérieurs. Deux possibilités s'offrent à nous.

La première est de considérer une approximation asymptotique de la densité postérieure. Il est alors possible d'approximer, voir plus haut et Tierney et Kadane (1986), tout moment a posteriori ou la densité marginale.

Nous avons vu que l'erreur d'approximation des moments est d'ordre $O(T^{-2})$ et que l'erreur d'approximation de la densité marginale est d'ordre $O(T^{-1})$. L'expérience suggère, pour les dimensions d'échantillon, *T*, généralement considérées dans la littérature, que l'approximation de Laplace de la densité marginale est satisfaisante[28].

La deuxième possibilité est d'évaluer les moments en recourant à des simulations par Monte-Carlo. L'intuition de cette approche repose sur la loi des grands nombres. Supposons, par exemple, que nous souhaitions évaluer l'espérance d'une variable aléatoire *X* de distribution *G*. Si l'on génère une suite de variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et distribuées selon *G*, alors une approximation de l'espérance est donnée par la moyenne empirique de ces variables[29] ; la loi des grands nombres assure que l'erreur d'approximation tend vers zéro presque sûrement lorsque le nombre de tirages n tend vers l'infini. Si on admet de plus que le moment du second ordre existe, la vitesse de convergence est $O(n^{1/2})$ par application du théorème *central limit*.

En pratique nous pouvons être intéressés par les moments postérieurs de θ. En notant que :

(23) $\mathrm{E}_t[\varphi(\theta)] = \int_\Theta \varphi(\theta)\, p_1(\theta | Y_T^*)\, d\theta$

il semble alors naturel d'utiliser la moyenne empirique de $(\varphi(\theta^{(1)})(\varphi(\theta^{(2)}),\ldots,(\varphi(\theta^{(n)}))$ où les $\theta^{(i)}$ sont des tirages indépendants dans la distribution postérieure, pour évaluer l'espérance de $\varphi(\theta)$. L'erreur d'approximation tend vers zéro lorsque le nombre de simulations (*n*) tend vers l'infini. il convient de noter que p_1 est généralement d'une forme inconnue et que l'on ne peut donc pas définir un générateur pseudo-aléatoire reproduisant la distribution a posteriori. Comme nous le verrons par la suite, la méthode de Monte Carlo dite de fonction d'importance, permet de remédier à cette difficulté sous certaines conditions. Ce principe de Monte Carlo se généralise au cas où les variables simulées ne sont pas indépendantes. Il est ainsi possible dans certains cas de construire, moyennant certaines conditions, une chaîne de Markov $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)}$ de loi stationnaire (ergodique) p_1 telle que la moyenne empirique des $\varphi(\theta^{(j)})$ (φ est la statistique d'intérêt) converge presque sûrement

vers la quantité d'intérêt comme dans le cas des tirages indépendants. Tout l'art de l'exercice est alors de déterminer une chaîne de Markov (et plus précisément son noyau de transition) telle que sa loi ergodique corresponde à la loi a posteriori désirée et d'évaluer le temps d'arrêt des simulations, *i.e.* de diagnostiquer la convergence de la chaîne de Markov[30].

L'échantillonnage bayésien par fonction d'importance

Idéalement, on souhaiterait générer les paramètres d'intérêt suivant la distribution *a posteriori*. Cependant, à l'exception de rares modèles, ceci n'est pas possible. On peut néanmoins exploiter le fait qu'il n'est pas nécessaire de générer une suite de tirage $\{\theta^{(s)}\}$, suivant la distribution a posteriori, pour obtenir une évaluation correcte des moments postérieurs. En effet, sous certaines conditions de régularité, on peut utiliser une densité de probabilité *h*, définie sur le même espace et appelée fonction d'importance, suffisamment proche de p_1 (dans un sens à préciser) et échantillonner à partir de cette dernière. Il est alors possible de montrer par la loi des grands nombres que l'intégrale (23) définissant le moment postérieur est approchée par :

$$(24)\quad \mathrm{E}_t[\varphi(\theta)] \approx \frac{\Sigma_{s=1}^{n} \varphi(\theta^{(s)})\, w(\theta^{(s)})}{\Sigma_{s=1}^{n} w(\theta^{(s)})}$$

avec

$$w(\theta^{(s)}) = \frac{p(Y_T^* \mid \theta^{(s)})\, p_0(\theta^{(s)})}{h(\theta^{(s)})}$$

Le choix de la fonction d'importance est crucial : elle doit être suffisamment proche de la loi *a posteriori*, ce qui n'est pas toujours simple en pratique. En effet, si *h* est une mauvaise approximation de p_1, alors les poids sont généralement faibles pour la plupart des valeurs échantillonnées de θ, la somme est alors dominée par quelques termes dont les poids sont très élevés. Il en résulte une estimation peu fiable, voir Casella et Robert (2004) pour plus de détails. L'algorithme se résume comme suit :

Algorithme 1.

(1) *Maximiser le noyau postérieur par rapport à* θ. *On obtient le mode de la densité postérieure,* θ^m, *et le hessien au mode qui caractérise la courbure de la densité postérieure au mode et dont l'inverse de l'opposé, noté* $\Sigma(\theta^m)$, *approxime la variance postérieure.*

(2) *Générer* $\theta^{(s)}$, *suivant une fonction d'importance, h, dont les moments du premier et second ordre dépendent de* θ^m *et* $\Sigma(\theta^m)$

(3) *Déterminer les poids* $w(\theta^{(s)})$ *selon (24).*

(4) *Effectuer (2-3) pour s* =1,...,*n*.

(5) *Calculer :*

$$\frac{\Sigma_{s=1}^{n} \varphi(\theta^{(s)})\, w(\theta^{(s)})}{\Sigma_{s=1}^{n} w(\theta^{(s)})}$$

La première étape n'est pas spécifique à l'algorithme par fonction d'importance : il s'agit de calculer les moments associés à la fonction d'importance, *h*. Cette « calibration » des moments de la fonction d'importance est généralement faite à partir de la maximisation du logarithme du noyau *a posteriori*. Étant données les propriétés asymptotiques de la distribution postérieure, ces choix sur les moments d'ordre un et deux associés à *h* sont d'autant plus satisfaisants que la taille de l'échantillon est importante.

Bien que très populaire en statistique, cette méthode est peu utilisée dans le cadre de l'estimation de modèles DSGE. À titre d'exemples, Dejong *et alii* (2000) estiment avec cette méthode un modèle de croissance stochastique linéarisée. An et Schorfheide (2007) comparent l'algorithme d'importance avec celui de Metropolis (à pas aléatoire) dans une version simplifiée du modèle de Smets et Wouters (2002). Pour ce faire, ils retiennent comme fonction d'importance une distribution de Student multivariée.

Les méthodes de Monte-Carlo à chaînes de Markov

Cette seconde classe d'algorithmes permet de générer des variables aléatoires suivant approximativement la loi *a posteriori*, lorsque cette dernière n'est pas disponible. Elle évite donc l'appel à une fonction d'importance, *h*, souvent difficile à déterminer pour les modèles DSGE[31]. On cherche ainsi à définir une chaîne de Markov dont la distribution ergodique est approximativement le noyau *a posteriori*. Si cette chaîne existe, la méthode d'échantillonnage est grossièrement définie comme suit : dans un premier temps, on initialise (arbitrairement) la chaîne de Markov. Dans un second temps, on génère les $\theta^{(s)}$ à partir de cette chaîne. À l'issue d'un certain nombre de tirages (disons n_0), on dispose de réalisations de variables aléatoires $\{\theta^{(s)}, s=n_0,\ldots,n\}$ approximativement distribuées comme la distribution *a posteriori*.

Chaînes de Markov

Une chaîne de Markov est une suite de variables aléatoires continues à valeurs dans Θ,$(\theta^{(0)},\ldots,\theta^{(n)})$, générée par un processus de Markov. Une suite de variables aléatoires est générée par un processus de Markov (d'ordre 1) si la distribution de $\theta^{(s)}$ ne dépend que de $\theta^{(s-1)}$. Une chaîne de Markov est caractérisée par un noyau de transition qui spécifie la probabilité de passer de $\eta \in \Theta$ à $S \subseteq \Theta$. Nous

noterons $P(\eta,S)$ le noyau de transition, il vérifie $P(\eta,\Theta)=1$ pour tout η dans Θ. Si la chaîne de Markov définie par le noyau P converge vers une distribution invariante π, alors le noyau doit satisfaire l'identité suivante :

$$\pi(S)=\int_{\Theta} P(\eta,S)\pi(d\eta)$$

pour tout sous ensemble mesurable S de Θ. Plus généralement, avant d'atteindre la distribution ergodique π, si nous notons $P^{(s)}(\eta,S)$ la probabilité que $\theta^{(s)}$ soit dans S sachant que $\theta^{(s-1)}=\eta$, nous avons :

$$P^{(s)}(\eta,S)=\int_{\Theta} P(\nu,S)P^{(s-1)}(\eta,dv)$$

la distribution de θ s'ajuste d'itération en itération puis rejoint la distribution ergodique, $\lim_{s\to\infty} P^{(s)}(\eta,S)=\pi(S)$. L'idée est alors de choisir le noyau de transition qui nous aménera vers la distribution invariante désirée.

Définissons $p(\eta,\mu)$ et $\tilde{\pi}$ les densités associées au noyau P et à la distribution π [32]. Tierney (1994) montre que si la densité $p(\eta,\mu)$ vérifie la condition de réversibilité[33] :

$$\tilde{\pi}(\eta)\,p(\eta,v)=\tilde{\pi}(v)\,p(v,\eta)$$

alors π est la distribution invariante associée au noyau P [34]. De façon équivalente :

$$\frac{\tilde{\pi}(\eta)}{\tilde{\pi}(v)}=\frac{p(v,\eta)}{p(\eta,v)}$$

Cette condition nous dit simplement que si la densité de $\theta=\eta$, $\tilde{\pi}(\eta)$, domine la densité associée à $\theta=v$, $\tilde{\pi}(v)$, alors il doit être plus « facile » de passer de v à η que de η à v.

Cette propriété nous aidera à construire une chaîne de Markov dont la distribution invariante est la distribution postérieure des paramètres θ dans le modèle DSGE. On comprend bien que le noyau de cette chaîne est difficile à définir. Supposons que l'on puisse choisir un noyau de transition $Q(\eta,S)$; alors il est presque sûr que la condition de réversibilité ne sera pas vérifiée, c'est-à-dire que nous aurons :

$$p_1(\eta|Y_T^*)\,q(\eta,v)\neq p_1(v|Y_T^*)\,q(v,\eta)$$

L'algorithme de Metropolis-Hastings est une approche générale qui permet de *corriger* ce noyau, de façon à respecter la condition de réversibilité.

L'algorithme de Métropolis-Hasting

Supposons que l'on puisse définir une densité instrumentale, qui permette d'approcher le noyau de transition de la chaîne de Markov dont la densité ergodique est la loi a posteriori de notre modèle. Cette densité est définie par $q(\eta,v)\equiv q(v|\eta)$.

Algorithme 2 (Metropolis-Hastings).

(1) *Se donner une condition initiale* $\theta^{(0)}$ *telle que* $K(\theta^{(0)}|Y_T^*)>0$ *et poser s =1.*

(2) *Générer un candidat (une proposition)* θ^* *à partir d'une densité* $q(\theta^{(s-1)},\theta^*)$.

(3) *Générer u dans une loi uniforme entre [0,1].*

(4) *Appliquer la règle suivante* :

$$\theta^{(s)}=\begin{cases}\theta^* & si\ \alpha(\theta^{(s-1)},\theta^*)>u\\ \theta^{(s-1)} & sinon.\end{cases}$$

où

$$\alpha(\theta^{(s-1)},\theta^*)=\min\left\{1,\frac{K(\theta^*|Y_T^*)}{K(\theta^{(s-1)}|Y_T^*)}\,\frac{q(\theta^{(s-1)}|\theta^*)}{q(\theta^*|\theta^{(s-1)})}\right\}$$

(5) *Effectuer (2-4) pour s* =2,...,*n.*

Notons qu'il suffit de pouvoir évaluer le noyau postérieur pour mettre en œuvre cet algorithme ; la connaissance de la densité postérieure à une constante près est suffisante. L'algorithme de Metropolis-Hasting requiert le choix d'une fonction instrumentale q à partir de laquelle on génère des transitions dans l'espace des paramètres. La densité conditionnelle q permet de générer un vecteur candidat θ^*. Puisqu'elle n'est pas nécessairement la densité conditionnelle associée au noyau de transition dont la distribution ergodique est la distribution *a posteriori* recherchée, la condition de réversibilité n'est pas vérifiée[35]. L'algorithme de MH corrige cette erreur[36] en n'acceptant pas systématiquement les propositions de q. En introduisant une probabilité d'acceptation de la transition proposée, α, on peut finalement vérifier la condition de réversibilité. Pour cela, la probabilité d'acceptation doit être telle que :

$$K(\eta|Y_T^*)\,q(\eta,v)\alpha(\eta,v)=K(v|Y_T^*)\,q(v,\eta)\alpha(v,\eta)$$

Soit :

$$\alpha(\eta,v)=\min\left\{1,\frac{K(v|Y_T^*)}{K(\eta|Y_T^*)}\,\frac{q(v,\eta)}{q(\eta,v)}\right\}$$

Il nous reste à déterminer (*i*) comment nous devons initialiser la chaîne et (*ii*) la longueur de la chaîne.

Nous reviendrons par la suite, lors de la présentation de l'algorithme de Metropolis à pas aléatoires, sur le premier point. Nous aborderons la question du nombre de simulations nécessaires, c'est-à-dire de la longueur de la chaîne, dans la section sur les diagnostics de convergence plus bas. Pour l'instant nous supposerons que pour tout $s>n_0$ les $\theta^{(s)}$ sont tirés dans la distribution ciblée. Afin de s'assurer que les résultats sont indépendants des conditions initiales, nous ne considèrons pas les simulations indicées par $s=0,\ldots,n_0$. Ainsi, pour évaluer $E[\varphi(\theta)]$ nous calculons :

$$(n-n_0)^{-1}\sum_{s=n_{0+1}}^{n}\varphi(\theta^{(s)})$$

qui converge vers le moment postérieur recherché lorsque le nombre de simulations, n, tend vers l'infini.

Deux variantes de l'algorithme MH

L'algorithme de MH à pas aléatoires. Comme nous l'avons expliqué plus haut, l'utilisation de l'algorithme de Metropolis-Hastings repose sur le fait qu'il est aisé d'échantillonner à partir de la densité instrumentale q. L'inconvénient est que cette dernière n'est pas toujours facile à déterminer. Dans cette perspective, l'algorithme de Metropolis à pas aléatoires est utile lorsqu'il est difficile d'obtenir une bonne approximation de la densité *a posteriori*. Une proposition à l'itération s est définie par :

$$\theta^{*}=\theta^{(s-1)}+z$$

où z est le pas aléatoire. Le choix de la densité de z détermine la forme précise de la densité instrumentrale, q.

Un choix standard est la distribution gaussienne multivariée : $z\sim N(0,\Sigma)$. Ainsi la densité de θ^{*} conditionnel à $\theta^{(s-1)}$ est gaussienne :

$$q\left(\theta^{(s-1)},\theta^{*}\right)\equiv q\left(\theta^{*}\,|\,\theta^{(s-1)}\right)\sim N\left(\theta^{(s-1)},\Sigma\right)$$

Par symétrie de la loi normale, la densité instrumentale vérifie $q(\eta,v)=q(v,\eta)$. Ainsi la probabilité d'acceptation ne dépend que du noyau postérieur :

$$\alpha(\eta,v)=\min\left\{1,\frac{K(v\,|\,Y_T^{*})}{K(\eta\,|\,Y_T^{*})}\right\}$$

Autrement dit, si $K(\theta^{*}\,|\,Y_T^{*})\geq K(\theta^{(s-1)}\,|\,Y_T^{*})$, la chaîne de Markov se déplace en θ^{*}. Si ce n'est pas le cas, la chaîne se déplace avec une probabilité égale au rapport des densités *a posteriori*. On accepte avec une probabilité unitaire la proposition dans une phase ascendante (c'est-à-dire lorsque la probabilité *a posteriori* croît) et avec une probabilité non nulle la proposition dans une phase descendante (si nous décidions de rejeter systématiquement ces propositions défavorables la chaîne ne Markov ne visiterait pas complétement l'espace des paramètres et *a fortiori* elle ne pourrait converger vers la distribution postérieure recherchée).

La probabilité d'acceptation peut être ajustée à partir de la matrice de variance-covariance Σ. En effet, si celle-ci est *grande* alors il y a de fortes chances pour que la transition proposée nous amène dans les queues de la distribution *a posteriori* c'est-à-dire dans une région où la densité est faible et où donc la probabilité d'acceptation est proche de zéro (si on vient d'une zone plus dense). Si la matrice Σ est *petite* alors les transitions proposées ne seront que des petits pas. Dans ce cas il n'y aura pas de grandes variations de la densité postérieure et donc la probabilité d'acceptation sera proche de 1. Nous écrivons Σ sous la forme $c\Omega$, où Ω est une estimation de la matrice de variance-covariance a posteriori et c un paramètre d'échelle qui permet de jouer sur le caractère plus ou moins diffus de la matrice de variance-covariance et donc d'ajuster le taux d'acceptation. On peut expérimenter différentes valeurs de c afin de déterminer une probabilité d'acceptation raisonnable(37).

Cette version de l'algorithme de MH est généralement utilisée dans la littérature concernée par l'estimation bayésienne des modèles DSGE.

Algorithme 3.

(1) *Maximiser le noyau postérieur par rapport à* θ. *On obtient le mode de la densité postérieure,* θ^{m}, *et le hessien au mode qui caractérise la courbure de la densité postérieure au mode et dont l'inverse de l'opposé, noté* $\Omega(\theta^{m})$, *approxime la variance postérieure. On pose* $\Sigma=c\Omega(\theta^{m})$ *avec* $c>0$, $s=1$ *et* $\theta^{(0)}=\theta^{m}$.

(2) *Générer* θ^{*} *à partir d'une gaussienne d'espérance* $\theta^{(s-1)}$ *et de variance* Σ

(3) *Générer u dans une loi uniforme entre* [0,1]

(4) *Appliquer la règle suivante* :

$$\theta^{(s)}=\begin{cases}\theta^{*} & \text{si } \alpha(\theta^{(s-1)},\theta^{*})>u\\ \theta^{(s-1)} & \text{sinon.}\end{cases}$$

où

$$\alpha(\theta^{(s-1)},\theta^{*})=\min\left\{1,\frac{K(\theta^{*}\,|\,Y_T^{*})}{K(\theta^{(s-1)}\,|\,Y_T^{*})}\right\}$$

(5) *Effectuer (2-4) pour* $s=2,\ldots,n$.

L'algorithme à chaînes indépendantes. Si la proposition est indépendante de l'état courant, l'algorithme est dit à chaines indépendantes (Tierney, 1994). La probabilité d'acceptation se simplifie alors comme suit :

$$\alpha(\theta^{(s-1)},\theta^{*}) = \min\left\{1, \frac{K(\theta^{*}\,|\,Y_T^{*})}{K(\theta^{(s-1)}\,|\,Y_T^{*})}\,\frac{q(\theta^{(s-1)})}{q(\theta^{*})}\right\}$$

Cet algorithme est particulièrement adapté au cas où il existe une approximation naturelle de la densité *a posteriori*. En effet, l'algorithme à chaînes indépendantes est alors similaire à l'algorithme par fonction d'importance. Pour s'en convaincre, il suffit de remarquer que l'on peut définir des poids analogues à ceux vus précédemment comme suit :

$$w(\theta) = \frac{p_1(\theta^{*}\,|\,Y_T, A)}{q(\theta)}$$

La probabilité d'acceptation est alors donnée par $\alpha(\theta^{(s-1)},\theta^{(s)}) = \min\{1, w(\theta^{*})/(\theta^{(s-1)})\}$. Autrement dit, il s'agit du ratio des poids d'échantillonnage par importance entre le vecteur candidat et le vecteur précédent.

Les diagnostics de convergence

Un certain nombre de résultats théoriques sont disponibles pour l'étude de la convergence des chaînes de Markov. Il est cependant extrêmement compliqué d'énoncer des règles pratiques. Ainsi, il n'existe aucune réponse simple à la question : Quel est le nombre optimal de simulations pour garantir la convergence de la chaîne de Markov vers la distribution ergodique ? Nous discutons brièvement quelques diagnostics de convergence. Le lecteur intéressé pourra consulter, par exemple, Casella et Robert (2004, chapitre 12).

La littérature bayésienne distingue généralement trois types de convergence : (*i*) la convergence vers la distribution stationnaire, (*ii*) la convergence des moments empiriques (ou approchés) vers les moments théoriques et (*iii*) la convergence vers un échantillonage *i.i.d*. Nous nous intéressons ici aux deux dernières formes de convergence[38]. Quelle que soit la convergence étudiée, les résultats reposent soit sur des méthodes d'évaluation graphiques dont il est difficile d'en déduire des règles générales, soit sur des tests statistiques formels.

Avant de présenter certaines de ces méthodes, il convient de noter qu'il est important de distinguer les méthodes qui font appel à *M* chaînes de Markov parallèles et indépendantes et celles basées sur une seule chaîne (*on-line Markov chain*). L'utilisation de plusieurs chaînes est coûteuse en temps mais elle réduit la dépendance aux conditions initiales et accroît la possibilité de parcourir efficacement l'espace des paramètres, Θ. En particulier, si une chaîne de Markov est mélangeante au sens faible – elle reste coincée dans des régions (de mesure dominée) de l'espace des paramètres – une explication possible est la présence d'une distribution a posteriori multimodale (notamment lorsque les densités *a priori* sont en conflit avec la vraisemblance du modèle). Dans ce cas, la mise en oeuvre de chaînes de Markov en parallèle et indépendantes, très dispersées, peut permettre de résoudre cette difficulté. L'argument des chaînes multiples est aussi avancé pour s'assurer de la convergence. Si on se donne plusieurs vecteurs initiaux de paramètres, suffisamment dispersés, et que l'on obtient les mêmes résultats, la convergence sera assurée. L'argument est valide si et seulement si on a bien vérifié que chaque chaîne a convergé. Nous revoilà donc revenu au point de départ... Il existe une littérature abondante discutant des avantages et inconvénients respectifs de chaque méthode[39].

La convergence des moments empiriques ou approchés vers les moments théoriques peut s'appréhender à partir de méthodes graphiques. Yu et Mykland (1998) se basent sur les sommes cumulatives des moments d'intérêt dans le cas d'une seule chaîne de Markov. Au contraire, Gelman et Rubin (1992) proposent un test formel qui repose sur des méthodes de chaînes de Markov en parallèle. La convergence est diagnostiquée si les différences entre *J* des *M* chaines de Markov restent dans un intervalle raisonnable. Gelman et Rubin formalisent cette idée en recourant à des statistiques de type ANOVA. Pour chaque statistique d'intérêt ϕ, ils déterminent la variance intra et inter-chaînes. L'intuition du test est alors la suivante. Si l'effet des valeurs initiales de chaque chaîne de Markov a été supprimé, les chaînes en parallèle doivent être relativement proches. En d'autres termes, la variance inter-chaîne ne devrait pas être trop grande par rapport à la variance intra-chaîne[40]. La statistique de test est alors définie à partir d'un estimateur de la variance *a posteriori* de ϕ.

Plus précisément, ce dernier est une moyenne pondérée de la variance intra-chaîne et de la variance inter-chaîne. Le critère de convergence est ainsi le rapport de cet estimateur à la variance intra-chaîne. En utilisant une approximation de ce critère, les auteurs montrent que si sa valeur excède 1, 2, on peut en conclure qu'il n'y a pas convergence. Dans une autre optique, Geweke (1992) propose de comparer la moyenne de deux sous-échantillons disjoints, S_1 et S_3, d'une chaîne de Markov (après avoir éliminé les n_0 premières valeurs). On choisit S_1 (resp. S_3) au début (resp. à la fin) de la chaîne de Markov. Si la chaîne de Markov a atteint la distribution stationnaire, la moyenne des deux sous-échantillons doit être égale. Une version modifiée de la statistique *z* est alors élaborée par Geweke[41]. Une valeur de la statistique de test supérieure à 2 indique qu'un nombre plus élevé

d'itérations est sans doute nécessaire. Le test de Raftery et Lewis (1992a) (voir aussi Raftery et Lewis (1992b)) est plus informatif. Il se base sur les quantiles de la statistique d'intérêt. L'idée est de construire une chaîne de Markov à deux états à partir d'un quantile (par exemple, 2,5% et 97,5%) qui permette d'estimer les probabilités de transition et ainsi d'estimer le nombre de simulations nécessaires pour approcher la stationnarité.

Finalement, les méthodes de Monte-Carlo présentées dans les sections précédentes ne sont valides que si les éléments de la chaîne de Markov sont *i.i.d*. Or, l'intuition suggère que les valeurs adjacentes d'une chaîne de Markov devraient être corrélées positivement. De manière plus générale, le fait que des autocorrélations d'ordre élevé puissent subsister est problématique si la taille de la chaîne de Markov n'est pas suffisamment grande. Dans cette perspective, il est nécessaire de vérifier cette propriété ou tout du moins d'éviter une corrélation trop élevée de la chaîne de Markov à partir de laquelle on déduit les quantités ou statistiques d'intérêt. Plusieurs procédures ont été suggérées dans la littérature. Nous mentionnons ici deux stratégies. La première repose sur un facteur correctif à appliquer à la dimension de la chaîne de Markov en présence d'un degré observé k d'autocorrélations(42). Une autre stratégie est de conserver seulement chaque $k^{\text{ième}}$ élément de la chaîne de Markov (après avoir supprimé les n_0 premiers éléments de la chaine de Markov). Cette technique est connue sous le nom de sous-échantillonnage, voir Schmeiser (1989), Raftery et Lewis (1992a) ou Raftery et Lewis (1992b).

Estimation de la densité marginale

Nous disposons d'une suite de vecteurs de paramètres $\{\theta^{(s)}\}_{s=n_{0+1}},\ldots,n$ où chaque $\theta^{(s)}$ est extrait de la distribution postérieure. À partir de cette suite nous pouvons estimer les moments postérieurs, les densités prédictives, et finalement la densité marginale de l'échantillon, $p(Y_T^*)$. Cette densité marginale, comme nous l'avons vu plus haut, permet de quantifier la capacité du modèle à expliquer l'échantillon à notre disposition et éventuellement de comparer différents modèles. Par exemple, Rabanal et Rubio Ramirez (2005) évaluent différentes spécifications des rigidités nominales sur les salaires et les prix dans le cadre d'un modèle DSGE, en comparant des densités marginales. Il existe de nombreuses méthodes pour estimer $p(Y_T^*)$. Dans cette partie nous présentons la méthode généralement utilisée pour les modèles DSGE.

L'estimateur par la moyenne harmonique est motivé par la propriété suivante de l'espérance postérieure :

$$\mathrm{E}\left[\frac{f(\theta)}{p_0(\theta)\,p(Y_T^*|\theta)}\right]=\int_\Theta \frac{f(\theta)\,p_1(\theta|Y_T^*)}{p_0(\theta)\,p(Y_T^*|\theta)}\,d\theta$$

où f est une fonction de densité quelconque et E est l'espérance postérieure. Le membre de droite de l'égalité, en utilisant la définition de la densité postérieure, s'écrit alternativement :

$$\int_\Theta \frac{f(\theta)}{p_0(\theta)\,p(Y_T^*|\theta)}\,\frac{p_0(\theta)\,p(Y_T^*|\theta)}{\int_\Theta p_0(\theta)\,p(Y_T^*|\theta)\,d\theta}\,d\theta$$

On obtient donc :

$$\mathrm{E}\left[\frac{f(\theta)}{p_0(\theta)\,p(Y_T^*|\theta)}\right]=\frac{\int_\Theta f(\theta)\,d\theta}{\int_\Theta p_0(\theta)\,p(Y_T^*|\theta)\,d\theta}$$

Puisque l'intégrale de f somme à un, nous obtenons finalement :

$$\mathrm{E}\left[\frac{f(\theta)}{p_0(\theta)\,p(Y_T^*|\theta)}\right]=\frac{1}{\int_\Theta p_0(\theta)\,p(Y_T^*|\theta)\,d\theta}$$

Ainsi, un estimateur de la densité marginale (l'intégrale du noyau postérieur qui apparaît au dénominateur du second membre), est l'inverse de l'espérance postérieure de $f(\theta)/K(\theta|Y_T^*)$ de la densité marginale :

$$\hat{p}(Y_T^*)=\frac{1}{n-n_0}\sum_{s=n_0+1}^{n}\frac{f(\theta^{(s)})}{K(\theta^{(s)}|Y_T^*)}$$

Ce résultat est valable pour toute densité f. Geweke (1992) propose d'utiliser une gaussienne tronquée centrée sur l'espérance postérieure. L'idée est d'accorder moins de poids, voire d'éliminer, les simulations éloignées du centre de la distribution postérieure. Cela permet de diminuer la variance de l'estimateur de la densité marginale.

Un DSGE pour le B du VAR

Dans cette partie, nous illustrons en quoi les modèles VAR et DSGE sont des outils complémentaires que l'on ne doit pas nécessairement chercher à opposer. Nous avons vu dans la section consacrée au modèle VAR que la spécification des croyances *a priori* sur la paramètrisation d'un VAR ne va pas de soi. En effet, dans la mesure où le contenu économique d'un modèle VAR est ténu, l'interprétation des paramètres du VAR est délicate, ce qui rend l'élicitation des *priors* ardue. Ingram et Whiteman (1994) proposent d'utiliser un modèle DSGE afin de construire le prior d'un modèle VAR. Ils montrent qu'en utilisant les restrictions définies par un modèle RBC pour définir le *prior* d'un modèle VAR, on peut produire avec ce dernier des prévisions comparables, en termes de précision, à celles que nous obtiendrions avec un *prior* Minnesota. Ce résultat est remarquable, car même si le modèle RBC canonique est mal spécifié dans de nombreuses directions, il impose des restrictions utiles pour améliorer les prévisions du VAR.

Plus récemment, Del Negro et Schorfheide (2004) ont repris cette idée sous une forme plus simple à mettre en oeuvre. Leur approche permet d'estimer simultanément les paramètres structurels du DSGE et les paramètres du modèle VAR. Nous présentons leur approche dans cette partie.

Les régressions mixtes

Dans la section consacrée au modèle VAR nous avons noté, au moins dans le cadre d'un modèle linéaire gaussien, une analogie entre les priors du paradigme bayésien et les contraintes linéaires sur les paramètres de l'approche classique.

Del Negro et Schorfheide utilisent cette analogie (voir Theil et Golberger (1961), Tiao et Zellner (1964) et Theil (1971)) pour mettre en œuvre le *prior* DSGE. Dans le modèle VAR, on peut définir un *prior* sur A en utilisant des observations artificielles, cohérentes avec nos croyances, et un *prior* diffus à la Jeffrey. Par exemple, si ces observations artificielles sont générées par un modèle DSGE, alors l'estimation sur la base de l'échantillon augmenté sera attirée vers la projection du DSGE dans l'espace des VAR.

Plus formellement supposons que nous disposions des observations artificielles $(\breve{Y}(\eta), \breve{Z}(\eta))$, où η est un vecteur de paramètres qui définit le processus générateur des données artificielles (*i.e.* les croyances *a priori*).

Comme l'échantillon artificiel est indépendant de Y_T^*, la vraisemblance de l'échantillon augmenté s'écrit de la façon suivante :

$$(25)\ p(\breve{Y}(\eta), Y_T^* | A,\Sigma) = p(\breve{Y}(\eta)|A,\Sigma)\times p(Y_T^* | A,\Sigma)$$

Le premier terme du membre de droite, si l'échantillon artificiel est de dimension $[\lambda T]$ où $\lambda \in R$, s'écrit :

$$(26)\ p(\breve{Y}(\eta)|A,\Sigma) \propto$$

$$|\Sigma|^{-\frac{[\lambda T]}{2}}\ e^{-\frac{1}{2}tr[\Sigma^{-1}(\breve{Y}'\breve{Y} - A'\breve{Z}'\breve{Y} - \breve{Y}'\breve{X}A + A'\breve{Z}'\breve{Z}A)]}$$

et, à la lumière de l'avant dernière équation, s'interprète comme un *prior* pour A et Σ. La croyance *a priori* est d'autant plus informative que l'échantillon artificiel est de grande taille. Quand λ, tend vers l'infini, le poids de la vraisemblance (le second terme sur le membre de droite de (25)) devient négligeable par rapport au *prior* (le premier terme sur le membre de droite de (25)). En complétant le *prior*, défini avec les données artificielles, par un *prior* diffus (ou plat) à la Jeffrey :

$$p_0(A,\Sigma) \propto |\Sigma|^{-\frac{m+1}{2}}$$

le *prior* est au final de type normal-Wishart, le *prior* conjugué dans un modèle linéaire gaussien. En particulier, A est *a priori* normalement distribué :

$$vecA|\Sigma \sim N(vec\,\tilde{A}(\eta), \Sigma \otimes (\breve{Z}'\breve{Z})^{-1})$$

où $\tilde{A}(\eta) = (\breve{Z}'\breve{Z})^{-1}\breve{Z}'\breve{Y}$, est l'estimateur des MCO (MV) des paramètres autorégressifs pour l'échantillon artifiel.

On voit immédiatement, en considérant la vraisemblance de l'échantillon augmenté (25), le *prior* diffus à la Jeffrey et les résultats obtenues dans la première partie, que la distribution a posteriori est du type normale-Wishart :

$$(27)\ A|\Sigma, Y_T^*, \breve{Y}(\eta) \sim MN_{k,m}(\hat{A}(\eta), \Sigma, (Z'Z + \breve{Z}'\breve{Z})^{-1})$$

$$\Sigma | Y_T^*, \breve{Y}(\eta) \sim iW_m(\hat{S}(\eta), \tilde{\nu})$$

où $\hat{A}(\eta)$ et $\hat{S}(\eta)$ sont respectivement les estimateurs du maximum de vraisemblance de A et $(T + [\lambda T])\Sigma$, pour l'échantillon augmenté des données artificielles, $\tilde{\nu} = [(1+\lambda)T] - k$. En intégrant la densité jointe postérieure par rapport à Σ, on montre que la distribution postérieure marginale de A est une distribution de Student matricielle, centrée en $\hat{A}(\eta)$.

Lorsque λ augmente, $\hat{A}(\eta)$ se rapproche de $\tilde{A}(\eta)$, en effet, on établit facilement que :

$$\hat{A}(\eta) = (\breve{Z}'\breve{Z} + Z'Z)^{-1}(\breve{Z}'\breve{Z}\tilde{A}(\eta) + Z'Z\hat{A})$$

Ainsi, lorsque le poids du *prior* augmente, la distribution postérieure du VAR se rapproche de la projection dans l'espace des VAR du modèle générateur des données cohérent avec nos croyances *a priori*.

Le modèle BVAR-DSGE

Del Negro et Schorfheide (2004) proposent, à la suite d' Ingram et Whiteman (1994), d'utiliser un modèle DSGE pour spécifier le *prior* d'un modèle VAR. Contrairement à ces derniers, Del Negro et Schorfheide utilisent les régressions mixtes décrites dans la section précédente, même si en pratique ils ne simulent pas des données.

Afin d'éviter que les résultats puissent varier, à cause des simulations, ils préfèrent remplacer les moments empiriques dans (26) par des moments théoriques calculés à partir d'une approximation de Taylor d'ordre 1 de la forme réduite (18) du modèle DSGE. Par exemple ils remplacent $\breve{Y}'\breve{Y}$ par la matrice de variance-covariance des endogènes observées, c'est-à-dire une sous matrice de $\Gamma_{yy}(\theta) = \mathrm{E}[(y_t - \mathrm{E}y_t)'(y_t - \mathrm{E}y_t)]$, multipliée par la taille de l'échantillon artificiel, $[\lambda T]$. Pour tout vecteur de paramètres structurels, $\theta \in \Theta$, la définition du *prior* du VAR est pratiquement immédiate, il suffit d'écrire la forme réduite du modèle DSGE et de calculer ses moments asymptotiques. Del Negro et Schorfheide ne se contentent pas d'estimer les paramètres du VAR, ils estiment simultanément les paramètres du modèle DSGE. Ils spécifient donc un *prior* joint sur les paramètres du modèle VAR et les paramètres structurels du modèle DSGE :

$$p_0(A,\Sigma,\theta|\lambda) = p_0(A,\Sigma|\theta,\lambda) \times p_0(\theta)$$

Le *prior* est conditionnel au paramètre λ, qui spécifie la taille de l'échantillon artificiel relativement à l'échantillon d'origine, c'est-à-dire la quantité relative d'information structurelle *a priori*. On peut alors utiliser l'algorithme de Metropolis-Hastings pour obtenir la distribution postérieure de θ (et indirectement de A et Σ) en utilisant la densité postérieure du modèle BVAR spécifiée par (27). Ici, la vraisemblance du modèle DSGE n'a pas à être calculée, ce qui simplifie considérablement l'estimation puisque le filtre de Kalman n'est plus nécessaire. Les paramètres du modèle DSGE sont identifiés grâce à la vraisemblance, plus exactement la densité postérieure, de son approximation VAR. Le modèle VAR joue ici en quelque sorte le même rôle qu'un modèle auxiliaire en inférence indirecte (voir Gouriéroux et Monfort, 1996).

L'estimation de θ (et donc de A et Σ) est conditionnelle aux choix de p, le nombre de retards dans le VAR, et λ, la quantité relative d'information structurelle *a priori* dans le VAR. Il convient de choisir un nombre de retards assez grand pour que le modèle VAR puisse être une approximation acceptable du modèle DSGE. En effet la forme réduite (18) approximée du modèle DSGE n'appartient pas à la famille des modèles VAR, il faudrait un nombre de retard infini pour approximer au mieux le modèle DSGE[43]. Del Negro et Schorfheide estiment un VAR décrivant l'inflation, le taux d'intérêt et le taux de croissance du produit. Ils affirment qu'un VAR(4) permet une approximation satisfaisante de leur modèle DSGE. Le choix de λ est plus délicat : en variant ce paramètre de zéro à l'infini, on passe d'un prior diffus (l'espérance postérieure de A est alors l'estimateur du MV) à un *prior* très informatif (l'espérance postérieur de A tend vers $\Gamma_{zz}(\theta)^{-1}\Gamma_{yz}(\theta)$, les contraintes DSGE sur les paramètres du modèle VAR). Del Negro et Schorfheide proposent d'estimer plusieurs modèles pour une grille de valeurs de λ. Ils choisissent alors le modèle, c'est-à-dire la valeur de λ, qui maximise la densité marginale. Ils sélectionnent le modèle dont la qualité d'ajustement est la meilleure.

Del Negro *et alii* (2007) utilisent le BVAR-DSGE pour estimer le modèle de Smets et Wouters (2002), ils obtiennent $\lambda = 0{,}75$. Ils montrent ainsi que les restrictions apportées par le modèle de Smets et Wouters sont utiles pour améliorer les performances du modèle VAR. Cette procédure est relativement compliquée à mettre en œuvre. Pour chaque valeur de λ il faut s'assurer de la convergence de l'algorithme de Metropolis- Hastings, afin d'estimer la densité marginale[44]. Plus haut nous avons noté l'analogie entre le choix d'un modèle dans une collection de modèles et l'estimation d'un paramètre dont les valeurs seraient discrètes. Une approche plus directe est d'associer une distribution *a priori* à λ puis d'estimer ce paramètre (avec les paramètres structurelles θ). Il faut alors définir un *prior* joint sur A,Σ,θ et λ :

$$p_0(\Sigma,\theta,\lambda) = p_0(A,\Sigma|\theta,\lambda) \times p_0(\theta) \times p_0(\lambda)$$

Adjemian et Darracq-Pariès (2007) estiment une version deux pays du modèle de Smets et Wouters, avec le modèle BVAR-DSGE, en posant un prior uniforme (entre 0 et 10) pour le paramètre λ. Ils obtiennent une distribution postérieure de λ centrée sur 2,5. Il n'est pas surprenant d'obtenir dans ce cas une valeur de λ largement supérieure. La version à deux pays du modèle de Smets et Wouters est estimée avec un VAR comprenant dix-huit variables observables, alors que Del Negro *et alii* (2007), pour la version à un pays, ne considèrent que sept variables. Avec dix-huit variables, les restrictions structurelles deviennent plus nécessaires, même si le modèle n'est pas mieux spécifié.

Usages et avantages du BVAR-DSGE

Del Negro et Schorfheide (2004) et surtout Del Negro *et alii* (2007) présentent le modèle BVAR-DSGE comme un outil d'évaluation de la qualité d'ajustement d'un modèle DSGE. Pour ces derniers la valeur de λ, le poids du *prior* structurel, donne une idée de l'intérêt empirique du modèle. Si les restrictions structurelles définies par le modèle DSGE sont pertinentes, alors la procédure sélectionne une valeur élevée de λ. Si le modèle apporte des informations totalement incohérentes avec les données alors la procèdure sélectionne une valeur proche de zéro. Malheureusement cette mesure n'a pas d'échelle et nous ne savons pas à partir de quelle valeur de λ on peut dire que le modèle apporte des informations pertinentes.

Un autre problème est que ce paramètre ne mesure pas la qualité d'ajustement du modèle DSGE ; il nous donne la quantité optimale, au sens du *fit* du modèle BVAR, d'information DSGE qu'il faut incorporer dans le *prior* du VAR. Del Negro *et alii* (2007) utilisent le BVAR-DSGE afin de dévoiler les éventuelles mauvaises, spécifications d'un modèle DSGE. Or le niveau optimal de λ ou la densité marginale, $p(Y_T^*)$, du modèle DSGE ne sauraient donner une idée précise des directions dans lesquelles le modèle est insatisfaisant puisque ces deux indicateurs donnent des informations trop agrégées. Les quatre auteurs recherchent les directions dans lesquelles le modèle DSGE est mal spécifié en comparant les fonctions de réponses (IRF) du modèle BVAR-DSGE avec celles du modèle DSGE. Ils identifient les chocs structurels dans le modèle BVAR-DSGE en se fondant sur le modèle DSGE (17). À partir de la forme réduite (18) il est possible de calculer l'impact instantanné de chaque choc structurel sur les variables observables :

$$Z\frac{\partial H_\theta}{\partial \varepsilon}$$

où Z est une matrice de sélection définie dans l'équation de mesure (19a). Del Negro et Schorfheide utilisent cette information pour indentifier les innovations structurelles dans le modèle BVAR (se reporter à Del Negro et Schorfheide, 2004 pour les détails). Même si le BVAR-DSGE est construit sur la base (au moins partiellement) d'une information structurelle provenant du DSGE, ce modèle est moins contraint que le modèle DSGE. Ainsi, l'observation d'une différence significative entre les IRFs du BVAR-DSGE et celles du modèle DSGE conduit Del Negro *et alii* (2007) à identifer les directions dans lesquelles le modèle DSGE est mal spécifié. Par exemple, les quatre auteurs observent que les réponses du produit, de la consommation et des heures face à un choc de préférence (sur la désutilité du travail) sont plus persistentes dans le BVAR-DSGE que dans le DSGE. Ils concluent alors que le modèle DSGE manque de mécanismes de propagation des chocs sur l'offre de travail. Cet exercice de comparaison entre BVAR-DSGE et DSGE peut être mis en oeuvre en utilisant des statistiques autres que des fonctions de réponse : décompositions de variance des variables observées, moments théoriques des variables observées,... La limite de l'exercice est que les conditions d'identification des chocs dans le BVAR sont directement dérivées du modèle DSGE. Si nous n'observons pas de grandes différences entre les IRFs du BVAR-DSGE et celles du DSGE c'est peut être parce que nous utilisons les mêmes conditions d'identifications. Ce problème ne se pose pas si on compare des statistiques qui ne reposent pas sur des conditions d'identification, par exemple si on compare des moments (variances, fonction d'autocorrélation,...).

L'avantage du modèle BVAR-DSGE est plus évident en termes de prévisions. Tout modèle est, par nature, mal spécifié dans une multitude de directions. Malgré cette limite intrinsèque, les modèles apportent souvent des informations utiles et pertinentes. L'expérience d'Ingram et Whiteman (1994) est, à cet égard, des plus éclairantes. Ils montrent que même le plus stylisé des modèles DSGE (le modèle de cycle réel canonique) est suffisament riche pour aider un BVAR à produire des prévisions plus précises et moins biaisées. Un modèle que personne ne voudrait utiliser pour produire des prévisions peut aider un modèle a-théorique (plus souple) à produire de meilleures prévisions. Cette idée pourrait être développée dans de nombreuses directions. Nous pourrions par exemple utiliser plusieurs modèles DSGE pour construire le *prior* d'un modèle VAR (ou de tout autre modèle a-théorique, par exemple un modèle à facteurs communs) et optimiser les parts de chaque modèle dans le prior du VAR.

Notes

(1) Voir les travaux de Smith (1993), Canova (1994), Dejong, Ingram, et Whiteman (1996), Geweke (1999), Dridi, Guay, et Renault (2007) et Bierens (2007).

(2) Le lecteur intéressé trouvera une introduction intéressante pour l'estimation des modèles non linéaires dans Andrieu, Doucet, et Robert (2004), Arulampalam *et alii* (2002), et Andrieu *et alii* (2004), ainsi que dans les contributions de Gordon *et alii* (1993) et Kitagawa (1996). Pour des applications en économie, voir Chopin et Pelgrin (2004), Fernández-Villaverde et Rubio-Ramírez (2005) et An et Schorfheide (2007).

(3) Le mot *croyance* suggère une dimension subjective de l'information *a priori*. Il convient néanmoins de signaler que parmi les économètres bayésiens il n'y a pas de consensus sur l'interprétation subjective ou objective des probabilités. Par exemple, l'approche bayésienne empirique utilise l'échantillon pour définir l'information *a priori* (voir par exemple le *prior* Minnesota de la section consacrée au modèle VAR.

(4) Pour une présentation des principales distributions utilisées dans cette littérature, le lecteur peut se reporter aux annexes de Zellner (1971). La distribution uniforme est un cas particulier de la bêta.

(5) Cette incertitude peut s'expliquer par une adéquation imparfaite entre le concept théorique et l'enquête microéconomique.

(6) Une condition nécessaire est que nous disposions d'une expression analytique de la vraisemblance.

(7) Nous supposons un instant qu'il n'y a qu'un paramètre dans le modèle.

(8) La comparaison est moins simple dans le cas de l'inférence classique.

(9) Se reporter à Zellner (Zellner, 1971 , chapitre 10), en particulier la première section pages 292 à 298. Le choix d'un modèle parmi une collection de modèles s'apparente à l'estimation d'un paramêtre dont la distribution est discrête. Il y a donc une analogie entre le choix d'un modèle et l'estimation ponctuelle de θ.

(10) Pour une description des méthodes de pondération *Bayesienne des modèles,* voir Koop (2003), chapitre 11.

(11) Dans le cas du modèle AR(1) l'information apportée par l'échantillon est résumée par la dernière observation y_T^*.

(12) Une expérience moins extrême serait de considérer des densités *a priori* plus générales. Supposons que notre *a priori* sur un paramètre μ soit caractérisé par une loi normale centrée en μ_0 et de variance σ_0^2. Nous pourrions évaluer la sensibilité des résultats à ce choix en reprenant l'estimation avec une densité *a priori* de Student : $p_0(\mu) \propto (vs+(x-\mu_0)^2)^{-\frac{v+1}{2}}$.

L'espérance *a priori* serait alors μ_0 mais la variance *a priori* serait $\frac{v}{v-2}s$ (pour v strictement supérieur à 2). En faisant varier le nombre de degré de liberté v on s'écarte ou se rapproche du prior gaussien.

(13) Cette propriété est indispensable pour représenter l'ignorance. Dans la littérature DSGE, la distribution inverse-gamma avec un moment d'ordre deux infini est souvent utilisée pour représenter le peu d'information dont nous disposons sur la variance des chocs structurels (voir par exemple Smets et Wouters, 2002). Cette distribution est informative dans le sens où, même si le moment d'ordre deux n'est pas défini, il est possible de comparer les probabilités qu'une variance soit supérieure ou inférieure à $c>0$.

(14) Si le logarithme de σ est uniformément distribué sur $]-\infty,\infty[$ alors le logarithme de σ^a (avec $a>0$) est aussi uniformément distribué sur $]-\infty,\infty[$ car $\log(\sigma^\alpha)=a\log(\sigma)$.

(15) Se reporter à Gouriéroux et Monfort (1989, chapitre 7).

(16) Notons c cette constante d'intégration, c'est-à-dire la constante telle que $\int c^{-1}K(\theta)d\theta=1$. Cette constante (voir les équations (3) et (4)) est une approximation de la densité marginale, $p(Y_T^*)$. Par définition de la densité d'une loi normale, on a :

$$c=K(\theta^*)(2\pi)^{\frac{q}{2}}|H(\theta^*)|^{\frac{1}{2}}$$

On dit que c est l'approximation de Laplace de la densité marginale. L'erreur d'approximation est d'ordre $O(T^{-1})$.

(17) Pour plus de détails, voir Carlin et Louis (2000), Poirier (1995), et Tierney et Kadane (1986).

(18) Se reporter, par exemple, à Kadiyala et Karlsson (1997) qui comparent différentes spécifications des croyances *a priori* et étudient les conséquences sur les prévisions.

(19) Nous pourrions choisir un *prior* conjugué, c'est-à-dire une densité *a priori* qui confrontée aux données *via* la vraisemblance induit une densité postérieure de la même forme. Les propriétés des densités gaussienne et Wishart, ainsi que l'équation (8), suggèrent la densité jointe *a priori* conjuguée suivante :

$$\begin{cases} A|\Sigma \sim MN_{k,m}(A_0,\Sigma,M_0^{-1}) \\ \Sigma \sim iW_m(S_0,v_0) \end{cases}$$

où A_0 est une matrice réelle de même dimension que A, Σ et M_0 sont des matrices symétriques définies postives respectivement de dimensions $m\times m$ et $p\times p$, S_0 est une matrice symétrique définie positive. On montre alors facilement que la densité postérieure est encore Normale-Wishart. Ce résultat est direct si on couple le prior non informatif de la section précédente et un pré-échantillon pour former le *prior* normal Wishart, voir Tiao et Zellner (1964) et la cinquième partie.

(20) La présence d'une racine unitaire ne ferait qu'accroître l'ordre de divergence, ce qui ne change pas qualitativement la conclusion.

(21) Voir Judge *et alii* (1985, pages 52-54) la section 3.2.1 intitulée " Exact Nonsample Information".

(22) Se reporter à Theil (1971, pages 670-673).

(23) Voir, par exemple, Litterman (1986) ou Kadiyala et Karlsson (1997). Pour d'autres priors on peut se reporter à Kadiyala et Karlsson.

(24) Nous devons aussi poser un *a priori* sur la matrice de variance-covariance de l'innovation du VAR, Σ. Litterman (1986) considère que celle-ci est diagonale et donnée (variance *a priori* nulle). Cela ne correspond pas à notre hypothèse de la section précédente, où nous avons supposé que cette matrice était pleine (égale à l'estimateur du maximum de vraisemblance). Nous pourrions, sans changer qualitativement les résultats, adopter une matrice diagonale dans la section précédente. Nous choisissons de poursuivre avec une matrice pleine, comme Phillips (1996), mais notre prior ne correspondra plus à des marches aléatoires indépendantes. En fait la motivation principale de Litterman était de justifier une estimation équation par équation, car à l'époque l'estimation d'un système était trop coûteuse numériquement, et ne reposait pas sur la croyance que les séries macroéconomiques sont réellement indépendantes.

(25) Voir Kim (1998).

(26) Quand Smets et Wouters établissent que leur modèle DSGE surpasse un modèle BVAR en terme de densité marginale de l'échantillon, on ne peut véritablement conclure à la bonne la qualité du DSGE puisque nous n'avons aucune idée des performances du modèle BVAR.

(27) La première catégorie correspond aux variables prédéterminées, les suivantes aux variables non prédéterminées.

(28) Plusieurs approches sont envisageables : quadrature, Monte Carlo, quasi Monte Carlo... voir Judd (1998).

(29) Par exemple, Smets et Wouters (2002, tableau 2) estiment la densité marginale de leur modèle DSGE à l'aide de l'approximation de Laplace et d'une méthode "exacte" (dans le sens où elle ne repose pas sur une approximation asymptotique, voir plus bas) basées sur des simulations. Avec l'approximation de Laplace ils obtiennent (en logarithme) -269,59 à comparer au -269,20 obtenu avec un exercice de Monte-Carlo. Ces deux évaluations sont très proches, on retrouve généralement cette proximité dès lors que l'échantillon est assez grand.

(30) En fait on peut montrer, dans certains cas, qu'il s'agit d'un estimateur du maximum de vraisemblance.

(31) Dans ce qui suit, nous omettons les méthodes d'échantillonnage de Gibbs. Cette méthode consiste à générer chaque paramètre conditionnellement à tous les autres paramètres. Il est donc nécessaire de pouvoir écrire toutes les distributions conditionnelles. C'est pourquoi cette méthode n'est généralement pas priviligiée pour l'estimation des modèles DSGE. Cependant, il est à noter que les algorithmes de Metropolis-Hasting et d'échantillonnage de Gibbs peuvent être combinés, on parle alors d'algorithme *Metroplis-Within-Gibbs*.

(32) Nous verrons par la suite qu'un algorithme à chaînes de Markov indépendantes peut s'interpréter comme un algorithme par fonction d'importance. Par ailleurs, les méthodes de Monte-Carlo à chaînes de Markov nécessitent de déterminer le noyau de transition de la chaîne de Markov, dont on sait seulement qu'il vérifie certaines propriétés d'ergodicité, etc. Le choix de la fonction définissant les changements d'état peut être assimilé, *toutes choses égales par ailleurs,* à celui de la fonction d'importance.

(33) Le noyau $P(\eta,S)$ spécifie la probabilité d'aller de η à S. Dans un cas favorable, θ est en S à l'itération suivante, nous pouvons envisager deux possibilités : (*i*) θ se déplace effectivement et rejoint la région S à l'itération suivante, (*ii*) θ ne se déplace pas mais η est déjà dans S. La densité associée au noyau est donc une densité continue - discrète. Tierney adopte la définition suivante :

$$P(\eta,dv) = p(\eta,v)dv+(1-r(\eta))\delta_\eta(dv)$$

où $p(\eta,v)\equiv p(v|\eta)$ est la densité associée à la transition de η à v, $r(\eta)=\int p(\eta,v)dv<1$, $1-r(\eta)$ s'interprète comme la probabilité de ne pas quitter la position $\theta=\eta$, $\delta_\eta(S)$ est une fonction indicatrice égale à un si et seulement si $\eta\in S$.

(34) Il ne s'agit pas à proprement parler de la condition de réversibilité, mais d'une implication de la propriété de réversibilité.

(35) (Suite de la note 32) Techniquement, il suffit de substituer la définition du noyau dans $\int_\Theta P(\eta,S)\pi(\eta)d\eta$ qui, si la chaîne est réversible, se réduit alors à $\pi(S)$.

(36) Le noyau de transition du *MH*, $Q(\eta,S)$, est défini de la même façon que *P* dans la section précédente et la note 32.

(37) Par exemple, nous pourrions avoir :

$$K(\eta|Y_T^*)q(\eta,v)>K(v|Y_T^*)q(v,\eta)$$

Dans ce cas, l'échantillonage à partir de *q* ne propose pas assez souvent des transitions de $\theta=v$ à $\theta=\eta$ ou trop souvent des mouvements de $\theta=\eta$ à $\theta=v$. L'algorithme de *MH* corrige cette erreur en n'acceptant pas systématiquement les propositions de *q*. En introduisant une probabilité d'acceptation de la transition proposée, α, élevée (faible) quand il s'agit de rejoindre une région où la densité *a posteriori* est élevée (faible), on peut rétablir l'équilibre et finalement vérifier la condition de réversibilité. Dans notre exemple, la probabilité d'acceptation de la transition de *v* à η devrait être la plus grande possible puisque *q* ne propose pas assez souvent cette transition, nous poserons donc $\alpha(v,\eta)=1$. À l'inverse la densité conditionnelle *q* propose trop de transitions de *v* vers η, la probabilité d'acceptation associée à cette proposition, $\alpha(v,\eta)$, doit donc être inférieure à 1. Pour équilibrer les deux transitions, elle doit être telle que :

$$K(\eta|Y_T^*)q(\eta,v)\alpha(\eta,v)=K(v|Y_T^*)q(v,\eta)\alpha(v,\eta)$$

soit, puisque $\alpha(v,\eta)=1$, de façon équivalente :

$$\alpha(\eta,v)=\frac{K(v|Y_T^*)}{K(\eta|Y_T^*)}\frac{q(v,\eta)}{q(\eta,v)}$$

On ne rejette donc pas systématiquement la transition proposée par *q*. En considérant l'exemple et en renversant l'inégalité, on comprend la règle donnée dans l'étape 4 de l'algorithme 2.

(38) Il n'existe pas une règle universelle. Un taux d'acceptation de l'ordre de 0,25-0,40 est généralement considéré comme approprié. Dans le même temps, il est important de noter que ce n'est pas tant le taux d'acceptation qui est crucial mais plutôt la garantie que la chaîne de Markov a effectivement convergé. Le taux d'acceptation peut néanmoins influencer le temps qu'il faudra à la chaîne de Markov pour rejoindre sa distribution invariante.

(39) Pour plus de détails sur la convergence vers la distribution stationnaire et l'hypothèse de stationnarité, voir Gelfand et Smith (1990), Roberts (1992) et Liu et al. (1992).

(40) Pour plus de détails, voir Raftery et Lewis (1996), Cowles et Carlin (1996), et Brooks et Roberts (1998).

(41) Plus formellement, notons $\hat{\phi}_{n_i}^{(i)}$ l'estimateur de $E[\phi(\theta)]$ obtenu à partir du vecteur initial θ^i lorsque les $n_1\equiv n-n_0$ dernières valeurs de la chaîne sont prises en compte. La variance *intra* d'une chaîne, obtenue à partir du vecteur initial θ^i, est définie par $s_i=\frac{1}{n_1-1}\sum_{s=n_0+1}^{n}[\phi(\theta^{(s,i)})-\hat{\phi}_{S_1}^{(i)}]^2$. La moyenne des variances *intra* est alors donnée par $W=\frac{1}{m}\sum_{i=1}^{m}s_i^2$, où *m* est le nombre de chaînes en parallèle ou de vecteurs initiaux. De la même manière, on peut montrer que la variance-inter est estimée par $B=\frac{n_1}{m-1}\sum_{i=1}^{m}(\hat{\phi}_{n_1}^{(i)}-\hat{\phi})^2$ où $\hat{\phi}$ est donnée par $\hat{g}=\frac{1}{m}\sum_{i=1}^{m}\hat{\phi}_{n_1}^{(i)}$. Un estimateur de la variance *a posteriori* de ϕ est alors défini comme $\frac{n_1-1}{n_1}W+\frac{1}{n_1}B$.

(42) Plus formellement, supposons que l'on dispose d'une chaîne de Markov $(\theta^s)_{s=1,\ldots,n}$ et que l'on subdivise cette chaîne en sous-ensembles,
$S_0=(\theta^s, s=1,\ldots,n_0)$, $S_1=(\theta^s, s=n_0+1,\ldots,n_0+n_a)$,
$S_2=(\theta^s, s=n_0+n_a+1,\ldots,n_0+n_a+n_b)$ et
$S_3=(\theta^s, s=n_0+n_a+1,\ldots,n_0+n_a+n_b+n_c)$
On choisit généralement, $n_a=0{,}1\,n_1$, $n_b=0{,}5\,n_1$ et $n_c=0{,}4\,n_1$, où $n_1\equiv n-n_0$. Le test de Geweke revient à déterminer la variance *a posteriori* de ϕ, $\hat{s}_1$ et $\hat{s}_3$, pour les sous-ensembles S_1 et S_3 et à évaluer $\hat{\phi}_{S_1}$ et $\hat{\phi}_{S_3}$. La statistique de test est alors

définie par $\frac{\hat{\phi}_{S_1} - \hat{\phi}_{S_3}}{\frac{\hat{s}_1}{\sqrt{n_a}} + \frac{\hat{s}_3}{\sqrt{n_c}}}$.

(42) L'intuition repose sur un théorème fondamental de l'analyse des séries temporelles qui nous indique que si les $\theta^{(s)}$ sont échantillonnées à partir d'un processus stationnaire et corrélé, les réalisations des tirages (qui sont donc corrélés) fournissent encore une information non biaisée de la distribution si la taille de l'échantillon est suffisamment grande.

(43) Par exemple, Campbell (1994) établit, en écrivant analytiquement la forme réduite du modèle RBC linéarisé, que le produit par tête est un processus ARMA(2,1). Ce modèle prédit donc que le produit par tête est un AR(1) que l'on pourrait approximer avec un AR(*p*) pour un nombre de retards, *p*, assez grand.

(44) On pourrait se contenter de l'estimation du mode postérieur et d'une approximation de Laplace, mais cette possibilité n'est pas évoquée par Del Negro et Schorfheide.

Bibliographie

Adjemian S. et Darracq-Pariès M. (2007). «Assessing the International Spillovers Between the US and Euro Area: Evidence from a two country DSGE-VAR», *miméo*, Cepremap.

An S. et Schorfheide F.(2007). «Bayesian Analysis of DSGE Models», *Econometric Reviews*, à paraître.

Andrieu C., Doucet A. et Robert P. (2004). «Computational Advances for and from Bayesian Analysis», *Statistical Science*, 19(1), pp.118–127.

Andrieu C., Doucet D., Singh S. et Tadic V. (2004). «Particle Methods for Change Detection, System Identification,and Control», *IEEE Transactions on Signal Processing*, 92(3), pp. 423–438.

Arulampalam S., Clapp T., Gordon N. et Maskell S. (2002). «Tutorial on Particle Filters», *IEEE Transactions on Signal Processing*, 50(2), pp. 174–188.

Bernanke B. (1986). «Alternative Explanations of the Money-Income Correlation», *Carnegie Rochester Conference Series on Public Policy*, 25(10), pp. 49–99.

Bierens H. J. (2007). «Econometric Analysis of Linearized Singular Dynamic Stochastic General Equilibrium Models», *Journal of Econometrics*, 136(2), pp.595–627.

Blanchard O. et Quah D. (1986). «The Dynamic Effects of Aggregate Demand and Supply Disturbances»,*The American Economic Review,* 79, pp. 655–673.

Brooks S. et Roberts G. (1998). «Assessing Convergence of Markov Chain Monte Carlo Algorithms», *Statistics and Computing*, 8, pp.319–335.

Campbell J. Y. (1994). «Inspecting the Mechanism: An Analytical Approach to the stochastic Growth Model», *Journal of Monetary Economics*, 33, pp.463–508.

Canova F. (1994). «Statistical Inference in Calibrated Models», *Journal of Applied Econometrics*, 9, pp.123–144.

Carlin B. et Louis T. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hill.

Casella G. et Robert C. (2004). *Monte Carlo Statistical Methods.* Springer.

Chopin N. et Pelgrin F. (2004). «Bayesian Inference and State Number Determination for Hidden Markov Models: An Application to the Information Content of the Yield Curve about Inflation», *Journal of Econometrics*,123(2), pp. 327–344.

Christiano L., Eichenbaum M. et Charles E. (2003). «Nominal Rigidities and the Dynamic Effects of a Schock to Monetary Policy», *Journal of Political Economy,* 113, pp. 1–45.

Cowles M. et Carlin B. (1996). «Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Study», *Journal of the American Statistical Association*, 91, pp. 883–904.

Dejong D., Ingram B. F. et Whiteman C. H. (1996). «A Bayesian Approach to Calibration», *Journal of Business and Economic Statistics*, 14(1), pp. 1–9.

Del Negro M. et Schorfheide F. (2004). «Priors from General Equilibrium Models for VARs», *International Economic Review*, 45(2), pp. 643–673.

Del Negro M., Schorfheide F., Smets F. et Wouters R. (2007). «On the Fit and Forecasting Performance of New Keynesian Models», *Journal of Business and Economic Statistics*, p. forthcoming.

Dridi R., Guay A. et Renault E. (2007). «Indirect Inference and Calibration of Dynamic Stochastic GeneralEquilibrium Models», *Journal of Econometrics*, 136(2), pp. 397–430.

Fernández-Villaverde J. et Rubio-Ramírez J. F. (2005). «Estimating Dynamic Equilibrium Economies : Linear versus NonLinear Likelihood», *Journal of Applied Econometrics*, 20(7), pp. 891–910.

Fernández-Villaverde J. et Rubio-Ramírez J. F. (2001). «Comparing Dynamic Equilibrium Economies toData»,*Working Paper* 2001-23, *Federal Reserve Bank of Atlanta.*

Gelfand A. et Smith A. (1990). «Sampling Based Approaches to Calculating Marginal Densities», *Journal ofthe American Statistical Association*, 85, pp. 398–409.

Gelman A. et Rubin D. B. (1992). «Inference from Iterative Simulations Using Multiple Sequences», *Statistical Science*, 7(4), pp. 457–472.

Geweke J. (1992). «Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments», *dans , édité par , Oxford University Press*, pp. 169–193.

Geweke J. (1999). «Using Simulation Methods for Bayesian Econometric Models : Inference, Development and Communication», *Econometric Reviews*, 18(1), pp. 1–126.

Gordon N., Salmond D. et Smith A. (1993). «Novel Approach to NonLinear and Non-Gaussian Bayesian State Estimation», *IEEE Transactions on Signal Processing,* 40(2), pp. 107–113.

Gouriéroux C. et Monfort A. (1989). *Statistique et Modèles Économétriques*, vol. 1 - Notions générales, Estimation, Prévisions, Algorithmes. Economica.

Gouriéroux C. et Monfort A. (1996). *Simulation Based Econometric Methods*. Oxford University Press.

Harvey A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

Ingram B. F. et Whiteman C. H. (1994). «Supplanting the Minnesota Prior. Forecasting Macroeconomic time series using real business cycle model.», *Journal of Monetary Economics*, 34, pp. 497–510.

Jeffrey H. (1961). *The Theory of Probability*. Clarendon Press.

Judd K. L. (1998). *Numerical Methods in Economics*. MIT.

Judge G. G., Griffiths W., Hill R. C., Lütkpohl H. et Lee T.-C. (1985). *The Theory and Practice of Econometrics*. JohnWiley & Sons.

Kadiyala K. R. et Karlsson S. (1997). «Numerical Methods for Estimation and Inference in Bayesian VARModels », *Journal of Applied Econometrics*, 12(2), pp. 99–132.

Kim J.-Y. (1998). «Large Sample Properties of Posterior Densities, Bayesian Information Criterion and the Likelihood Principle in Nonstationary Time Series Models», *Econometrica*, 66(2), pp. 359–380.

Kitagawa G. (1996). «Monte Carlo Filter and Smoother for Non-Gaussian NonLinear State Space Models», *Journal of Computational and Graphical Statistics*, 5(1), pp. 1–25.

Koop G. (2003). *Bayesian Econometrics*. John Wiley & Sons.

Kydland F. et Prescott E. (1982). «Time to Build and Aggregate Fluctuations», *Econometrica*, 50, pp. 1345–1370.

Litterman R. B. (1986). «Forecasting with Bayesian Vector Autoregressions – five years of experience», *Journal of Business & Economic Statistics*, 4(1), pp. 25–38.

Liu C., Liu J. et Rubin D. B. (1992). «A Varational Control Variable for Assessing the Convergence of the Gibbs Sampler», *Proceedings of the American Statistical Association*, pp. 74–78.

Phillips P. C. (1991a). «Bayesian Routes and Unit Roots: De rebus prioribus semper est disputandum», *Journal of Applied Econometrics*, 6(4),pp. 435–73.

Phillips P. C. (1991b). «To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends», *Journal of Applied Econometrics*, 6(4), pp. 333–64.

Phillips P. C. B. (1996). «Econometric Model Determination», *Econometrica*, 64(4), pp. 763–812.

Poirier D. (1995). *Intermediate Statistics and Econometrics : A Comparative Approach. Cambridge*, The MIT Press.

Rabanal P. et Rubio Ramirez J. F. (2005). «Comparing New Keynesian Models of the Business Cycle: A Bayesian approach», *Journal of Monetary Economics*, 6, pp. 1151–1166.

Raftery A. et Lewis S. (1992a). «How Many Iterations in the Gibbs Sampler ?», *dans Bayesian Statistics,* édité par J.M. Bernando, J.O. Berger, A.P. David et A.F.M. Smith, Oxford University Press, pp. 763–773.

Raftery A. et Lewis S. (1992b). «The Number of Iterations, Convergence Diagnostics and Generic Metropolis Algorithms», *Document de Travail*, Department of Statistics, University of Washington.

Raftery A. et Lewis S. (1996). «Implementing MCMC», *dans Markov Chain Monte Carlo in Practice,* édité par W.R. Gilks, S.T.Richardson et D.J. Spiegelhalter, Chapman & Hall, pp. 115–130.

Robert C. (2006). *Le Choix Bayesien*. Springer.

Roberts G. (1992). «Convergence Diagnostics of the Gibbs Sampler», *dans Bayesian Statistics,* édité par J.M.Bernando, J.O. Berger, A.P. David et A.F.M. Smith, Oxford University Press, pp. 775–782.

Rotemberg J. et Woodford M. (1997). «An Optimization-Based Econometric Framework for the Evaluation of Monetary Policy», *NBER Macroeconomics Annual*, 12, pp. 297–346.

Schmeiser B. (1989). «Simulation Experiments», *Working Paper* SMS 89-23, Purdue University.

Sims C. (1980). «Macroeconomics and Reality», *Econometrica*, 48(1), pp.1–48.

Sims C. (1986). «Are Forecasting Models Usable for Policy Analysis», *Federal Reserve Bank of Minneapolis Quarterly Review*, 10(1), pp. 2–16.

Sims C. (1991). «Comment on 'To Criticize the Critics,' by Peter C.B. Phillips», *Journal of Applied Econometrics*, 6(4), pp. 423–34.

Sims C. (2003). «Probability Models for Monetary Policy Decisions», *miméo*, Princeton University.

Sims C. A. et Uhlig H. (1991). «Understanding Unit Rooters : a Helicopter Tour», *Econometrica*, 59(6), pp. 1591–1599.

Smets F. et Wouters R. (2002). «An Estimated Stochastic Dynamic General Equilibrium Model of the Euro Area»,*Working Paper* Series 171, European Central Bank.

Smith A. (1993). «Estimating NonLinear Time-Series Models Using Simulated Vector Autoregressions», *Journal of Applied Econometrics*, 8, pp. 63–84.

Theil H. (1971). *Principles of Econometrics*. John Wiley & Sons.

Theil H. et GolbergerA. S. (1961). «On Pure and Mixed Statistical Estimation in Economics», *International Economic Review*, 2(1), pp. 65–78.

Tiao G. C. et Zellner A. (1964). «Bayes Theorem and the Use of Prior Knowledge in Regression Analysis»,*Biometrika*, 51(162), pp. 219–230.

Tierney L. (1994). «Markov Chains for Exploring Posterior Distributions», *The Annals of Statistics*, 22(4), pp. 1701–1762.

Tierney L. et Kadane J. B. (1986). «Accurate Approximations for Posterior Moments and Marginal Density», *Journal of the American Statistical Association*, 81(393), pp. 82–86.

Tierney L., Kass R. et Kadane J. (1989). «Fully Exponential Laplace Approximations to Expectations and Variances of Non Positive Functions», *Journal of the American Statistical Association*, 84, pp.710–716.

Woodford M. (2003). *Interest and Prices*. Princeton university press.

Yu B. et Mykland P. (1998). «Looking at Markov Samplers Through Cusum Path Plots: A Simple Diagnostic Idea», *Statistics and Computing*, 8(3), pp. 275–286.

Zellner A. (1971). *An Introduction to Bayesian Inference in Econometrics*. John Wiley & Sons.

Annexe : densités pour le modèle BVAR

Distribution normale matricielle

Définition 4. La matrice $p \times q$ aléatoire $\mathbf{X}$ est distribuée conformément à une loi normale matricielle :

$$\mathbf{X} \sim MN_{p,q}(\mathbf{M}, \mathbf{Q}, \mathbf{P})$$

où $\mathbf{M}$ est une matrice $p \times q$, $\mathbf{Q}$ et $\mathbf{P}$ sont respectivement des matrices $q \times q$ et $p \times p$ symétriques et définies positives, si et seulement si $vec\,(\mathbf{X})$ est distribué comme une v.a. normale multivariée :

$$vec(\mathbf{X}) \sim N_{pq}(vec(\mathbf{M}), \mathbf{Q} \otimes \mathbf{P})$$

Ainsi, la fonction de densité associée à $\mathbf{X}$ est donnée par :

$$f_{MN_{p,q}}(X; \mathbf{M}, \mathbf{P}, \mathbf{Q}) = (2\pi)^{-\frac{pq}{2}} |\mathbf{Q}|^{-\frac{p}{2}} |\mathbf{P}|^{-\frac{q}{2}} e^{-\frac{1}{2} tr\{\mathbf{Q}^{-1}(X-M)' P^{-1}(X-M)\}}$$

Distributions deWishart

La loi deWishart est une version multivariée de la loi du χ^2. Soit $\{X_i\}_{i=1}^{v}$ une suite de variables aléatoires gaussiennes indépendantes et identiquement distribuées $N(0,Q)$, avec Q une matrice symétrique définie positive $q \times q$. Par définition $Y = \sum_{i=1}^{v} X_i X_i'$ est distribué selon une loi de Wishart. Les définitions suivantes caractérisent cette loi et la densité de l'inverse d'une v.a. de Wishart.

Définition 5. *La matrice aléatoire, de dimension $q \times q$, symétrique et semi définie positive $\mathbf{Y}$ est distribuée selon une loi de Wishart, $\mathbf{Y} \sim W_q(\mathbf{Q}, v)$, si et seulement si sa densité est donnée par* :

$$f(Y; \mathbf{Q}, v) = \frac{|\mathbf{Q}|^{-\frac{v}{2}} |Y|^{\frac{v-q-1}{2}}}{2^{\frac{vq}{2}} \pi^{\frac{q(q-1)}{4}} \Pi_{i=1}^{q} \Gamma\left(\frac{v+1-i}{2}\right)} e^{-\frac{1}{2} tr\{Y \mathbf{Q}^{-1}\}}$$

pour Q une matrice symétrique semie définie positive, et $v \leq q$ le degré de liberté.

Définition 6. Une matrice aléatoire, de dimension $q \times q$, $\mathbf{X}$ est distribuée selon une loi inverse Wishart,

$$\mathbf{X} \sim iW_q(\mathbf{Q}, v)$$

si et seulement si : $\mathbf{X}^{-1} \sim W_q(\mathbf{Q}^{-1}, v)$

Ainsi la fonction de densité associée à $\mathbf{X}$ est définie par :

$$f_{iW_q}(X; \mathbf{Q}, v) = \frac{|\mathbf{Q}|^{\frac{v}{2}} |X|^{-\frac{v+q+1}{2}}}{(2\pi)^{\frac{vq}{2}} \pi^{\frac{q(q-1)}{4}} \Pi_{i=1}^{q} \Gamma\left(\frac{v+1-i}{2}\right)} e^{-\frac{1}{2} tr\{X^{-1} \mathbf{Q}\}}$$