STATISTIQUE BAYÉSIENNE

Dating and forecasting turning point by Bayesian clustering with dynamic structure

Romain Lesauvage et Alain Quartier-la-Tente

1 Introduction

Dans son ouvrage *Les vagues longues de la conjoncture*, Nikolai Kondratiev mettait en évidence l'existence de cycles économiques formés de deux périodes, une phase ascendante puis une phase descendante. Bien que contestée et complétée par la suite par d'autres analyses, cette découverte des cycles économiques a très vite mené les chercheurs a essayé de savoir dans quelle phase l'économie se trouvait et, de fait, savoir déterminer et prévoir le moment le cycle s'inverse est un enjeu majeur. C'est sur ce sujet que nous allons travaillé ici, à partir de l'article intitulé "Dating and forecasting turning points by Bayesian clustering with dynamic structure: a suggestion with an application to Austrian data. Journal of Applied Econometrics" Sylvia Kaufmann (2010).

Ce papier poursuit un premier travail introduit par Frühwirth-Schnatter S, Kaufmann S. (2008). L'idée est de travail sur un ensemble de plusieurs séries temporelles que l'on va essayer de regrouper. Nous chercherons à les regrouper au mieux afin d'étudier ensemble des séries qui possèdent la même dynamique. Nous travaillerons alors sur deux périodes, selon que l'on est inférieur ou supérieur à la moyenne de croissance et on cherchera à déterminer le point de retournement dans le cycle.

Dans ce travail, nous commencerons par détailler les spécifications théoriques du modèle et l'estimation bayésienne associée. Nous tenterons ensuite d'appliquer cette méthode à la manière de ce qui a été fait par Sylvia Kaufmann (2010) mais sur des données françaises.

2 Cadre théorique

2.1 Spécification du modèle

Dans la suite, nous travaillerons avec des séries centrées, réduites et corrigées des variations saisonnières. Notons y_{it} la série correspondant à un taux de croissance, avec $i \in \{1, \dots, N\}$ représentant les différentes séries à notre disposition et $t \in \{1, \dots, T\}$ représentant les différentes périodes d'études. Nous supposerons que nos séries suivent des processus autorégressifs, c'est-à-dire que l'on a

$$y_{it} = \mu^i_{I_{it}} + \phi^i_1 y_{i,t-1} + \cdots + \phi_p y_{i,t-p} + \epsilon_{it}$$

avec $\epsilon_{it} \sim \mathcal{N}\big(0, \frac{\sigma}{\lambda_i^2}\big)$. L'indice I_{it} précisera à quel moment du cycle nous nous trouverons, ainsi on aura $I_{it} \in \{1, 2\}$.

Théoriquement, les séries étudiées sont supposées être indépendantes les unes des autres. En pratique, cependant, il existe des séries qui suivent des évolutions proches et il peut donc être utile d'exploiter cette information pour notre estimation. Nous allons donc chercher à créer des groupes de séries de manière intelligente. Dans ce but, nous introduisons la variable $S_i \in \{1, \dots, K\}$ qui indique dans quel groupe est classée la variable i. Dans la suite, nous supposons que toutes les séries d'un groupe vons avoir des paramètres distribués selon une même loi mais différente selon la position dans le cycle :

$$\mu^i_{I_{S_i,t}} \sim \begin{cases} \mathcal{N}(\mu^k_1, q^k_1) \text{ si } S_i = k \text{ et } I_{kt} = 1 \\ \mathcal{N}(\mu^k_2, q^k_2) \text{ si } S_i = k \text{ et } I_{kt} = 2 \end{cases}$$

$$(\phi^i_1, \ldots, \phi^i_p) \sim \mathcal{N}(\phi^k, Q^k_\phi) \text{ si } S_i = k$$

Pour la suite, nous considérons que la période 1 ($I_{kt} = 1$) comme celle où l'on se trouve la croissance est supérieure à la moyenne (en haut du cycle) et la période 2 comme celle où la croissance est inférieure à la moyenne (en bas de cycle). Nous ferons également l'hypothèse que I_{kt} suit un processus de Markov, c'est-à-dire $\mathbb{P}(I_{k,t} = j | I_{k,t-1} = j) = \xi^k_{j,l}$ avec $j, l \in \{1, 2\}$. Par propriété sur les chaînes de Markov, on sait donc que $\forall j \in \{1, 2\}, \sum_{l=1}^{2} \xi^k_{j,l} = 1$. Cette spécification permet de repérer en temps réel les points où le cycle change de période.

Nous supposons que la variable S_i suit un modèle logit multinomial, c'est-à-dire $\mathbb{P}(S_i = k | \gamma_1, \ldots, \gamma_K, \gamma_{z,1}, \ldots, \gamma_{z,K-1}) = \frac{\exp(\gamma_k + Z_k \gamma_{z,k})}{1 + \sum_{l=1}^{K-1} \exp(\gamma_l + Z_l \gamma_{z,l})}$ où l'on prendra le groupe K comme référence ($\gamma_K = \gamma_{z,K} = 0$), Z_i est un vecteur permettant de repérer l'appartenance à un groupe et les paramètres γ sont inconnus mais spécifiques aux groupes.

Enfin, il reste à ajouter une spécification pour en compte la dynamique du cycle. Pour cela, nous supposons avoir deux groupes spécifiques pour la suite de l'analyse et $K - 2$ groupes indépendants. Les séries telles que $S_i = 2$ seront les séries qui mènent le cycle (indicateurs avancés), tandis que les séries pour lesquelles $S_i = 1$ seront les séries qui coïncident avec le cycle (tels que le PIB). Nous définissons donc une nouvelle variable I^* de la façon suivante:

$$\begin{cases} I^* = 1 := (I_{1t} = 1, I_{2t} = 1) \\ I^* = 2 := (I_{1t} = 1, I_{2t} = 2) \\ I^* = 3 := (I_{1t} = 2, I_{2t} = 1) \\ I^* = 4 := (I_{1t} = 2, I_{2t} = 2) \end{cases}$$

En cas d'incertitude sur la structure dynamique des nos séries, c'est-à-dire sur l'identification des groupes 1 et 2, on peut introduire un paramètre de structure ρ^*, caractérisant la structure entre les groupes. Il sera donc la réalisation d'une des $K(K-1)$ permutations de $\{1, 2, 0_{K-2}\}$. On considéra que $\rho^* = 1$ pour le groupe des séries coïncidentes, $\rho^* = 2$ pour celles qui mènent le cycle, et toute autre valeur renverra aux séries aux comportements autonomes.

2.2 Estimation bayésienne

2.2.1 Notations

Dans la suite, notons $\forall i \in \{1, \ldots, N\}, y^t_i = \{y_{i,t}, y_{i,t-1}, \ldots, y_{i,1}\}, Y_t = \{y_{1,t}, \ldots, y_{N,t}\}$ et $Y^t = \{Y_t, \ldots, Y_1\}$. De même, notons $S^N = \{S_1, \ldots, S_N\}$ l'ensemble des indicatrices d'appartenances aux groupes, $I^T = \{I^T_1, \ldots, I^T_K\}$ où $I^T_k = \{I_{k,T}, \ldots, I_{k,1}\}$ les indicatrices d'états. Enfin, nous notons $\lambda^N = \{\lambda_1, \ldots, \lambda_N\}$ qui regroupe les poids qui seront utilisés.

Par commodité, nous noterons également $\theta = (\mu^1_1, \mu^1_2, \ldots, \mu^K_1, \mu^K_2, \phi^1, \ldots, \phi^K, Q^1, \ldots, Q^K, \sigma^2, \xi^*, \xi^{\rho^*(k)=0}, \gamma, \gamma_z)$ où $Q^k = \begin{pmatrix} q^k_1 & 0 & 0 \\ 0 & q^k_2 & 0 \\ 0 & 0 & Q^k_\phi \end{pmatrix}$ représente l'hétérogénéité intra-groupe, $\xi^{\rho^*(k)=0} = \{\xi^k_{11}, \xi^k_{12}, \xi^k_{21}, \xi^k_{22}\}$, $\gamma = (\gamma_1, \ldots, \gamma_{K-1})$ et $\gamma_z = (\gamma_{z,1}, \ldots, \gamma_{z,K-1})$.

Nous pouvons passons à l'estimation par MCMC pour obtenir une posteriori sur le paramètre $\psi = (\theta, S^N, I^T, \lambda^N, \rho^*)$.

2.2.2 Estimation par MCMC

Nous partons de $\pi(\psi|Y^T) \propto L(Y^T|\psi)\pi(\psi)$ où $L(Y^T|\psi)$ est la vraisemblance, que l'on peut réécrire. En effet, $L(Y^T|\psi) = \prod_{t=p+1}^{T}\prod_{i=1}^{N} f(y_{i,t}|y_i^{t-1}, \mu^{S_i}_{I_{S_i,t}}, \phi^{S_i}, Q^{S_i}, \lambda_i, \sigma^2)$, avec

$$f(y_{i,t}|y_i^{t-1}, \mu^{S_i}_{I_{S_i,t}}, \phi^{S_i}, Q^{S_i}, \lambda_i, \sigma^2) = \frac{1}{\sqrt{2\pi\nu^{S_i}_{it}}} \exp\Big(-\frac{1}{2\nu^{S_i}_{it}}\big(y_{it} - \mu^{S_i}_{I_{S_i,t}} - \sum_{j=1}^{p}\phi^{S_i}_j y_{i,t-1}\big)^2\Big)$$

Avec $y_{it} = X^{S_i}_{it}\beta^{S_i} + \epsilon^*_{it}$, où $\epsilon^*_{it} \sim N(0, \nu^{S_i}_{it})$, $X^{S_i}_{it} = (D^{I(S_i)}_{1t}, D^{I(S_i)}_{2t}, y_{i,t-1}, \ldots, y_{i,t-p})$, $D^{I(S_i)}_{jt} = \mathbb{1}(I_{S_i,t} = j)$, $\beta^{S_i} = (\mu^{S_i}_1, \mu^{S_i}_2, \phi^{S_i})$., et $\nu^{S_i}_{it} = X^{S_i}_{it} Q^{S_i} X^{S'_i}_{it} + \frac{\sigma^2}{\lambda_i}$.

Nous supposerons dans la suite que la structure ρ^* est connue (déterminée par d'autres méthodes), nous pouvons alors réécrire la distribution a priori sur ϕ ainsi :

$$\pi(\phi) = \pi(I^{*T}|\rho^*, \xi^*) \prod_{\rho^*(k)=0} \pi(I^T_k|\rho^*, \xi^*)\pi(S^N|\gamma, \gamma_z, Z^N)\pi(\lambda^N)\pi(\theta)$$

où les densités $\pi(I^{*T}|\rho^*, \xi^*)$,$\pi(I^T_k|\rho^*, \xi^*)$ et $\pi(S^N|\gamma, \gamma_z, Z^N)$ sont connues. Nous supposerons que les poids (λ^N) sont indépendants et suivent une loi de Gamma. Il reste à spécifier l'a priori sur $\pi(\theta)$: pour cela, nous décomposons θ en blocs sur lesquels nous faisons des a priori standards.

Nous cherchons ensuite à générer à partir de la loi a posteriori $\pi(\phi|Y^T)$, en se basant sur le procédé détaillé dans Frühwirth-Schnatter S, Kaufmann S. (2008). Nous pouvons raisonner en 4 étapes différentes :

1. $\pi(S^N|Y^T, I^T, \rho^*, \lambda^N, \theta)$
2. $\pi(I^T|Y^T, S^N, \rho^*, \lambda^N, \theta)$
3. $\pi(\lambda^N|Y^T, I^T, S^N, \theta)$
4. $\pi(\theta|Y^T, S^N, I^T)$

Dans la première étape, nous générerons les indicatrices de groupes pour chaque série puis, à partir de ρ^*, nous obtenons une réalisation pour l'indicatrice d'état de l'étape 2. Dans l'étape 4, toutes les distributions a posteriori sont conjugées aux a priori, sauf pour γ et γ_z, nous utiliserons alors l'aglorithme de Metropolis-Hastings pour générer selon leur loi.

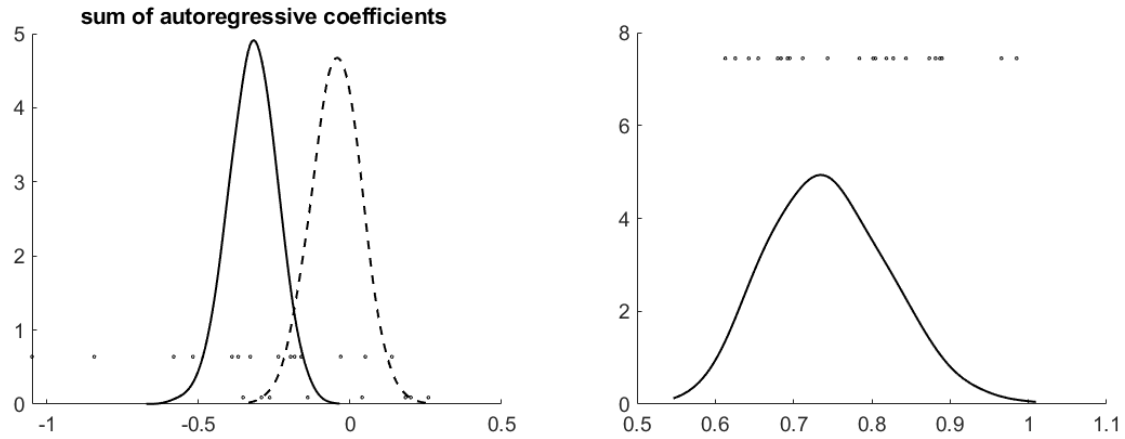
3 Application aux données françaises

L'article Sylvia Kaufmann (2010) a utilisé la spécification théorique présentée dans la partie précédente sur des données autrichiennes. Nous avons décidé dans le cadre de ce travail d'appliquer cela sur des données françaises. Nous avons donc récupéré un certain nombre de données liées au PIB ou aux enquêtes conjonctures [1], dont le détail est expliqué dans le tableau 1.

Nous avons retenu ici $K = 2$ groupes pour les séries : celles qui coïncident avec le cycle ($S_i = 1$) et celles qui le mène ($S_i = 2$). Pour l'analyse, nous nous limitons aux hypothèses suivantes :

1. La structure (ρ^*) est connue en avance : nous avons réparti chaque série dans un des groupes.
2. La classification est totale : on suppose qu'il n'y a pas de groupes de séries indépendantes.
3. L'*a priori* sur S_i est de type logit.

1. Nous avons décidé de ne retenir que les soldes d'opinion utilisés pour construire l'indicateur de retournement France publié par l'Insee (voir section 4.2) en supposant que, pour sa construction, les soldes les plus pertinents pour détecter les points de retournement avaient été retenus par l'Insee. Nous avons également retenu deux indicateurs synthétiques publiés par l'Insee : le climat des affaires France, qui retrace le cycle conjoncturel, et le climat de l'emploi France, qui retrace le cycle conjoncturel spécifique de l'emploi salarié.



(a) Somme des coefficients de l'AR.

(b) $\sigma^2, \sigma^2/\lambda_i$.

FIGURE 1 – Hétérogénéité des séries - groupe 1 (trait plein) et 2 (pointillés).

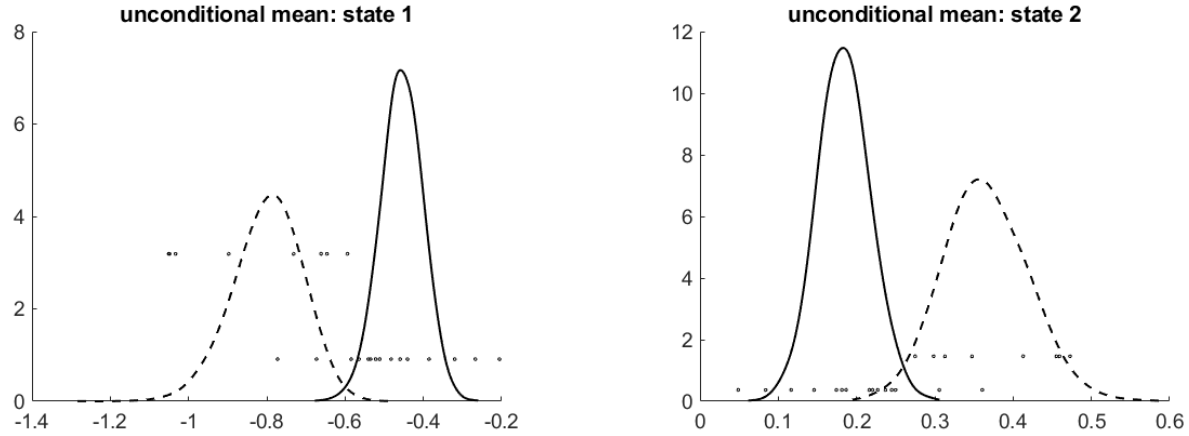


FIGURE 2 – Moyenne inconditionnelle dans les différents états

Nous avons adapté les programmes de Sylvia Kaufmann (2010) à nos données et nous analysons dans ce rapport les résultats obtenus.

Le modèle étudié ici se caractérise entre autres par la prise en compte de la spécification de l'hétérogénéité entre séries. Ceci est illustré par le graphique 1. La figure 1a représente la moyenne de la somme des paramètres autorégressifs. Dans l'article de Sylvia Kaufmann (2010), l'importance de la prise en compte de l'hétérogénité était observée, ce qui semble s'observer sur les données française. La figure 1b représente la moyenne des variances $\frac{\sigma^2}{\lambda_i}$. Cette fois-ci, on note une différence avec ce que donnaient les données autrichiennes. En effet, notre distribution est plus étirée et plus applatie que celle de l'article. Cela montre que la dispersion de l'erreur de variance est plus faible, ce qui s'explique sans doute par le processus de standardisation des séries préalables et par le fait que nous travaillons sur moins de séries. Le graphique 2 représente la moyenne inconditionnelle de chaque groupe de série dans les deux états, selon qu'ils sont en-dessous de la croissance moyenne ou au-dessus. Cela permet d'illustrer que, contrairement à l'article, nos deux groupes de séries n'ont pas exactement le même comportement puisque les séries sont décalées les unes par rapport aux autres.

Le graphique 3 permet de juger la significativité de l'information *a priori* utilisée pour la classification. Nous obtenons des conclusions légèrement différentes de celles de l'article : pour le groupe 1, la distribution *a posteriori* des effets du PIB est éloignée de zéro, tandis que pour le groupe 2, on n'observe cela que pour les carnets de commandes (variable manuf-oscd). Cela s'explique par la façon dont la classification *a priori* a été effectuée, selon que la série allait être en avance ou non sur le PIB.

Le graphique 4 permet de se rendre compte de l'effet de la corrélation avec le PIB et le carnet de commandes entre les

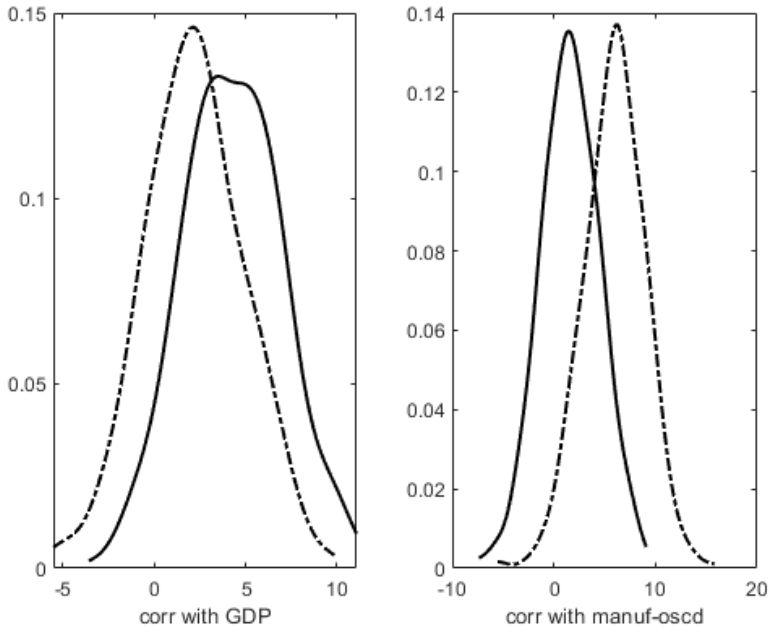


FIGURE 3 – $\pi(\gamma, \gamma_z|Y^T)$ - groupe 1 (trait plein) et 2 (pointillé)

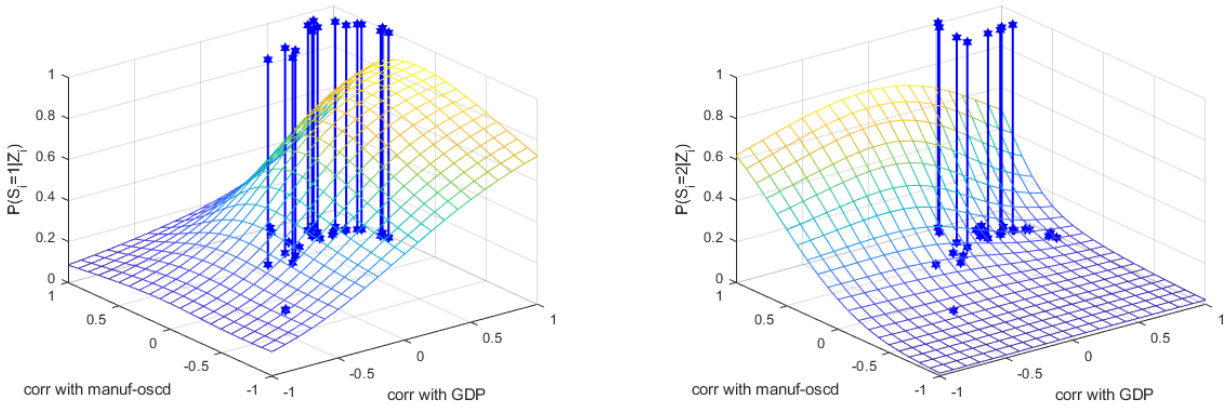


FIGURE 4 – Probabilités *a priori* et *a posteriori* $P(S = j|Z, Y^T)$ et $P(S_i = j|Z_i, Y^T)$

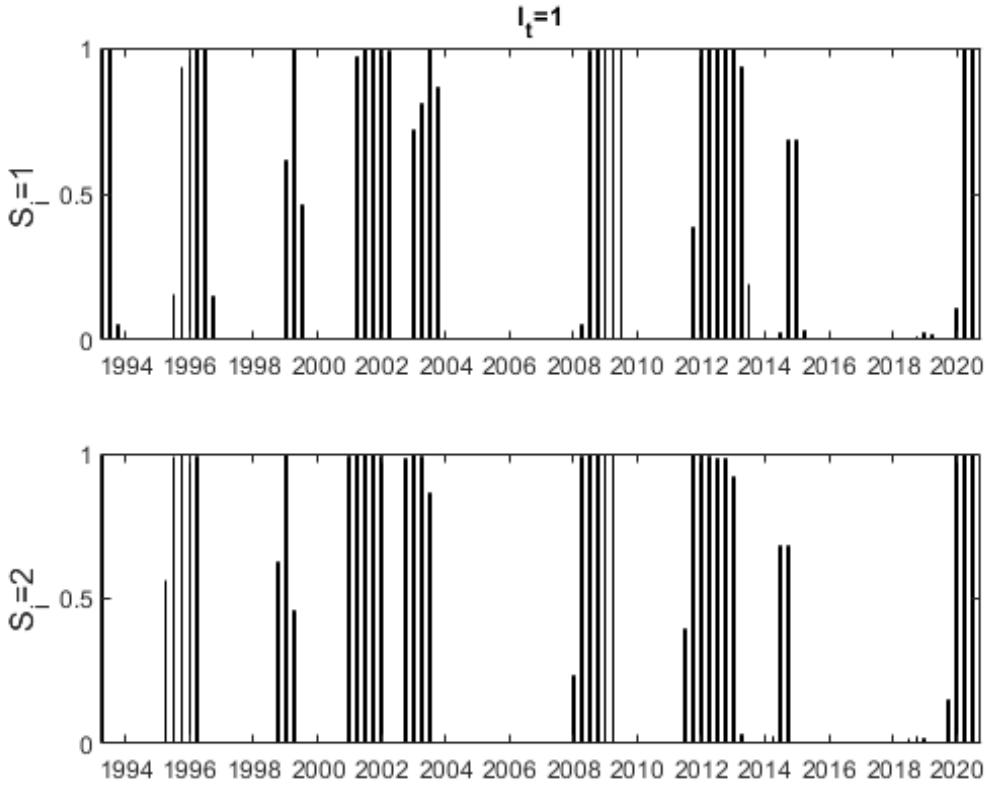


FIGURE 5 – Probabilités *a posteriori* $P(I_{kt} = 1|Y^T)$ des séries coïncidantes ($S_i = 1$) et qui mènent le cycle ($S_i = 2$) sur la période 1992-2020 avec une classification a priori.

groupes 1 (à gauche) et 2 (à droite). La courbe représente $P(S = j|Z, Y^T) = \int P(S = j|Z, \gamma, \gamma_z)\pi(\gamma, \gamma_z|Y^T)\mathrm{d}\gamma\mathrm{d}\gamma_z$, la probabilité conditionnelle sur le PIB et le carnet de commande. Cette probabilité est la plus élevée pour les séries du groupe 1 qui sont fortement corrélées avec le PIB et corrélées positivement avec le carnet de commande, la corrélation avec le PIB semblant jouer un rôle plus important puisque la probabilité varie plus selon la valeur de cette dernière. Nous retrouvons ici la définition du premier groupe de variables qui contient celles qui sont coincidentes avec le cycle et donc corrélées avec le PIB. Pour le groupe 2, la valeur de la probabilité est maximale pour les séries fortement corrélées avec le carnet de commandes et cette variable semble être plus importante. Cela conforte les analyses tirées du graphique 3.

La hauteur des lignes verticales du graphique 4 représente la probabilité $P(S_i = j|Z_i, Y^T)$ a posteriori pour une série d'appartenir aux groupes 1 et 2. On remarque que dans la grande majorité des cas, cette probabilité a été mise à jour à 0 ou à 1, grâce à l'information supplémentaire obtenue. On peut vérifier la qualité de notre pré-classification grâce à cela : elle semble cohérente.

Le graphique 5 est central puisqu'il epermet de repérer les points de retournement de notre cycle. Il représente les probabilités $P(I_{k,t} = 1|Y^T)$ pour les séries des deux groupes, obtenues en moyennant les résultats sur M simulations. On remarque que les résultats semblent significatifs puisque les probabilités ont des valeurs très proches de 0 ou de 1 dans une majorité de cas. À partir de ce graphique, on définit le temps t comme un point de retournement si $P(I_{k,t-1} = 1, I_{k,t} = 1|Y^T) < 0,5$ et $P(I_{k,t+1} = 1, I_{k,t+2} = 1|Y^T) > 0,5$, c'est-à-dire une faible probalité de ne pas changer d'état à ce moment et, une fois qu'on a passé ce moment, une forte probabilité de rester dans le même état. On peut donc repérer ici différents points de retournement qui correspondent à des réels évènements sur les marchés financiers en 1996 et 1999 (crise financière asiatique), 2002 et 2003 (krach boursier), 2009 (crise économique mondiale) ou encore 2013 (fin de la crise de 2008 avec encore des répercutions en France). Une fois des données disponibles pour la période entière du Covid-19, il sera intéressant de voir ce qui a été capté par notre modèle.

4 Discussion

4.1 Prévision

Nous l'avons vu, notre modèle permet de repérer les points de retournement de l'économie. Cependant, s'il est toujours intéressant de pouvoir le remarquer après coup, une utilisation plus immédiate serait appréciée. Le modèle permet cela puisqu'il peut être utilisé dans une optique de prévision. L'idée est d'utiliser la densité prédictive *a posteriori* $\pi(I^*_{T+h}|Y^T) = \pi(I^*_{T+h|T})$ et de simuler récursivement les valeurs futures à partir de cela. En moyennisant nos résultats sur un grand nombre de simulations, on peut alors déterminer la probabilité d'être dans les différents états.

4.2 Comparaison avec l'indicateur de retournement de l'Insee

Afin de pouvoir juger de l'efficacité de notre modèle, il peut être intéressant de le comparer à des méthodes déjà existantes et utilisées par les instituts nationaux de statistiques. En France, l'Insee publie un indicateur de retournement de conjoncture estimé à partir d'un modèle markovien à variables cachées sur des soldes d'opinion recodés en deux modalités (voir Bortoli *et al* (2015)).

La valeur de cet indicateur correspond à la différence entre la probabilité que la phase conjoncturelle soit favorable et la probabilité qu'elle soit défavorable. Il évolue donc entre $+1$ et -1 : un point très proche de $+1$ (respectivement de -1) signale que l'activité économique est en période de nette accélération (respectivement de nette décélération). Les moments où l'indicateur est proche de 0 sont assimilés à des phases de stabilisation, c'est-à-dire de retour du rythme de croissance de l'activité vers sa moyenne de long terme.

TODO.

5 Conclusion

TODO.

6 Annexes

Variable	**Commentaire**	S_i
YER	PIB	1
MTR	Imports	1
PCR	Consommation	1
ITR	Investissements	1
XTR	Exports	1
manuf-oscd	Enquête de conjoncture industrie - carnet de commandes généraux	2
manuf-oscde	Enquête de conjoncture industrie - carnet de commandes étrangers	2
manuf-ossk	Enquête de conjoncture industrie - niveau des stocks de produits finis	1
manuf-tppre	Enquête de conjoncture industrie - tendance passée de production	1
manuf-tppre	Enquête de conjoncture industrie - tendance prévue de production	2
manuf-pgp	Enquête de conjoncture industrie - perspectives générales	2
ser-capa	Enquête de conjoncture services - évolution passée du chiffre d'affaires	1
ser-capre	Enquête de conjoncture services - évolution prévue du chiffre d'affaires	2
ser-dem	Enquête de conjoncture services - demande prévue	2
bat-jcc	Enquête de conjoncture bâtiment - niveau des carnets de commandes	2
bat-epa	Enquête de conjoncture bâtiment - évolutions passée des effectifs	1
bat-tuc	Enquête de conjoncture bâtiment - taux d'utilisation des capacités de production	1
bat-apa	Enquête de conjoncture bâtiment - évolution passée de l'activité	1
bat-apre	Enquête de conjoncture bâtiment - évolution prévue de l'activité	2
CLIMAT-FR	Climat des affaires France	1
CLIMAT-FR-EMPL	Climat de l'emploi France	1
HIPC-FO	Indice des prix - nourriture	1
HICP-IG	Indice des prix - produits industriels	1
HICP-E	Indice des prix - énergie	1
HICP	Indice des prix	1
IPI-CZ	Indice de production industrielle - industrie manufacturière	1

Table 1 – Variables utilisées dans l'analyse

Références

Kaufmann S. (2010). Dating and forecasting turning points by Bayesian clustering with dynamic structure: a suggestion with an application to Austrian data. Journal of Applied Econometrics, **25**(2): 309-344

Frühwirth-Schnatter S, Kaufmann S. (2008). Model-based clustering of multiple time series. Journal of Business and Economic Statistics **26**(1): 78

Bortoli C., Gorin Y., Olive P.-D. et Renne C. (2015), « De nouvelles avancées dans l'utilisation des enquêtes de conjoncture de l'Insee pour le diagnostic conjoncturel », *Note de Conjoncture*, mars, p. 21-41.