

Projets MODAL MAP 474D
Simulation Numérique Aléatoire (SNA)
autour des évènements rares
Promotion X2015

Equipe pédagogique :
F. Benaych-Georges, M. Bompaire, S. De Marco,
G. Fort, E. Gobet, I. Kortchemski

Période : Avril 2017 - Juin 2017

(Version préliminaire du 14 avril 2017)

Table des matières

1 Matrices aléatoires	4
2 Identification de communautés	7
3 Risque de défaut d'entreprises	12
4 Saturation de réseaux	15
5 Réaction chimique	19
6 Évolution de produits concurrents	23
7 Risque systémique	25
8 Barrages et risques d'inondation	30
9 Risque de collision spatiale	34
10 Championnat de football	39

Objectifs. Il s'agit de réaliser un projet de modélisation et simulation sur un des thèmes applicatifs décrits ci-dessous, dans le contexte des évènements rares.

Des références bibliographiques sont données dans le descriptif de chaque projet, ces documents sont accessibles sur le moodle du cours ou via la BCX. Prendre connaissance de ces références est important et fait partie du travail d'expérimentation et d'investigation du MODAL. Par cette démarche on attend des étudiants qu'ils :

- se familiarisent avec le thème proposé,
- en comprennent les enjeux de sorte à se poser les *bonnes questions*,
- développent des outils stochastiques de simulation pour mieux appréhender les évènements rares (quantile, distribution conditionnelle à l'évènement, statistique dans de tels évènements, scénario rare typique...). Ils auront en particulier à adapter à leur contexte les méthodes abordés en cours ou en TP.

Cela implique de l'autonomie et des initiatives.

Il est demandé de tester différentes approches, de les *comparer quantitativement et qualitativement*, avec un regard critique.

Encadrement : une permanence de suivi de projets est assurée chaque vendredi après-midi, de 13h30 à 15h30.

Livrables : un rapport écrit, avec les codes de simulation en annexe.

Outils informatiques : il est recommandé de choisir **Python** comme langage de développement.

L'utilisation de **Scilab** est aussi possible. En cas de besoin en calculs particulièrement intensifs, on pourra recourir à **Java** ou **C/C++**.

Soutenance : elle est prévue **vendredi 9 juin 2017**, d'une durée de 45' (30' de présentation, 15' de questions).

Barême de notation du projet : le projet est noté sur 20 pts.

- (4pts) Capacités d'initiatives/autonomie face aux difficultés/problèmes soulevés par le projet.
- (4pts) Capacités à décomposer le problème en plusieurs étapes intermédiaires, avant de parvenir à une version aboutie (démarche d'expérimentation).

- (5pts) Capacités à mesurer les améliorations, les apports, comparaison numérique.
- (4pts) Rapport écrit.
- (3pts) Soutenance orale.

Les 3 premiers items sont principalement notés pendant les séances de TPs.

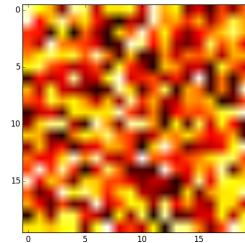
1 Matrices aléatoires

1.1 Contexte

La théorie des matrices aléatoires a de nombreuses applications en analyse statistique multivariée, surtout à l'ère des données massives. En particulier, l'étude des matrices de covariance empirique est fondamentale.

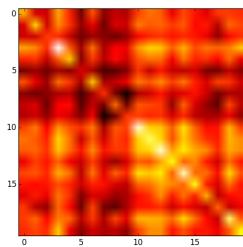
Considérons un signal aléatoire $\mathbf{x} \in \mathbb{R}^p$ de dimension p , que l'on supposera gaussien, centré (quitte à estimer au préalable l'espérance et à recentrer) et de matrice de covariance notée C . Cette matrice donne les corrélations entre les composantes x_i du signal, information essentielle dans de nombreux contextes applicatifs.

Par exemple dans le cas où les x_i représentent les rendements mensuels de divers actifs boursiers et où ces corrélations expriment la corrélation, positive ou négative, entre divers secteurs d'activité économique (d'autres exemples sont donnés par exemple dans [Joh01], en études climatiques, où p est le nombre de stations météorologiques, ou pour un moteur de recherches où p est le nombre de features).



Matrice de taille 20×20
avec entrées aléatoires
indépendantes

1.2 Estimation de covariance



Matrice de covariance
empirique \widehat{C} avec $n = 30$ et
 $p = 20$

L'estimation de C est donc une question fondamentale, élémentaire, et pourtant non encore épousée. Supposons que l'on dispose de n copies indépendantes $\mathbf{x}_1, \dots, \mathbf{x}_n$ du signal \mathbf{x} (obtenues par exemple par des mesures régulières dans le cas de rendements mensuels d'actifs boursiers). Comme $C = \mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ (où \mathbf{x} est assimilé à un vecteur colonne), par la LGN, une approximation naturelle de C est donnée par

$$\widehat{C} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top = \frac{1}{n} X X^\top$$

pour $X := [\mathbf{x}_1 \cdots \mathbf{x}_n]$. Le problème est alors que l'approximation $\widehat{C} \approx C$ ne fonctionne bien que si non seulement n est grand, mais aussi p est petit face à $n!$ (voir par exemple [Ver16, Th. 3.1.2]).

- Par exemple, si $C = Id$ (cas correspondant dans la suite à l'*hypothèse nulle* H_0 où les composantes du signal forment un bruit blanc) lorsque $p \leq n$ sont grands, le spectre

$$\hat{l}_1 \geq \dots \geq \hat{l}_p$$

de \widehat{C} ne se concentre pas autours de 1 mais se distribue approximativement selon une loi à densité explicite, s'étalant de $(1 - \sqrt{p/n})^2$ à $(1 + \sqrt{p/n})^2$, c'est le théorème de Marcenko-Pastur (voir [Joh06, p. 316]). D'autre part, la plus grande valeur propre \hat{l}_1 converge en loi, lorsque $n, p \rightarrow \infty$ (voir [Joh06, eq. (6)] ou [BBP05, eq. (8)])

$$\frac{\hat{l}_1 - (1 + \sqrt{p/n})^2}{\sigma(n, p)} \xrightarrow{\text{loi}} W_1$$

où W_1 a la loi de Tracy-Widom et

$$\sigma(n, p) = \frac{(1 + \sqrt{n/p})^{4/3}}{\sqrt{n/p} n^{2/3}}.$$

- Un autre exemple intéressant est celui où $C \neq Id$ mais où le spectre de C comporte r valeurs propres égales à $1 + \theta_1 \geq \dots \geq 1 + \theta_r$ (pour des $\theta_i > 0$) et $p - r$ valeurs propres égales à 1 (cas correspondant dans la suite à l'*hypothèse non nulle* H_1 dudit *spiked model de rang r*). On a alors alors la transition de phase suivante (voir [BGR11, Sect. 3.2]) : pour $n \geq p$ grands et du même ordre, pour tout $i \geq 1$ très faible devant p ,

$$\hat{l}_i \approx \begin{cases} (1 + \sqrt{p/n})^2 & \text{si } i > r \text{ ou } \theta \leq \sqrt{p/n}, \\ (1 + \theta_i)(1 + \frac{p/n}{\theta_i}) & \text{si } i \leq r \text{ et } \theta_i > \sqrt{p/n}. \end{cases}$$

Par ailleurs, dans le deuxième cas, pour $r = 1$, le vecteur propre de \widehat{C} associé à \hat{l}_1 s'exprime en partie en fonction de celui de C associé à $1 + \theta_1$ (voir [BGR11, Sect. 3.2]).

Application statistique : si on veut tester si \mathbf{x} est un bruit blanc, un test suffisant consiste à examiner la valeur de \hat{l}_1 . Si $\hat{l}_1 \geq x$ avec x "grand" alors on rejette l'hypothèse H_0 . Le seuil x est relié à l'erreur de première espèce (probabilité de rejeter H_0 à tort), qu'on fixe par exemple à $\alpha = 1\%$, 0.1% ou 0.01% . Vue la relation de dominance stochastique, il suffit de calculer le seuil $x(\alpha)$ avec la distribution de \hat{l}_1 .

1.3 Objectifs

Ce projet va se concentrer sur quelques-unes des propriétés de \hat{l}_1 et des applications statistiques.

- 1) PREMIÈRES QUESTIONS À ABORDER :
 - Vérifier la distribution asymptotique de \hat{l}_1
 - Expérimenter la fiabilité du test

- 2) QUESTIONS PLUS AVANCÉES : on souhaite, sous l'hypothèse nulle, calculer des statistiques extrêmes de \hat{l}_1 , pour être en mesure de calculer des probabilités de 1ère espèce α très faibles, disons à 10^{-6} .
 - Mettre en œuvre des méthodes d'échantillonage préférentiel pour simuler les queues de l_1 .
 - Pour implémenter une méthode particulière, proposer une chaîne de Markov $\mathbf{y}_1, \dots, \mathbf{y}_k$ avec k étapes, tels que \mathbf{y}_i a la loi de \mathbf{x} à chaque date i . En déduire un calcul de α par systèmes de particules en interaction.

Références

- [AGZ09] G.W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*. Cambridge University Press, 2009.
- [BBP05] J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, 33(5) :1643–1697, 2005.
- [BGR11] F. Benaych-Georges, R.N. Rao *The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices*, Adv. Math. (2011), Vol. 227, no. 1, 494–521.
- [Joh01] I.M. Johnstone. *On the distribution of the largest eigenvalue in principal components analysis*. Annals of Statistics, 29(2) :295–327, 2001.
- [Joh06] I. M. Johnstone, *High Dimensional Statistical Inference and Random Matrices*, Proc. International Congress of Mathematicians 2006, 307–333.
- [Ve16] R. Vershynin *Four lectures on probabilistic methods for data science*, 2016 PCMI Summer School, AMS.

2 Identification de communautés

2.1 Introduction

L'identification de communautés est un problème standard dans des contextes très divers : en analyse économique, où l'on cherche à regrouper les entreprises ou des actifs financiers selon des caractéristiques communes, en linguistique en biologie, où l'on cherche à en faire autant avec les langues ou les espèces, etc... Ce problème est aussi actuellement au centre de nombreuses attentions avec l'avènement des réseaux sociaux : la mise en évidence de groupes d'utilisateurs homogènes d'un réseau social donné, basée, par exemple sur l'étude des liens "d'amitié" qu'ils y ont tissés, est utile dans la mise en place de la publicité ciblée (*targeted advertising* en anglais) mais aussi, par exemple, dans le cadre d'études sociologiques.

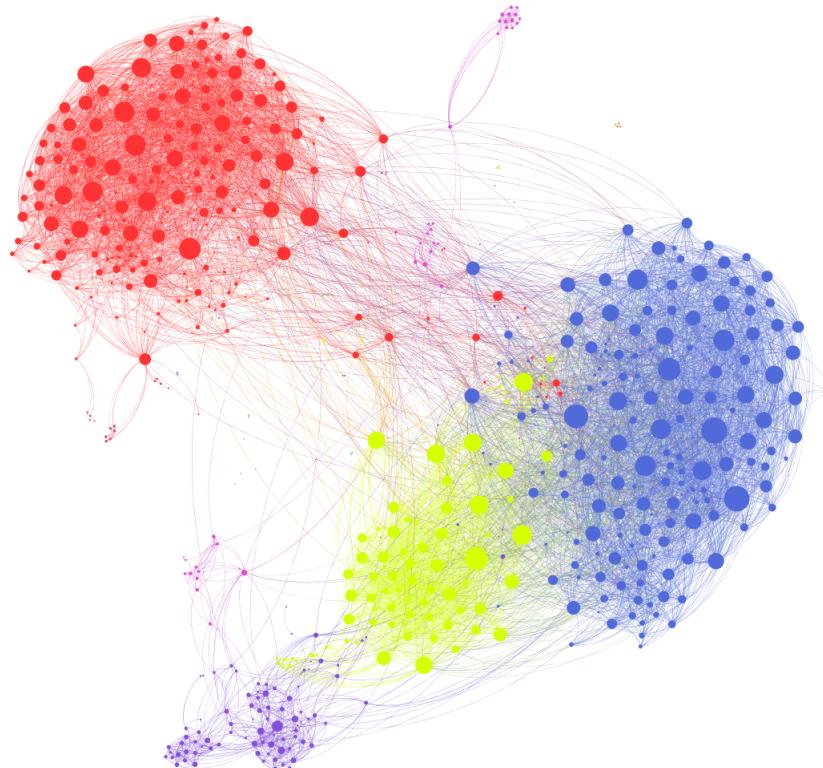


FIGURE 1 – Le réseau Facebook clusterisé à une date t

Les algorithmes sous-jacents s'appellent des algorithmes de *partitionnement de données* (ou *data clustering* en anglais) : méthode d'analyse des données qui vise à diviser un ensemble de *données* ou *observations* en différents

sous-ensembles, que l'on souhaite individuellement les plus homogènes possibles, et deux à deux les plus différents possibles, en ce sens que les données de chaque sous-ensemble partagent le plus de caractéristiques communes et que les données de sous-ensembles différents sont les plus dissemblables possibles. Les contextes d'utilisation du clustering étant très variés, comme on l'a vu plus haut, les types de données que l'on est amené à partitionner peuvent avoir des caractéristiques très diverses (taille des échantillons, nombre de *features* (i.e. dimension) de chaque observation, etc...), ce qui amène à des choix d'algorithmes tout aussi variés. Le problème auquel on s'intéresse ici est celui du clustering d'un ensemble $V = \{v_1, \dots, v_n\}$ d'individus (qui peuvent aussi être des entreprises, des langues, des espèces animales, etc...) intégrés dans un réseau d'*amitiés* : on observe donc un graphe $G = (V, E)$, où E est l'ensemble des liens d'amitié (i.e. $E \subset V \times V$). Le graphe est supposé non orienté, i.e. $(v, w) \in E \iff (w, v) \in E$.

Le choix d'un algorithme de clustering pour traiter un tel problème doit intégrer la difficulté essentielle suivante : dès que n est un peu grand, l'observation associée à chaque sommet v est un vecteur de $\frac{n-1}{2}$ booléens (donc, essentiellement, de dimension $\approx n$), ce qui rend les algorithmes de clustering classiques (EM, *k-means*, clustering hiérarchique [HTB09, JWHT13]) inadaptés, car ils fonctionnent bien avec des données de dimension raisonnable. Un des algorithmes alors couramment utilisés, qui repose sur une opération de *réduction de la dimension*, est le *clustering spectral* (voir [VL07]), dont nous allons utiliser une variante dans ce projet.

2.2 Présentation du modèle

On suppose qu'il existe une partition¹ de l'ensemble $V = \{v_1, \dots, v_n\}$ en k classes $\mathcal{C}_1, \dots, \mathcal{C}_k$ et une matrice symétrique $B = [B_{i,j}]_{i,j=1}^k$ à coordonnées dans $[0, 1]$ telle que le graphe G est obtenu en reliant les éléments de V de la façon suivante :

- pour tout $v, w \in V$, lorsque $(v, w) \in \mathcal{C}_i \times \mathcal{C}_j$, la probabilité qu'il existe une arête entre v et w est $B_{i,j}$,
- les différentes arêtes existent de façon indépendante.

Ce modèle, très couramment utilisé, s'appelle le *stochastic block model* (voir, par exemple, [BN10] ou l'introduction de [RCY11]).

Toute l'information contenue dans le graphe G est contenue dans la *matrice d'adjacence* $A = [\mathbb{1}_{(i,j) \in E}]_{i,j=1}^n$. On tentera donc d'identifier les classes $\mathcal{C}_1, \dots, \mathcal{C}_k$ par observation de A .

1. On rappelle qu'une partition d'un ensemble V est la décomposition de V en une *union* de sous-ensembles *deux à deux disjoints*, appelés ici *classes*.

On se focalisera ici sur le cas où les termes diagonaux de B sont tous égaux à un nombre p_{in} et où les termes non diagonaux de B sont tous égaux à un nombre $p_{\text{out}} < p_{\text{in}}$, ce qui signifie que l'on a une probabilité d'amitié intra-classe et une probabilité d'amitié inter-classe, et que celle-ci est plus faible que la première. On se placera dans le régime $n \gg 1$ et $p_{\text{in}}, p_{\text{out}}$ de l'ordre de $1/n$, plus précisément

$$p_{\text{in}} = c_{\text{in}}/n \quad p_{\text{out}} = c_{\text{out}}/n,$$

ce qui signifie que le nombre d'amis de chaque individu reste raisonnable, même lorsque n est très grand. Enfin, on supposera que les classes contiennent toutes un nombre d'individus de l'ordre de n : pour tout i ,

$$\text{Card}(\mathcal{C}_i) \approx n a_i,$$

avec $a_i \in]0, 1[$.

L'algorithme appelé *clustering spectral* est le suivant : on considère la matrice U de taille $n \times k$ constituée des k vecteurs propres de A (*option 1*) (ou, *option 2*, de $M = D^{-1/2}AD^{-1/2}$, avec D la matrice diagonale définie par $D_{i,i} = \sum_{j=1}^n A_{i,j}$) associés aux plus grandes valeurs propres sur lesquels on applique l'algorithme k -means², et on obtient une approximation des classes de départ.

Il est expliqué, dans [Ver16, Sect. 2.3], que si $k = 2$ et

$$c_{\text{in}} - c_{\text{out}} \gg \sqrt{\log(n)(c_{\text{in}} + c_{\text{out}})}, \quad (1)$$

alors cet algorithme fonctionne bien (et cela se généralise à d'autres valeurs de k). Ce seuil est dépassable : d'autres algorithmes, qui permettent d'aller un peu au delà de la condition (1), ont été développés dans [Mas13, MNS14, MNS13].

2.3 Travail des élèves

1) PREMIÈRES QUESTIONS À ABORDER :

- Simuler le stochastic block model, essayer de visualiser les résultats (en utilisant par exemple la bibliothèque `igraph`³ ou tout simplement la fonction `imshow`⁴)

2. L'algorithme k -means est présenté, par exemple, dans [JWHT13] et il a été écrit en Python, voir <http://scikit-learn.org/stable/>.

3. Voir <http://igraph.org/python/>, mais ne pas essayer de l'installer sur Mac.

4. Voir http://matplotlib.org/users/image_tutorial.html et <http://stackoverflow.com/questions/9707676/defining-a-discrete-colormap-for-imshow-in-matplotlib>

- Dans le cas $k = 2$ (auquel on pourra se limiter pour la suite), vérifier le fonctionnement de l'algorithme (options 1 et 2) sous la condition (1).
 - Expérimenter la fiabilité du clustering, comparer les fiabilités des options 1 et 2.
- 2) QUESTIONS PLUS AVANCÉES :
- Estimer, à l'aide de l'importance sampling, la probabilité qu'un ensemble fixé de sommets, de cardinal faible (commencer par 2), soit mal clusterisé.
 - Lorsque la condition (1) n'est presque plus satisfaite, estimer, à l'aide de l'importance sampling, la probabilité qu'un grand nombre de sommets soit mal clusterisé.

Références

- [HTB09] T. Hastie, R. Tibshirani, J. Friedman *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*, disponible en téléchargement libre à <http://statweb.stanford.edu/~tibs/ElemStatLearn/>
- [JWHT13] G. James, D. Witten, T. Hastie, R. Tibshirani *An Introduction to Statistical Learning with Applications in R*, disponible en téléchargement libre à <http://www-bcf.usc.edu/~gareth/ISL/getbook.html>
- [BN10] B. Karrer, M. E. J. Newman *Stochastic blockmodels and community structure in networks*, <http://arxiv.org/abs/1008.3926>
- [VL07] U. Von Luxburg *A Tutorial on Spectral Clustering*. Statistics and Computing, vol. 17(4), p. 395–416, 2007. http://www.informatik.uni-hamburg.de/ML/contents/people/luxburg/publications/Luxburg07_tutorial.pdf
- [Mas13] L. Massoulié *Community detection thresholds and the weak Ramanujan property*, <http://arxiv.org/abs/1311.3085>
- [MNS14] E. Mossel, J. Neeman, A. Sly *Stochastic Block Models and Reconstruction*, Probability Theory and Related Fields, to appear, 2014, <http://arxiv.org/abs/1202.1499>
- [MNS13] E. Mossel, J. Neeman, A. Sly *A proof of the block model threshold conjecture*, <http://arxiv.org/abs/1311.4115>
- [NN12] R. Nadakuditi, M. Newman *Graph spectra and the detectability of community structure in networks* Phys. Rev. Lett. 108 (2012), 188701. <http://web.eecs.umich.edu/~rajnrao/community12.pdf>

- [RCY11] K. Rohe, S. Chatterjee, B. Yu *Spectral clustering and the high-dimensional stochastic blockmodel*. Ann. Statist. 39 (2011), no. 4, 1878–1915. <http://arxiv.org/abs/1007.1684>
- [Ver16] R. Vershynin *Four lectures on probabilistic methods for data science*, 2016 PCMI Summer School, AMS.

3 Risque de défaut d'entreprises

Depuis une vingtaine d'années, sous l'impulsion du Comité de Bâle, se développe une vision globale des risques au niveau bancaire : risque de marché, risque opérationnel, risque de crédit.

Cette vision globale a été adoptée dès 1998, avec la Value At Risk (VaR), indicateur de risque mesurant le seuil de *pertes potentielles* à un horizon donné et pour un quantile donné. Dans ce projet, nous considérons le risque associé à la défaillance d'entreprises (dit risque de crédit ou risque de défaut).

Source Wikipedia : *Le risque de crédit est le risque qu'un emprunteur ne rembourse pas tout ou partie de son crédit aux échéances prévues par le contrat signé entre lui et l'organisme prêteur (généralement une banque). La maîtrise du risque de crédit est au cœur du métier du banquier car il détermine la rentabilité des opérations effectuées. Si l'établissement financier sous-évalue ce risque, le montant prêté et les intérêts dus ne seront pas perçus et viendront s'inscrire en perte. L'évaluation du risque de crédit passe par une bonne connaissance du client et, si c'est une entreprise, par une bonne évaluation de son projet et du secteur dans laquelle elle exerce son activité.*

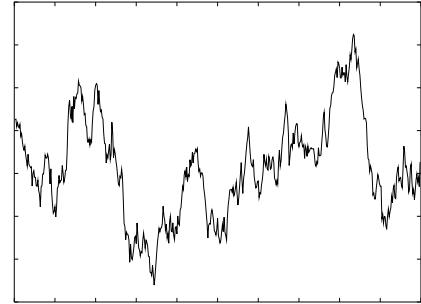
3.1 Ensemble d'entreprises

Nous considérons un ensemble de N entreprises dont les valeurs économiques sont des variables aléatoires qui peuvent évoluer au cours du temps, c'est-à-dire des processus aléatoires. On supposera que la valeur de l'entreprise i est modélisée par un mouvement brownien géométrique

$$S_t^i = S_0^i e^{-\frac{1}{2}\sigma_i^2 t + \sigma_i W_t^i}.$$

Précisons les notations :

- $W^i = (W_t^i)_{t \geq 0}$ est un mouvement brownien, c'est-à-dire un processus continu à temps continu, partant de $W_0^i = 0$, à accroissements indépendants et stationnaires, dont les accroissements $W_{t+h}^i - W_t^i$ sont de loi gaussienne $\mathcal{N}(0, h)$.
- σ_i mesure l'écart-type des rendements (appelé volatilité).



L'entreprise i est considérée en faillite si sa valeur est inférieure à une valeur B_i fixée, très inférieure à S_0^i .

Nous cherchons à évaluer la distribution du nombre de faillites à une date T future fixée (disons $T = 1$ an). Notons par D_T l'ensemble des indices des entreprises en défaut à la date T :

$$D_T = \{1 \leq i \leq N : S_T^i \leq B_i\}.$$

Nous nous intéressons à la distribution de

$$L = \#D_T.$$

Avoir beaucoup de faillites est un évènement rare, c'est-à-dire $\mathbb{P}(L \geq k) \ll 1$ pour k grand.

Pour les tests, on pourra prendre $N = 125$, $S_0^i = 100$, $\sigma_i = 40\%$ et $B_i = B$ pour un certain seuil $B > 0$.

3.2 Travail demandé

Evolution indépendante. Supposons d'abord que les mouvements browniens modélisant les entreprises sont indépendants.

▷ QUESTIONS À TRAITER :

1. Evaluer numériquement $\mathbb{P}(L \geq k)$ pour tous les k pour un seuil de faillite $B = 36$.
2. Analyser l'influence du seuil B .
3. On définit la perte associée au défaut à l'instant T par

$$P_T = \sum_{i \in D_T} R_i S_T^i$$

où R_i est un taux de recouvrement. Evaluer numériquement $\mathbb{E}(P_T \mid L \geq k)$ pour tous les k .

On pourra considérer d'abord le cas où $R_i = 30\%$, puis le cas de recouvrements indépendants et aléatoires de loi $\text{Beta}(a, b)$ de moyenne 30% et d'écart-type autour de 15%.

4. Mêmes questions lorsque $\sigma_i = 20\%$ pour $1 \leq i \leq 25$, $\sigma_i = 25\%$ pour $26 \leq i \leq 50$, $\sigma_i = 30\%$ pour $51 \leq i \leq 75$, $\sigma_i = 35\%$ pour $76 \leq i \leq 100$, et $\sigma_i = 50\%$ pour $101 \leq i \leq 125$.

Evolution dépendante. Supposons maintenant que les mouvements browniens sont corrélés positivement. Typiquement, la loi de $W_t = (W_t^1, \dots, W_t^N)$ est gaussienne, centrée, de matrice de covariance

$$\begin{pmatrix} 1 & \rho & \cdots & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \ddots & 1 & \ddots & \vdots \\ \rho & \cdots & \rho & 1 & \rho \\ \rho & \cdots & \cdots & \rho & 1 \end{pmatrix} t \quad \text{avec } \rho \in [0, 1].$$

▷ Reprendre les questions précédentes, tout en identifiant l'impact de la corrélation.

Extensions. Dans cette question, la faillite est observée à la première date t_k à laquelle la valeur de l'entreprise descend sous le seuil B_i :

$$\begin{aligned} \tau_i &= \inf \{t_k \text{ tq } S_{t_k}^i \leq B_i\}, \\ D'_T &= \{1 \leq i \leq N \text{ tq } \tau_i \leq T\}, \\ L' &= \#D'_T, \\ P'_T &= \sum_{i \in D'_T} R_i S_{\tau_i}^i. \end{aligned}$$

On prendra typiquement les dates t_k égales à une date par mois.

▷ Mêmes questions qu'avant.

Références

- [CC10] R. Carmona and S. Crépey. Particle methods for the estimation of credit portfolio loss distributions. *Int. J. Theor. Appl. Finance*, 13(4) :577–602, 2010.
- [CFV09] R. Carmona, J.P. Fouque, and D. Vestal. Interacting particle systems for the computation of rare credit portfolio losses. *Finance Stoch.*, 13(4) :613–633, 2009.
- [GHS00] P. Glasserman, P. Heidelberger, and P. Shahabuddin. Variance reduction techniques for estimating value-at-risk. *Management Science*, 46 :1349–1364, 2000.
- [Gla03] P. Glasserman. *Monte-Carlo methods in Financial Engineering*. Springer Verlag, New York, 2003.

4 Saturation de réseaux

4.1 Le modèle

Considérons l'émetteur d'un réseau de télécommunication fonctionnant de la façon suivante : des paquets (les messages) lui arrivent à des instants aléatoires, ils sont alors acheminés dans leur ordre d'arrivée. Le temps nécessaire à l'acheminement de chaque paquet est non négligeable. Il est donc possible qu'un paquet arrive sans pouvoir être traité tout de suite : avant d'être envoyés, les paquets sont alors stockés dans une mémoire, appelée *mémoire tampon* de l'émetteur. La mémoire tampon n'est pas de taille infinie, elle doit être dimensionnée de façon à ce que sa saturation soit très rare.

On modélise les instants d'arrivée des paquets avec un processus ponctuel de Poisson sur \mathbb{R}^+ d'intensité $\lambda > 0$: les paquets arrivent à des instants $T_1 < T_2 < T_3 < \dots$ tels que, en posant $T_0 = 0$, les intervalles de temps $(T_i - T_{i-1})_{i \geq 1}$ sont des v.a.i.i.d. de loi $\mathcal{E}(\lambda)$ ⁵. Les durées d'envoi E_1, E_2, \dots (E_i est la durée d'envoi du $i^{\text{ème}}$ message) des messages sont des v.a.i.i.d. indépendantes des instants d'arrivées de loi $\mathcal{E}(\mu)$, où $\mu > 0$. On note X_t l'encombrement de la mémoire tampon (message en cours d'envoi inclus) à l'instant t . On a donc

$$X_t = \begin{cases} 0 & \text{si } t < T_1 \\ 1 & \text{si } T_1 \leq t < \min\{T_2, T_1 + E_1\} \\ 0 & \text{si } T_1 + E_1 < T_2 \text{ et } T_1 + E_1 \leq t < T_2 \\ 2 & \text{si } T_1 + E_1 > T_2 \text{ et } T_2 \leq t < \min\{T_1 + E_1, T_3\} \\ \vdots & \end{cases}$$

Le processus $(X_t)_{t \in \mathbb{R}^+}$ est connu dans la littérature sous le nom de file d'attente de type M/M/1 (voir par exemple [Bou02] ou même [Wik]). Il pourra être simulé, selon le contexte, de façon exacte ou en discréétisant le temps. Afin de discréétiser le temps, il convient de bien comprendre l'évolution du processus (X_t) : il découle de la propriété d'*absence de mémoire* des lois exponentielles que connaissant la valeur de X_t à un certain instant $t = t_0$, on

5. Pour tout $\alpha > 0$, on note $\mathcal{E}(\alpha)$ la loi exponentielle de paramètre $\alpha > 0$ (*i.e.* la loi sur \mathbb{R}_+ de fonction de répartition $1 - e^{-\alpha t}$ et d'espérance $1/\alpha$). Une v.a. Y distribuée selon $\mathcal{E}(\alpha)$ est *sans mémoire* : pour tout $t > 0$, la loi de $X - t$ sachant que $X > t$ est encore la loi $\mathcal{E}(\alpha)$.

peut simuler son devenir $(X_t)_{t \in [t_0, +\infty[}$ de la façon suivante :

1er cas : $X_{t_0} = 0$ (aucun message en attente d'envoi) : on tire une variable aléatoire T de loi $\mathcal{E}(\lambda)$, X_t garde la valeur 0 pour $t_0 \leq t < t_0 + T$, et prend la valeur 1 en $t = t_0 + T$.

2ème cas : $X_{t_0} \geq 1$ (au moins un message en attente d'envoi) : on tire deux variables aléatoires T et E , indépendantes et de lois respectives $\mathcal{E}(\lambda)$ et $\mathcal{E}(\mu)$, X_t garde la valeur de X_{t_0} pour $t_0 \leq t < t_0 + \min\{T, E\}$, et, en $t = t_0 + \min\{T, E\}$, X_t prend la valeur $X_{t_0} + 1$ si $T < E$ et la valeur $X_{t_0} - 1$ si $T \geq E$. Mais il est simple de voir que la loi de $\min\{T, E\}$ est $\mathcal{E}(\lambda + \mu)$ et que, connaissant la valeur de $\mathcal{E}(\lambda + \mu)$, la probabilité que ce minimum soit T (resp. E) est $\frac{\lambda}{\lambda+\mu}$ (resp. $\frac{\mu}{\lambda+\mu}$) : on peut donc ne tirer qu'une variable aléatoire Y de loi $\mathcal{E}(\lambda + \mu)$ ainsi qu'une variable aléatoire ε qui vaut 1 ou -1 avec probabilités respectives $\frac{\lambda}{\lambda+\mu}$ et $\frac{\mu}{\lambda+\mu}$ et laisser X_t garder la valeur de X_{t_0} pour $t_0 \leq t < t_0 + Y$, et, en $t = t_0 + Y$, poser $X_t = X_{t_0} + \varepsilon$.

On réitère ensuite le même procédé pour déterminer l'évolution de X_t au delà.

4.2 Objectifs

Il s'agit tout d'abord de comprendre comment le processus X_t évolue (tendance à la stabilité ou non, valeur moyenne) en fonction des paramètres λ et μ . Au delà de ce préliminaire, on choisit $\lambda < \mu$ et, pour une taille $N \gg 1$ de la mémoire tampon et un seuil de temps $S \gg 1$, on essaye d'estimer l'espérance du *temps de saturation* :

$$T_N := \min\{t \geq 0 ; X_t = N\}$$

et la probabilité de l'événement de *saturation avant S* :

$$T_N \leq S.$$

On dimensionnera ensuite la taille N de la mémoire tampon de façon à ce que

$$\mathbb{E}[T_N] > S$$

ou

$$\mathbb{P}(T_N \leq S) \leq \alpha$$

avec S un seuil élevé et α un niveau de risque faible.

4.3 Feuille de route

a) Les deux premiers enjeux de ce projet sont :

1. La détermination d'une condition sur λ et μ pour la *stabilité* du processus (*i.e.* pour que la valeur de X_t n'ait pas tendance à exploser lorsque t devient grand).
2. La maîtrise de la *discrétisation en temps* du processus. Il va être en effet profitable de le remplacer par une chaîne de Markov à temps discret, mais il sera nécessaire de calibrer le pas de temps choisi de façon à ce que le processus à temps discret garde le même comportement que le processus continu (ce qui impose un pas de temps relativement petit), sans impliquer trop de calculs (ce qui exclut un pas de temps excessivement petit).

On pourra donc commencer par simuler le processus en temps continu et en temps discret et conjecturer un critère de stabilité, pour lequel on pourra aussi se documenter *via* [Bou02] ou [Wik]. On calculera $\mathbb{E}[T_N]$ et $\mathbb{P}(T_N \leq S)$ pour des valeurs raisonnables de N et S et on en déduira une calibration de la discrétisation du temps de façon à ce que le modèle discret soit proche du modèle continu. On rappelle que l'analogue discret d'une loi exponentielle est la loi géométrique de même espérance. Pour des exemples de valeurs numériques pour λ et μ , on pourra consulter la Table II de l'article [PW89].

b) De façon à aborder le cas où N est relativement élevé, faire un changement de probabilité pour le modèle discret en modifiant de façon intuitive les paramètres λ et μ , estimer $\mathbb{E}[T_N]$ et $\mathbb{P}(T_N \leq S)$ (on pourra éventuellement s'inspirer de la partie II de [PW89]).

c) Toujours pour le modèle discret, afin d'estimer $\mathbb{E}[T_N]$ et $\mathbb{P}(T_N \leq S)$ pour N élevé, optimiser son changement de probabilité.

4.4 Extensions

Pour la suite du projet, au moins une direction parmi les suivantes devra être prise :

d) On peut modifier le modèle de façon à tenir compte de la *taille* des paquets : X_t est alors la somme des tailles des paquets arrivés et non envoyés jusqu'à l'instant t . La taille d'un paquet sera une variable aléatoire, éventuellement dépendante de sa durée d'envoi. Il existe de multiples façons de générer de la dépendance entre deux v.a. T et D . En voici quelques exemples :

- on peut choisir D comme étant une fonction de T ,
 - on peut choisir D comme étant une fonction de T et d'une autre variable aléatoire indépendante de T (par exemple $D = \alpha cT + (1 - \alpha)Y$, avec $\alpha \in]0, 1[$ et $c > 0$ des coefficients et Y une v.a. indépendante de T de loi exponentielle),
 - On peut choisir D comme étant distribuée selon une loi exponentielle dont le coefficient est une fonction de la valeur de T ,
 - ...
- d') Supposons maintenant que, avant d'être envoyés, les messages doivent être traités (par exemple compressés) : avant son envoi, chaque message arrivant fait l'objet d'un traitement nécessitant un temps de loi $\mathcal{E}(\tau)$ indépendant des autres aléas et ne pouvant être réalisé sur deux paquets en même temps. En s'inspirant de la théorie des réseaux de Jackson à guichets multiples (voir par exemple le dernier chapitre de [Bou02]), on essayera de mettre en évidence un critère de stabilité (pour l'ensemble des deux files d'attente) et de calibrer la taille de la mémoire tampon.
- d") On pourra essayer d'utiliser la théorie des grandes déviations (voir [Buc04] ou [SW95]) pour proposer un autre échantillonnage préférentiel. On pourra notamment utiliser Theorem 11.3 dans [SW95] pour identifier la trajectoire typique menant X_t à l'état $N \gg 1$.

Références

- [Bou02] P. Bougerol. *Processus de sauts et files d'attente*. Cours de M1 en téléchargement libre à <http://www.proba.jussieu.fr/supports.php>, 2002.
- [Buc04] J.A. Bucklew. *Introduction to Rare Event Simulation*. Springer, 2004.
- [PW89] S. Parekh and J. Walrand. A quick simulation method for excessive backlogs in networks of queues. *IEEE Trans. Automat. Control*, <http://www.ece.ucdavis.edu/~chuah/classes/EEC273/refs/PW89-quick-simulation.pdf>, 34 :54–66, 1989.
- [SW95] A. Schwartz and A. Weiss, editors. *Large Deviations for Performance Analysis*. Chapman and Hall, London, 1995.
- [Wik] Wikipedia. M/M/1 queue. http://en.wikipedia.org/wiki/M/M/1_queue.

5 Réaction chimique

5.1 Description

La simulation aléatoire occupe un grand rôle en chimie (ou biochimie) où elle est couramment utilisée notamment pour explorer des conformations de moindre énergie (comme par exemple le repliement des protéines ; voir aussi [PotM1]). On se propose de modéliser au niveau microscopique le taux d'une réaction chimique.

La transformation d'un réactif R en un produit P , écrite $R \rightarrow P$, peut se voir comme la transition d'une molécule (ou d'un ensemble de molécules) de l'état R dans un état P . On suppose ces différents états caractérisés par une énergie potentielle E . Cette énergie dépend a priori d'un vecteur $x \in \mathbb{R}^d$ décrivant la position des éléments (distances entre atomes, angles relatifs...). \triangleright **Modélisation.** Le modèle mis en œuvre (qui est purement classique et ne tient aucun compte des effets quantiques) est le suivant [Kra40] : chaque degré de liberté de la molécule subit une agitation thermique modélisée par un mouvement aléatoire et subit l'énergie potentielle.

On se place tout d'abord dans le cas 1D ($d = 1, x \in \mathbb{R}$) i.e. l'énergie E ne dépend que d'un seul paramètre. On se donne un pas de temps $h > 0$. L'évolution de ce vecteur x est décrite par la suite aléatoire $\{X_{nh}, n \geq 0\}$ définie par

$$X_{(n+1)h} = X_{nh} - E'(X_{nh}) h + \sqrt{2\varepsilon h} W_n,$$

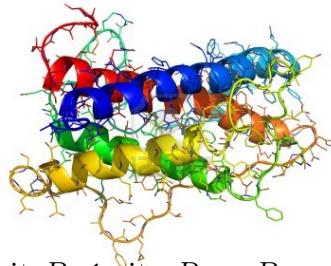
où $\varepsilon > 0$ est un paramètre petit (identifié à la température car il décrit l'agitation thermique) et $\{W_n, n \geq 0\}$ est une suite de vecteurs aléatoires i.i.d. de loi normale centrée réduite.

Dans les situations réelles, il est délicat de calculer l'énergie E ; des méthodes existent (voir [PotM1] et [SJ96] notamment sur l'énergie d'une liaison alcane $C - C - C - C$). On va supposer que E n'a que deux minima, qui correspondent aux états R et P , et qu'entre P et R , il y a alors un unique maximum T (T comme état de transition). On prendra dans un premier temps

$$E(x) = \frac{x^4}{4} - \frac{x^2}{2} + \frac{x^3}{3}. \quad (2)$$

5.2 Objectifs

\triangleright **Temps de transition.** Pour commencer, on étudie une seule molécule, et notamment son temps de transition τ nécessaire pour passer de l'état



R à l'état P , ou vice-versa. On peut avoir deux cas, soit $E(R) > E(P)$ soit $E(R) < E(P)$; on suppose que $E(R) > E(P)$. On suppose aussi que $E''(T) < 0$.

On prend $\varepsilon < \frac{1}{2}(E(R) - E(P))$. On s'intéresse à la probabilité $\mathbb{P}(\tau < t_0)$ qu'une transition ait lieu avant le temps t_0 , pour t_0 fixé.

1. Donner les états P , R et T pour l'énergie E donnée par (2).
2. Calculer numériquement la probabilité $\mathbb{P}(\tau < t_0)$ pour une molécule réagissant de R à P , ou de P à R . Expliquer de manière heuristique la différence entre les deux valeurs.
3. Quand ε tend vers 0, la variable aléatoire $\frac{\tau}{\mathbb{E}[\tau]}$ semble-t-elle converger en loi ? Identifier la loi.

On fera attention à l'influence de h sur les résultats.

▷ **Taux de réaction.** On va maintenant s'intéresser à un ensemble important de particules. On considère que les particules sont dans l'état R ou P si elles sont dans le bassin d'attraction de l'énergie correspondant : dans le cas du modèle (2), la particule X_{nh}^1 est dans l'état R si $X_{nh}^1 > 0$ et dans l'état P si $X_{nh}^1 < 0$. Dans d'autres cas de potentiel (voir les extensions), on pourra choisir de petits voisinages centrés autour des minima.

On appelle le taux de réaction, la valeur κ définie empiriquement par

$$\frac{d[P]}{dt} = \kappa[R] \quad (3)$$

où $[A]$ désigne la concentration de l'état A .

1. Calculer numériquement le taux de réaction κ ; comment évolue-t-il avec ε ? Comparer avec $\frac{1}{\mathbb{E}[\tau]}$. Expliquer heuristiquement l'équation (3) en fonction de la question 3 précédente.
2. On pourra comparer la valeur obtenue de $\mathbb{E}[\tau]$ (ou de κ) avec la valeur suivante (formule d'Eyring-Kramers [BEGK04, HTB90, Kra40])

$$T_0 = \frac{2\pi}{\sqrt{|E''(T)| E''(R)}} e^{(E(T)-E(R))/\varepsilon}$$

et justifier numériquement que ces valeurs coïncident lorsque ε tend vers 0.

5.3 Extensions : impact du potentiel

Le potentiel joue un rôle important dans ces problèmes. On cherchera à traiter au moins l'une des extensions suivantes.

- a) Si E ne vérifie pas les conditions précédentes, notamment si E est quartique en T i.e $E''(T) = 0, E^{(4)}(T) < 0$ ou si E non dérivable en T , analyser empiriquement comment évolue la vitesse de réaction en fonction de ε . La réaction a-t-elle lieu plus vite ou moins vite ?
- b) Que se passe-t-il si E possède un intermédiaire de réaction supplémentaire (voir par exemple figure 2) ? On pourra regarder plusieurs cas

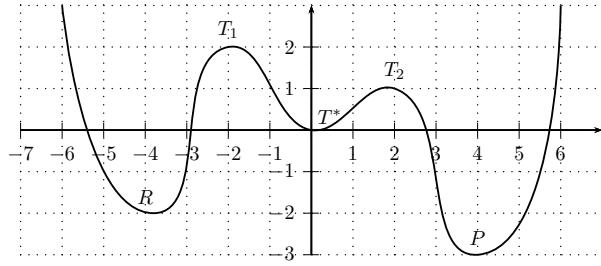


FIGURE 2 – Exemple de potentiel à trois puits : T^* produit intermédiaire de réaction.

dont notamment $E(T_1) = E(T_2)$. La vitesse de réaction est-elle plus ou moins favorisée par l'apparition de cet état intermédiaire ? Regarder aussi la stabilité de l'état T^* : est-il plus ou moins stable ? à quelle vitesse se transforme-t'il ? vers quels états ? On pourra pour cela estimer la probabilité $\mathbb{P}_{T^*}[\tau(R) < \tau(P)]$ qui représente la probabilité d'atteindre l'état R avant l'état P en partant de l'état T_* .

Enfin, on pourra étendre l'analyse en dimension 2 avec un potentiel bi-dimensionnel du type de la figure 3, c'est-à-dire $E(x, y) = \frac{x^4+y^4}{4} - \frac{x^2+y^2}{2} + \gamma \frac{(x-y)^2}{2}$ avec $\gamma < 1/2$.

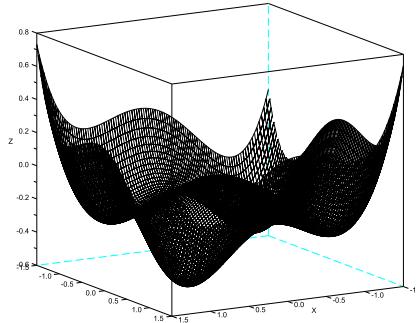


FIGURE 3 – Exemple de potentiel.

Références

- [BEGK04] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes. I. Sharp asymptotics for capacities and exit times. *J. Eur. Math. Soc. (JEMS)*, 6(4) :399–424, 2004.
- [HTB90] P. Hanggi, P. Talkner, and M. Borkovec. Reaction-rate theory : fifty years after Kramers. *Rev. Modern Phys.*, 62(2) :251–341, 1990.
- [Kra40] H.A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7, 1940.
- [OV05] E. Olivieri and M.E. Vares. *Large deviations and metastability*, volume 100 of *Encyclopedia of Mathematics and its Applications*. Chapitre 5. Cambridge University Press, Cambridge, 2005.
- [PotM1] R. Poteau. *Modélisation moléculaire*. Cours de M1 en téléchargement libre à http://www.ressources-pedagogiques.ups-tlse.fr/cpm/MODMOL/slides_mod_mol-C.pdf, M1.
- [SJ96] G.D. Smith and R.L. Jaffe. Quantum chemistry study of conformational energies and rotational energy barriers in n-alkanes. *The Journal of Physical Chemistry*, 100(48) :18718–18724, 1996.

6 Évolution de produits concurrents

6.1 Contexte

Le but de ce projet est d'étudier l'évolution de produits en concurrence, où les individus n'utilisent qu'un seul type de produit, et convainquent progressivement leurs connaissances de l'utiliser (par exemple A : avoir un téléphone Android, B : avoir un iphone). De manière équivalente, on peut reformuler cette question en termes de propagation de rumeurs ou d'épidémies. Plus formellement, on modélise les individus et leurs connaissances par un graphe $G = (V, E)$, où les sommets représentent les individus et les arêtes les relations de connaissance.

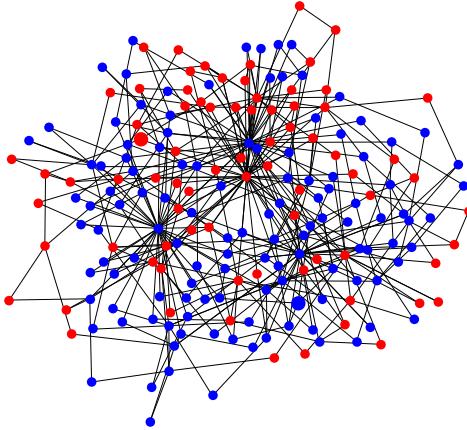


FIGURE 4 – Dans cet exemple, 78 sommets sont de type A (en rouge) et 122 sont de type B (en bleu).

6.2 Première partie (monopole)

On s'intéresse d'abord à l'évolution d'un produit A dans une situation de monopole avec n individus, et on suppose que ρ représente le temps moyen (en jours) qu'un individu utilisant A met à convaincre une de ses connaissances à également utiliser A . On suppose ici que le graphe G est le graphe complet à n sommets.

- (1) En fonction de ρ , estimer et étudier le temps au bout duquel tout le monde utilise A (partant d'une personne qui utilise A).

On pourra prendre $n = 6 \cdot 10^7$ (population de la France), $n = 7 \cdot 10^9$ (population mondiale).

- (2) Étudier le comportement du nombre de personnes utilisant A lorsque le temps évolue.

6.3 Deuxième partie (deux produits concurrents)

On suppose maintenant que dans une population avec n individus, à l'état initial, un sommet utilise A , un autre sommet utilise B (les autres sommets sont dits vides). Soient $\lambda, \mu > 0$ deux nombres réels strictement positifs (qui sont liés par exemple à l'efficacité de A et B). On suppose que A se propage à taux λ et que B se propage à taux μ , et que quelqu'un utilisant A n'utilisera jamais B (et réciproquement). Cela signifie que dès qu'un sommet x utilise A , on attribue à chaque arête reliant ce sommet x à un sommet vide une horloge qui sonne au bout d'un temps qui est une variable aléatoire exponentielle de paramètre λ (toutes les variables sont supposées indépendantes), qu'on interprète comme les temps nécessaires pour A pour se propager (de même pour B avec des variables aléatoires exponentielles de paramètre μ). Dès qu'une horloge sonne, il y a une propagation (à condition que le sommet vide cible n'ait pas été rempli entre temps). Le but est d'étudier la proportion de sommets utilisant A lorsque le processus s'arrête. On note \mathcal{A}_n (resp. \mathcal{B}_n) le nombre de sommets qui utilisent A (resp. B) lorsque le processus s'arrête.

- (3) Dans cette question, on considère le graphe "en ligne" avec n sommets u_1, u_2, \dots, u_n avec u_1 relié à u_2 , u_2 relié à u_3 , etc. jusqu'à u_{n-1} relié à u_n . On suppose qu'à l'instant initial u_1 utilise A , u_n utilise B et que tous les autres sommets sont vides. En fonction de λ, μ , étudier \mathcal{A}_n , et en particulier $\mathbb{P}(\mathcal{A}_n > \mathcal{B}_n)$ dans le régime des événements rares.

- (4) Dans cette question, on considère le graphe complet avec les n sommets u_1, u_2, \dots, u_n (c'est-à-dire que u_i est relié à u_j pour tout $i \neq j$). On suppose qu'à l'instant initial u_1 utilise A , u_n utilise B et que tous les autres sommets sont vides. En fonction de λ, μ , étudier \mathcal{A}_n , et en particulier $\mathbb{P}(\mathcal{A}_n > \mathcal{B}_n)$ dans le régime des événements rares.

On pourra remarquer que si (A_i, B_i) est le nombre de sommets utilisant respectivement A et B lorsqu'il y a eu i propagations, alors $(A_i, B_i)_{0 \leq i \leq n-2}$ est une chaîne de Markov et $A_{n-2} = \mathcal{A}_n$.

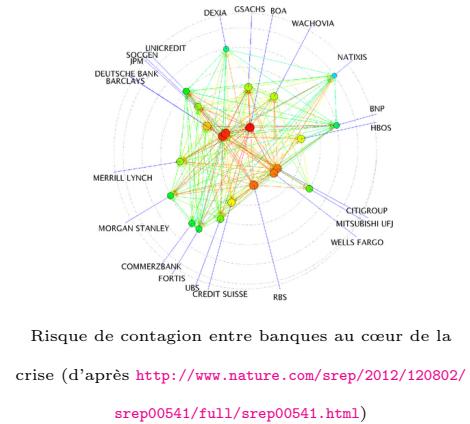
- (5) Que se passe-t-il pour d'autres graphes ?

7 Risque systémique

7.1 Réseau financier et risque systémique d'un défaut

La crise financière que nous avons traversé depuis 10 ans est un évènement d'une rare ampleur et lourd en conséquence. Elle a été qualifiée de systémique car elle s'est manifestée à l'échelle du système bancaire mondiale. Toutes les banques ont du faire face à de grandes difficultés de liquidités ayant conduit certaines d'entre elles à la faillite (Lehman Brothers en 2008). Depuis le début de la crise financière en 2007, plus de 370 des banques américaines (sur près de 8000 banques assurées par la Federal Deposit Insurance Corporation) ont fait faillite. Entre 2000 et 2004, seulement 30 banques ont fait faillite et aucune entre 2005 et le début de l'année 2007.

Le risque systémique est le risque d'effondrement d'un système suite à un choc sur certaines institutions financières qui entraînent par un effet *domino* la dégradation brutale ou la faillite de beaucoup d'autres. Ce n'est que très récemment que les institutions financières se sont intéressées à la modélisation mathématique de ces épisodes de contagion par défaut où un choc économique causant des pertes initiales et le défaut de quelques institutions sont amplifiés en raison de liens financiers complexes, pour finalement conduire à des faillites à plus grande échelle.



7.2 Modélisation du bilan d'une banque et réseaux financiers

Un système d'institutions financières est naturellement modélisé par un réseau de relations ou contreparties (graphe) : un ensemble de n noeuds et des liens pondérés représentés par une matrice $E = (e_{i,j})_{1 \leq i,j \leq n}$ où $e_{i,j}$ est (la valeur marché de) l'exposition de l'institution financière i à l'institution financière j .

A la date t , l'institution i dispose d'un capital (propre) $X_i(t)$, c'est-à-dire un matelas de sécurité pour les créanciers de l'entreprise, pour absorber les pertes potentielles. Le bilan d'une banque se présente sous la forme d'un équilibre entre ses actifs⁶, c'est-à-dire les biens qu'elle possède qui ont une

6. on ne considère ici que des actifs interbancaires : $A_i(t) \equiv 0$.

Actifs	Passifs
Actifs interbancaires $\sum_{j=1}^n e_{i,j}$	Passifs interbancaires $\sum_{j=1}^n e_{j,i}$
Autres actifs $A_i(t) = 0$	Capital $X_i(t)$

TABLE 1 – Bilan Actifs-Passifs de l’institution financière i à la date t

valeur économique positive, et ses passifs, c’est-à-dire l’ensemble de ses dettes ou de ses biens qui ont une valeur économique négative, c.f. Table 1.

Entre les temps t_k et t_{k+1} ($t_k := k\delta$ avec $\delta > 0$), le capital d’une institution subit des fluctuations dues au marché dont la dynamique peut s’écrire

$$X(t_{k+1}) = e^{-\lambda\delta} X(t_k) + \mu (1 - e^{-\lambda\delta}) + \sigma\sqrt{\delta}W_k \quad (4)$$

où $(W_k)_{k \geq 0}$ est une suite de v.a.i.i.d. gaussiennes centrées réduites. Le paramètre μ s’interprète comme l’équilibre moyen du capital, c’est-à-dire que lorsque $X(t)$ s’écarte de cette valeur le paramètre λ agit comme une force de rappel vers μ , et le paramètre $\sigma > 0$ est l’écart-type des fluctuations du capital. Le processus $(X_{t_k})_{k \geq 0}$ est alors un processus gaussien.

▷ **Bilan à l’horizon T .** A la date $T = N\delta$, un bilan de solvabilité a lieu permettant de conclure quant à la solidité financière de l’institution i . Si le capital de l’institution i est en dessous d’un seuil critique déterministe c_i celle-ci n’est plus *solvable*. Une banque est solvable si son capital restant à la date T est supérieur au seuil critique, c’est-à-dire si $X_i(T) > c_i$. Une banque insolvable fait défaut et est liquidée. Chacun de ses créanciers perd une fraction $1-R$ de l’exposition à la banque faisant défaut. Cette perte vient alors se soustraire au capital et peut entraîner à son tour l’insolvabilité des créanciers. Cette cascade de défaut dépend fortement du taux de récupération R (recovery rate) de la banque faisant défaut : pour simplifier, on suppose que ce taux de récupération est le même pour toutes les institutions (on prendra par exemple $R = 5\%$).

On définit l’ensemble $D_0^T = \{i \in \{1, \dots, n\} : X_i(T) < c_i\}$ des institutions financières faisant initialement défaut dans le réseau à la date T , la cascade de défaut est la séquence d’ensemble $D_0^T \subset D_1^T \subset \dots \subset D_{n-1}^T$, définie par

$$D_k^T = D_{k-1}^T \cup \left\{ j \notin D_{k-1}^T : X_j(T) - \sum_{p \in D_{k-1}^T} (1-R)e_{j,p} < c_j \right\}, \quad k \geq 1.$$

Dans un réseau financier de taille n , cette cascade finit après au plus $n-1$ étapes. A l’étape k , D_k^T représente l’ensemble des institutions financières

insolvables (et donc faisant défaut) suite à l'exposition de contrepartie à des banques de l'ensemble D_{k-1}^T qui viennent de faire défaut à l'étape précédente.

Pour quantifier le risque systémique et l'effet de contagion, on définit *l'impact de défaut* $I(T)$ à la date T du à la cascade de défaut à l'instant T ,

$$I(T) = \sum_{j \in D_{n-1}^T} \left(X_j(T) + \sum_{p \notin D_{n-1}^T} (1 - R)e_{p,j} \right)$$

qui est la somme des pertes générées par la contagion du défaut des banques à la date T .

▷ **Capitaux indépendants.** On supposera dans un premier temps que les capitaux évoluent de manière indépendante : les suites de variables $(W_k)_{k \geq 0}$ apparaissant dans (4) pour chaque institution sont mutuellement indépendantes.

L'horizon T est 1 an et les dates $(t_k)_k$ d'observation d'évolution du capital sont trimestrielles ou mensuelles par exemple. Les autres paramètres sont $X_i(0) = \mu = 15$, $\sigma = 8$, $\lambda = 20$, $c_i = 10$, $i = 1, \dots, 5$ avec un réseau simple constitué de 5 institutions financières, dont les liens interbancaires pourront être donnés par la matrice

$$E = \begin{pmatrix} 0 & 3 & 0 & 0 & 6 \\ 3 & 0 & 0 & 0 & 0 \\ 3 & 3 & 0 & 0 & 0 \\ 2 & 2 & 2 & 0 & 2 \\ 0 & 2 & 3 & 3 & 0 \end{pmatrix}.$$

On abordera les points suivants.

1. Estimation de la probabilité que le nombre de banques insolubles soit 1, 2, 3, 4 ou 5.
2. Estimation de la Value-at-Risk et de la Conditional Value-at-Risk de la distribution $I(T)$ pour différents seuils α proches de 1 (par exemple $\alpha = 99.9999\%$), ainsi que des intervalles de confiance associés aux estimateurs.
3. Estimation de la cascade d'insolubilité la plus probable, selon le nombre de banques en défaut.
4. Identification des liaisons de contrepartie dangereuses, des scénarios critiques menant à l'effondrement du système ainsi que des configurations (raisonnables) permettant d'améliorer la sûreté du système.
5. Influence de la taille du réseau ($n = 10, 50$) et des configurations dangereuses.
6. Evaluation de la distribution de l'impact de défaut conditionnellement à l'évènement : "le réseau entier d'institutions financières s'est effondré".

7.3 Extensions

On cherchera à aborder l'un des points suivants.

- ▷ **Contagion dynamique.** En réalité, l'insolvabilité des institutions peut avoir lieu à tout moment entre la date 0 et la date T . Il est donc nécessaire de prendre en compte ce risque de manière dynamique. Cela nous conduit à considérer des impacts de défaut dynamique $I(t_k)$, $k = 1, \dots, N$. On définit l'impact de défaut total à la date T comme la somme de tous les impacts de défaut entre la date 0 et la date T

$$\sum_{k=1}^N I(t_k).$$

On reprendra les questions précédentes en intégrant le risque de contagion dynamique.

- ▷ **Modèle d'équilibre long terme.** Un modèle couramment utilisé pour les simulations de risque de défaut est un modèle d'équilibre à 1 facteur idiosyncratique et 1 facteur systémique. Les capitaux de toutes les institutions ont un facteur commun $(Z_{t_k})_{k \geq 0}$ régissant l'équilibre long terme du réseau financier. Ce capital long terme est un processus gaussien de dynamique

$$Z(t_{k+1}) = e^{-\lambda_e \delta} Z(t_k) + \sigma_e \sqrt{\delta} \bar{W}_k,$$

où $(\bar{W}_k)_{k \geq 0}$ est une nouvelle suite indépendante de v.a.i.i.d. gaussiennes centrées réduites et $\lambda_e = 10$, $\sigma_e = 3$. Ce capital d'équilibre vient s'ajouter à chaque capital $X_i(t)$ de l'institution i . On reprendra les questions précédentes avec ce nouveau modèle.

Références

- [FS02] H. Föllmer and A. Schied. *Stochastic finance*, volume 27 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 2002. An introduction in discrete time.
- [GHS00] P. Glasserman, P. Heidelberger, and P. Shahabuddin. Variance reduction techniques for estimating value-at-risk. *Management Science*, 46 :1349–1364, 2000.
- [Gla03] P. Glasserman. *Monte-Carlo methods in Financial Engineering*. Springer Verlag, New York, 2003.
- [Min11] A. Minca. *Modélisation mathématique de la contagion de défaut*. PhD thesis, Université Pierre et Marie Curie, 2011. <https://tel.archives-ouvertes.fr/tel-00624419>

[Mou11] A. Moussa. *Contagion and Systemic Risk in Financial Networks*.
PhD thesis, Columbia University, 2011. <https://academiccommons.columbia.edu/catalog/ac:131474>

8 Barrages et risques d'inondation

Le risque d'inondation est classé premier risque naturel en France par le Ministère de l'Ecologie et du Développement durable⁷. Il concerne 13.300 communes sur le territoire français, dont 300 grandes agglomérations, pour un total de 5 à 6 millions de personnes concernées, directement ou indirectement.⁸ Dans le mécanisme de contrôle des risques d'inondation, un rôle est joué par les bassins artificiels, qui ont une fonction de régulation des crues (en plus de la production d'énergie hydroélectrique). En même temps, lorsqu'un bassin artificiel est alimenté par un volume d'eau trop important, pour éviter que le barrage qui le retient ne cède, on est obligé d'ouvrir les vannes d'évacuation de façon importante, ce qui entraîne un fort débit d'eau dans la vallée sous-jacente et donc un risque d'inondation est toujours présent. Il existe en France plusieurs lacs artificiels de grande taille, le lac de Serre-Ponçon en tête, avec ses $1.2 \cdot 10^9 m^3$ d'eau retenue.



8.1 Modélisation



Vidange du barrage, ouverture des vannes.

Dans une région montagneuse, deux vallées sont occupées par deux lacs artificiels créés par deux barrages B^1 et B^2 , dont on modélise leur volume contenu respectif $\{X_t^i, t \geq 0\}$, ($i = 1, 2$) au cours du temps $t \geq 0$. Chaque lac est alimenté par une rivière principale et plusieurs torrents, qui apportent un volume cumulé A_t^i ($i = 1, 2$) au temps t . En même temps, chaque bassin se vide à un taux instantané $r_i(X_t^i)$ qui

est une fonction du volume d'eau X_t^i stocké dans le bassin (en cas de crue, le lac de Serre-Ponçon peut évacuer jusqu'à $1.24 \cdot 10^5 m^3$ par heure). L'équation qui gouverne l'évolution temporelle du contenu de chaque réservoir (unité= $10^5 m^3$) s'écrit donc

$$X_t^i = x_0^i + A_t^i - \int_0^t r_i(X_s^i) ds, \quad (5)$$

7. http://catalogue.prim.net/43_dppr-livretrisqmajeurs-v7.pdf

8. voir la conférence http://www.coriolis.polytechnique.fr/Confs/Leleu_poster.pdf

où x_0^i est le contenu à l'instant initial. On fait l'hypothèse de débit proportionnel en le volume du barrage : $r_i(x) = r_i x$.

▷ **Volumes surcritiques.** Chaque barrage $i = 1, 2$ est construit pour supporter un volume contenu maximal x_{crit}^i (la géométrie de chaque bassin artificiel étant fixée, le niveau critique pourrait aussi bien être exprimé par la hauteur maximale de l'eau à contact du barrage). Au-delà de ce niveau critique, on est obligé d'ouvrir les évacuateurs du barrage à tel point que cela entraîne un risque d'inondation pour la vallée sous-jacente. Pour réduire le plus possible ce risque, ce niveau critique x_{crit}^i est défini comme le quantile $Q_{X_T^i}(\alpha)$ d'ordre $\alpha = 10^{-6}$ à un horizon T .

On souhaite estimer à l'aide de simulations la valeur de ce niveau critique, et évaluer ensuite s'il est envisageable, du point de vue financier et technique, de construire un barrage qui puisse tolérer un tel niveau de remplissage. On cherche aussi à quantifier les risques d'inondation dus à un des deux barrages ou aux deux.

▷ **Modélisation de l'apport d'eau.** On suppose que l'apport d'eau $\{A_t^i, t \geq 0\}$ est essentiellement déterminé par des chutes de pluie intenses et de courte durée (on néglige les ruissellements continus). Les modèles que l'on peut utiliser pour les précipitations atmosphériques sont typiquement fondés sur l'usage de processus stochastiques à sauts : on décrit l'arrivée d'un épisode de pluie par un processus de Poisson (le début de l'événement correspondant à un instant de saut du processus), la durée de l'épisode par une variable aléatoire positive, et on décrit aussi une structure fine de l'événement entre son début et sa fin (voir [CIO07]). On pourra retenir ici cette modélisation aléatoire de l'arrivée des pluies : on fait donc l'hypothèse que, pour chaque bassin, le débit entrant est modélisé par un processus de Poisson composé

$$A_t^i = \sum_{n=1}^{N_t^i} U_n^i, \quad i = 1, 2 \tag{6}$$

où pour tout $i = 1, 2$, $N^i = \{N_t^i, t \geq 0\}$ est un processus de Poisson d'intensité $\lambda_i > 0$, et $\{U_n^i, n \geq 1\}$ sont des variables aléatoires indépendantes et identiquement distribuées, et indépendantes de N^i . Pour la loi de U_1^1 et U_1^2 , une combinaison de lois exponentielles permettra de modéliser séparément des pluies de grande et petite intensité :

$$\nu(u) = b\delta_1 e^{-\delta_1 u} \mathbf{1}_{\{u>0\}} + (1-b)\delta_2 e^{-\delta_2 u} \mathbf{1}_{\{u>0\}}, \quad 0 < b < 1.$$

Pour modéliser une différence importante entre les chutes de pluie intenses et les faibles, on pourra prendre δ_2/δ_1 de l'ordre de 10 et $\delta_2 = 0.7$; on pourra

choisir λ de façon à exprimer le déclenchement de plusieurs dizaines à une centaine de pluies par an en moyenne ; enfin, le taux de vidange annuel r_i sera de l'ordre de l'unité.

8.2 Objectifs.

▷ **Etude d'un seul barrage.** On étudie tout d'abord les barrages séparément. Les questions à aborder sont les suivantes.

- Q1. (a) Estimer le niveau critique x_{crit}^i pour chaque bassin $i = 1, 2$, à un horizon $T = 1$ an. Déterminer la distribution du volume d'eau dans chaque bassin lorsqu'il y a un dépassement du niveau critique en T .
- (b) Pour comprendre si l'évaluation des risques que l'on vient d'effectuer à un horizon fixé est pertinente, on reconside le critère de définition du niveau de remplissage critique, en se basant maintenant sur le volume maximum $\max_{t_j} X_{t_j}$ atteint par le bassin tout au long de l'année. Estimer par simulation le nouveau volume critique, ainsi que la distribution de dépassement de ce niveau au cours de l'année. Comparer avec les résultats précédents.

▷ **Prise en compte des deux barrages, inondation en aval.** Les chaînes de montagne en amont des deux bassins étant suffisamment distantes, on fait l'hypothèse que les précipitations qui alimentent les deux lacs sont indépendantes : les processus de Poisson composés définis par $\{N^i, U_n^i, n \geq 0\}$ pour $i = 1, 2$, sont indépendants.

Les deux vallées occupées par les bassins, raides et inhabitées, se rejoignent à l'embouchure d'une vallée plus large qui est occupée par des villages. Au point de jonction O entre les deux vallées, une rivière est formée par l'eau venant des deux bassins, qui atteint le point O après avoir parcouru la première vallée en un temps T_1 , et la deuxième en un temps T_2 . On supposera que ces retards d'arrivée d'eau en O sont fixes, et indépendants du volume d'eau qui se déplace. Au point de jonction des vallées, le débit observé au temps t est donc

$$D_t = r_1 X_{t-T_1}^1 + r_2 X_{t-T_2}^2.$$

La rivière qui coule dans la vallée et qui est alimentée par les deux barrages, peut tolérer un débit limite, au-delà duquel la rivière déborde et entraîne des inondations des zones habitables. On traitera les questions suivantes.

- Q2. (a) Quelle est la probabilité de dépasser le débit de seuil dans la vallée, soit $\mathbb{P}(D_T \geq \text{seuil})$?

- (b) Quelle est la probabilité de dépassement du débit de seuil lorsque l'un des deux bassins est rempli au-dessus du niveau critique ?
- (c) A partir de l'observation du débit instantané au point de jonction O , on souhaite remonter à de l'information sur l'état de remplissage des bassins. Estimer le volume d'eau dans chaque bassin lorsqu'on observe un niveau limite de débit à la jonction.

8.3 Extensions

On cherchera à aborder l'un des points suivants.

- **Dépendance** : On souhaite relaxer l'hypothèse d'indépendance des précipitations en amont des deux bassins (i.e. l'indépendance des deux processus de comptage des pluies). Proposer une modélisation qui tienne compte d'une dépendance entre les phénomènes atmosphériques. On reprendra l'analyse précédente dans ce cadre.
- **D'autres sources d'aléa** : La vallée habitée est alimentée en eau pas seulement par les deux bassins, mais aussi par l'arrivée de pluies. Prendre en compte ce phénomène dans la modélisation et l'analyse du risque d'inondation.
- **Une analyse en temps long** : Proposer une analyse du système (5)-(6) en temps long ($t \rightarrow \infty$), en étudiant notamment la possibilité d'obtenir des formules explicites pour la loi de X_t^i dans ce régime, et donc pour ses quantiles. Comparer avec les simulations de la question 8.

Références

- [Asm03] S. Asmussen. *Applied probability and queues*, volume 51 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.
- [CIO07] P. Cowpertwait, V. Isham, and C. Onof. Point process models of rainfall : developments for fine-scale structure. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 463(2086) :2569–2587, 2007.
- [Sam04] S. Sambou. Modèle statistique des hauteurs de pluies journalières en zone sahélienne : exemple du bassin amont du fleuve sénégal. *Hydrological Sciences – Journal des Sciences Hydrologiques*, 49(1) :115–129, 2004.

9 Risque de collision spatiale

9.1 Contexte

Depuis le lancement de Spoutnik 1 le 4 octobre 1957, le nombre de satellites et de débris en orbite autour de la Terre n'a cessé d'augmenter. Un des mécanismes principaux à l'origine de cette augmentation de débris spatiaux est l'auto génération suite à des collisions en orbite. En attendant de pouvoir diminuer le nombre de débris en orbite, les trajectoires de près de 20000 objets de taille supérieure à 10 cm sont surveillées. Il est possible de manœuvrer un satellite dans le but d'éviter une collision. Mais il faut aussi prendre en compte le fait que ces manœuvres d'évitement diminuent substantiellement la durée de vie du satellite. Dans ce contexte, nous avons besoin d'estimer le risque que représentent les objets spatiaux pour les satellites opérationnels.

Des exemples de destruction de satellites suite à une collision en orbite existent. Par exemple, les débris créés par la destruction d'un satellite chinois par une arme anti-satellite, ont atteint un satellite russe en janvier 2013 (voir le site <http://www.space.com/20138-russian-satellite-chinese-space-junk.html>)



FIGURE 6 – Trajectoire d'Iridium 33 et Cosmos 2251

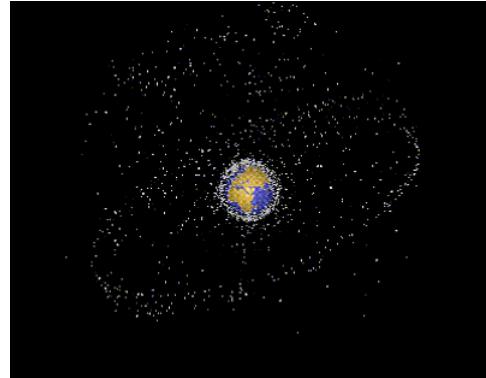


FIGURE 5 – Débris autour de la Terre

Un satellite de télécommunications américain en fonctionnement, Iridium 33, a été détruit lors de l'impact avec un satellite de télécommunications militaires russe retiré du service, Cosmos 2251. La plupart du temps, le NORAD (North American Aerospace Defense Command) prévoit ce genre d'incident et prévient les opérateurs de satellites, Boeing en l'occurrence pour Iridium, qui peuvent faire manœuvrer le satellite pour éviter la collision. Ainsi, cette organisation préconise

régulièrement de modifier l'orbite de la station spatiale internationale.

La taille des débris spatiaux peut varier de quelques millimètres à la taille d'un bus ; les plus gros d'entre eux sont des morceaux de lanceurs spatiaux ou des satellites inutilisés. Les débris vont à une grande vitesse (de l'ordre de 8 km/s dans le référentiel orbital décrit ci-dessous, pour les débris en orbite basse) et peuvent entraîner des dégâts considérables.

Un catalogue des gros objets en orbite, le Two Lines Elements, est établi et mis à la disposition des opérateurs spatiaux par l'USSPACECOM, un organisme militaire américain. Il présente cependant deux défauts. Il est, d'une part, incomplet, avec seulement 13 000 gros objets sur les 18 000 actuellement dans l'espace, car les objets « sensibles » ne sont pas répertoriés ; d'autre part, à l'instar du signal GPS public, il est brouillé, la précision n'étant que d'une dizaine de kilomètres. Nous nous intéressons donc aux méthodes d'estimation de probabilité de collisions spatiales. Heureusement, ces risques de collisions sont faibles ; leur estimation nécessite donc de techniques spécifiques aux événements rares.



FIGURE 7 – Collision spatiale

9.2 Modélisation

La première loi de Kepler affirme que la trajectoire d'un objet spatial autour de la Terre est une conique dont un des foyers est la Terre. Dans ce projet on considère des orbites elliptiques. Une orbite elliptique est décrite au moyen de deux plans (le plan de l'orbite et un plan de référence : ici, le plan de l'équateur céleste) et de six paramètres appelés *éléments orbitaux* (voir Figures 8 et 9). Deux de ces paramètres caractérisent la forme de l'orbite de l'objet

- a : demi-grand axe
- e : excentricité ($0 < e < 1$)

et les quatre autres, la position de cette orbite par rapport à la Terre (voir Figure 8)

- i : inclinaison par rapport à l'équateur terrestre

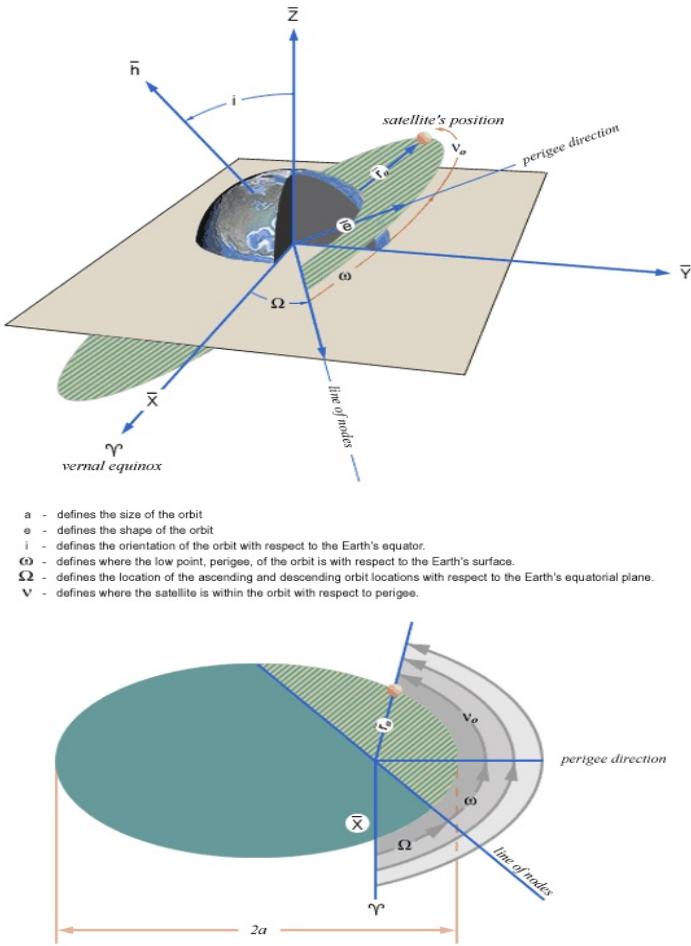


FIGURE 8 – Eléments orbitaux – voir [BMW71, p.59]

- Ω : longitude du nœud ascendant. La ligne des nœuds est l'intersection entre le plan de l'orbite de l'objet et le plan équatorial ; le nœud ascendant est le point de l'orbite en correspondance duquel l'objet passe de l'hémisphère sud à l'hémisphère nord. La longitude est mesurée par rapport au point vernal.
- ω : angle entre le nœud ascendant, la Terre et le périgée (le point de l'orbite le plus proche du foyer de l'ellipse)
- ν : angle donnant la position de l'objet sur son orbite (mesuré à partir du périgée, et dans le sens du mouvement de l'objet).

Ces 6 paramètres orbitaux définissent une orbite et la position d'un objet sur cette orbite. Plutôt que de décrire le mouvement d'un objet spatial par des

coordonnées cartésiennes classiques, on va utiliser le fait que le mouvement se déroule sur une ellipse dans l'espace.

Dans le *plan périfocal* (T, \vec{p}, \vec{q}) (Figure 9), le vecteur \vec{r} donnant la position de l'objet s'écrit

$$\vec{r} = r(\nu) \cos(\nu) \vec{p} + r(\nu) \sin(\nu) \vec{q}$$

où

$$— r(\nu) = \frac{l}{1+e \cos(\nu)}$$

$$— l = a(1 - e^2).$$

En utilisant la deuxième loi de Kepler, il est possible de montrer la relation $r(\nu_t) \nu'_t = \text{constante} = \sqrt{l\mu}$ (voir [BMW71, p.72]), où ν_t est l'argument de la position de l'objet au cours du temps, et $\mu = 398600 \text{ km}^3/\text{s}^2$. En utilisant cette relation, on montre que la vitesse de l'objet dans le même référentiel est donné par

$$\vec{v} = -\sqrt{\frac{\mu}{l}} \sin(\nu) \vec{p} + \sqrt{\frac{\mu}{l}} (e + \cos(\nu)) \vec{q}.$$

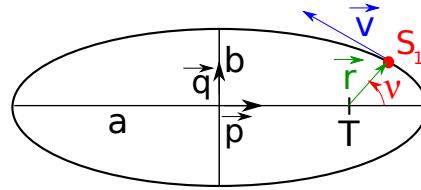


FIGURE 9 – Orbite elliptique

9.3 Objectifs

- 1) On considère un satellite S de position initiale donnée par les 6 éléments orbitaux suivants (en degrés et kilomètres) :

$$\begin{bmatrix} a \\ e \\ i \\ \Omega \\ \omega \\ \nu_0 \end{bmatrix} = \begin{bmatrix} 20000 \\ 0.6 \\ 3 \\ 45 \\ 90 \\ -10 \end{bmatrix}.$$

Donner, à l'aide des formules de passage du repère périfocal au repère cartésien, les coordonnées cartésiennes (position et vitesse) $[x_0, y_0, z_0, \dot{x}_0, \dot{y}_0, \dot{z}_0]$ correspondantes.

- 2) Intégrer la trajectoire du satellite S et la représenter sur une période.
 3) On considère un débris D , ayant les coordonnées orbitales suivantes :

$$\begin{bmatrix} a \\ e \\ i \\ \Omega \\ \omega \\ \nu_0 \end{bmatrix} = \begin{bmatrix} 20000 \\ 0.6 \\ 6 \\ 45 \\ 90 \\ 0 \end{bmatrix}.$$

Ecrire un programme donnant la distance minimale entre S et D sur une période.

- 4) On considère qu'il y a un risque de collision quand la distance entre deux objets spatiaux est inférieure à un certain seuil critique. La position du débris sur une orbite n'étant pas parfaitement connue, on introduit une

incertitude sur sa position initiale : $\begin{bmatrix} a \\ e \\ i \\ \Omega \\ \omega \\ \nu_0 \end{bmatrix} = \begin{bmatrix} 20000 \\ 0.6 \\ 3 \\ 45 \\ 90 \\ W \end{bmatrix}$ où $W \sim \mathcal{N}(0, 1)$.

Estimer le risque de collision entre le débris et le satellite considérés. On prendra comme seuil critique $s = 650$ kilomètres.

Références

- [BMW71] R.R. Bate, D.D. Mueller, and J.E. White. *Fundamentals of astrodynamics*. Dover publications, 1971. Partiellement disponible sur https://books.google.fr/books?id=UtJK8cetqGkC&printsec=frontcover&hl=fr&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

10 Championnat de football

L'industrie des paris sportifs, très florissante dans certains pays, a suscité le développement de modèles et de techniques visant à déterminer les cotes auxquelles les bookmakers proposent leurs paris. L'objet de ce projet est l'étude d'un modèle de championnat (où N équipes se rencontrent toutes les unes les autres), l'estimation des probabilités d'événements rares pour ce modèle, et, éventuellement, son amélioration. Ils étudieront en particulier le cas de la *Premier League*⁹ 2015/16 ainsi que les régimes asymptotiques mentionnés dans l'article [Esc17].

10.1 Modèle de Bradley-Terry

On considère ici un modèle de compétition entre N équipes (qui pourraient aussi être des agents économiques, des traitements médicaux,...) dans lequel les équipes ont des valeurs intrinsèques V_1, \dots, V_N (pour tout i , $V_i > 0$ est la valeur de l'équipe i) et se confrontent deux à deux dans des "matchs" ayant un vainqueur et un perdant. L'issue d'un match est aléatoire et distribuée de la façon suivante : pour tous $i \neq j \in \{1, \dots, N\}$, lors qu'un match entre i et j ,

$$\mathbb{P}(i \text{ gagne}) = \frac{V_i}{V_i + V_j}. \quad (7)$$

On suppose par ailleurs que les matchs successifs, même si ils impliquent les mêmes équipes, sont indépendants. Chaque match rapporte un point au vainqueur (et zéro au perdant), et à la fin du championnat, le score S_i de l'équipe i est la somme des points qu'elle a gagnés lors de ses matchs. Si les équipes se rencontrent deux à deux une et une seule fois, en notant

$$X_{i,j} = 1 - X_{j,i} = \begin{cases} 1 & \text{si } i \text{ est le vainqueur du match entre } i \text{ et } j, \\ 0 & \text{si } j \text{ est le vainqueur du match entre } i \text{ et } j, \end{cases}$$

on a donc

$$S_i = \sum_{j \neq i} X_{i,j}. \quad (8)$$

Si, comme dans les championnats de football, les équipes se rencontrent toutes 2 fois, on a

$$S_i = \sum_{j \neq i} X_{i,j}^{(1)} + X_{i,j}^{(2)}, \quad (9)$$

9. La *Premier League* est la première division du championnat masculin anglais de football.

où $X_{i,j}^{(1)}$ (resp. $X_{i,j}^{(2)}$) désigne l'issue du premier (resp. deuxième) match entre i et j . Les variables aléatoires

$$(X_{i,j})_{1 \leq i < j \leq N} \quad (\text{ou } (X_{i,j}^{(k)})_{1 \leq i < j \leq N, k=1,2})$$

sont des variables aléatoires de Bernoulli indépendantes telles que

$$\mathbb{P}(X_{i,j} = 1) = \frac{V_i}{V_i + V_j} \quad (\text{ou } \mathbb{P}(X_{i,j}^{(k)} = 1) = \frac{V_i}{V_i + V_j}).$$

La loi de l'issue de la compétition est entièrement déterminée par la valeur des paramètres V_1, \dots, V_N (et invariante par multiplication de ces nombres par un même nombre strictement positif).

10.2 Premier League 2015/16

La *Premier League 2015/16* a vu un événement particulièrement surprenant se produire : la petite équipe de Leicester a gagné le championnat, reléguant les grosses cylindrées de Premier League au second plan. Si les exploits de "petits poucets" lors des coupes (compétitions avec élimination directe, dont l'issue repose sur un petit nombre de matchs) sont courants, ce genre de surprises est rare dans les championnats.

a) Les élèves implémenteront le modèle décrit à l'équation (9). Pour cela, il est nécessaire de choisir des valeurs pour les V_i . La littérature fournit des méthodes d'estimation de ces paramètres (basées sur les matchs passés, voir par exemple [Hun04, YYX12]), mais nous nous contenterons ici des V_i donnés par une des deux possibilités suivantes (κ, b, θ sont des paramètres qui ne dépendent pas de i) :

- pour tout i , $V_i = (P_i)^\kappa$ où P_i est le nombre de points obtenus par l'équipe i en Première League l'année précédente, 2014/15 (pour les équipes promues de la seconde division, on prend le nombre de points obtenus en seconde division l'année précédente et on divise par 1.8),

ou bien :

- pour tout i , $V_i = (b - (R_i)^\theta)$ où R_i est le classement de l'équipe i en Première League l'année précédente, 2014/15 (pour les équipes promues de la seconde division, on prend $R_i = 15$).

Les P_i, R_i , ainsi que des suggestions de valeurs pour les paramètres κ, b, θ , sont disponibles dans le fichier [Data].

Les élèves compareront les probabilités de victoires des équipes de Chelsea, Manchester City, Arsenal, Manchester United et Tottenham données par ce modèle avec celles choisies, avant le début du championnat, par les bookmakers, telles que présentées par exemple en [Sch16] (qui sont donc, pour ces équipes, respectivement, $8/(8 + 13)$, $2/(5 + 2)$, $2/(7 + 2)$, $1/(5 + 1)$ et $1/(100 + 1)$).

b) Les élèves estimeront alors, avec ce modèle et ces V_i , la probabilité que Leicester gagne le championnat. Ils pourront aussi estimer les probabilités d'autres événements plus rares qui se sont réalisés, comme la probabilité que Leicester gagne avec les deux équipes de Manchester hors du podium et Liverpool et Chelsea hors des *7 places européennes*. Ils pourront aussi estimer la probabilité de l'événement (non réalisé, celui-ci) de la victoire de l'équipe la plus faible selon ces V_i . Ils commenteront les probas obtenues. Ils devront pour cela utiliser l'échantillonnage préférentiel.

10.3 Régimes limites

Dans l'article [CDL16] (voir aussi la note de vulgarisation [Esc17]), les auteurs considèrent, pour des valeurs très élevées de N , l'effet de la répartition des V_i sur la question de savoir si l'équipe favorite (celle ayant le plus grand V_i) gagne toujours, dans le modèle de l'équation (8). De façon très surprenante, ils montrent que l'équipe favorite ne gagne pas toujours : ils modélisent la répartition des V_i en les supposant choisis au hasard selon certaines lois et mettent en évidence plusieurs régimes différents, certains d'entre eux donnant en fait assez rarement la victoire à l'équipe favorite.

a) Les élèves pourront implémenter cette sophistication du modèle et mettre en évidence les tendances caractéristiques des différents régimes de l'article.

b) Ils pourront ensuite estimer, au moyen de méthodes vues en cours, la probabilité d'événements rares comme celui où une équipe ayant un petit V_i gagne finalement ou bien celui où l'équipe ayant le plus grand V_i ne gagne pas, alors que celui-ci est très au dessus des autres V_i .

10.4 Perfectionnement du modèle

Le modèle de championnat décrit par les équations (7) et (9) est assez rudimentaire : il ne tient compte ni de la possibilité de matchs nuls, ni de l'évolution des équipes dans le temps, ni de la différence entre les matchs à l'extérieur et les matchs à domicile, etc... Les élèves pourront essayer d'affiner le modèle, en gardant en tête que l'augmentation du nombre de paramètres augmente le nombre de données à estimer avant toute implémentation pratique.

Références

- [CDL16] R. Chetrite, R. Diel, M. Lerasle *The number of potential winners in Bradley-Terry model in random environment*, Ann. Appl. Probab., 2016. <https://arxiv.org/abs/1509.07265>
- [Esc17] S. Escalón. *Le ballon rond à l'épreuve des probabilités*, Le Journal du CNRS, 2017. <https://lejournal.cnrs.fr/articles/le-ballon-rond-a-lepreuve-des-probabilites>.
- [Sch16] A. Schooler. *Premier League 2015/16 : How odds changed as Leicester claimed title* <http://www.skysports.com/football/news/11712/10261535/premier-league-201516-how-the-odds-changed-as-leicester-claimed-the-title>
- [Data] DataPL201516.py Données disponibles sur le Moodle du cours.
- [Hun04] D. Hunter. *MM algorithms for generalized Bradley-Terry models*. Ann. Statist. 32, 384–406, 2004.
- [YYX12] T. Yan, Y. Yang, J. Xu *Sparse paired comparisons in the Bradley-Terry model*. Statist. Sinica 22 1305–1318, 2012.