

# 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 à essayer 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 travailler 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 travailler 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 supposons que nos séries suivent des processus autorégressifs, c'est-à-dire que l'on a

$$y_{it} = \mu_{I_{it}}^i + \phi_1^i y_{i,t-1} + \dots + \phi_p^i y_{i,t-p} + \epsilon_{it}$$

avec  $\epsilon_{it} \sim \mathcal{N}\left(0, \frac{\sigma}{\lambda_i^2}\right)$ . 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 chemins proches et on peut donc vouloir exploiter cette information pour notre estimation. On va donc chercher à créer des groupes de séries de manière intelligente. Dans ce but, on va donc introduire une indicatrice d'appartenance à un groupe que l'on va noter  $S_i \in \{1, \dots, K\}$ . Nous allons donc dans la suite supposer que pour les séries appartenant à un même groupe, nous allons avoir des paramètres distribués au sein d'une moyenne et ajouter un paramètre d'hétérogénéité, tel que :

$$\mu_{I_{S_i,t}}^i \sim \begin{cases} \mathcal{N}(\mu_1^k, q_1^k) & \text{si } S_i = k \text{ et } I_{kt} = 1 \\ \mathcal{N}(\mu_2^k, q_2^k) & \text{si } S_i = k \text{ et } I_{kt} = 2 \end{cases}$$

$$(\phi_1^i, \dots, \phi_p^i) \sim \mathcal{N}(\phi^k, Q_\phi^k) \text{ si } S_i = k$$

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

Il faut également proposer une spécification pour la variable  $S_i$ , nous supposons qu'elle suit un modèle logit multinomial, c'est-à-dire  $\mathbb{P}(S_i = k | \gamma_1, \dots, \gamma_K, \gamma_{z,1}, \dots, \gamma_{z,K-1}) = \frac{\exp(\gamma_k + Z_k \gamma_{z,k})}{1 + \sum_{l=1}^{K-1} \exp(\gamma_l + Z_l \gamma_{z,l})}$  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  $\gamma$  sont inconnus mais spécifiques aux groupes.

Enfin, il reste à ajouter une spécification pour en compte la dynamique du cycle. Pour cela, on va supposer 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, tandis que les séries pour lesquelles  $S_i = 1$  seront les séries qui coïncident avec le cycle. On va donc définir une nouvelle variable  $I^*$ .

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

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  $\rho^*$ , 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 considérera 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, on va noter  $\forall i \in \{1, \dots, N\}, y_i^t = \{y_{i,t}, y_{i,t-1}, \dots, y_{i,1}\}, Y_t = \{y_{1,t}, \dots, y_{N,t}\}$  et  $Y^t = \{Y_t, \dots, Y_1\}$ . De même, on notera  $S^N = \{S_1, \dots, S_N\}$  l'ensemble des indicatrices d'appartenances aux groupes,  $I^T = \{I_1^T, \dots, I_K^T\}$  où  $I_k^T = \{I_{k,T}, \dots, I_{k,1}\}$  les indicatrices d'états. Enfin, on notera  $\lambda^N = \{\lambda_1, \dots, \lambda_N\}$  qui regroupe les poids qui seront utilisés.

Par commodité, on notera également  $\theta = (\mu_1^1, \mu_2^1, \dots, \mu_1^K, \mu_2^K, \phi^1, \dots, \phi^K, Q^1, \dots, Q^K, \sigma^2, \xi^*, \xi^{\rho^*(k)=0}, \gamma, \gamma_z)$  où  $Q^k = \begin{pmatrix} q_1^k & 0 & 0 \\ 0 & q_2^k & 0 \\ 0 & 0 & Q_\phi^k \end{pmatrix}$  représente l'hétérogénéité intra-groupe,  $\xi^{\rho^*(k)=0} = \{\xi_{11}^k, \xi_{12}^k, \xi_{21}^k, \xi_{22}^k\}$ ,  $\gamma = (\gamma_1, \dots, \gamma_{K-1})$  et  $\gamma_z = (\gamma_{z,1}, \dots, \gamma_{z,K-1})$ .

Nous pouvons passer 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,  $L(Y^T|\psi) = \prod_{t=p+1}^T \prod_{i=1}^N f(y_{i,t}|y_i^{t-1}, \mu_{I_{S_i,t}}^{S_i}, \phi^{S_i}, Q^{S_i}, \lambda_i, \sigma^2)$ , avec

$$f(y_{i,t}|y_i^{t-1}, \mu_{I_{S_i,t}}^{S_i}, \phi^{S_i}, Q^{S_i}, \lambda_i, \sigma^2) = \frac{1}{\sqrt{2\pi\nu_{it}^{S_i}}} \exp\left(-\frac{1}{2\nu_{it}^{S_i}}\left(y_{it} - \mu_{I_{S_i,t}}^{S_i} - \sum_{j=1}^p \phi_j^{S_i} y_{i,t-1}\right)^2\right)$$

Avec  $y_{it} = X_{it}^{S_i} \beta^{S_i} + \epsilon_{it}^*$ , où  $\epsilon_{it}^* \sim N(0, \nu_{it}^{S_i})$ ,  $X_{it}^{S_i} = (D_{1t}^{I(S_i)}, D_{2t}^{I(S_i)}, y_{i,t-1}, \dots, y_{i,t-p})$ ,  $D_{jt}^{I(S_i)} = \mathbb{1}(I_{S_i,t} = j)$ ,  $\beta^{S_i} = (\mu_1^{S_i}, \mu_2^{S_i}, \phi^{S_i})$ , et  $\nu_{it}^{S_i} = X_{it}^{S_i} Q^{S_i} X_{it}^{S_i'} + \frac{\sigma^2}{\lambda_i}$ .

Nous supposons dans la suite que la structure  $\rho^*$  est connue (déterminée par d'autres méthodes), on peut alors réécrire la distribution a priori sur  $\phi$  ainsi :

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

où les densités  $\pi(I^{*T}|\rho^*, \xi^*)$ ,  $\pi(I_k^T|\rho^*, \xi^*)$  et  $\pi(S^N|\gamma, \gamma_z, Z^N)$  sont connues. On supposera que les poids  $(\lambda^N)$  sont indépendants et suivent une loi de Gamma. Il reste à spécifier l'a priori sur  $\pi(\theta)$  : pour cela, nous décomposons  $\theta$  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  $\rho^*$ , on obtient une réalisation pour l'indicatrice d'état de l'étape 2. Dans l'étape 4, toutes les distributions a posteriori sont conjuguées aux a priori, sauf pour  $\gamma$  et  $\gamma_z$ , nous utiliserons alors l'agorithme 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, dont le détail est expliqué dans le tableau 1.

Nous avons retenu ici deux 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  $(\rho^*)$  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.

Nous nous sommes basés sur les programmes mis à disposition par [Sylvia Kaufmann \(2010\)](#) et allons essayer d'analyser les résultats obtenus.

Le modèle étudié ici se caractérise entre autres par sa spécification de l'hétérogénéité entre séries, qui est prise en compte. Ceci est illustré par le graphique 1. Il représente, à gauche, la moyenne de la somme des paramètres

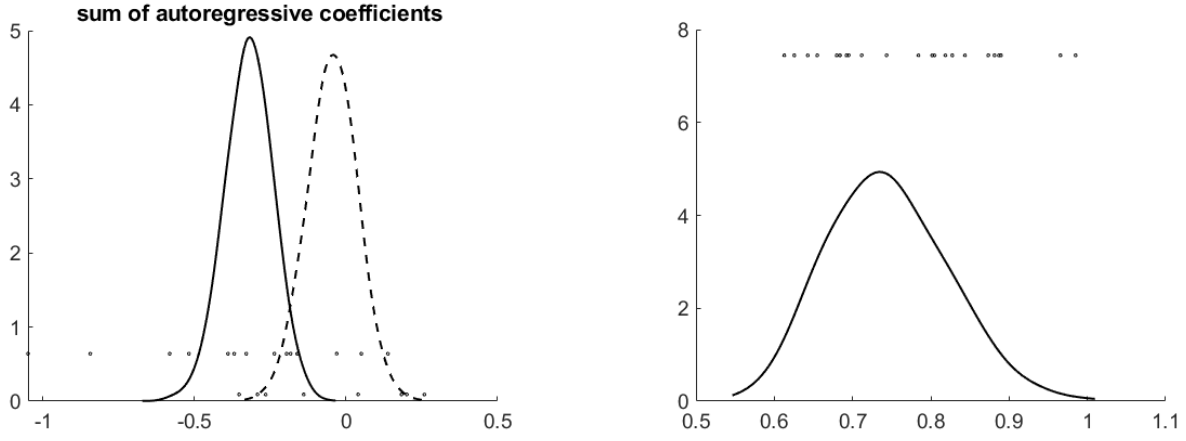


FIGURE 1 – Hétérogénéité des séries - groupe 1 (trait plein) et 2 (pointillés). A gauche, somme des coefficients de l'AR. A droite,  $\sigma^2, \sigma^2/\lambda_i$

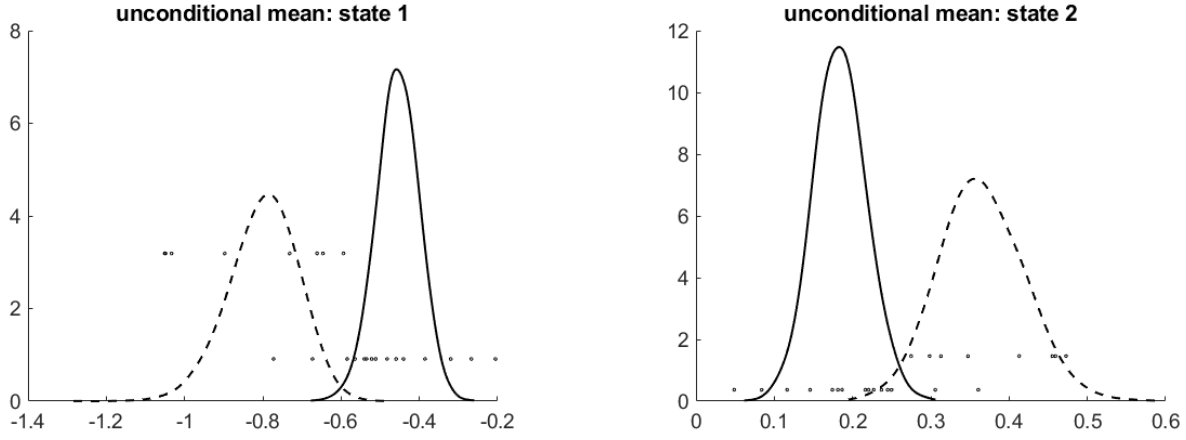


FIGURE 2 – Moyenne inconditionnelle dans les différents états

autorégressifs. Dans l'article, ils observaient sur leurs données une importance hétérogénéité, dans notre cas, cela semble être la même chose. A droite, on 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 moins haute que celle de l'article. Cela montre que la dispersion de l'erreur de variance est plus faible, sans doute explicable par le processus de standardisation des séries préalables et parce qu'on travaille sur moins de données. 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 graphi 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 groupes 1 (à gauche) et 2 (à droite). La courbe représente  $P(S = j|Z, Y^T) = \int P(S = j|Z, \gamma, \gamma_z)\pi(\gamma, \gamma_z|Y^T)d\gamma d\gamma_z$ , 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. 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.

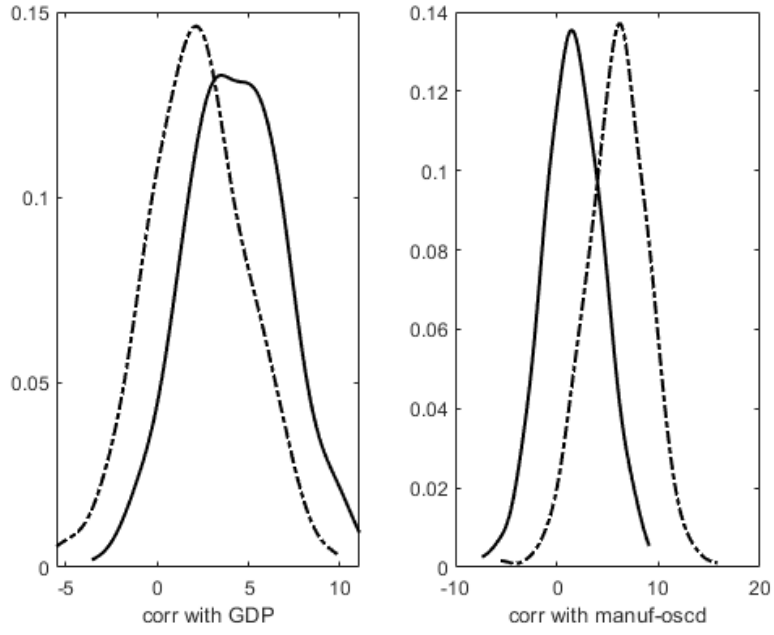


FIGURE 3 –  $\pi(\gamma, \gamma_z | Y^T)$  - groupe 1 (trait plein) et 2 (pointillé)

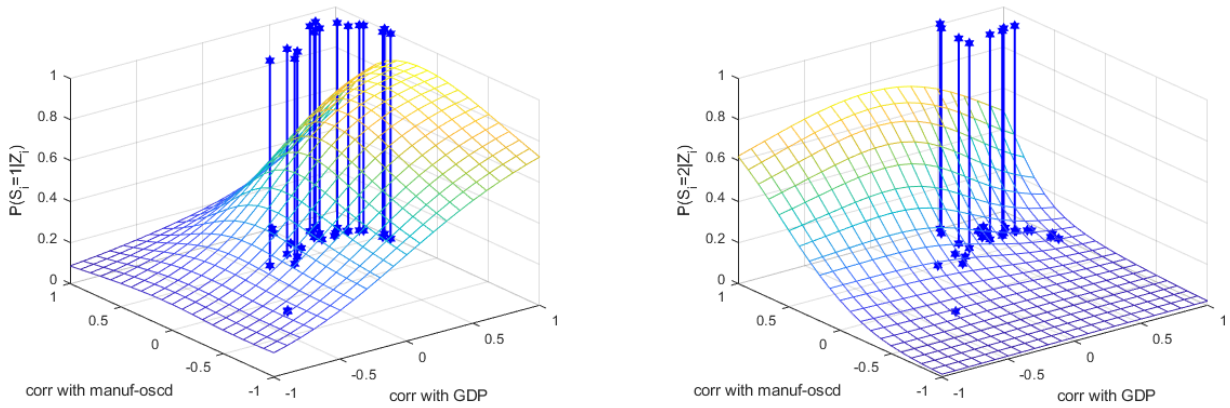


FIGURE 4 – Probabilités *a priori* et *a posteriori*  $P(S = j | Z, Y^T)$  et  $P(S_i = j | Z_i, Y^T)$

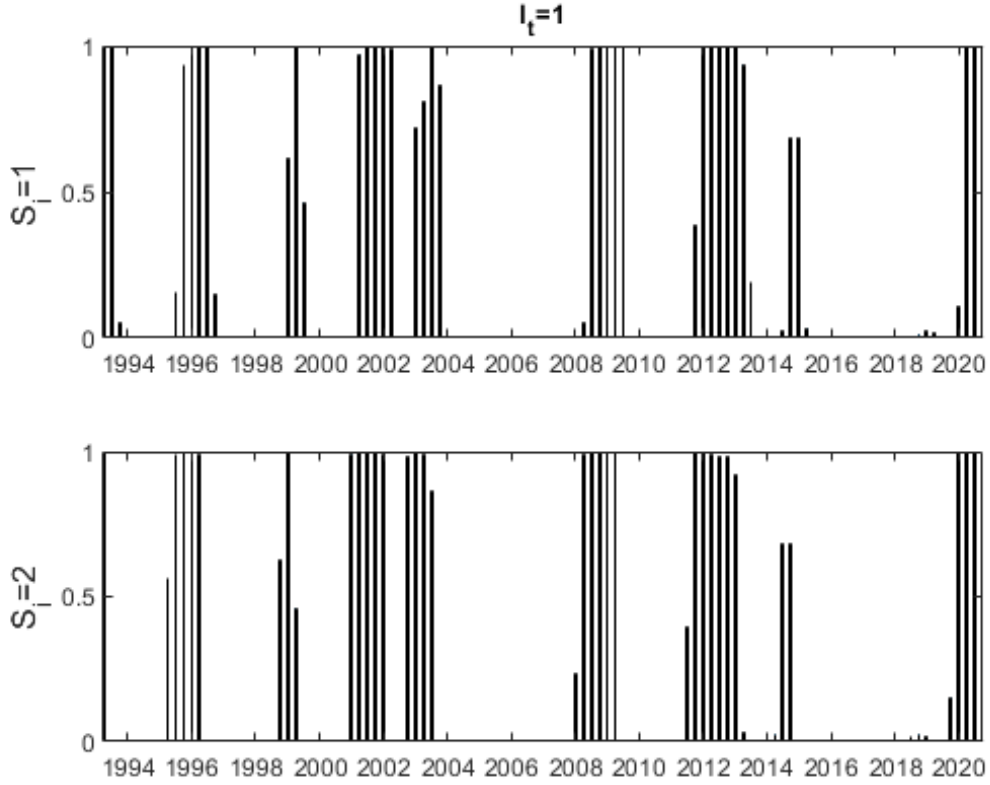


FIGURE 5 – Probabilités *a posteriori*  $P(I_{kt} = 1|Y^T)$  des séries coïncidentes ( $S_i = 1$ ) et qui mènent le cycle ( $S_i = 2$ ) sur la période 1992-2020 avec une classification *a priori*.

La hauteur des lignes verticales du graphique 4 représentent la probabilité  $P(S_i = j|Z_i, Y^T)$  la probabilité *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 est celui permettant 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. A 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 probabilité 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 événements 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 répercussions 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*  $\pi(I_{T+h}^*|Y^T) = \pi(I_{T+h}^*|I_T)$  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 calcule un indicateur de retournement de conjoncture.

TODO.

## 5 Conclusion

TODO.

## 6 Annexes

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

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