Liste des projets - Langage R en Actuariat

Nicolas Baradel

16 avril 2024

Introduction

Les projets se découpent en quatre grands thèmes

- Finance (2 projets),
- Assurance dommage (2 projets),
- Assurance vie (1 projet),
- Solvabilité II (1 projet).

Chaque projet se fait à 3 ou 4. Le format du rendu du projet doit être l'un des suivants

- Fichier de code R (.R) avec un document pdf (10 pages maximum). Il doit y avoir un seul fichier R, mais très bien architecturé / commenté. Il doit être portable et s'exécuter sur d'autres machines. Il y a bien deux fichiers : en aucun cas le code doit figurer dans le rapport.
- Fichier Rmarkdown (.Rmd) avec un document pdf (10 pages maximum). Le fichier Rmarkdown ne se substitue pas au rapport pdf, il est considéré comme un autre format de rendu du code R, et doit être portable.

La portabilité se définit sous réserve d'avoir installé les éventuels paquets R nécessaires, qui seront chargés en tout premier dans le .R ou .Rmd, même si leur utilisation est ultérieure.

1 Finance

1.1 Evaluation par Monte Carlo et Réduction de variance

Soit $(W_t)_{0\leq t\leq T}$ un $\mathbb{Q}$ - mouvement brownien sur $[0,T]$ et

$$S_t = S_0 e^{(r-\frac{\sigma^2}{2})t+\sigma W_t}, \quad 0\leq t\leq T.$$

On pose

$$C_T^a := \left(\frac{1}{T}\int_0^T S_t dt - K\right)^+.$$

On prendra $S_0 = 1, r = 0.01, \sigma = 0.2$, $T = 1$ et $K \in \{0.8, 1, 1.2\}$. On fixera $n = 10^5$ et $\Delta t = 0.01$. Quand on donnera un estimateur, on donnera sa précision via son erreur potentielle à un niveau de confiance de 95%.

- Proposez une méthode de Monte Carlo simple afin d'estimer

$$C_0^a := e^{-rT}\mathbb{E}^{\mathbb{Q}}\left(C_T^a\right).$$

  On pourra approximer l'intégrale via des sommes de Riemann le long du pas de temps.
- Proposez une méthode d'estimation de C_0^a avec réduction de variance qui s'appuie sur les variables antithétiques et comparez la précision. On pourra se référer à [2] Section 3.1 et à la Remarque 6 de la Section 4.1.
- On introduit la variable aléatoire :

$$X := \left(\exp\left(\frac{1}{T}\int_0^T \log(S_t)dt\right) - K\right)^+.$$

  Proposez une méthode d'estimation de C_0^a avec réduction de variance utilisant la méthode de variable de contrôle et comparez la précision. On pourra également se référer à [2, Section 3.2]. On rappelle que $\int_0^T W_t dt \sim \mathcal{N}(0, \frac{T^3}{3})$.

1.2 Gestion de portefeuille dynamique

Soit $(W_t^1)_{0\leq t\leq T}$ un mouvement brownien sur $[0,T]$ et l'actif

$$dY_t = \mu Y_t dt + \sqrt{V_t}Y_t dW_t^1, \quad 0 \leq t \leq T.$$

Soit $(W_t^2)_{0\leq t\leq T}$ un second mouvement brownien indépendant de W^1. Le processus de variance qui apparait ci-dessus, V, est défini par

$$dV_t = -a(V_t - \sigma^2)dt + \xi\sqrt{V_t}\left[\rho dW_t^1 + \sqrt{1-\rho^2}dW_t^2\right], \quad 0 \leq t \leq T.$$

Pour plus de généralité, on introduit la notation $(Y_s^{t,y,v})_{t\leq s\leq T}$ qui est le processus Y sur $[t,T]$, partant de $Y_t = y > 0$ avec $V_t = v > 0$. De même on introduit $(V_s^{t,v})_{t\leq s\leq T}$ partant de $v > 0$ (ce processus est autonome et ne dépend pas de y).

On note $(X_s^{t,x,y,v})_{t\leq s\leq T}$ la valeur d'un portefeuille autofinançant partant de la richesse initiale $X_t^{t,x,y,v} = x > 0$. On peut acheter et vendre des actions Y. On note α_u la richesse investie en actions en date u. On considère les taux nuls, et par construction, la valeur du porteufeille ne varie que par l'investissement dans l'actif risqué. Le montant investi en actions est α_u et, comme une action a un prix $Y_u^{t,y,v}$ en u, le nombre d'actions détenues en u est $\frac{\alpha_u}{Y_u^{t,y,v}}$.

La valeur du portefeuille $(X_s^{t,x,y,v,\alpha})_{t\leq s\leq T}$ est, par la condition d'autofinancement, régie par

$$\begin{aligned} X_s^{t,x,y,v,\alpha} &= x + \int_t^s \frac{\alpha_u}{Y_u^{t,y,v}} dY_u^{t,y,v}, \\ &= x + \int_t^s \alpha_u(\mu du + \sqrt{V_u^{t,v}}dW_u^1). \end{aligned} \tag{1}$$

Nous remarquons que $X^{t,x,y,v,\alpha}$ ne dépend pas de la condition initiale y. Nous écrirons plus simplement $X^{t,x,v,\alpha}$. Nous posons la fonction d'utilité U de la valeur terminale du portfeuille $x \in \mathbb{R}$:

$$U(x) = -\exp(-\eta x).$$

Partant de (t,x,v), on note $\mathcal{A}^t$ l'ensemble des processus progressifs $(\alpha_u)_{t\leq u\leq T}$ à valeurs dans $\mathbb{R}$ tels que $X^{t,x,v,\alpha}$ soit uniformément borné.

Soit $\alpha \in \mathcal{A}^t$, on note

$$J(t,x,v,\alpha) = \mathbb{E}\left[U(X_T^{t,x,v,\alpha})\right].$$

L'objectif est de choisir $\alpha \in \mathcal{A}^t$ afin de maximiser J. Cela permet de définir la fonction valeur

$$\mathrm{v}(t,x,v) := \sup_{\alpha\in\mathcal{A}^t} J(t,x,v,\alpha) = \sup_{\alpha\in\mathcal{A}^t} \mathbb{E}\left[U(X_T^{t,x,v,\alpha})\right].$$

Pour une introduction au contrôle stochastique, on pourra se référer à [4, Section 2].
Pour la suite, on supposera que v est de classe $C^{1,2}([0,T[,\mathbb{R}\times\mathbb{R}) \cap C([0,T],\mathbb{R}\times\mathbb{R})$.

- Montrez que

$$\mathrm{v}(t,x,v) = e^{-\eta x}\mathrm{v}(t,0,v) \tag{2}$$

  et que le contrôle optimal ne dépend pas de la richesse x. On pose

$$\mathrm{v}_\circ(t,v) := \mathrm{v}(t,0,v).$$

- Pour résoudre ce problème, on a besoin d'un principe de programmation dynamique, qui dit ici que pour tout temps d'arrêt (qui incluent les temps constants) θ à valeurs dans $[t,T]$, on a

$$\mathrm{v}(t,x,v) = \sup_{\alpha\in\mathcal{A}^t} \mathbb{E}\left[\mathrm{v}(\theta, X_\theta^{t,x,v,\alpha}, V_\theta^{t,v})\right]. \tag{3}$$

  Ce dernier permet de découper l'optimisation sur $[t,T]$ en une optimisation sur $[t,\theta]$ et $[\theta,T]$. En vous référant à [4, Théorème 2.3 p25], montrez (3). En déduire que

$$\mathrm{v}_\circ(t,v) = \sup_{\alpha\in\mathcal{A}^t} \mathbb{E}\left[e^{-\eta X_\theta^{t,0,v,\alpha}}\mathrm{v}_\circ(\theta, V_\theta^{t,v})\right]. \tag{4}$$

Approche heuristique par Monte Carlo

Pour résoudre le problème numériquement en s'appuyant directement sur (3), on propose de découper les intervalles en $[t,t+\Delta t]$ où on fixe $\alpha = a \in \mathbb{R}$ constant sur cet intervalle. Le problème approché est

$$\mathrm{v}_\circ(t,v) = \sup_{a\in\mathbb{R}} \mathbb{E}\left[e^{-\eta X_{t+\Delta t}^{t,0,v,a}}\mathrm{v}_\circ(t+\Delta t, V_{t+\Delta t}^{t,v})\right]. \tag{5}$$

- En utilisant la précédente équation (5) avec la condition terminale $\mathrm{v}(T,x,v) = U(x)$, résolvez numériquement le problème discret associé, c'est-à-dire, calculez pour tout $t \in [0,T]$ et pour tout $\sqrt{v} \in \sqrt{\mathbf{V}}$ (défini ci-après), la fonction $\mathrm{v}_\circ(t,v)$ et le contrôle optimal α (qui est une fonction de $[t,v]$). On utilisera
  - $\eta = 0.2$,
  - $\rho = -0.7$,
  - $\mu = 0.07$,
  - $\sigma^2 = 0.04$,
  - $a = 2$,
  - $\xi = 0.25$,
  - $T = 1$,
  - $\Delta t = 0.004$ (251 points),

— $\sqrt{\mathbf{V}} = \{0.01, 0.01 + \Delta v, 0.01 + 2\Delta v, \ldots, 0.51\}$ avec $\Delta v = 0.005$ (101 points).

On procédera par Monte Carlo : on testera une allocation $a \in [0, 100]$ puis on calculera, avec au moins 250 simulations, $J(t, v, a)$, et on en déduira $a[t, v]$ optimal et $\mathrm{v}_\circ(t, v) = J(t, v, a[t, v])$. On se restraindra à $a \in \{0, 1, 2, \ldots, 100\}$, on pourra tous les tester et en déduire celui qui atteint le maximum.

Remarque 1. Les simulations de $V_{t+\Delta t}^{t,v}$ ne tomberont pas dans $\mathbf{V}$. Dans ce cas là, on utilise le barycentre pour calculer la valeur (interpolation linéaire). Ceci est fondamental et prendre le point le plus proche biaiserait le résultat : si le pas de temps est très petit, il se peut que $V_{t+\Delta t}^{t,v}$ soit toujours très proche de v ce qui ferait considérer au problème V comme constant (ou quasi constant).

En cas de sortie des bornes du maillage, on prendra la valeur la plus proche.

Remarque 2. Pour améliorer l'estimation de $J(t, v, a)$, on pourra regarder du côté des variables antithétiques pour Monte Carlo, voir [2], Section 3.1.

- Illustrez en
  - — Affichant la fonction de contrôle optimal $\alpha[t, v]$;
  - — Simulant (Y_t^{0,Y_0,V_0}) et (V_t^{0,V_0}) et en affichant les trajectoires de (Y_t^{0,Y_0,V_0}), de (V_t^{0,V_0}), de $\alpha[t, V_t^{0,V_0}]$ et de (X_t^{0,X_0,V_0}) en partant d'une richesse inititale de $X_0 = 100$, d'un prix initial de $Y_0 = 10$ et d'une variance initiale de $V_0 = 0.04$.
  - — Donnant l'histogramme de répartition empirique des performances du portefeuille.
  - — Produisant tout graphique qui vous parraît pertinant.

2 Assurance dommage

2.1 Tarification en assurance dommage

Nous utilisons le paquet `CASdatasets` [1], et plus précisément les données

- `freMTPLfreq` qui sont l'observation du nombre de sinistres de 413 169 contrats d'assurance de responsabilité civile automobile, avec des caractéristiques sur les contrats ;
- `freMTPLsev` qui sont les montants des sinistres associés aux nombres de `freMTPLfreq`.

Les données se chargent avec la fonction `data`, et l'aide de **R** associée à ces données fourni des informations utiles.

- Explorez les données avec des statistiques descriptives, en proposant des graphiques, qui permettent d'appréhender les données et de voir des premiers effets.
- On prend au hasard [2] 3/4 des données pour estimer le modèle, le 1/4 restant servira à la validation. Testez plusieurs modèles linéaires généralisés pour le nombre de sinistres. On veillera à ne pas conserver de variables explicatives inutiles. Proposez un modèle en vous appuyant sur l'AIC et un autre (qui peut être le même) en vous appuyant les données de validation. On n'oubliera pas de supprimer les variables explicatives sans effet.
- Faites de même pour le coût des sinistres.

1. Le package n'est pas sur le CRAN car son volume est trop imporant, il est disponible à `http://dutangc.perso.math.cnrs.fr/RRepository/`

2. On fixera l'aléa avec `set.seed(A)` où `A` est votre numéro de groupe, à demander. Pour tirer l'aléa, on utilisera `sample(1:n)` où les 75% de premiers numéros seront les numéros de ligne de l'échantillon pour estimer le modèle.

- Donnez un tarif et vérifiez celui-ci avec les données de validation.
- Estimer le modèle validé sur l'ensemble des données.
- Grâce au paquet `glmnet`, estimer le modèle glm avec pénalisation de type Lasso. Pour la validation croisée, on pourra prendre 3/4 du portefeuille pour estimer et 1/4 pour valider. Une fois le coefficient de pénalisation fixé, estimer sur l'ensemble des données. Comparer et conclure.

2.2 Sinistres tardifs et estimation des provisions IBNR

Nous utilisons le paquet `CASdatasets` [3], et plus précisément les données `freclaimset2motor`. Il s'agit des déclarations de sinistres individuels et des évolutions de la charge estimée par année. L'objectif est de calculer le montant des provisions pour sinistres tardifs IBNR. On consultera l'aide associée aux données pour bien comprendre ce que représente chaque colonne.

Les données se chargent avec la fonction `data`, et l'aide de **R** associée à ces données fournit des informations utiles. Il s'agit d'une liste qui comprend deux `data.frame` : les évolutions des sinistres indivuduels et quelques quantités aggrégées par année d'occurrence.

L'objectif est d'estimer les provisions sur les montants *ExpectCharge* en utilisant la méthode de Mack et la méthode de Schnieper, on pourra se référer à [1, Deuxième partie - Sinitres tardifs]. Pour Mack, on utilisera le paquet `ChainLadder`.

- Après avoir aggrégé vos données, appliquez le modèle de Mack - ChainLadder pour estimer le montant des provisions et le quantile à 99.5% en utilisant les paramètres de Mack (moyenne et variance des provisions) avec une loi normale. On tracera la distribution des provisions.
- Appliquez la méthode du bootstrap au modèle de Mack - ChainLadder pour en déduire à nouveau la provision et le quantile à 99.5% associé. On tracera la distribution empirique des provisions avec la méthode des noyaux.
- Après avoir aggrégé vos données, appliquez le modèle de Schnieper pour estimer le montant des provisions. Appliquez la méthode du bootstrap pour estimer le quantile à 99.5% du montant des provisions. On tracera la distribution des provisions avec la méthode des noyaux. L'exposition aggrégée est la colonne *Exposure* dans la `data.frame` aggrégée de `freclaimset2motor`.

3 Assurance-vie

3.1 Tables de mortalité

Nous utilisons le paquet `CASdatasets` [4], et plus précisément les données `canlifins`. Il s'agit de données de mortalité de couple, avec la donnée individuelle (mortalité de l'homme et de la femme). On note τ_H la date de survenance du décès de l'homme et τ_F la date du décès de la femme. L'objectif est de tarifer des contrats d'assurance qui versent un capital lors de la survenance

- Du premier décès dans le couple, à la date $\tau_* := \min(\tau_H, \tau_F)$,

3. Le package n'est pas sur le CRAN car son volume est trop imporant, il est disponible à `http://dutangc.perso.math.cnrs.fr/RRepository/`

4. Le package n'est pas sur le CRAN car son volume est trop imporant, il est disponible à `http://dutangc.perso.math.cnrs.fr/RRepository/`

- Du dernier décès dans le couple, à la date $\tau^* := \max(\tau_H, \tau_F)$.

La connaissance de la seule table de mortalité des hommes et de celle des femmes n'est pas suffisante a priori, si τ_H et τ_F ne sont pas indépendants.

- Estimer la table de mortalité des hommes et des femmes entre 50 et 95 ans (indépendamment). On tiendra compte de la censure et on appliquera un lissage de Whittaker-Henderson.
- Implémentez une fonction qui, de l'âge de l'homme, de l'âge de la femme, du taux d'actualisation, et de la durée d'une garantie décès, évalue la prime pure dans le cas où
  - l'assurance décès porte sur τ_*, le premier décès dans le couple,
  - l'assurance décès porte sur τ^*, le second décès dans le couple,

  sous hypothèse d'indépendance des décès. Pour les deux types de garantie, donnez la prime pure pour un couple où l'homme a 65 ans, la femme a 63 ans, en fonction de la durée de garantie (on tracera un graphique).
- Pour modéliser la dépendance, on propose d'utiliser une copule de Clayton, on pourra utiliser le package `copula`. On gardera comme lois marginales les deux tables de mortalité sur τ_H et τ_F et la copule à estimer est celle du couple (τ_H, τ_F). Estimez par maximum de vraisemblance le paramètre de dépendance dans la copule de Clayton. On pourra se référer à [3].
- En déduire une nouvelle fonction de prix pour la prime pure de la garantie décès pour les deux types de garantie. À nouveau, donnez la prime pure pour un couple où l'homme a 65 ans, la femme 63 ans, en fonction de la durée de garantie (on tracera un graphique) et comparez au cas indépendant.

4 Solvabilité II

4.1 Aggrégation des risques

Partie 1 - Agrégation simple Nous disposons de deux risques, S_1 et S_2, tous deux suivants la loi (identique pour simplifier, sans conséquence sur l'interprétation) :

$$S_i \sim \mathcal{LN}(\mu_{\log}, \sigma_{\log}), \quad i = 1; 2. \tag{6}$$

On suppose que S_1 et S_2 ne sont pas indépendantes, et leur dépendance est caractérisée par une copule en dimension 2 qui pourra être

- Copule 1 : gaussienne de paramètre $-1 \leq \rho_C \leq 1$,
- Copule 2 : de Clayton de paramètre inversé $\alpha_C > 0$.

En supposant que le modèle sous-jacent exact est celui énoncé : marginales selon (6) et dépendance avec la copule 1 ou 2, l'objectif est de comparer le Solvency Capital Requirement (SCR) obtenu dans la formule standard avec celui exact (qui sera approximé par Monte Carlo pour $S = S_1 + S_2$). Pour une variable aléatoire X, le Best Estimate (BE) et le SCR seront définis par

$$\mathrm{BE}(X) = \mathbb{E}(X) \qquad \mathrm{SCR}(X) = VaR_{99.5\%}(X) - \mathrm{BE}(X).$$

Dans ce modèle simple, le BE de S_1, S_2, S et le SCR de S_1, S_2 peuvent être calculés de manière exacte, indépendamment de la copule. Seul $SCR(S)$ dépend de la copule et ne peut pas être calculé de manière directe.

On rappelle que la formule standard agrège les SCR de S_1 et de S_2 avec la formule :

$$\mathrm{SCR}(S_1+S_2)=\sqrt{\mathrm{SCR}(S_1)^2+\mathrm{SCR}(S_2)^2+2\rho(S_1,S_2)\mathrm{SCR}(S_1)\mathrm{SCR}(S_2)},$$

où ρ est la fonction de corrélation linéaire. On fixe les paramètres

$$\mu_{\log}:=16,97;\quad \sigma_{\log}=0,08398;\quad \rho_C=0,25;\quad \alpha_C=0,35.$$

Enfin, la copule de Clayton inversée, si on note C_{Clayton} celle de Clayton, est définie par

$$C(1-u_1,1-u_2):=C_{\text{Clayton}}(u_1,u_2),\quad (u_1,u_2)\in[0,1]^2.$$

On utilisera le package `copula` pour les copules.

- Avec $n=10^7$ simulations, simulez (S_1,S_2), estimez $\rho(S_1,S_2)$ pour en déduire $\mathrm{SCR}(S)$ par la formule standard pour les deux copules. On fournira un intervalle de confiance pour l'estimation de la corrélation linéaire en utilisant

  $$\sqrt{n}(\hat{r}_n-\hat{r})\to\mathcal{N}(0,1),$$

  où $\hat{r}_n:=\frac{1}{2}\log\left(\frac{1+\hat{\rho}_n}{1-\hat{\rho}_n}\right)$ avec $\hat{\rho}_n$ le coefficient de corrélation linéaire empirique et $r:=\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$. Cette transformation du coefficient de corrélation s'appelle la transformation de Fisher. On en déduira un intervalle de confiance pour $\mathrm{SCR}(S)$.
- Avec $n=10^7$ simulations, simulez (S_1,S_2) pour en déduire $\mathrm{SCR}(S)$ en prenant directement le quantile empirique. On fournira un intervalle de confiance pour l'estimation de $SRC(S)$ en utilisant le fait que

  $$\sqrt{n}(\hat{q}_n^\alpha-q_\alpha)\to\mathcal{N}\left(0,\frac{\alpha(1-\alpha)}{f_S(q_\alpha)}\right)$$

  où q_α est le quantile d'odre $\alpha\in]0,1[$, $\hat{q}_n^\alpha$ est le quantile empirique, f_S est la densité de probabilité de S qui pourra être approximée par méthode des noyaux.
- Comparez les différences entre la formule standard et l'approche exacte pour les deux copules, permise car on a supposé le modèle connu à l'erreur de Monte Carlo près, contrôlée par l'intervalle de confiance.

Partie 2 - Agrégation de sommes aléatoires Un assureur possède un portfeuille d'assurance automobile séparé en deux sous branches : une pour les sinistres *ordinaires* et une pour les sinitres *exceptionnels*. Chaque sous branche est représentée par :

$$S_i=\sum_{n=1}^{N_i}X_n^i,\quad i=1;2. \tag{7}$$

S_1 est le coût total des sinistres *ordonnaires*, S_2 celui des sinistres *exceptionnels*. $N_i\sim\mathcal{NB}(n_i,p_i)$ avec $n_i>0,\ 0<p_i<1$ pour $i=1;2$. Les $(X_n^i)_{n\geq1}$ sont une suite i.i.d. de variables aléatoires positives et indépendantes des N_i, pour $i=1;2$.

On suppose que

— Le couple (N_1,N_2) est dépendant, par une copule définie ci-après ;

— Les suites $(X_n^1)_{n\geq1}$ et $(X_n^2)_{n\geq1}$ sont indépendantes ;

— $\forall n \geq 1,\ X_n^1 \sim \mathcal{LN}(\mu_{\log}, \sigma_{\log})$ et $X_n^2 \sim \mathcal{GP}(k, s, \xi)$, loi de pareto généralisée, définie par $X_n^2 \overset{loi}{=} k + \frac{s(U^{-\xi}-1)}{\xi}$ où $U \sim \mathcal{U}([0,1])$,

Pour la copule du couple (N_1, N_2), nous reprendrons les deux copules précédentes : gaussienne et Clayton inversée.

On prendra pour la suite les paramètres suivants, tirés de données réelles :

— $(n_1, p_1) = (8477, 0.3679)$,

— $(n_2, p_2) = (8, 0.2191)$,

— $(\mu_{\log}, \sigma_{\log}) = (6.8102, 1.0786)$,

— $(k, s, \xi) = (50000, 67117, 0.4270)$,

— $(\rho_C, \alpha_C) = (0.5, 0.9)$.

- Avec $n = 10^5$ simulations, calculez, idéalement sans boucle, $SCR(S)$ en formule standard, en estimant par Monte Carlo les quantités qui ne peuvent pas être facilement connues. On ne fera pas d'intervalle de confiance.
- Avec $n = 10^5$ simulations, calculez, idéalement sans boucle, $SCR(S)$ par Monte Carlo directement sur le modèle aggrégé, on fournira un intervalle de confiance de l'estimateur.
- Comparez et concluez.

Références

[1] Nicolas Baradel. Assurance dommage. `https://nicolasbaradel.fr/enseignement/ressources/cours_assurance_dommage.pdf`.

[2] Nicolas Baradel. Méthodes numériques en finance. `https://nicolasbaradel.fr/enseignement/ressources/cours_methodes_numeriques_finance.pdf`.

[3] Edward W Frees, Jacques Carriere, and Emiliano Valdez. Annuity valuation with dependent mortality. *Journal of Risk and Insurance*, pages 229–261, 1996.

[4] Nizar Touzi. Stochastic control and application to finance. `http://www.cmap.polytechnique.fr/~touzi/Master-LN.pdf`.