Société de Calcul Mathématique, S. A.
Outils d'aide à la décision
depuis 1995

Régularisation probabiliste de données irrégulières

Bernard Beauzamy

Juillet 2015

I. Introduction

Dans l'un de nos contrats, relatif à l'enfoncement d'un terrain, on constate que les données enregistrées en chaque point sont irrégulières ; du fait des erreurs de mesure, il est localement difficile de constater un enfoncement. La situation est la suivante : on dispose (en un point donné) d'une série temporelle représentant la position d'un point (en fait, sa cote z) et cette série devrait montrer un accroissement de z (en valeur absolue), ce qui n'est pas le cas en pratique du fait des erreurs de mesure. La solution retenue pour les éliminer en partie est d'avoir recours à des moyennes glissantes, mais ceci reste artificiel.

Nous indiquons ici une méthode purement probabiliste qui va permettre de régulariser la série de données. Elle est inspirée de la méthode présentée dans mon livre [GRE] pour "régulariser" les données relatives aux phénomènes extrêmes.

Nous travaillons sur un point donné ; nous disposons d'une série z_k indexée par le temps (ici les années, pour simplifier). Dans le cas du contrat, il n'y a qu'une seule mesure à chaque fois. Mais, pour bien comprendre la méthode, nous allons supposer que nous travaillons sur deux années seulement (mettons 2011 et 2012), et que, pour chaque année, nous avons un grand nombre de relevés. Ceci n'a rien d'impossible.

II. Présentation de la méthode

Nous discrétisons l'échelle des enfoncements, par centimètre, de 0 à 100 cm (mettons). L'échelle sera la même pour les deux années : ceci est important.

Notons λ_0 la probabilité que l'enfoncement en 2011 soit de 0 cm, λ_1 la probabilité qu'il soit de 1 cm, et ainsi de suite jusqu'à λ_{100} ; ces probabilités sont inconnues et seront traitées comme des taux de risque (voir [NMP]). On a $\lambda_0 + \cdots + \lambda_{100} = 1$; de même, soient $\mu_0, ..., \mu_{100}$ pour 2012.

Soit n_k le nombre de mesures en 2011 qui auront donné un enfoncement de k centimètres. Nous savons que la loi conjointe de l'ensemble des taux de risque (voir le livre [NMP]) est, en tenant compte des observations :

$$f(\lambda_0, ...\lambda_{99}) = \lambda_0^{n_0} ... \lambda_{99}^{n_{99}} (1 - \lambda_0 - ... - \lambda_{99})^{n_{100}}$$

et la même chose pour 2012 avec des nombres m_k .

Maintenant, nous voulons rendre compte du fait que le mouvement ne peut se faire que dans un seul sens : le parking s'enfonce. L'échelle de mesure correspond à des courbes de niveau. Pour tout niveau k, il est plus probable de voir le niveau dépassé en 2012 qu'en 2011. La probabilité de rester au-dessous du niveau k en 2011 est $\lambda_0 + ... + \lambda_k$ et en 2012 c'est $\mu_0 + ... + \mu_k$.

Cela se traduit par les inégalités :

$\lambda_0 \geq \mu_0$

$\lambda_0 + \lambda_1 \geq \mu_0 + \mu_1$

etc.,

$\lambda_0 + ... + \lambda_k \geq \mu_0 + ... + \mu_k$ (1)

III. Un exemple numérique

Voici un exemple avec quatre classes seulement :

λ	μ
0.15	0.1
0.25	0.2
0.20	0.21
0.40	0.49

La loi conjointe des taux de risque $\lambda_0, ..., \lambda_{99}, \mu_0, ..., \mu_{99}$ sera donc donnée par la densité :

$$\varphi(\lambda_0, ..., \lambda_{99}, \mu_0, ..., \mu_{99}) = 1_S \lambda_0^{n_1} \cdots \lambda_{99}^{n_{99}} (1 - \lambda_0 - ... - \lambda_{99})^{n_{100}} \mu_0 .... \mu_{99}^{n_{99}} (1 - \mu_0 - ... - \mu_{99})^{m_{100}} \quad (2)$$

où S est le simplexe caractérisé par $\lambda_k \geq 0,\ \sum \lambda_k = 1$, la même chose pour les μ_k, et toutes les inégalités (1). Bien sûr, il faut normaliser la fonction φ en divisant par l'intégrale sur tout le simplexe.

Dans la pratique, avec les données du contrat, la fonction φ est très simple, puisque tous les n_k sont nuls, sauf 1 qui prend la valeur 1, et de même pour tous les m_k.

La valeur moyenne attribuée à λ_0 sera alors l'espérance de (2) par rapport à la première variable, c'est-à-dire :

$$E(\lambda_0) = \int \lambda_0 \varphi(\lambda_0, ..., \lambda_{99}, \mu_0, ..., \mu_{99})\, d\lambda_0 \cdots d\lambda_{99} d\mu_0 \cdots d\mu_{99} \qquad (3)$$

et de même pour tous les λ_k et tous les μ_k

Une fois que nous avons une valeur moyenne pour les λ_k et les μ_k, il est facile de donner une valeur moyenne de l'enfoncement chaque année ; pour 2011 ce sera :

$$Enfoncement_{2011} = \sum_{k=0}^{100} k\lambda_k$$

$$Enfoncement_{2012} = \sum_{k=0}^{100} k\mu_k$$

IV. Méthode pratique de calcul

Les intégrales (2) et (3) ne sont généralement pas faciles à calculer, même si la fonction dépend de peu de variables, parce que le domaine est défini par un grand nombre d'inégalités. Il sera sans doute préférable d'utiliser une méthode de type Monte-Carlo : on tire au hasard un 100-uple de λ_k et un 100-uple de μ_k et on ne s'en sert pour estimer les intégrales que s'il vérifie les inégalités (1). Voir le livre [GRE].

1. Cas de deux années, deux classes de mesure à chaque fois.

On a simplement λ, μ avec les inégalités : $1 \geq \lambda \geq \mu \geq 0$

Si on fait une intégration, on fixe d'abord λ et on intègre par rapport à μ entre 0 et λ, puis par rapport à λ entre 0 et 1.

Si on utilise la méthode aléatoire, il faut respecter l'ordre, selon la méthode de Peter Robinson, voir [GRE].

2. Cas de trois classes de mesure

On a $\lambda_1, \lambda_2, \mu_1, \mu_2$ avec les inégalités :

$1 \geq \lambda_1 \geq \mu_1 \geq 0$

$1 \geq \lambda_1 + \lambda_2 \geq \mu_1 + \mu_2 \geq 0$

3. Calcul par intégration

Fixons $\lambda_1, \mu_1, \lambda_2$ avec $\lambda_1 \geq \mu_1$; intégrons par rapport à μ_2 entre 0 et $\lambda_1 + \lambda_2 - \mu_1$, puis par rapport à μ_1 entre 0 et λ_1, puis par rapport à λ_2 entre 0 et $1 - \lambda_1$, puis par rapport à λ_1 entre 0 et 1 (note : il n'y a pas d'ordre entre λ_1 et λ_2, ni entre μ_1 et μ_2).

4. Calcul aléatoire

On génère aléatoirement $\lambda_1, \lambda_2, \lambda_3$ de somme 1, μ_1, μ_2, μ_3 de somme 1, et on garde seulement les runs qui vérifient les relations précédentes. Mais cette méthode, lorsque le nombre de classes est élevé, sera très coûteuse : presque tous les candidats seront éliminés.

Rappel : pour générer des variables aléatoires de somme 1, il faut générer des variables selon la loi exponentielle, et ensuite diviser par la somme.

V. Références

[NMP] Bernard Beauzamy : Nouvelles Méthodes Probabilistes pour l'évaluation des risques. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA. ISBN 978-2-9521458-4-8. ISSN 1767-1175, avril 2010.

[GRE] Bernard Beauzamy : Méthodes probabilistes pour la gestion des risques extrêmes. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA. ISBN : 978-2-9521458-9-3, ISSN : 1767-1175. Relié, 208 pages, juin 2015.