Comment partitionner automatiquement des marches aléatoires ? Avec application à la finance quantitative

Gautier MARTI[1,2], Frank NIELSEN[2], Philippe VERY[1], Philippe DONNAT[1]

[1]Hellebore Capital Management
63, avenue des Champs-Elysées, 75008 Paris, France

[2]Laboratoire d'Informatique de l'Ecole Polytechnique
1, rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France

gautier.marti@polytechnique.edu, nielsen@lix.polytechnique.fr
philippe.very@helleborecapital.com, philippe.donnat@helleborecapital.com

Résumé – Nous présentons dans cette communication une approche non paramétrique pour regrouper automatiquement des séries temporelles suivant une marche aléatoire. Nous introduisons d'abord une étape de pré-traitement qui consiste à transformer les réalisations indépendantes et identiquement distribuées des incréments du processus de Markov en un vecteur représentant sans perte toute l'information disponible de ces séries temporelles, et la factorisant en une composante dépendance et une composante distribution. Nous définissons ensuite une distance entre ces représentations tenant compte des deux types d'information et permettant d'en controler l'importance pour le partitionnement automatique à l'aide d'un seul paramètre. Ce paramètre de mélange peut être appris ou manipulé par un expert à des fins exploratoires comme illustré par l'étude des séries temporelles financières. Des expériences, implémentations et résultats sont disponibles sur `http://www.datagrapple.com`.

Abstract – We present in this paper a novel non-parametric approach useful for clustering Markov processes. We introduce a pre-processing step consisting in mapping multivariate independent and identically distributed samples from random variables to a generic non-parametric representation which factorizes dependency and marginal distribution apart without losing any. An associated metric is defined where the balance between random variables dependency and distribution information is controlled by a single parameter. This mixing parameter can be learned or played with by a practitioner, such use is illustrated on the case of clustering financial time series. Experiments, implementation and results obtained on public financial time series are online on a web portal `http://www.datagrapple.com`.

1 Introduction

Les marches aléatoires peuvent être utilisées pour partitionner les données, elles constituent par exemple un point de vue de la classification spectrale [7]. Dans cette communication, nous nous intéresserons au problème inverse : partitionner des marches aléatoires. Ces processus stochastiques sont un important outil de modélisation des séries temporelles financières, savoir les regrouper dans des groupes homogènes statistiquement peut permettre d'établir de meilleurs indicateurs de risque que la simple « valeur à risque ». Pour effectuer ce partitionnement automatique des marches aléatoires, nous devons disposer d'une représentation de celles-ci ainsi que d'une distance entre les représentations. En général, représentation et distance idoines ne sont pas connues et des heuristiques sont utilisées comme les deux décrites en légende de la Figure 1. Dans le cas restreint des séries temporelles s'écrivant comme la somme $\sum_i X_i$ de variables aléatoires X_i indépendantes et identiquement distribuées (i.i.d.), nous proposons en Section 2 distance et représentation adaptées et mathématiquement fondées. Celles-ci travaillent sur la série temporelle des incréments X_i portant toute l'information des marches aléatoires considérées. Finalement, en Section 3 nous présentons brièvement une application aux séries temporelles financières. Pour une étude plus approfondie et davantage d'expériences, le lecteur pourra se référer à `http://www.datagrapple.com`, portail se consacrant au partitionnement automatique des séries temporelles, notamment issues du marché des couvertures de défaillance.

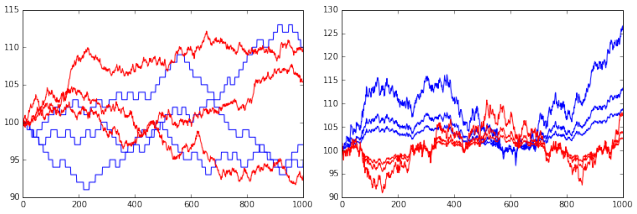

FIGURE 1 – Pour regrouper ces exemples de marches aléatoires, deux critères sont utilisés : pour celles de gauche, la forme du signal ; celles de droite sont similaires à transformations homothétiques près.

2 Une représentation non paramétrique des marches aléatoires

Soit $(\Omega, \mathcal{F}, \mathbf{P})$ un espace de probabilité. Soit $\mathcal{V}$ l'espace des variables aléatoires réelles continues définies sur $(\Omega, \mathcal{F}, \mathbf{P})$. Soient $\mathcal{U}$ l'espace des variables aléatoires suivant une loi uniforme sur $[0,1]$ et $\mathcal{G}$ l'espace des fonctions de répartitions absolument continues. Nous définissons maintenant une représentation non paramétrique des vecteurs aléatoires qui capture et sépare sans perte la partie comportement joint des variables de leur distribution propre. Soit $\mathcal{T}$ l'application qui associe à un vecteur aléatoire $X = (X_1, \ldots, X_N)$ sa représentation non paramétrique, élément de $\mathcal{U}^N \times \mathcal{G}^N$, définit comme suit :

$$\begin{aligned} \mathcal{T} : \mathcal{V}^N &\rightarrow \mathcal{U}^N \times \mathcal{G}^N \\ X &\mapsto (G_X(X), G_X) \end{aligned} \tag{1}$$

où $G_X = (G_{X_1}, \ldots, G_{X_N})$, G_{X_i} étant la fonction de répartition de X_i.

$\mathcal{T}$ est une bijection et ainsi préserve la totalité de l'information. La Figure 2 illustre cette projection sur un exemple concret issu de la finance. On peut remarquer que ce résultat réplique le théorème de Sklar [6], résultat fondateur de la théorie des copules. Néanmoins, nous n'utilisons pas ici le cadre générique de cette théorie et nous verrons par la suite où cette analogie s'arrête. Nous exploitons ensuite cette représentation pour définir une distance d_θ entre les variables aléatoires qui prend en compte à la fois la distribution des marginales et leur comportement joint.

Soit $(X, Y) \in \mathcal{V}^2$. Soient G_X, G_Y leur fonction de répartition. Nous définissons la distance suivante, dépendante du paramètre $\theta \in [0,1]$:

$$d_\theta^2(X,Y) = \theta d_1^2(G_X(X), G_Y(Y)) + (1-\theta) d_0^2(G_X, G_Y),$$

avec

$$d_1^2(G_X(X), G_Y(Y)) = 3\mathbf{E}[|G_X(X) - G_Y(Y)|^2], \tag{2}$$

et

$$d_0^2(G_X, G_Y) = \frac{1}{2} \int_{\mathbf{R}} \left(\sqrt{\frac{dG_X}{d\lambda}} - \sqrt{\frac{dG_Y}{d\lambda}} \right)^2 \mathrm{d}\lambda. \tag{3}$$

En particulier, nous obtenons d_0 la distance d'Hellinger, f-divergence qui quantifie la similarité entre deux distributions et qui garantit la monotonicité de l'information, propriété qui assure que la distance entre des histogrammes grossiers est moindre que la distance entre des histogrammes plus précis ; $d_1 = \sqrt{(1-\rho_S)/2}$ est une distance de corrélation mesurant la dépendance statistique entre deux variables aléatoires à l'aide de ρ_S, corrélation de Spearman entre X et Y. Remarquons que pour $\theta \in [0,1]$, $0 \leq d_\theta \leq 1$ et pour $0 < \theta < 1$, d_θ est une distance métrique. Pour $\theta = 0$ ou $\theta = 1$, l'axiome de séparation n'est pas vérifié. Cette distance est également invariante par transformations monotones, propriété désirable car elle affranchit de l'arbitraire du choix des unités ou de la méthode de mesure (que ce soit l'appareillage ou la modélisation mathématique) du signal.

Pour appliquer la distance proposée sur des données échantillonnées, nous définissons alors une estimation de d_θ. La distance d_1 travaillant avec des distributions uniformes continues peut être approximée de manière discrète par des statistiques de rang qui en sus d'être robustes aboutissent à une analogie avec le formalisme des copules : la statistique de rang utilisée correspond à une coordonnée de la copule empirique de Deheuvels [1] qui est un estimateur non paramétrique et non biaisé convergeant uniformément [2] vers la copule sous-jacente au processus. La distance d_0 peut être approximée par sa forme discrète travaillant sur une estimation des densités marginales obtenues par histogrammes, par exemple. Pour calculer d_1, nous avons besoin d'une fonction de rang bijective et puisque nous considérons l'application aux séries temporelles, il est naturel de privilégier l'ordre d'arrivée pour départager les égalités.

Soient $(X_i)_{i=1}^M$ les M réalisations de $X \in \mathcal{V}$. Soit S_M le groupe des permutations de $\{1, \ldots, M\}$ et $\sigma \in S_M$ une permutation quelconque, disons $\sigma = Id_{\{1,\ldots,M\}}$. Une fonction de rang bijective pour $(X_i)_{i=1}^M$ peut être définie comme une fonction

$$\begin{aligned} \mathrm{rk}^X : \{1, \ldots, M\} &\rightarrow \{1, \ldots, M\} \\ i &\mapsto \#\{k \in \{1, \ldots, M\} \mid \mathcal{P}_\sigma\} \end{aligned} \tag{4}$$

avec $\mathcal{P}_\sigma \equiv (X_k < X_i) \vee (X_k = X_i \wedge \sigma(k) \leq \sigma(i))$.

Soient $(X_i)_{i=1}^M$ et $(Y_i)_{i=1}^M$ les M réalisations des variables aléatoires $X, Y \in \mathcal{V}$. Une distance empirique entre les réalisations de ces variables aléatoires peut être définie par

$$\tilde{d}_\theta^2 \left((X_i)_{i=1}^M, (Y_i)_{i=1}^M \right) \overset{a.s.}{=} \theta \tilde{d}_1^2 + (1-\theta) \tilde{d}_0^2, \tag{5}$$

avec

$$\tilde{d}_1^2 = \frac{3}{M^2(M-1)} \sum_{i=1}^{M} \left(\mathrm{rk}^X(i) - \mathrm{rk}^Y(i) \right)^2 \tag{6}$$

et

$$\tilde{d}_0^2 = \frac{1}{2} \sum_{k=-\infty}^{+\infty} \left(\sqrt{g_X^h(hk)} - \sqrt{g_Y^h(hk)} \right)^2, \tag{7}$$

le paramètre h étant un paramètre de lissage approprié, et $g_X^h(x) = \frac{1}{M} \sum_{i=1}^M \mathbf{1}\{\lfloor \frac{x}{h} \rfloor h \leq X_i < (\lfloor \frac{x}{h} \rfloor + 1)h\}$ étant un histogramme de densité estimant la fonction de densité de probabilité g_X à partir des $(X_i)_{i=1}^M$, les M réalisations de la variable aléatoire $X \in \mathcal{V}$.

3 Application au partitionnement automatique de séries temporelles financières

Nous illustrons notre approche sur les séries temporelles des volumes traités sur le marché des couvertures de défaillance [3] (CDS). Nous prenons en compte les $N = 658$ actifs ayant des volumes reportés depuis juillet 2010. En sus d'être des données accessibles publiquement (fournies par DTCC - `http:`

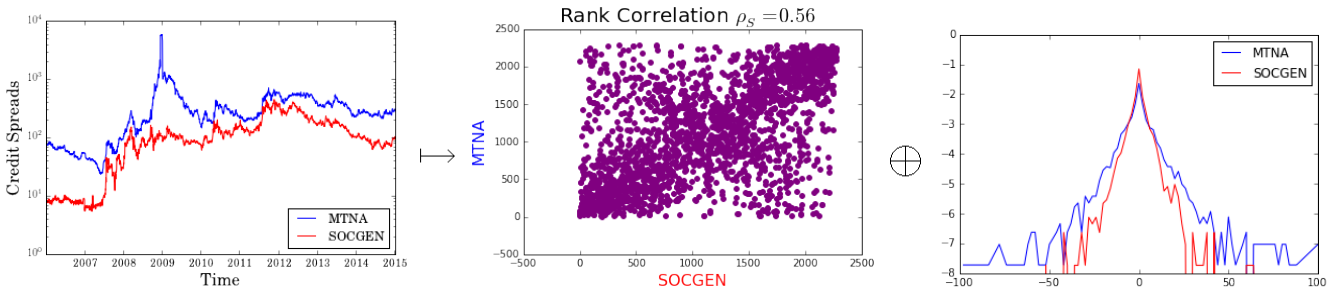

FIGURE 2 – L'approche présentée en résumé : deux séries temporelles sont projetées sur l'espace dépendance $\oplus$ distribution.

//www.dtcc.com/) contrairement aux prix des CDS, ces séries temporelles sont très bruitées et font montre de moins de corrélations évidentes que les séries de prix [4] (cf. Figure 3 et Figure 4 pour une comparaison), ce qui rend ce jeu de données intéressant pour notre méthode. A notre connaissance, il s'agit de la première fois qu'un papier s'intéresse au regroupement automatique de séries temporelles des volumes traités sur un marché financier.

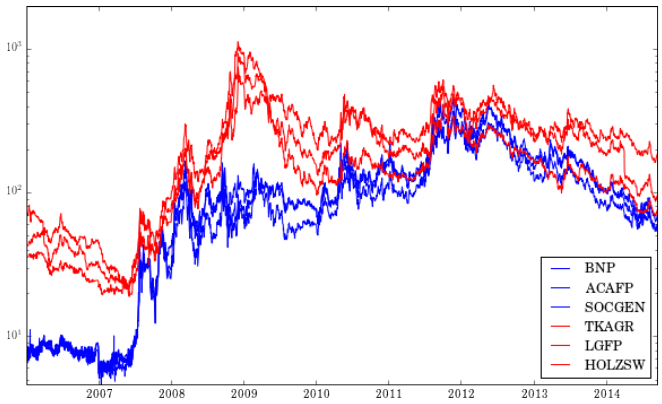

FIGURE 3 – Les prix de CDS de deux industries entre janvier 2006 et janvier 2015 : les entreprises financières françaises (en bleu) et les cimentiers (en rouge) ; observez la corrélation importante à l'intérieur de chaque secteur industriel.

Notre but est de comprendre comment ces séries temporelles se regroupent lorsque nous considérons uniquement leur comportement joint (notre approche avec $\theta = 1$) ou en se concentrant seulement sur la proximité de la distribution de leurs volumes traités (notre approche avec $\theta = 0$), et finalement lorsque nous prenons en compte la totalité de l'information (notre approche avec $\theta = 0.5$). Nous estimons d'abord le nombre de groupes dans chaque cas grâce à un critère de stabilité [5] et nous trouvons $K_1 = 3$, $K_0 = 5$ et $K_{0.5} = 7$ respectivement.

La Table 1 affiche quelques caractéristiques (espérance et quantiles) de la distribution des $K_{0.5} = 7$ groupes trouvés en utilisant la totalité de l'information. Nous pouvons remarquer

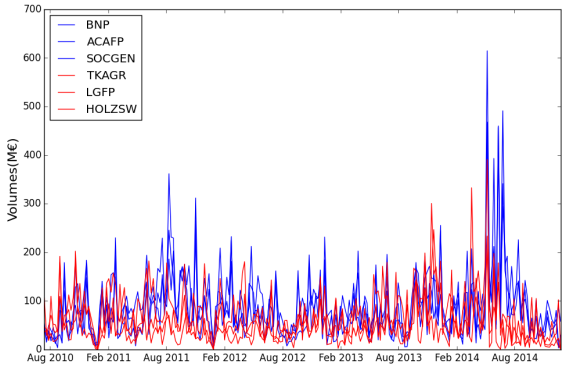

FIGURE 4 – Les volumes de CDS traités selon DTCC ; En bleu, les entreprises financières françaises et en rouge les volumes traités sur les cimentiers tels que reportés entre juillet 2010 et janvier 2015.

que ces groupes correspondent en fait aux $K_0 = 5$ groupes trouvés en utilisant uniquement l'information de distribution dont les espérances et quantiles sont reportés dans la Table 2. Cependant, ces indicateurs sur la distributions ne permettent pas d'expliquer les différences entre les groupe 3 et 4 qui se ressemblent pour ces mesures, idem pour les groupes 5 et 6.

Concernant $\{C_1^{0.5}, C_2^{0.5}, C_7^{0.5}\}$, nous pouvons d'ores et déjà constater que $C_1^{0.5}$ est composé des CDS ayant un important volume traité, notamment les CDS sur la dette souveraine de pays tels que le Brésil, la Chine, l'Allemagne, la France, l'Italie, la Russie et l'Espagne. $C_2^{0.5}$ est constitué des entreprises financières ainsi que de quelques fournisseurs d'énergie qui représentent les entités les plus activement traitées sur le marché des couvertures de défaillance, en dehors des dettes souveraines. $C_7^{0.5}$ se compose des entreprises asiatiques, notamment japonaises, dont les CDS sont relativement peu traités, les rendements étant très faibles. Pour comprendre les différences entre les groupes $C_3^{0.5}$, $C_4^{0.5}$ et $C_5^{0.5}$, $C_6^{0.5}$, nous étudions les résultats du regroupement automatique en utilisant seulement les comportements joints, c'est-à-dire les $K_1 = 3$ groupes $\{C_1^1, C_2^1, C_3^1\}$. C_1^1 est essentiellement composé d'entités ayant une liquidité croissante, c'est-à-dire une tendance

TABLE 1 – Les $K_{0.5} = 7$ groupes obtenus avec $\theta = 0.5$

	$C_1^{0.5}$	$C_2^{0.5}$	$C_3^{0.5}$	$C_4^{0.5}$	$C_5^{0.5}$	$C_6^{0.5}$	$C_7^{0.5}$
Mean	441	84	32	29	17	17	8
Quantile 10%	116	46	18	17	8	5	4
Quantile 90%	924	141	50	44	29	36	15
Size	13	89	169	79	161	90	57

TABLE 2 – Les $K_0 = 5$ groupes obtenus avec $\theta = 1$

	C_1^0	C_2^0	C_3^0	C_4^0	C_5^0
Mean	458	92	40	22	10
Quantile 10%	196	60	29	16	4
Quantile 90%	924	139	51	29	15

haussière des volumes traités, et correspond au groupe $C_6^{0.5}$. C_2^1 contient les CDS des entreprises européennes considérées comme étant sûres par les agences de notations, ce marché est connu pour être très fortement corrélé en comparaison de ses équivalents américain et asiatique. C_3^1 semble rassembler le reste des actifs ne partageant pas de points communs évidents.

Nous pensons que ces volumes traités constituent un jeu de données intéressant pour illustrer l'usage de notre méthode car cela montre le gain qu'on obtient à exploiter l'information totale disponible dans ces marches aléatoires. En sus, nous trouvons que le regroupement automatique optimal (d'un point de vue de la stabilité des groupes par rapport à des petites perturbations) est constitué des groupes qui sont eux-mêmes résultats optimaux des regroupements automatiques lorsque l'algorithme travaille seulement sur la partie « dépendance » de l'information ou seulement sur la partie « distribution » : les CDS sont regroupés en 5 groupes pouvant être expliqués par le volume moyen traité et qui résume approximativement l'information de distribution, cependant deux groupes supplémentaires émergent à cause de l'information sur les comportements joints qui raffine cette partition en 5 groupes : un groupe émerge à cause des fortes corrélations présentes dans le marché européen des actifs sûrs, et l'autre rassemble les entités dont le volume des transactions est en augmentation (Figure 5).

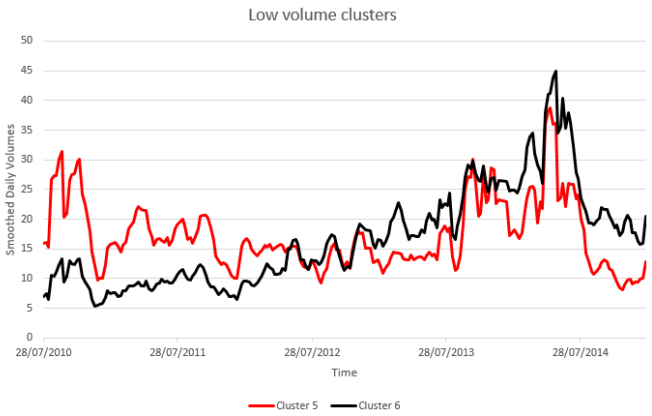

FIGURE 5 – Des dynamiques inverses pour $C_5^{0.5}$ et $C_6^{0.5}$

4 Discussion

Dans cette communication, nous avons présenté une nouvelle représentation, mathématiquement fondée, des séries temporelles suivant une marche aléatoire. Cette représentation peut être utilisée pour le partitionnement automatique des séries temporelles comme illustré en Section 3 par l'exemple des volumes traités, mais est également adaptée à l'apprentissage supervisé. Dans cette communication, nous avons montré son utilité sur des données réelles, néanmoins nous avons également validé l'approche sur des cas tests engendrés par des modèles de corrélations hiérarchiques se subdivisant en groupes de distribution. Nous nous concentrons maintenant à prouver la consistance statistique d'une telle approche. Les résultats expérimentaux, des données ainsi que des implémentations, sont disponibles sur `http://www.datagrapple.com` se consacrant au partitionnement automatique de séries temporelles.

Remerciements

Merci à Valentin Geffrier et Benjamin d'Hayer pour leur relecture attentive, et Laurent Beruti pour son retour et son expertise sur le marché des CDS.

Références

[1] P. Deheuvels. *La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance*. Acad. Roy. Belg. Bull. Cl. Sci.(5), 1979.

[2] P. Deheuvels. *An asymptotic decomposition for multivariate distribution-free tests of independence*. Journal of Multivariate Analysis, 1981.

[3] J. Hull. *Options, futures, and other derivatives*. Pearson Education, 2006.

[4] D. Kane. *Modelling single-name and multi-name credit derivatives*. John Wiley & Sons, 2011.

[5] O. Shamir et T. Naftali. *Model selection and stability in k-means clustering*. Conference on Learning Theory, 2008.

[6] A. Sklar. *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8, 1959.

[7] U. Von Luxburg. *A tutorial on spectral clustering*. Statistics and computing, 2007.