

Modélisation Charge Sinistre M2 Actuariat

Chapitre IV: Inférence d'une distribution composée

Pierre-Olivier Goffard

Université de Lyon 1
ISFA

`pierre-olivier.goffard@univ-lyon1.fr`

ISFA

November 29, 2021

I. Problème

Soit une variable aléatoire de la forme

$$X = \sum_{i=1}^N U_i$$

où

- N est une variable aléatoire de comptage de fonction de masse $p_N(\cdot, \theta_N)$
- (U_i) est une suite de variables aléatoires iid de fonction de densité $f_U(\cdot, \theta_U)$

Soient x_1, \dots, x_t un échantillon d'observations iid de X sur t période d'exercice.

Problem 1

Comment inférer la valeur des paramètres $\theta = (\theta_N, \theta_U)$?

Remark 1

Il s'agit d'un problème de données incomplètes, les données complètes comprennent les fréquences et montants individuels de sinistre

$$(n_s, s) = \{n_s, (u_{1,s}, \dots, u_{n_s, s})\}, \quad s = 1, \dots, t.$$

Ce problème correspond à la décomposition d'une somme aléatoire



Boris Buchmann and Rudolf Grübel.

Decomposing: an estimation problem for Poisson random sums.

Ann. Statist., 31(4):1054–1074, 08 2003.

Il est étudié en actuariat



Pierre-Olivier Goffard and Patrick J. Laub.

Approximate bayesian computations to fit and compare insurance loss models.

Insurance: Mathematics and Economics, 100:350–371, sep 2021.

et dans la modélisation des précipitations.



Peter K. Dunn.

Occurrence and quantity of precipitation can be modelled simultaneously.

International Journal of Climatology, 24(10):1231–1239, jul 2004.

Example 1

Supposons que $N \sim \text{Geom}(p)$ et $U \sim \text{Exp}(\lambda)$ avec

$$p_N(k) = (1-q)q^k, \quad k \geq 0, \quad \text{et} \quad f_U(x) = e^{-x/\delta}/\delta, \quad x \geq 0.$$

II. Approche fréquentiste

1. Méthode des moments

La méthode des moments consiste à faire matcher les moments théopriques et les moments théoriques en résolvant le système

$$\mathbb{E}(X^k) = \mu_k, \quad k \geq 1$$

où les μ_k désignent les moments empiriques

$$\mu_k = \frac{1}{n} \sum_{j=1}^n x_j^k$$

Dans notre exemple $X \sim \text{Geom}(p) - \text{Exp}(\lambda)$, seuls les moments jusqu'à l'ordre 2 sont nécessaires. On résout donc

$$\begin{cases} \bar{X} = \mathbb{E}(X), \\ S_n^2 = \mathbb{V}(X), \end{cases} \quad (1)$$

où $\bar{X} = \frac{1}{n} \sum_{j=1}^n x_j$ et $S_n^2 = \frac{1}{n-1} \sum (x_j - \bar{X})^2$. On a

$$\mathbb{E}(X) = \frac{q\delta}{(1-q)} \quad \text{et} \quad \mathbb{V}(X) = \frac{q\delta^2}{1-q} + \frac{q\delta^2}{(1-q)^2}.$$

Le système (1) est équivalent à

$$\begin{cases} \bar{X} = \frac{q\delta}{(1-q)} \\ S_n^2 = \frac{q\delta^2}{1-q} + \frac{q\delta^2}{(1-q)^2} \end{cases} \quad (2)$$

Le système (2) d'inconnu q et δ est non-linéaire.

- Solution pas forcément unique
- Il faut prendre en compte les contraintes $0 < q < 1$ et $\delta > 0$
- Variance empirique assez volatile

Estimation alternative en exploitant le nombre de zéros dans les données:

$$\hat{q} = 1 - \frac{t_0}{t} \quad \text{et} \quad \hat{\delta} = \frac{(1 - \hat{q})\bar{X}}{\hat{q}},$$

où t_0 correspond au nombre de zéros dans les données.

Remark 2

Les estimateurs sont consistants, ils ne sont pas viables si les données ne comprennent pas de zéro ou que des zéros.

2. Maximum de vraisemblance

La distribution de X est mixte au sens ou

$$d\mathbb{P}_X(x) = \mathbb{P}(N=0)d\delta_0(x) + \sum_{n=1}^{+\infty} f_U^{*n}(x)\mathbb{P}(N=n)d\lambda(x) = \mathbb{P}(N=0)d\delta_0(x) + f_X^+(x)d\lambda(x).$$

Dans le cadre de notre modèle, nous avons

$$\mathbb{P}(N=0) = 1 - q, \text{ et } f_X^+(x) = \frac{q(1-q)}{\delta} \exp\left(-x \frac{1-q}{\delta}\right)$$

La vraisemblance s'écrit

$$L(\mathbf{x}; \theta) = (1-q)^{t_0} \left(\frac{q(1-q)}{\delta}\right)^{t-t_0} \exp\left[-\frac{1-q}{\delta} \sum_{i=1}^{t-t_0} x_i^+\right],$$

où $x_1^+, \dots, x_{t-t_0}^+$ sont les observations strictement positives. On cherche les valeurs de δ et q qui maximise la log vraisemblance

$$l(\mathbf{x}; \theta) = t \ln(1-q) + (t-t_0) \ln(q) - (t-t_0) \ln(\delta) - \frac{1-q}{\delta} \sum_{i=1}^{t-t_0} x_i^+,$$

avec

$$(\hat{q}, \hat{\delta}) = \operatorname{argmax}_{q \in (0,1), \delta > 0} l(\mathbf{x}; q, \delta)$$

On résout le système

$$\begin{cases} \frac{\partial l}{\partial q} = 0 \\ \frac{\partial l}{\partial \delta} = 0 \end{cases} \quad (3)$$

On a

$$\frac{\partial l}{\partial q} = -\frac{t}{1-q} + \frac{t-t_0}{q} + \frac{1}{\delta} \sum_{s=1}^{t-t_0} x_s^+, \text{ et } \frac{\partial l}{\partial \delta} = -\frac{t-t_0}{\delta} + \frac{1-q}{\delta^2} \sum_{s=1}^{t-t_0} x_s^+.$$

et

$$\hat{q} = \frac{t-t_0}{t}, \text{ et } \hat{\delta} = \frac{t_0}{t} \frac{1}{t-t_0} \sum_{s=1}^{t-t_0} x_s^+.$$

III. Approche Bayésienne

L'inférence Bayésienne repose sur le calcul de la loi a posteriori

$$p(\theta|\mathbf{x}) = \frac{L(\mathbf{x};\theta)p(\theta)}{\int L(\mathbf{x};\theta)p(\theta)d\theta} = \frac{L(\mathbf{x};\theta)p(\theta)}{L(\mathbf{x})}, \quad (4)$$

du paramètre θ sachant les données via la mise à jour de la loi a priori $p(\theta)$ par la fonction de vraisemblance $L(\mathbf{x};\theta)$.

Problem 2

La constante de normalisation $L(\mathbf{x})$, aussi appelée vraisemblance marginale, ne prend que très rarement une forme analytique.

Remark 3

L'inférence se fait généralement sur la base d'un échantillon tiré depuis la loi a posteriori en utilisant un algorithme de type Markov Chain Monte Carlo (MCMC).

1. Metropolis-Hasting

Pour $i = 1, \dots, I$

- ① Initialisation ($i = 1$) par exemple $\theta_1 \sim p(\theta)$ sinon Etape 2
- ② Perturbation $\theta^* = \theta_i + \epsilon$, avec $\epsilon \sim M\text{-Normal}(0, \Sigma)$
- ③ Acceptation/Rejet, soit $U \sim \text{Unif}([0,1])$
 - Si $\min\left(1, \frac{L(\mathbf{x}|\theta^*)p(\theta^*)}{L(\mathbf{x}|\theta_i)p(\theta_i)}\right) > U$ acceptation et $\theta_{i+1} = \theta^*$
 - Sinon rejet et $\theta_{i+1} = \theta_i$

Le résultat est un échantillon $\theta_1, \dots, \theta_I$ distribué suivant la loi a posteriori.

Remark 4

Il est nécessaire de choisir judicieusement le paramètre Σ de la perturbation pour obtenir un algorithme efficace.

Prenons des lois a priori uniformes et indépendantes

$$p(\theta) = p(q) \cdot p(\delta) = \frac{1}{b_q - a_q} \mathbb{1}_{[a_q, b_q]}(q) \times \frac{1}{b_\delta - a_\delta} \mathbb{1}_{[a_\delta, b_\delta]}(\delta)$$

2. Echantillonneur de Gibbs

Lorsque la dimension de l'espace des paramètres est plus grand que 1 (deux dans notre cas), il est possible de simuler la loi a posteriori, soit la loi jointe $q, \delta | x$ composante par composante via l'algorithme suivant:

Pour $i = 1, \dots, l$

- ① Si $i = 1$ alors $(q_1, \delta_1) \sim p(q, \delta)$, sinon étape 2
- ② Simuler $q_{i+1} \sim p(q | \delta_i, x)$
- ③ Simuler $\delta_{i+1} \sim p(\delta | q_{i+1}, x)$

Le résultat est une suite de réalisations $(q_1, \delta_1), \dots, (q_l, \delta_l)$ d'une chaîne de Markov dont la loi stationnaire est la loi a posteriori.

Remark 5

Cette méthode d'échantillonnage, dite de Gibbs, requiert la connaissance des lois conditionnelles de $q | \delta, x$ et $\delta | q, x$ qui sont potentiellement plus facile à déterminer que la loi a posteriori $p(q, \delta | x)$.

Exemple 2

Supposons que $\delta \sim \text{Inverse-Gamma}(\alpha, \beta)$ de densité

$$p(\delta) = \frac{\beta^\alpha e^{-\beta/\delta}}{\Gamma(\alpha)\delta^{\alpha+1}}, \text{ pour } \delta > 0.$$

La loi conditionnelle de $\delta|q, \mathbf{x}$ est une loi inverse gamma avec

$$\text{Inverse-Gamma}(t - t_0 + \alpha, (1 - q)\sum x_i^+ + \beta)$$

La loi conditionnelle $q|\delta, \mathbf{x}$ n'admet pas une forme explicite la simulation de $q|\delta, \mathbf{x}$ passe par un schéma de Metropolis Hasting avec

$$p(q|\delta, \mathbf{x}) = \frac{L(\mathbf{x}|q, \delta)p(q)}{\int L(\mathbf{x}|q, \delta)p(q)dq}.$$

3. Estimation Bayésienne approchée

Lorsque la vraisemblance n'admet pas de forme analytique. On s'affranchit de la fonction de vraisemblance via des simulations. L'algorithme est le suivant:

Tant que $i < I$

- 1 $\theta^* \sim p(\theta)$
- 2 $\mathbf{x}^* \sim p(\mathbf{x}|\theta^*)$
- 3
 - Si $D(\mathbf{x}, \mathbf{x}^*) < \epsilon$ alors $\theta_i = \theta^*$ et $i = i + 1$
 - Sinon Etape 1

où $D(\cdot, \cdot)$ est une mesure de dissimilarité entre les données observées \mathbf{x} et simulées \mathbf{x}^* .

Remark 6

Le résultat est un échantillon distribué suivant la loi a posteriori approchée

$$p_{abc}(\theta|\mathbf{x}) = \frac{\int_{\mathbb{R}^t} L(\mathbf{x}^*|\theta) \mathbb{1}_{D(\mathbf{x}, \mathbf{x}^*) < \epsilon} d\mathbf{x}^* p(\theta)}{\int_{\Theta} \int_{\mathbb{R}^t} L(\mathbf{x}^*|\theta) \mathbb{1}_{D(\mathbf{x}, \mathbf{x}^*) < \epsilon} d\mathbf{x}^* p(\theta) d\theta}$$

La précision est liée au choix de ϵ , il s'agit d'un compromis précision/temps de calcul.

Le paramètre important est la distance D permettant la comparaison des données observées et simulées.

Remark 7

Le choix d'une distance euclidienne

$$D(x, x^*) = \sqrt{\sum_{s=1}^t (x_s - x_s^*)^2}$$

conduit à ne jamais sélectionner de paramètre eu égard à une trop grande variance (pour de grands échantillons).

Une solution consiste à définir des résumés statistiques

$$S : x \mapsto S(x) \in \mathbb{R}^d, \text{ avec } d \leq t$$

puis de mesurer la dissimilarité entre données observées et données simulées par $D[S(x), S(x^*)]$.

Remark 8

L'utilisation de statistiques rajoute une approximation puisque la loi a posteriori approchée converge vers la loi des paramètres sachant $S(x)$ ce qui correspond à une perte d'information.

L'utilisation de statistique exhaustive permettent d'assurer la convergence vers la vraie loi a posteriori.

Definition 1

Pour rappel, une statistique S est exhaustive la vraisemblance se décompose en deux fonctions h et g telles que

$$L(x|\theta) = h(x)g(\theta, S(x)).$$

Exemple 3

Le modèle étudié admet les statistiques exhaustives suivantes

$$t_0 \text{ et } \sum_{i=1}^t x_i.$$

Références bibliographiques I



Boris Buchmann and Rudolf Grübel.

Decompounding: an estimation problem for Poisson random sums.

Ann. Statist., 31(4):1054–1074, 08 2003.



Peter K. Dunn.

Occurrence and quantity of precipitation can be modelled simultaneously.

International Journal of Climatology, 24(10):1231–1239, jul 2004.



Pierre-Olivier Goffard and Patrick J. Laub.

Approximate bayesian computations to fit and compare insurance loss models.

Insurance: Mathematics and Economics, 100:350–371, sep 2021.