Dow Jones et test de Lo-MacKinlay

P. Gazzano – N. Lygeros

Soit $X(t)$ le prix d'un indice boursier. Nous notons $X(t) = \log(X(t))$, le rendement logarithmique. Le but de cette note est de tester l'hypothèse de marche aléatoire pour le Dow Jones avec des données intraday dont le nombre de tick est égale à 1048576. Le test est issu du livre *A Non-Random Walk Down Wall Street* de Andrew Lo et A.Craig MacKinlay.

Pour un cours boursier, il est possible d'avoir deux types de comportements : l' Homoscédasticité qui correspond à la propriété d'i.i.d (independant and identical distribution) dont la distribution est gaussienne, et l'Hétéroscédasticité, où nous n'avons pas d'autocorrélation, car la loi n'est pas obligatoirement une loi normale, et la variance de celle-ci peut varier avec le temps.

Sous l'hypothèse d'Homoscédasticité, les variances de $X_t - X_{t-1}$ et de $X_t - X_{t-q}$ seront proportionnelles. Pour tout entier $q > 1$, si nous sélectionnons une série d'observations $X_0, ..., X_{nq}$ de longueur $nq+1$, nous pouvons les généraliser aux différences $q^{ièmes}$ et écrire les nouveaux estimateurs :

$$\hat{\mu} = \frac{1}{nq}(X_{nq} - X_0)$$

$$\hat{\sigma}_a^2 = \frac{1}{nq}\sum_{k=1}^{nq}\left(X_k - X_{k-1} - \hat{\mu}\right)^2$$

$$\hat{\sigma}_b^2 = \frac{1}{nq}\sum_{k=1}^{nq}\left(X_{kq} - X_{qk-q} - q\hat{\mu}\right)^2$$

Nous considérons alors le ratio des variances :

$$\Psi_{ratio} = \frac{\sigma_b^2}{\sigma_a^2} - 1 \quad \text{avec } \sqrt{2n}\Psi_{ratio} \sim N(0,2)$$

Le test peut-être affiné en utilisant le fait que sous l'hypothèse, les distributions asymptotiques s'écrivent $\sqrt{nq}\psi_{ratio} \sim N(0, 2(q-1))$

Nous pouvons donc créer une variance de l'estimateur $\hat{\sigma}_b^2$:

$$\hat{\sigma}_c^2(q) = \frac{1}{nq^2}\sum_{k=q}^{nq}\left(X_k - X_{k-1} - q\hat{\mu}\right)^2$$

Et à partir de cet estimateur, nous avons la possibilité de créer des estimateurs pour le ratio des variances :

$$M_{ratio}(q) = \frac{\hat{\sigma}_c^2(q)}{\hat{\sigma}_a} - 1$$

Cela permet d'avoir une version non biaisée des estimateurs $\hat{\sigma}_c^2(q)$ et $\hat{\sigma}_a$, que nous notons $\bar{\sigma}_c^2(q)$ et $\bar{\sigma}_a^2$.

$$\bar{\sigma}_a^2 = \frac{1}{nq-1}\sum_{k=1}^{nq}\left(X_k - X_{k-1} - \hat{\mu}\right)^2$$

$$\bar{\sigma}_c^2(q) = \frac{1}{q(nq-q+1)(1-\frac{q}{nq})}\sum_{k=q}^{nq}\left(X_k - X_{k-q} - q\hat{\mu}\right)^2$$

Par conséquent, nous avons $\bar{M}_{ratio}(q) = \frac{\bar{\sigma}_c^2(q)}{\bar{\sigma}_a^2(q)} - 1$

Sous l'hypothèse d'homoscédasticité, les statistiques se comportent de la façon suivante :

$$\sqrt{nq}\bar{M}_{ratio}(q) \sim N(0, \frac{2(2q-1)(q-1)}{3q})$$

Ceci peut se réécrire de la façon suivante :

$$z_1(q) = \sqrt{nq}.\bar{M}_{ratio}.\left[\frac{2(2q-1)(q-1)}{3q}\right]^{-1/2} \sim N(0,1)$$

Pour $q = 2$, la statistique peut se simplifier comme ceci :

$$M_{ratio}(2) = \hat{\rho}_1 - \frac{1}{4n\hat{\sigma}_a^2}\left[\left(X_1 - X_0 - \hat{\mu}\right)^2 + \left(X_{2n} - X_{2n-1} - \hat{\mu}\right)^2\right] \approx \hat{\rho}_1$$

Ceci provient du fait qu'il est possible de négliger le terme entre crochet. Ainsi, nous pouvons considérer que $\bar{M}_{ratio}$ peut s'écrire comme une combinaison linéaire des coefficients d'autocorrélation et de q de la manière suivante :

$$M_{ratio}(q) = \frac{2(q-1)}{q}\hat{\rho}_1 + \frac{2(q-2)}{q}\hat{\rho}_2 + ... + \frac{2}{q}\hat{\rho}_{q-1}$$

Les $\hat{\rho}_k$ s'écrivent comme ceci :

$$\hat{\rho}_i = \frac{\sum_{k=i+1}^{nq} \left(X_k - X_{k-1} - \hat{\mu}\right)\left(X_{k-i} - X_{k-i-1} - \hat{\mu}\right)}{\sum_{k=1}^{nq} \left(X_k - X_{k-1} - \hat{\mu}\right)^2}$$

Sous l'hypothèse d'hétéroscédasticité, le ratio doit toujours tendre vers 1 pour obtenir la relation de proportionnalité entre les variances. Il est possible d'adapter le test pour qu'il y ait non-correlation. La statistique $\bar{M}_{ratio}(q)$ tend encore vers zéro sous cette hypothèse, et il est seulement nécessaire de calculer la variance asymptotique $\theta(q)$ de $\bar{M}_{ratio}(q)$ pour arriver au résultat. Si nous pouvons obtenir les variances asymptotiques δ_i de chaque $\hat{\rho}_i$ sous l'hypothèse d'Homoscédasticité, il est possible de calculer les variances asymptotiques $\theta(q)$ et $\bar{M}_{ratio}(q)$ comme une somme pondérée des δ_i. Ainsi, si nous notons

$$\delta_i = Var\left[\hat{\rho}_i\right]$$
$$\theta(q) = Var\left[\bar{M}_{ratio}(q)\right]$$

Sous l'hypothèse d'hétéroscédasticité, nous obtenons les résultats suivants :

-les statistiques $\bar{M}_{ratio}(q)$ convergent vers 0.
-un estimateur consistant de l'hétéroscédasticité de la variance δ_i :

$$\hat{\delta}_i = \frac{nq \sum_{k=i+1}^{nq} \left(X_k - X_{k-1} - \hat{\mu}\right)^2 \left(X_{k-i} - X_{k-i-1} - \hat{\mu}\right)^2}{\left[\sum_{k=1}^{nq} \left(X_k - X_{k-1} - \hat{\mu}\right)^2\right]^2}$$

-un estimateur consistant de l'hétéroscédasticité de la variance $\theta(q)$:

$$\hat{\theta}(q) = \sum_{i=1}^{q-1} \left[\frac{2(q-i)}{q}\right]^2 \hat{\rho}_i$$

Malgré la présence d'hétéroscédasticité, la statistique suivante suit encore une loi normale :

$$z_2(q) = \sqrt{nq}\, \frac{\bar{M}_{ratio}(q)}{\hat{\theta}(q)^{1/2}} \sim N(0,1)$$

Il est donc possible de l'évaluer à partir de q et n. En répétant ce test, nous obtenons une liste de valeurs $z_2(q)$. En calculant la moyenne et la variance de cette liste, nous montrons que pour un mouvement brownien la variable suit effectivement une loi normale.

Pour tester la robustesse de cette approche, nous la testons sur un mouvement brownien. Nous remarquons que pour des faibles valeurs de Q, il existe une bonne ergodicité, puisque nous trouvons une variance proche de 1, et une moyenne proche de 0.

Q=2

N	1000	2000	3000	4000	5000	6000	10000	15000
Moyenne	-0.02793458	0.00065997	0.03110849	-0.01092805	-0.01119984	0.00634354	0.04442413	0.02862193
Variance	1.02214314	0.96340183	0.98855375	0.9863631	0.95843202	0.97204862	0.98154402	0.97171375

Q=9

N	1000	2000	3000	4000	6000	10000
Moyenne	-0.01433867	-0.01402628	-0.05100081	0.00330666	0.0157248	-0.03076282
Variance	1.02052538	0.98799055	0.98536514	1.01296259	1.01232676	0.98600104

Q=10

N	1000	2000	3000	4000	5000	6000	10000	15000
Moyenne	-0.05178577	0.03413217	0.00267348	0.0185927	-0.00169503	0.02612549	0.10220093	-0.14945782
Variance	1.00250919	1.02048022	0.9600497	0.96060502	1.01245163	0.90492088	1.03459751	1.03284921

Q=20

N	1000	2000	3000	4000	5000	6000	10000	15000
Moyenne	-0.02737818	0.01964599	-0.08488547	-0.16764725	-0.07892226	0.06324067	0.00215912	0.12176926
Variance	1.03611291	1.0463133	1.02438933	1.02423018	0.91131377	0.9356086	1.04157279	1.11141091

Pour Q=20 et Q=50, nous observons un éloignement de la valeur 1 lorsque N=15000, mais le test reste consistant.

N	1000	2000	4000
Moyenne	-0.08274741	-0.08968219	-0.12456404
Variance	1.00184305	1.00351036	0.96610859

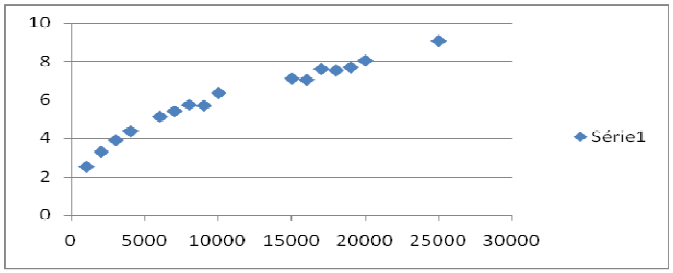
Q=50

N	500	1000	10000
Moyenne	0.00222836	-0.05458075	-0.24510791
Variance	0.99592872	0.9673307	1.05230643

Voici enfin, une application du test aux valeurs du Dow Jones.

Q=2

N	1000	2000	3000	4000	6000	7000	8000	9000	10000	15000	16000	17000	18000	19000	20000	25000
Moyenne	2.51490017	3.29008133	3.89173546	4.35825194	5.12264548	5.40459718	5.73831825	5.69920804	6.35183101	7.10792561	7.04146562	7.59823445	7.53078369	7.68890786	8.03715763	9.0605163
Variance	5.34294336	6.83551549	7.53387671	8.21837136	9.36749113	9.45323522	10.0167424	10.2313303	10.3781688	11.5335723	11.7838464	11.907157	12.1110925	12.1201971	12.4089082	13.1875544



Q=5

N	1000	2000	3000	4000
Moyenne	4.09726511	5.11806553	5.98583417	7.18708969
Variance	8.34597365	10.329443	11.3263272	13.2385126

Q=10

N	1000	2000	5000
Moyenne	4.5807594	5.35106038	9.10684365
Variance	9.44716981	11.2317567	15.3820776

Q=50

N	500	1000
Moyenne	4.60446972	5.73302805
Variance	8.85481597	11.9679381

Mais malgré le faible nombre de tests quand Q augmente, les résultats montrent que la variable z(q) ne suit pas une loi normale.