Estimation d'un quantile extrême sur données dichotomiques
Application à l'étude de la résistance de matériau

Émilie Miranda[1] [2]

[1] *LSTA, Université Pierre et Marie Curie, Paris VI*
4 place Jussieu, 75252 Paris Cedex 05, France
emilie.miranda@upmc.fr
[2] *Safran Aircraft Engines*
Rond-point René Ravaud, 77550 Moissy-Cramayel

Résumé. L'objectif de cette étude est l'estimation d'un quantile extrême dans le cas de données dichotomiques de dépassement de seuil. La problématique est issue d'un cas industriel : l'étude de la résistance d'un matériau pour des probabilités de rupture cibles très faibles. La méthode proposée est séquentielle et consiste à décomposer la probabilité de l'évènement rare en un produit d'évènements conditionnels. Elle se fonde sur l'utilisation de résultats sur les lois limites de dépassements de seuil.

Mots-clés. Méthode séquentielle, Valeurs extrêmes, Estimation de quantile, Données incomplètes

Abstract. The objective of this study is the estimation of an extreme quantile based on a sample of dichotomic data corresponding to peaks over a threshold. It comes from an industrial case : what stress can be applied to a material to guarantee a long lifetime with high probability. The approach is sequential. It consists in decomposing the target probability level, which is very low, into a product of probabilities of conditional events of higher order.

Keywords. Sequential analysis, Extreme value, Quantile estimation, Incomplete data

1 Cadre industriel : détermination de la contrainte admissible en fatigue des matériaux

Les pièces de turboréacteurs subissent différents types de sollicitations qui peuvent conduire à leur endommagement. Cette étude s'intéresse à la résistance de pièces soumises à l'application cyclique d'une charge (cf. figure 1). On parle alors d'endommagement en fatigue.

Afin d'étudier la résistance d'un matériau, des études expérimentales sont menées, reproduisant les conditions de vol. Les résultats de ces plans d'essai permettent de caractériser le comportement en fatigue d'un matériau.

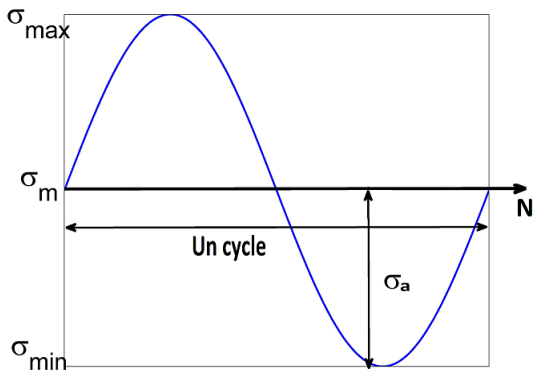

FIGURE 1 – Application de cycles d'effort sur un matériau

$\sigma_{\min}$ et $\sigma_{\max}$ sont les niveaux de contrainte maximale et minimale ; σ_m, le niveau moyen et $\sigma_a = \frac{\sigma_{\max} - \sigma_{\min}}{2}$, la contrainte alternée, définie comme la demi amplitude de la variation du chargement.

Ces travaux ont pour objet la caractérisation du niveau de contrainte s_α qui peut être appliqué sur un matériau, afin de garantir une durée de vie n_0 très grande avec une probabilité $1 - \alpha$ où $\alpha \approx 10^{-3}$. Formellement, notons N la durée de vie d'un matériau et S le niveau de contrainte appliqué, la quantité cible s_α est le seuil vérifiant :

$$s_\alpha = \underset{s}{\operatorname{arginf}}\ \mathbb{P}(N \leq n_0 | S = s) \leq \alpha \tag{1}$$

La finalité est de proposer une méthode d'estimation robuste de cette contrainte admissible à partir du plus petit nombre d'essais possibles.

2 Modélisation

La modélisation retenue fait intervenir une variable R_{n_0}, correspondant à la résistance interne du matériau pour un horizon de vie n_0 fixé. R est définie comme le niveau de contrainte en dessus duquel le matériau rompt avant n_0.

La résistance est ainsi la variable d'intérêt de l'étude. Elle est homogène à la contrainte et sa loi $\mathbb{P}_0$ est telle que :

$$\mathbb{P}(N \leq n_0 | S = s) = \mathbb{P}(R \leq s) \tag{2}$$

Ainsi, le seuil admissible recherché est le quantile d'ordre α de la distribution de la variable d'intérêt R.

Cependant, au cours des essais de fatigue, la résistance n'est pas observée. En effet, les résultats sont les couples formés par les temps à rupture et les sollicitations appliquées. R est une variable latente dans le modèle et l'information pertinente dont on

dispose est définie par (2) : $Y = \mathbb{1}_{R \leq s} = \mathbb{1}_{N \leq n_0 | S=s}$, indiquant si l'éprouvette a rompu sous une contrainte s à une durée inférieure à n_0, cela signifie que sa résistance était inférieure à s.

3 Méthodologie d'estimation d'un quantile extrême

La principale difficulté consiste à étudier la survenue d'un évènement qui n'est pas observé dans des conditions standards d'essai, ce qui oblige à extrapoler à partir de la loi estimée. Or l'estimation de la loi se fonde sur un échantillon réduit de données censurées à droite et à gauche (indicatrices de dépassement de seuil).

3.1 Splitting

Afin de se ramener à des problèmes d'estimation moins difficiles, la probabilité α est réécrite comme le produit de probabilités de niveaux plus élevés et donc plus facilement évaluables à partir d'un nombre limité d'observations. Formellement, l'évènement $\{R \leq s_\alpha\}$ peut se décomposer en l'intersection d'évènements conditionnels. Soit une série de m évènements inclusifs $\{R \leq s_\alpha\} = \{R \leq s_m\} \subset \cdots \subset \{R \leq s_1\}$, avec $s_\alpha = s_m < s_{m-1} < ... < s_1$, la probabilité de rupture sous s_α se réécrit alors :

$$\mathbb{P}_0(R \leq s_\alpha) = \mathbb{P}_0(R \leq s_1) \prod_{j=1}^{m-1} \mathbb{P}_0(R \leq s_{j+1} \mid R \leq s_j) \tag{3}$$

La détermination des seuils $(s_j)_{j=1,\ldots,m}$ est fonction de l'ordre des probabilités $(\mathbb{P}_0(R \leq s_{j+1} \mid R \leq s_j))_j$, qui est lui-même le résultat d'un arbitrage entre le nombre d'essais à réaliser et le nombre de niveaux de contraintes nécessaires à la reconstitution de α. En pratique, p est fixé entre 20 et 30 %.

La décomposition (3) a pour but de faire apparaître une méthode séquentielle de détermination de s_α.

On pose une hypothèse paramétrique sur la loi de R.

1. Un premier seuil s_1 est fixé et les premiers essais sont réalisés : $Y_1, \ldots, Y_n$, avec $Y_i = \mathbb{1}_{R_i < s_1}$;
2. Estimation de la loi de R, et détermination de s_2, le quantile à $p \approx 20\%$ de la loi $\mathbb{P}_0$;
3. Pour $j = 2, \ldots, m-1$, n essais sont réalisés en s_j.
4. Détermination du quantile s_{j+1} d'ordre p de la loi conditionnelle de $R \mid R \leq s_j$;

3.2 Modélisation de la loi de la résistance

La loi de la résistance est modélisée par une transformation d'une loi de Pareto Généralisée. La motivation est la suivante : la loi GDP correspond au régime limite de probabilités de dépassements de seuil renormalisés d'une variable aléatoire positive.

Soit $\widetilde{R} = \frac{1}{R}$,

$$\lim_{\substack{s\to 0\\ x<s}} \mathbb{P}\left(R < x \mid R < s\right) = \lim_{\substack{\frac{1}{s}\to\infty\\ \frac{1}{x}>\frac{1}{s}}} \mathbb{P}\left(\widetilde{R} > \frac{1}{x} \mid \widetilde{R} > \frac{1}{s}\right) = G(x)$$

où, G est de la forme :

$$1 - G(x) = \begin{cases} (1+\frac{c}{a}x)^{-1/c} & \text{si } c \neq 0 \\ \exp(-\frac{x}{a}) & \text{si } c = 0 \end{cases}$$

avec $\begin{cases} x \geq 0 & \text{si } c \geq 0 \\ 0 \leq x \leq -\frac{a}{c} & \text{si } c < 0 \end{cases}$

La GDP possède une propriété de stabilité qui facilite la décomposition en produit de probabilités conditionnelles. En effet, si $\widetilde{R} \sim GPD(c, a_0)$, de f.d.r G, alors et la loi conditionnelle de $\widetilde{R} - s|\widetilde{R} > s \sim GPD(c, a_s)$, avec $a_s = a_0 + cs$, de f.d.r G_s.

On suppose donc $\widetilde{R} \sim GPD(c, a_0)$. L'équation (3) se réécrit avec $\widetilde{R}$: pour une séquence de niveaux $\widetilde{s}_m > \widetilde{s}_{m-1} > \cdots > \widetilde{s}_1$,

$$\begin{aligned} \mathbb{P}_0(\widetilde{R} > \widetilde{s}_m) &= \mathbb{P}(\widetilde{R} > \widetilde{s}_1) \prod_{j=1}^{m-1} \mathbb{P}_0(\widetilde{R} > \widetilde{s}_{j+1} | \widetilde{R} > \widetilde{s}_j) \\ &= \bar{G}(\widetilde{s}_1) \prod_{j=1}^{m-1} \bar{G}_{\widetilde{s}_j}(\widetilde{s}_{j+1} - \widetilde{s}_j) \\ &= \left(1 + \frac{c\widetilde{s}_1}{a_0}\right)^{-1/c} \prod_{j=1}^{m-1} \left(1 + \frac{c(\widetilde{s}_{j+1} - \widetilde{s}_j)}{a_j}\right)^{-1/c} \end{aligned}$$

où $a_j = a_0 + c\widetilde{s}_j$

3.3 Procédure d'estimation

La nature dégradée des données rend l'estimation difficile : les méthodes classiques d'estimation (maximum de vraisemblance ou minimum de divergence) ne permettent pas à elles seules d'obtenir une estimation correcte des paramètres. En effet, il existe tout un ensemble valeurs des paramètres quasiment équivalentes du point de vue des critères d'estimation considérés. Pour pallier ce problème, l'approche retenue consiste à ajouter un critère de discrimination.

La procédure ayant pour but l'estimation d'un quantile, la solution retenue consiste à intégrer un critère de stabilité sur ce dernier.

Ainsi, à chaque étape j,

1. Une série de n essais est réalisée au seuil en cours s_j. À partir de la proportion de dépassements observés (ruptures) $\widehat{p}_j = \frac{1}{n}\sum_i Y_i$, on construit un intervalle de confiance d'ordre β, noté $I_\beta(p_j)$.

2. L'ensemble des paramètres assurant que la proportion théorique, définie par le modèle de Pareto, appartient à I_β est obtenu en utilisant un algorithme d'optimisation ensembliste (SAFIP [5]) :

$$\mathcal{S}_j = \left\{(\widehat{c}_k, \widehat{a}_k)_k : p_{j,k} = \bar{G}_{\widehat{c}_k,\widehat{a}_k} \in I_\beta\right\}$$

3. Parmi $\mathcal{S}_j$, le couple de paramètres retenu est celui qui satisfait au mieux un critère de stabilité sur les quantiles précédents, soit :

$$(\widehat{c}^*, \widehat{a}^*) = \underset{k}{\operatorname{argmin}} \, |G^{-1}_{(\widehat{c}_k,\widehat{a}_k)}(\widehat{p}_{j-1}) - s_{j-1}|$$

ou bien, en utilisant un critère prenant en compte tous les résultats obtenus au cours des étapes précédentes :

$$(\widehat{c}^*, \widehat{a}^*) = \underset{k}{\operatorname{argmin}} \sum_{\ell=1}^{j} \lambda_\ell |G^{-1}_{(\widehat{c}_k,\widehat{a}_k)}(\widehat{p}_\ell) - s_\ell|$$

avec $\sum_{\ell=1}^{j} \lambda_\ell = 1$.

La procédure a été testée sur données simulées en comparaison avec d'autres méthodes séquentielles d'estimation. De plus, les résultats ont été comparés à ceux obtenus sur données complètes (i.e. à partir d'observations directes de la résistance). Ces derniers fournissent une borne supérieure des performances de la méthode d'estimation.

Références

[1] Balkema A. A. and De Haan L. Residual life time at great age. *Ann. Prob.*, 2(5) :762–804, 1974.

[2] De Valk C. Approximation of high quantiles from intermediate quantiles. *Extremes*, 4 :661–684, 2016.

[3] Beirlant J., Goegebeur Y., and Teugels J. *Statistics of Extremes, Theory and Applications*. Wiley, 2004.

[4] Pickands J. Statistical inference using extreme order statistics. *Ann. Stat.*, 3(1) :119–131, 1975.

[5] Biret M. and Broniatowski M. Safip : a streaming algorithm for inverse problems. 2016.

[6] Feller W. *An introduction to probability theory and its applications*, volume 2. Wiley, 1971.