Estimation des paramètres pour des modèles adaptés aux séries de records

Hoayek Anis[1] & Ducharme Gilles[2] & Khraibany Zaher[3] & Zeineddine Hassan[4]

[1] *IMAG : anis.hoayek@univ-montp2.fr*

[2] *IMAG : gilles.ducharme@univ-montp2.fr*

[3] *Université Libanaise : zaher.khraibani@gmail.com*

[4] *Université Libanaise : doyenfs@ul.edu.lb*

Résumé. Soit $X_1, ..., X_n$ une suite de variables aléatoires. Le record (supérieur) dans cette suite est la valeur M_n telle que $M_n = \max(X_1, \ldots, X_n)$. Un des modèles populaires en théorie de records est le *Linear Drift Model* (*LDM*), qui suppose que $X_n = Y_n + cn$, où c est le paramètre de dérive. Ce modèle a été étudié sur le plan probabiliste où on lui a trouvé de nombreuses propriétés intéressantes. Mais sur le plan statistique, peu de travaux ont porté sur l'estimation du paramètre c et le comportement de cet estimateur. Le but de ce travail est de présenter quelques estimateurs de ce paramètre de dérive et d'étudier leur comportement afin de produire des énoncés d'inférence statistique dont les risques d'erreur sont quantifiés avec une bonne précision.

Mots-clés. Record, *LDM*, Estimateurs, Inférence statistique.

Abstract. Let $X_1, ..., X_n$ a sequence of random variables. In such a time series, a record (upper) is the value M_n such that $M_n = \max(X_1, \ldots, X_n)$. One of the popular models in record theory is the *Linear Drift Model* (*LDM*), which assumes that $X_n = Y_n + cn$, where c is the drift parameter. This model has been studied on probabilistic level where we found many interesting properties. But statistically, little work has been done on the estimation of the parameter c and the behavior of this estimator. The aim of this work is to present some estimators of the parameter c and to study their behavior in order to produce statements of statistical inference where the risks of error are accurately calculated.

Keywords. Record, *LDM*, Estimators, Statistical inference.

1 Introduction

Soit $X_1, ..., X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées (*iid*). Le record (supérieur) dans cette suite est la valeur M_n telle que $M_n = \max(X_1, \ldots, X_n)$. Intuitivement, un record est un résultat dans une chaîne d'événements qui dépasse tout ce qui a été rencontré auparavant. De ce fait, les records attirent l'attention et revêtent souvent une importante particulière.

En faisant évoluer n on obtient une suite de records. L'étude du comportement stochastique d'une telle suite a commencé à se distinguer des méthodes classiques d'analyse des valeurs extrêmes quand on

a rencontré des problèmes où les seules données disponibles étaient liées aux valeurs de ces records et où l'observation des variables sous-jacentes X_n s'avérait difficile, voire impossible. De telles données sont fréquentes avec les compétitions sportives et dans des problèmes hydrologiques. Les champs d'application se sont progressivement étendus et couvrent maintenant des problèmes dans l'analyse des changements climatique, des risques d'émergence d'une pathologie, en finance, etc.

La théorie de records s'est d'abord développée (Chandler 1952, Arnold et al. 1998) dans le contexte où les X_n sont *iid* et où la loi de X_n est d'un type donné. Comme souvent les X_n sont eux-mêmes des extrêmes de valeurs extrêmes, (performance du médaillé d'or d'une épreuve olympique contre des compétiteurs eux-mêmes champions nationaux), il est fréquent de supposer cette loi comme étant de type Gumbel, Weibull ou Fréchet.

Par la suite, on s'est rendu compte que le cas *iid* ne collait pas bien à de nombreux jeux de données. En effet, dans le cas *iid* les records se concentrent parmi les premières observations et peu de records apparaissent après un certain temps, ce qui ne correspond pas à la réalité de compétitions sportives notamment. D'où l'intérêt de modèles qui dépassent l'hypothèse *iid*.

Un des modèles populaires en théorie de records est le *Linear Drift Model* (*LDM*), qui suppose que $X_n = Y_n + cn$, où les Y_n sont *iid* et où c est le paramètre de dérive. Ce modèle a été introduit par Ballerini et Resnick (1985 – 1987) et étudié par Borovkov (1999) et Nevzorov (2001) sur le plan probabiliste où on lui a trouvé de nombreuses propriétés intéressantes. Mais sur le plan statistique, peu de travaux ont porté sur l'estimation du paramètre c. A ce chapitre, on peut signaler les travaux de Smith (1988) qui apparente les données de records aux valeurs non censurées d'une suite X_n. Si l'idée apparaît raisonnable, la censure est cependant informative et les méthodes développées par Smith (1988), qui n'incorporent pas cette information cruciale, sont suspectes sur le plan de leur pertinence et de leur efficacité.

Dans cette présentation, nous introduisons quelques estimateurs du paramètre de dérive et étudierons leur comportement afin de produire des énoncés d'inférence statistique dont les risques d'erreur sont quantifiés. Ces estimateurs sont grossiers dans la mesure où ils n'utilisent pas la pleine richesse de l'information quand celle-ci est disponible, mais ils s'appliquent dans des contextes où une information tronquée ou peu fiable est la seule disponible. Ils sont aussi relativement faciles à calculer. Leur intérêt dans d'autres contextes n'est pas négligeable, en particulier leur facilité de calcul en fait d'excellents points de départ d'algorithmes d'optimisation.

2 Contexte général et notation

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité suffisamment riche pour supporter une suite $\{X_n\}_{n\geq 1}$ de copies indépendantes d'une variable aléatoire réelle X définie sur Ω. L'observation X_j est dite un *record (supérieur) au temps* j si sa valeur est supérieure à celle de toutes les observations précédentes :

$$X_j = \max\{X_1, \ldots, X_{j-1}\}.$$

On définit la variable aléatoire δ_j appelée *l'indicatrice d'un record* et N_n le nombre total de records parmi $X_1, \ldots, X_n$ par les relations :

$$\delta_j = \begin{cases} 1 & \text{si } X_j > \max(X_1, \ldots, X_{j-1}) \\ 0 & \text{sinon} \end{cases} \quad \text{et } N_n = \sum_{j=1}^{n} \delta_j.$$

L'utilité de la théorie de records apparaît quand les seules données disponibles sont les deux suites a) $\{R_n : n \geq 1\}$ où $R_1 = X_1$ est le record trivial et R_n est la suite des records suivants de la suite et b) $\{\delta_n : n \geq 1\}$. Alternativement, on peut remplacer la suite $\{\delta_n : n \geq 1\}$ par la suite $\{L_n : n \geq 1\}$ dite des *indices des records* où $L_1 = 1$ (record trivial) et L_n est l'indice du $n^{ième}$ record. Avec cette notation, on a

$$R_n \quad = \quad X_{L_n} = \text{valeur du } n^{ième} \text{ record.}$$

3 Modèle LDM

Dans le modèle *LDM*, la $n^{ième}$ observation s'écrit $X_n = Y_n + cn$. Comme on considère les records supérieur, la dérive $c > 0$ et les Y_n sont des variables aléatoires *iid* de fonction de répartition $F(\cdot)$ ayant pour densité $f(\cdot)$. Ici, nous considérons le cas où la loi des Y_n est entièrement spécifiée. Soit $P_n(c)$ la probabilité que la $n^{ième}$ observation soit un record. Il est bien connu que :

$$P_n(c) = \mathbb{P}_c[\delta_n = 1] = \int f(y) \left(\prod_{k=1}^{n-1} F(y + ck) \right) \mathrm{d}y.$$

Si $\mathbb{E}(Y_n) < \infty$, on peut définir un *taux de record asymptotique* :

$$P(c) = \lim_{n \to \infty} P_n(c).$$

Ces quantités dépendent de c, ce qui montre que l'on peut extraire de la suite $\{\delta_n\}_{n \geq 1}$ des informations sur ce paramètre. Par ailleurs, comme les δ_n indiquent si X_n a été observée ou si elle a été censurée, on a donc, en analogie avec les modèles rencontrés en analyse de suivie, une censure informative.

Dans de nombreux problèmes, les données disponibles pour l'estimation de c est la série des couples $\{R_n, \delta_n\}_{n \geq 1}$. Nous allons considérer ici le cas où seule la suite $\{\delta_n\}_{n \geq 1}$ est exploitable. En effet, dans certaines applications, il se peut que les R_n ne soient pas (entièrement) connus ou que leur fiabilité puissent être mise en doute. De plus, nous supposons le cas où la distribution des Y_n suit la loi de Gumbel, $G(\mu, \beta)$ et, sans perte de généralité, posons $\mu = 0$, $\beta = 1$. La plausibilité de cette supposition vient du fait que :

1. Dans un modèle *LDM*, la distribution $G(\mu, \beta)$ est l'unique loi caractérisée par l'indépendance des δ_n (Borovkov 1999) .
2. Gumbel est le domaine d'attraction de nombreuses densités populaires (Normale, Exponentielle, ...).

4 Estimation ponctuelle de la dérive c

Pour estimer ponctuellement la dérive c avec la suite $\{\delta_n\}_{n \geq 1}$, nous allons présenter plusieurs approches. La plus simple utilise la distribution de probabilité du nombre de records N_n :

$$\mathbb{P}_c[N_n = m] = \frac{\exp(-nc)\, \mathcal{S}(n, m \mid \vec{u})}{\prod_{j=1}^{n} u_j}, \tag{1}$$

où le vecteur $\vec{u} = (u_0, \ldots, u_n)$ est défini par $u_j = \frac{e^{-c}\left(1 - e^{-jc}\right)}{\left(1 - e^{-c}\right)}$, avec $u_0 = 0$, et $\mathcal{S}(n, m \mid \vec{u})$ est la généralisation du nombre de Stirling de première espèce (Khraibani 2015). Ayant observé $N_n = m$, il suffit de déterminer la valeur $\widehat{c}_1$ qui maximise (en c) l'expression (1).

Une autre approche simple utilise le premier moment de N_n. Comme $\frac{N_n}{n} \longrightarrow P(c) = 1 - e^{-c}$ presque surement (Ballerini et Resnick 1985), on définit un deuxième estimateur $\widehat{c}_2 = -\log\left(1 - \frac{N_n}{n}\right)$. Une amélioration de cette idée sera aussi présentée.

Nous montrerons cependant que la statistique N_n n'est pas exhaustive pour c. En utilisant uniquement cette valeur, on perd de l'information. Un troisième estimateur $\widehat{c}_3$ de c peut être obtenu par le maximum de vraisemblance en utilisant la distribution de probabilité des indicatrices de records. On montrera que cet estimateur, noté $\widehat{c}_3$, est la solution de l'équation :

$$\frac{\mathrm{d}\log(L(\tau))}{\mathrm{d}\tau} = \frac{-m}{1-\tau} + \frac{n-m}{\tau} + \frac{n\tau^{n-1}}{1-\tau^n} + \sum_{j=2}^{n} \delta_j \frac{(j-1)\,\tau^{j-2}}{1-\tau^{j-1}} = 0, \tag{2}$$

avec $\tau = \exp(-c)$.

Pour la calculer, on devra recourir à une méthode numérique en prenant $\tau = \exp(-\widehat{c}_2)$ comme point de départ de l'algorithme. Ce qui du coup justifie le travail accompli en utilisant N_n.

5 Comportement des estimateurs.

On montrera que les estimateurs $\widehat{c}_2$ et $\widehat{c}_3$ sont asymptotiquement de loi normale et on donnera l'expression de leur variance. Ceci permettra de construire des intervalles de confiance et des tests d'hypothèses concernant la valeur de c dont les risques d'erreur découleront des approximations asymptotiques.

Ensuite, pour évaluer la qualité de ces estimateurs pour des tailles d'échantillons finis, des résultats de simulations seront présentés. Ceux-ci permettront d'apprécier leur biais, leur variance et la proximité de cette variance avec celle des approximations asymptotiques. Ces résultats seront complétés par des simulations sur la probabilité de couverture réelle des intervalles de confiance basés sur les approximations asymptotiques.

De ces résultats théoriques et empiriques, on concluera que l'estimateur $\widehat{c}_3$ est le meilleur selon le critère MSE, surtout pour les petites valeurs de c, ce qui est le cas habituellement rencontré dans les applications. Ainsi son utilisation en pratique pourra s'appuyer sur un risque calculé avec une bonne précision.

Références

[1] Arnold, B. C. Balakrishnan, N. Nagaraja, H. N. (1998) : Records. Wiley.

[2] Ballerini, R. Resnick, S. (1985) : Records from Improving Populations. Journal of Applied Probability.

[3] Ballerini, R. Resnick, S. (1987) : Records in the Presence of a Linear Trend. Advances in Applied Probability, Vol. 19, No. 4, pp. 801-828.

[4] Borovkov, K. (1999) : On records and related processes for sequences with trends. Journal of Applied Probability, Vol. 36, No. 3, pp. 668-681.

[5] Chandler, K. N. (1952) : The distribution and frequency of records values. J. Roy. Statist. Soc. Ser. B. 14, pp. 220-228.

[6] Khraibani, Z. Jacobb, C. Ducrotc, C. Charras-Garridoc, M. Salad, C. (2015) : A Non Parametric Exact Test Based on the Number of Records for an Early Detection of Emerging Events : Illustration in Epidemiology. Communications in Statistics : Theory and Methods, Volume 44, Issue 4, pages 726-749.

[7] Leroy, F. Dauxois, J. Y. Tubert-Bitter, P. (2013) : On the parametric maximum likelihood estimator for independent but non-identically distributed observations with application to truncated data. HAL Id : hal-00865962.

[8] Nevzorov, V. B. (2001) : Records : mathematical theory. American Mathematical Society.

[9] Smith, R. L. (1988) : Forecasting Records by Maximum Likelyhood. Journal of the American Statistical Association, Vol 83, No. 402, pp. 331-338.