



Estimation de quantiles extrêmes et de probabilités d'événements rares

Arnaud Guyader, Nicolas Hengartner, Eric Matzner-Løber

► To cite this version:

Arnaud Guyader, Nicolas Hengartner, Eric Matzner-Løber. Estimation de quantiles extrêmes et de probabilités d'événements rares. 42èmes Journées de Statistique, 2010, Marseille, France, France. inria-00494788

HAL Id: inria-00494788

<https://inria.hal.science/inria-00494788>

Submitted on 24 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ESTIMATION DE QUANTILES EXTRÊMES ET DE PROBABILITÉS D'ÉVÉNEMENTS RARES¹

Arnaud GUYADER ^a, Nicolas W. HENGARTNER ^b et Eric MATZNER-LØBER ^a

^a Université Rennes 2 – Haute Bretagne
Campus Villejean
Place du Recteur Henri Le Moal, CS 24307
35043 Rennes Cedex, France
arnaud.guyader@uhb.fr, eml@uhb.fr

^b Information Sciences Group, MS B256
Los Alamos National Laboratory
NM 87545, USA
nickh@lanl.gov

Mots-clés Extrêmes, Méthodes bayésiennes.

Résumé Etant donné une probabilité μ sur $\mathbb{R}^d$ (d grand), on note X un vecteur aléatoire générique de loi μ et $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ une application “boîte noire”. Un réel q étant fixé, le but est de générer un échantillon i.i.d. $(X_1, \dots, X_N)$ tel que pour tout $i : X_i \sim \mathcal{L}(X|\Phi(X) > q)$. Lorsque q est grand comparé aux valeurs typiques de la variable $\Phi(X)$, la méthode Monte-Carlo classique devient trop coûteuse. Dans ce travail nous présentons et analysons une nouvelle approche pour ce problème. Celle-ci procède en plusieurs étapes, s’inspirant de l’algorithme de Metropolis-Hastings et des méthodes dites multi-niveaux en estimation d’événements rares. Deux problèmes peuvent être traités très facilement via cette nouvelle méthode : estimation de quantiles extrêmes et estimation d’événements rares. Les idées présentées seront illustrées sur un problème de tatouage numérique.

Abstract Let X denote a generic random vector with probability distribution μ on $\mathbb{R}^d$, and let Φ be a mapping from $\mathbb{R}^d$ to $\mathbb{R}$. That mapping can be a black box, e.g., the result from some computer experiments for which no analytical expression is available. Our goal is to generate an i.i.d. sample $(X_1, \dots, X_N)$ with $X_i \sim \mathcal{L}(X|\Phi(X) > q)$ together with the probability $\mathbb{P}[\Phi(X) > q]$ for any arbitrary real number q . When q lies far out in the right-hand tail of the distribution of the random variable $\Phi(X)$, a naive Monte-Carlo

¹Travail effectué dans le cadre des projets Nebbiano, ANR-06-SETI-009, et LANL LDRD 20080391ER.

simulation becomes computationally intractable. In this article we present and analyse a novel simulation algorithm for this problem. It proceeds by successive elementary steps, each one being based on Metropolis-Hastings algorithm. Our technique is useful for both estimating the probability of events and estimating extreme quantiles. We demonstrate the practical usefulness of our method by applying it to a problem in watermarking.

1 Rappels sur les méthodes Monte-Carlo

Notons X un vecteur aléatoire générique de dimension d et de loi μ sur l'espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$. On suppose pouvoir simuler des réalisations i.i.d. selon μ . On considère alors une fonction « boîte noire » $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$, c'est-à-dire que l'on sait seulement évaluer $\Phi(x)$ en tout point x de $\mathbb{R}^d$. Le dernier paramètre est un nombre réel q . Dans ce contexte, notre but est de simuler un échantillon i.i.d. $(X_1, \dots, X_N)$ de N particules distribuées suivant la restriction de la loi de X à l'événement $\mathcal{R} = \{\Phi(X) > q\}$. La difficulté vient du fait qu'on s'intéresse typiquement au cas où la probabilité $p = \mathbb{P}(\mathcal{R}) = \mathbb{P}(\Phi(X) > q)$ est très petite, typiquement inférieure à 10^{-9} .

Face à cette situation, la méthode Monte-Carlo classique, ou naïve, consiste à simuler un échantillon i.i.d. $(X_1, \dots, X_n)$ selon la loi μ et à ne garder que les N éléments, notés $(X_1, \dots, X_N)$ pour simplifier, tel que $\forall i = 1, \dots, N : \Phi(X_i) > q$. Il est clair que l'échantillon $(X_1, \dots, X_N)$ a la propriété voulue, à savoir que $X_i \sim \mathcal{L}(X | \Phi(X) > q)$. Néanmoins, la loi forte des grands nombres implique que la proportion d'intérêt N/n tend presque sûrement vers p lorsque n tend vers l'infini. Cette méthode nécessite donc de simuler un échantillon initial de taille $n \approx N/p$ pour aboutir à un échantillon de taille N . De fait, lorsque p est très petit, ceci devient numériquement déraisonnable.

Supposons maintenant qu'on veuille estimer p , avec toujours p très petit. La méthode Monte-Carlo classique consiste à simuler un échantillon i.i.d. $(X_1, \dots, X_N)$ selon la loi μ , et à considérer l'estimateur

$$\hat{p}_{mc} = \frac{\hat{N}_{mc}}{N} = \frac{\text{Card}\{i \in \{1, \dots, N\} : \Phi(X_i) > q\}}{N}.$$

Puisque $\hat{N}_{mc}$ suit une loi binomiale de paramètres (N, p) , il est clair que $\hat{p}_{mc}$ est non biaisé et de variance relative :

$$\frac{V(\hat{p}_{mc})}{p^2} = \frac{1-p}{Np} \approx \frac{1}{Np}.$$

Par conséquent, la taille N de l'échantillon à simuler doit être au moins de l'ordre de $1/p$ pour atteindre une précision acceptable, ce qui devient rédhibitoire lorsque l'événement d'intérêt est très rare.

Face à cette situation, une astuce classique consiste à effectuer un échantillonnage préférentiel (ou Importance Sampling), lequel revient à simuler l'échantillon $(X_1, \dots, X_N)$ selon une autre loi ν , pour laquelle l'événement $\mathcal{R} = \{\Phi(X) > q\}$ n'est plus rare. On estime alors la probabilité comme ci-dessus et on la corrige grâce au lien entre ν et μ . Ceci suppose néanmoins que l'on ait de l'information a priori sur le modèle afin de proposer une nouvelle loi pertinente ν d'échantillonnage. Mais puisqu'on suppose ne rien connaître sur Φ (qui est une boîte noire), cette méthode est ici exclue.

Supposons maintenant le problème inverse : l'estimation d'un quantile extrême. On suppose p très petit fixé le but est de trouver le quantile q tel que $\mathbb{P}(\Phi(X) > q) = p$. La méthode Monte-Carlo classique marche comme suit : simuler un échantillon i.i.d. $(X_1, \dots, X_N)$ selon la loi μ , les trier de sorte que $\Phi(X_{(1)}) < \dots < \Phi(X_{(N)})$ et considérer l'estimateur :

$$\hat{q}_{mc} = \Phi(X_{(\lfloor (1-p)N \rfloor)}),$$

où $\lfloor y \rfloor$ désigne la partie entière de y . Si la fonction de répartition $F(y) = \mathbb{P}(\Phi(X) \leq y)$ admet une dérivée non nulle au point q , alors on peut montrer (voir par exemple Arnold *et al.* [1], p.128) que le biais est en $O(1/N)$ et l'écart quadratique de la forme suivante :

$$\text{Var}(\hat{q}_{mc}) = \frac{p(1-p)}{N+2} \cdot \frac{1}{f^2(q)} + o(1/N) \approx \frac{1}{N} \cdot \frac{p}{f^2(q)}.$$

2 L'algorithme

Plutôt que la méthode Monte-Carlo classique, nous proposons l'algorithme suivant :

- Simuler un échantillon i.i.d. $(X_1, X_2, \dots, X_N)$ selon μ et initialiser

$$X_1^1 = X_1, \dots, X_N^1 = X_N.$$

- Pour $m = 1, 2, \dots$, noter

$$L_m = \min(\Phi(X_1^m), \dots, \Phi(X_N^m)),$$

et définir

$$X_j^{m+1} = \begin{cases} X_j^m & \text{si } \Phi(X_j^m) > L_m \\ X^* \sim \mathcal{L}(X | \Phi(X) > L_m) & \text{si } \Phi(X_j^m) = L_m. \end{cases}$$

- En notant

$$M = \max\{m : L_m \leq q\},$$

ce maximum étant par convention égal à 0 si $L_1 > q$, nous verrons que $L_M \approx q$ avec une très bonne précision. Ainsi l'échantillon

$$(X_1, \dots, X_N) = (X_1^{M+1}, \dots, X_N^{M+1})$$

est i.i.d. de loi approximativement égale à $\mathcal{L}(X | \Phi(X) > q)$.

Le point délicat de l'algorithme ci-dessus est de réussir à simuler une nouvelle particule X^* distribuée selon la loi $\mathcal{L}(X|\Phi(X) > L_m)$. L'idée est d'utiliser un noyau de transition K qui soit μ -réversible, c'est-à-dire tel que :

$$\forall (x, x') \in \mathbb{R}^d \times \mathbb{R}^d \quad \mu(dx)K(x, dx') = \mu(dx')K(x', dx).$$

En particulier, la loi μ est alors stationnaire pour le noyau K . Partant d'un noyau instrumental q supposé μ -irréductible, l'algorithme de Metropolis-Hastings (voir [4], [3] et [5]) permet précisément d'en déduire un noyau K ayant cette propriété.

A chaque étape m , notons alors $A_m = \{x \in \mathbb{R}^d : \Phi(x) > L_m\}$, μ_m la restriction normalisée de μ à A_m , c'est-à-dire $\mu_m(dx) = \frac{1}{\mu(A_m)} \mathbb{1}_{A_m}(x) \mu(dx)$, et considérons le noyau de transition K_m défini par :

$$K_m(x, dx') = \mathbb{1}_{A_m^c}(x) \delta_x(dx') + \mathbb{1}_{A_m}(x) (K(x, dx') \mathbb{1}_{A_m}(x') + K(x, A_m^c) \delta_x(dx')).$$

Le principe de construction de ce noyau est très simple : partant de $x > L_m$, une transition $x \rightsquigarrow x'$ est proposée par le noyau K . Si $\Phi(x') > L_m$, la transition est acceptée, sinon elle est rejetée et x reste à la même place. Il est alors facile de voir que la loi μ_m est stationnaire pour le noyau de transition K_m .

Ainsi, à l'étape m , le but est de simuler $X^* \sim \mathcal{L}(X|\Phi(X) > L_m)$, avec X^* indépendant des $(N-1)$ particules X_j^m vérifiant déjà $\Phi(X_j^m) > L_m$. L'idée est tout simplement d'appeler V_0 un vecteur aléatoire correspondant à l'une de ces particules (tirée au hasard) et de considérer la chaîne de Markov $(V_n)_{n \geq 0}$ de noyau de transition K_m . Pour n « assez grand », la particule $X^* = V_n$ est distribuée suivant la loi $\mathcal{L}(X|\Phi(X) > L_m)$ et est indépendante des autres particules.

3 Applications

3.1 Estimation d'événement rare

Avec les mêmes notations que précédemment, supposons qu'on veuille estimer la probabilité $p = \mathbb{P}(\Phi(X) > q)$. Il suffit alors de considérer l'estimateur :

$$\hat{p} = \left(1 - \frac{1}{N}\right)^M.$$

On montre que la variable aléatoire M suit une loi de Poisson, ce qui permet d'en déduire la loi de $\hat{p}$ ainsi qu'une expression simple pour tous ses moments.

Proposition 3.1 *La variable aléatoire M suit une loi de Poisson de paramètre $-N \log p$. Il s'ensuit que pour tout $k \in \mathbb{N}$:*

$$\mathbb{E}[\hat{p}^k] = p^{N(1-(1-1/N)^k)}.$$

Ainsi $\hat{p}$ est un estimateur sans biais de p , de variance :

$$\text{Var}(\hat{p}) = p^2 \left(p^{-\frac{1}{N}} - 1 \right).$$

Lorsque p est très petit, $\hat{p}$ est donc un estimateur bien plus précis que celui obtenu par l'algorithme de Monte-carlo naïf :

$$\frac{V(\hat{p})}{p^2} \approx -\frac{\log p}{N} \ll \frac{1-p}{Np} = \frac{\text{Var}(\hat{p}_{mc})}{p^2}.$$

D'autre part, puisque le nombre N de particules est au moins de l'ordre de la centaine, l'approximation de la loi de Poisson $\mathcal{P}(-N \log p)$ par la loi normale $\mathcal{N}(-N \log p, -N \log p)$ est légitime. Un intervalle de confiance de niveau 95% pour p est ainsi donné par :

$$\hat{p}e^{-2\sqrt{\frac{-\log p}{N}}} \leq p \leq \hat{p}e^{+2\sqrt{\frac{-\log p}{N}}}.$$

3.2 Estimation de quantile extrême

Réciproquement, une probabilité p très petite étant fixée, supposons que l'on cherche à déterminer le quantile q tel que $\mathbb{P}(\Phi(X) > q) = p$. Avec l'algorithme proposé, il suffit de noter :

$$m = \left\lceil \frac{\log(p)}{\log(1 - N^{-1})} \right\rceil \approx \lceil -N \log p \rceil,$$

et d'estimer le quantile q par $\hat{q} = L_m$. On montre que le biais de cet estimateur est en $O(1/N)$ et que sa variance est contrôlable.

Proposition 3.2 *Si la variable $\Phi(X)$ admet une densité $f(q)$ non nulle en q , alors l'erreur quadratique est asymptotiquement bornée comme suit :*

$$\frac{p^2(-\log p - 2)}{f(q)^2} \leq \lim_{N \rightarrow \infty} N \cdot \mathbb{E}[(\hat{q} - q)^2] \leq \frac{p^2(-\log p + 2)}{f(q)^2}.$$

Ici encore, l'estimation proposée par $\hat{q}$ est bien plus précise que celle offerte par l'estimateur Monte-Carlo classique $\hat{q}_{mc}$, puisque :

$$\text{Var}(\hat{q}) \approx \frac{1}{N} \cdot \frac{-p^2 \log p}{f(q)^2} \ll \text{Var}(\hat{q}_{mc}) \approx \frac{1}{N} \cdot \frac{p}{f^2(q)}.$$

Comme pour la comparaison entre $\hat{p}$ et $\hat{p}_{mc}$, ceci montre que pour atteindre la même précision, l'estimateur Monte-Carlo classique nécessite environ $(-p \log p)^{-1}$ fois plus de particules que la méthode proposée dans ce papier.

Notons au passage que l'écart quadratique de notre estimateur fait intervenir la valeur de la densité f au point q , laquelle est bien sûr inconnue. Ceci n'empêche néanmoins nullement la construction d'intervalles de confiance pour $\hat{q}$.

Proposition 3.3 *Notons :*

$$\begin{cases} m^+ = \left\lceil -N \log p + 2\sqrt{-N \log p} \right\rceil \\ m^- = \left\lfloor -N \log p - 2\sqrt{-N \log p} \right\rfloor \end{cases}$$

Alors l'intervalle $I_{1-\alpha}(q) = [L_{m^-}, L_{m^+}]$ est un intervalle de confiance à 95% pour le quantile q .

4 Comparaison et illustration

Lors de la présentation, la méthode proposée ici sera comparée à l'état de l'art en estimation d'événements rares (notamment au récent travail de Cérou *et al.* [2]) et illustrée sur une application en tatouage numérique (watermarking).

Références

- [1] B.C. Arnold, N. Balakrishnan, and H.N. Nagaraja. *A first course in order statistics*. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1992. A Wiley-Interscience Publication.
- [2] F. Cérou, P. Del Moral, T. Furon, and A. Guyader. Rare event simulation for a static distribution. Technical Report RR-6792, Inria, January 2009. Submitted.
- [3] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1) :97–109, April 1970.
- [4] N. Metropolis and S. Ulam. The Monte Carlo method. *J. Amer. Statist. Assoc.*, 44 :335–341, 1949.
- [5] C.P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, 1999.