Projet de Statistique Bayésienne

Blanchard Eléonore, Ly Yannick

Regime Switching and Technical Trading with Dynamic Bayesian Networks in High-Frequency Stock Markets - *Luis Damiano, Brian Peterson, Michael Weylandt*

Partie 1 : Le sujet

a. Problématique

Un des plus grands défis de la finance est de prédire les variations du marché, c'est-à-dire s'il sera haussier ou baissier). A partir de certaines données présentes, la problématique de notre projet est donc de prédire si le marché sera haussier ou baissier, en utilisant exclusivement des données présentes (processus de Markov).

On se base sur l'hypothèse de l'analyse technique suivante afin de construire le modèle : des *patterns* de prix et de volume (quantités traités) sont récurrents et donnent une information sur les mouvements futurs des prix. Ces patterns sont constitués d'une tendance définie par une série de zig-zag (extrema locaux de la courbe de prix qui alternent entre minima et maxima). Les minima et les maxima sont croissants (resp. décroissants) lorsque la tendance est haussière (resp. baissière).

Figure 1 *: Courbe de prix composée de zig-zag positifs et négatifs*

Outre les prix, nous allons également utiliser les volumes pour prédire les tendances, car un volume élevé va « pousser » les prix (ils vont augmenter lorsque la tendance est haussière et diminuer lorsqu'elle est baissière).

b. Données et features

Nos données sont des séries temporelles financières en intraday (plusieurs données par seconde) représentant les prix et volume traités d'un actif à un instant t.

Pour construire notre modèle et représenter l'impact des prix et des volumes présents sur les prix futurs, il nous faut construire les indicateurs suivants : f_n^0 représente le sens du zig-zag, f_n^1 représente le sens de la tendance des prix et f_n^2 indique si le volume se renforce ou s'affaiblie. Ils nous permettent d'en déduire les features :

Zig-zag descendant				**Zig-zag montant**			
	Tendance	**Volume**	**Market State**		**Tendance**	**Volume**	**Market State**
$\boldsymbol{D_1}$	Up +1	Weak -1	Bull	$\boldsymbol{U_1}$	Up +1	Strong +1	Bear
$\boldsymbol{D_2}$	Down -1	Weak -1	Bull	$\boldsymbol{U_2}$	Down -1	Strong +1	Bear
$\boldsymbol{D_3}$	Up +1	Indeterminant 0	Bull	$\boldsymbol{U_3}$	Up +1	Indeterminant 0	Bear
$\boldsymbol{D_4}$	No trend 0	Weak -1	Bull	$\boldsymbol{U_4}$	No trend 0	Strong +1	Bear
$\boldsymbol{D_5}$	No trend 0	Indeterminant 0	Local vol.	$\boldsymbol{U_5}$	No trend 0	Indeterminant 0	Loval vol.
$\boldsymbol{D_6}$	No trend 0	Strong +1	Bear	$\boldsymbol{U_6}$	No trend 0	Weak -1	Bull
$\boldsymbol{D_7}$	Down -1	Indeterminant 0	Bear	$\boldsymbol{U_7}$	Down -1	Indeterminant 0	Bull
$\boldsymbol{D_8}$	Up +1	Strong +1	Bear	$\boldsymbol{U_8}$	Up +1	Weak -1	Bull
$\boldsymbol{D_9}$	Down -1	Strong +1	Bear	$\boldsymbol{U_9}$	Down -1	Weak -1	Bull

c. La pertinence d'un processus bayésien

Les données financières dont nous disposons sont bruitées. Un cadre statistique permettant d'estimer des probabilités conditionnelles et d'inférer des états cachés est intéressant pour traiter ces données.

Il est possible d'utiliser un réseau bayésien pour effectuer de l'inférence statistique. Il s'agit d'un modèle graphique qui permet de présenter des indépendances conditionnelles entre un ensemble de variables aléatoires ($S_{t+1} \perp S_{t-1}|S_t$ (propriété de Markov) et $Y_t \perp Y_{t'}|S_t$, pour $t' = t$).

L'inférence dans un réseau bayésien est le calcul de probabilités a posteriori en tenant compte simultanément des a priori et de l'expérience contenues dans les données ; mais il n'a pas vocation à déterminer les relations de cause à effet entre les variables. L'observation d'une cause n'entraine pas systématiquement l'effet escompté mais modifie seulement la probabilité de l'évènement.

En résumé, à partir des informations observées (prix et volume), nous calculons la probabilité que le marché sera haussier ou baisser, qui sont des données non observées.

L'utilisation d'un réseau bayésien dynamique (DBN) permet d'étendre le cas d'un réseau bayésien (qui est estimé à un instant t) en incorporant une composante stochastique, et permet de traiter des données financières.

Dans le cadre de notre exemple on cherche à déterminer si le marché est haussier ou baissier en fonction des caractéristiques du marché. Ce système peut ensuite être rendu dynamique en intégrant le fait que la probabilité d'être dans un marché haussier (ou baissier) au temps $t + 1$ dépend des données en t.

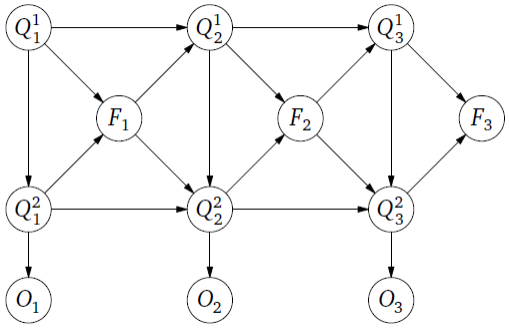

***Figure 2** : Réseau bayésien dynamique pour notre problème (pour 3 premiers pas de temps t) où Q_t^1 représente l'état du marché (haussier ou baissier), Q_t^2 le zig-zag (négatif ou positif), F_t une indicatrice qui vaut 1 si on change d'état de marché et O_t les output ($\{D_1, \dots D_9\}$ ou $\{U_1, \dots U_9\}$) en t*

Les réseaux bayésiens dynamiques sont une généralisation des modèles probabilistes de times series comme les modèles de Markov cachés.

Un modèle de Markov caché (HMM) est un modèle statistique dans lequel les observations sont générées par un système modelé comme un processus de Markov avec des états cachés (inobservables). A chaque pas de temps t, le système fait une transition d'un état à un autre (ou reste dans le même état) et sort comme output une observation suivant la distribution de probabilité spécifique à l'état dans lequel il se trouve.

Appliqué à notre exemple, les états cachés sont un marché haussier ou un marché baissier. Les données observables pour estimer la probabilité d'être dans un état ou un autre, ainsi que la matrice de transition, sont des features créés à partir du prix et du volume, comme indiqué précédemment. L'utilisation d'un HMM semble donc être pertinent pour représenter notre problème.

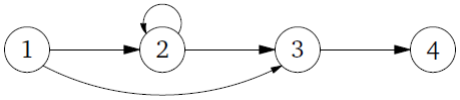

Figure 3 : Exemple d'un diagramme de transition pour un HMM avec 4 états cachés. Les flèches représentent les transitions avec probabilités non nulles.

Une extension du HMM est le modèle de Markov caché hiérarchique (HHMM), qui nous permet de développer le modèle décrit précédemment de manière plus précise. En effet, en finance nous considérons qu'il existe des tendances haussières et baissières définies par des micro-tendances modélisées par des zig-zag.

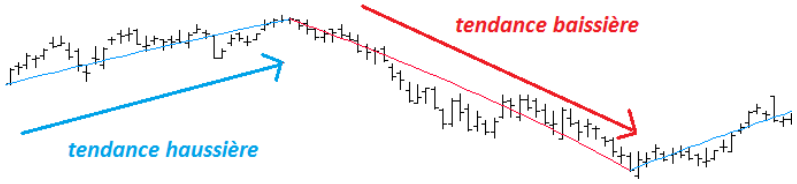

Figure 4 : Exemple de de tendances haussière et baissière définies par des zig-zag positifs et négatifs

Ci-dessous notre problème modélisé par un HHMM.

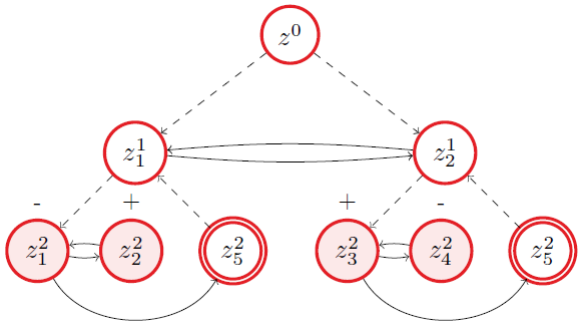

Figure 5 : HHMM pour les prix et volume

On peut résumer le HHMM comme un HMM dont les états (z_1^1 et z_2^1) sont les états de production (z_5^2) du HHMM correspondant. Chaque nœud active un HMM avec deux états latents (z_1^2 et z_2^2 pour le nœud z_1^1 ; et z_3^2 et z_4^2 pour le nœud z_2^1).

Dans le cadre d'un HHMM, les états (marché haussier ou baissier) ont une distribution de probabilité jointe. On fait donc l'hypothèse qu'un état influe sur l'autre, ce qui n'est pas forcément vrai, et peut induire une baisse de la précision de notre prédiction.

Dans un DBN, l'état actuel du modèle en t ainsi que les données en $t+1$ influents sur l'état en $t+1$, ce qui est plus représentatif de la réalité.

On va donc modéliser les prix et volume avec un HHMM, puis le transformer en DBN pour effectuer nos inférences. L'utilisation du DBN nous permettra de réduire notre erreur de prédiction, car comme on l'a vu le HHMM fait baisser la précision de la prédiction.

d. Lien avec le cours

Dans le cadre du projet nous sommes amenés à faire de l'inférence bayésienne. En effet, nous estimons les paramètres du modèle suivant grâce à des données empiriques, ces paramètres sont mis à jour de manière stochastique.

Vecteur des probabilités initiales : $\pi = [\pi_1 \quad 1-\pi_1]$ où π_1 (resp. $1-\pi_1$) est la probabilité initiale que le marché soit haussier (resp. baissier).

Matrice de transition des nœuds du niveau supérieur : $A^{q^0} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ qui correspond aux probabilités de passer d'un marché haussier à un marché baissier, et inversement. Dans le modèle nous supposons qu'on ne peut pas rester dans un marché haussier (resp. baissier) et qu'on ne peut que passer à un état où le marché est baissier (resp. haussier) avec probabilité 1.

Matrices de transition des nœuds du niveau inférieur : il y a une matrice de transition dans le cas où nous sommes dans un marché haussier ($A^{q_1^1}$) et une matrice de transition dans le cas où nous sommes dans un marché baissier ($A^{q_2^1}$).

$$A^{q_1^1} = \begin{pmatrix} 0 & p_1 & 0 & 0 & 1-p_1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \qquad A^{q_2^1} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_2 & 1-p_2 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Avec p_1 (resp. p_2) la probabilité de passer d'un zig-zag négatif (resp. positif) à un zig-zag positif (resp. positif) et $1-p_1$ (resp. $1-p_2$) la probabilité de passer d'un zig-zag négatif (resp. positif) au nœud de production qui renvoie l'information au nœud de la couche supérieure. Si nous passons d'un zig-zag négatif (resp. positif) à un zig-zag positif (resp. négatif), la probabilité de repasser à un zig-zag négatif (resp. positif) est de 1.

Les lignes 3 et 4 (resp. 1 et 2) de la matrice de transition $A^{q_1^1}$ (resp. $A^{q_2^1}$) sont égales à 0 car les nœuds des zig-zag du marché baissier (resp. haussier) ne peuvent pas être activés quand nous sommes dans un marché haussier (resp. baissier).

La ligne 5 correspondant au nœud de production, il renvoie nécessairement une information au niveau supérieur et donc n'a pas de probabilité de transition. Ce nœud de production renvoie une observation parmi un ensemble fini de possibilités ($U = \{U_1, \dots, U_9\}$ et $D = \{D_1, \dots D_9\}$).

Conditionnellement au zig-zag (z_1^2 et z_4^2 les zig-zag négatifs et z_2^2 et z_3^2 les zig-zag positifs), la valeur renvoyée par le nœud de production peut être différente. Si le zig-zag est négatif (resp. positif), la valeur prise par le nœud de production est parmi U (resp. D). D'où les probabilités conditionnelles suivantes :

$$B^1 = p(x_t|z_1^2) = [b_{D_1}^1, \dots b_{D_9}^1] \qquad B^3 = p(x_t|z_3^2) = [b_{U_1}^3, \dots b_{U_9}^3]$$

$$B^2 = p(x_t|z_2^2) = [b_{U_1}^2, \dots b_{U_9}^2] \qquad B^4 = p(x_t|z_4^2) = [b_{D_1}^4, \dots b_{D_9}^4]$$

Finalement, nous devons estimer 32 paramètres : $\theta = (\pi_1, p_1, p_2, b_{D_l}^1, b_{U_l}^2, b_{U_l}^3, b_{D_l}^4)$ avec $l \in \{1, \dots 8\}$.

Partie 2 : Applications

a. Code et difficultés computationnelles

L'implémentation du modèle a été effectuée sous R.

Stan est un logiciel probabiliste permettant de faire de l'inférence statistique écrit en C++. Il permet d'estimer les paramètres du modèle en utilisant des méthodes MCMC (Monte-Carlo par chaînes de Markov) et de calculer la fonction de densité des variables aléatoires qu'on cherche à estimer. Le package « RStan » permet l'utilisation du logiciel Stan sous R.

Nous n'avons pas eu le temps de nous intéresser à l'implémentation et à la résolution d'un HHMM sous Stan, c'est pourquoi nous avons décidé de reprendre le fichier Stan des auteurs du papier (« Config.stan » dans le dossier).

Toutefois, nous nous sommes intéressés à la méthode d'inférence utilisée par les auteurs : l'algorithme de Baum Welch (ou *forward-backard algorithm*). Cet algorithme permet d'estimer les paramètres du modèle qui maximisent la probabilité d'une séquence observable, ce qui est pertinent dans notre cas car nous avons déjà un modèle et cherchons les paramètres les plus optimaux.

L'algorithme Baum-Welch fonctionne de la manière suivante :

Pour chaque observation t de l'échantillon :

1- Calcul des probabilités dites forward avec le *forward algorithm*
2- Calcul des probabilités dites backward avec le *backaward algorithm*
3- Calcul de la contribution (gradients) de l'observation sur les 35 paramètres à estimer
4- Calcul des nouveaux paramètres en prenant en compte les contributions
5- Calcul de la nouvelle log-vraisemblance

On stop lorsque la variation de la log-vraisemblance est plus petite qu'un seuil donné
Ou lorsqu'on atteint un nombre maximum d'itérations (ici, il vaut 500)

Attention, l'utilisation du package « RStan » nécessite une installation particulière précisée ici.

b. Résultats

Les 35 paramètres estimés par l'algorithme sont les suivants :

Vecteur des probabilités initiales : $\pi = [0{,}54 \quad 0{,}46]$

On estime qu'il y a presque autant de chance d'être dans un marché haussier ou baissier à l'initialisation ce qui est représentatif d'un marché financier qui présente beaucoup d'incertitude.

Matrice de transition des nœuds du niveau supérieur : $A^{q^0} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

Matrices de transition des nœuds du niveau inférieur :

$$A^{q_1^1} = \begin{pmatrix} 0 & 0{,}73 & 0 & 0 & 0{,}27 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \qquad A^{q_2^1} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0{,}55 & 0{,}45 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Dans un marché haussier, lorsqu'il y a un zig-zag négatif, il y a de fortes chances de passer à un zig-zag positif (probabilité égale à 0,73), tandis que dans un marché baissier, lorsqu'il y a un zig-zag positif, il y a presque autant de chances de revenir à un zig-zag positif qu'à un zig-zag négatif. Il y a donc plus de chance de retournement de la courbe lorsqu'on est dans un marché haussier que baissier.

Vecteurs des probabilités conditionnelles :

$B1 = [0{,}0014 \quad 0{,}0152 \quad 0{,}00168 \quad 0{,}012 \quad 0{,}888 \quad 0{,}028 \quad 0{,}0286 \quad 0{,}00136 \quad 0{,}0173]$

$B_2 = [0{,}001 \quad 0{,}02386 \quad 0{,}00074 \quad 0{,}399 \quad 0{,}072 \quad 0{,}406 \quad 0{,}07473 \quad 0{,}00084 \quad 0{,}0206]$

$B_3 = [0{,}0092 \quad 0{,}00218 \quad 0{,}04331 \quad 0{,}014 \quad 0{,}887 \quad 0{,}017 \quad 0{,}00261 \quad 0{,}0125 \quad 0{,}0021]$

$B_4 = [0{,}0325 \quad 0{,}00076 \quad 0{,}07436 \quad 0{,}39 \quad 0{,}146 \quad 0{,}315 \quad 0{,}00068 \quad 0{,}03854 \quad 0{,}0010]$

Ces 35 paramètres ont été estimés sur les domaines d'un actif financier spécifique (G.TO). L'exercice est à réeffectuer pour estimer les paramètres sur les données d'un autre actif.

Pour aller plus loin, nous aurions pu tout d'abord tester notre modèle sur un échantillon afin de voir si nos probabilités correspondent à la réalité hors échantillon. Dans le papier, les auteurs appliquent par la suite une stratégie de trading afin de passer des ordres d'achat et de vente selon les prédictions renvoyées par le modèle (probabilité d'être dans un marché haussier ou baissier à la prochaine période).