Hétérogénéité inobservable, volumétrie limitée et mutualisation

Xavier Milhaud

Test manuscript No.
(will be inserted by the editor)

Hétérogénéité inobservable, volumétrie limitée et mutualisation

Inspiré de l'article Milhaud et al. (2020).

Xavier Milhaud*

Received: 03-12-2020 / Accepted: date

Abstract Nous abordons dans cet article la problématique de la fusion de divers portefeuilles, a priori hétérogènes, pour un risque donné. Dans notre cas, les caractéristiques des assurés et de leur contrat demeurent inconnues, rendant la prise en compte et la modélisation de l'hétérogénéité délicate. A travers l'usage de modèles mélanges, nous construisons un test statistique permettant de répondre à cette problématique. Une application à des portefeuilles de mortalité est conduite afin d'illustrer l'intérêt de la méthode, particulièrement dans le cas où les expositions au risque par portefeuille demeurent limitées et où la mutualisation de ces portefeuilles devient une condition nécessaire à une bonne gestion des risques.

1 Contexte et problématique

En assurance IARD, la gestion des risques est un enjeu crucial et se manifeste notamment par le développement de modèles sophistiqués en tarification et en réassurance, domaines dans lesquels la bataille des modèles fait rage et le principe de mutualisation fait foi. En effet, une population assurée étant par essence hétérogène, une des questions principales est de savoir sélectionner les informations pertinentes parmi l'ensemble des informations recueillies sur les assurés. Le but est d'aboutir à des modélisations plus individualisées, adaptées aux profils des assurés, tout en fournissant un résultat segmenté robuste qui permette à la fois une gestion des risques maitrisée et une bonne compétitivité. La robustesse du tarif provient essentiellement de la volumétrie des sous-populations créées, induites par les critères de modélisation conservés par l'assureur (exemple: les critères tarifaires). Pour ce type de tâches, la boîte à outil d'un actuaire comporte essentiellement des techniques basées sur la

X. Milhaud* (correspondance)
ISFA, Université Lyon 1, LSAF EA2429
E-mail: xavier.milhaud@univ-lyon1.fr

prise en compte d'informations individuelles observables tels que l'âge, ou la catégorie socio-professionnelle. Les deux grandes familles de modèles utilisés sont les modèles linéaires généralisés (GLM) et les modèles d'apprentissage statistique, aussi connus sous le nom de Machine Learning (ML) par le grand public. Mais quid de la modélisation d'une population hétérogène lorsque ces informations individuelles sont indisponibles, et donc les sources d'hétérogénéité "inobservables"? Et que dire de la mutualisation si l'exposition au risque reste faible sur un portefeuille donné?

2 Modélisation et test statistique

Pour la prise en compte de l'hétérogénéité sans disposer des caractéristiques individuelles des assurés, nous considérons dans la suite de cet article des modèles mélanges à deux composantes. Pour rappel, ces modèles s'écrivent sous la forme

$$h(x) = (1-p)g(x) + pf(x), \quad x \in \mathbb{R},$$

où h est la densité de probabilité du mélange (qui correspond à la loi sous-jacente à nos observations), f et g sont les densités composantes du mélange, et $p \in]0,1[$ détermine le poids des composantes dans le mélange. Dans notre contexte, la densité g est considérée connue, alors que les autres paramètres p et f sont inconnus. Ce modèle a donné lieu à plusieurs publications, dont Bordes and Vandekerkhove (2010) ou Celisse and Robin (2010).

Pour notre application, la fonction h fait référence à la distribution des âges x au décès (mortalité) dans un portefeuille d'assurés, et la composante g représente le profil national (connu) de distribution des âges au décès. Ainsi, la spécificité réelle de la population h à l'étude provient de la composante f inconnue, et de l'importance de cette composante dans la population via le poids p. Concrètement, un assureur commercialisant des garanties décès sur des segments de profession très différents (cadre supérieur et ouvrier du BTP par exemple) est confronté à deux profils de mortalité variables, ayant pourtant une composante commune g liée à la mortalité nationale.

Imaginons maintenant que nous disposons de K portefeuilles avec chacun leur profil de mortalité h_k $(k = 1, ..., K)$. Parmi ces portefeuilles, certains ont une petite volumétrie. De ce fait, la gestion des risques liée à une garantie y est peu fiable, et la question de la mutualisation avec d'autres portefeuilles s'avère dès lors pertinente. Pour simplifier, prenons $K = 2$ dans la suite de l'exposé. La mutualisation des portefeuilles 1 et 2, de densités respectives h_1 et h_2, n'a de sens que si les populations observées ont un même profil de mortalité, c'est à dire si f_1 ressemble à f_2 (puisque $g_1 = g_2 = g$ ici). Rappelons que même si $f_1 = f_2$, h_1 et h_2 n'ont pas le même aspect car p_1 est différent de p_2 en toute généralité. D'un point de vue purement statistique, la possibilité de mutualiser ces portefeuilles revient à tester l'égalité des composantes inconnues f_1 et f_2.

Pour ce faire, nous proposons un test de type χ^2 pénalisé, capable de comparer par paire (jusqu'à un certain rang) les coefficients obtenus lors de la décomposition des densités f_1 et f_2 dans une base orthonormale de

polynômes. Les données à disposition sont deux échantillons d'observations *i.i.d.* $X = (X_1, \ldots, X_{n_1})$ et $Y = (Y_1, \ldots, Y_{n_2})$, de densités respectives

$$\begin{cases} h_1(x) = (1-p_1)g_1(x) + p_1 f_1(x), & x \in \mathbb{R}, \\ h_2(x) = (1-p_2)g_2(x) + p_2 f_2(x), & x \in \mathbb{R}, \end{cases} \tag{1}$$

par rapport à une mesure de référence ν.
Etant donné ce modèle, notre but est de répondre au test statistique suivant:

$$H_0 : \quad f_1 \text{ est égale à } f_2 \qquad \textit{versus} \qquad H_1 : \quad f_1 \text{ est différente de } f_2 \ , \tag{2}$$

sans assigner les f_i $(i = 1, 2)$ à une famille paramétrique donnée.

3 Statistique de test et résultats théoriques

3.1 Définition de la statistique de test

La procédure de test est basée sur la comparaison des coefficients d'ordre k, notés $h_{i,k}$ $(i = 1, 2)$, obtenus lors de la décomposition des densités h_i dans une base orthonormale de polynômes adaptée (Ledwina (1994)). Ces coefficients sont intimement liés aux moments des distributions étudiées. Le but est de détecter (asymptotiquement, *i.e.* lorsque n_1, n_2 tendent vers de grandes valeurs) les écarts entre $h_{1,k}$ et $h_{2,k}$, pour chaque ordre k.

En utilisant cette décomposition et la modélisation (1), nous déduisons

$$h_{i,k} = (1-p_i)g_{i,k} + p_i f_{i,k}.$$

L'hypothèse H_0 peut donc être écrite autrement. En l'occurence $f_{1,k} = f_{2,k}$, ou encore

$$H_0 : \ p_2(h_{1,k} - (1-p_1)g_{1,k}) = p_1(h_{2,k} - (1-p_2)g_{2,k}), \qquad k \geq 1.$$

Puisque les densités g_1 et g_2 sont connues, les coefficients $g_{i,k}$, $i = 1, 2$, sont automatiquement connus. L'estimation des proportions p_i se fait grâce à l'estimateur présenté dans Bordes and Vandekerkhove (2010). Les coefficients $h_{i,k}$ correspondent à la moyenne empirique des coefficients obtenus pour chaque observation lors de la décomposition de h_i dans la base orthonormale.

Pour répondre à la question (2), nous considérons les différences

$$\widehat{R}_k := \hat{p}_2(\widehat{h}_{1,k} - (1-\hat{p}_1)g_{1,k}) - \hat{p}_1(\widehat{h}_{2,k} - (1-\hat{p}_2)g_{2,k}), \quad k \geq 1,$$

qui permettent de détecter les écarts à l'hypothèse nulle, où la notation $\hat{z}$ désigne l'estimateur de z. Pour tout $k \geq 1$, on définit $\widehat{U}_k = (\widehat{R}_1, \ldots, \widehat{R}_k)$, et

$$\widehat{T}_k = \frac{n_1 n_2}{n_1 + n_2} \widehat{U}_k^\top \widehat{D}_k^{-1} \widehat{U}_k, \tag{3}$$

où $\widehat{D}_k = \mathrm{diag}(\widehat{d}_1, \ldots, \widehat{d}_k)$ est une matrice diagonale d'estimateurs convergents des variances des $\widehat{R}_k$.

Pour sélectionner l'ordre k du développement dans la base orthonormale de polynômes, nécessaire à la formulation d'une réponse pour tester H_0, nous appliquons une procédure *data-driven* (voir Ledwina (1994) et Kallenberg and Ledwina (1995)). Plus précisément, nous utilisons la règle de pénalisation:

$$S(n_1,n_2) = \min\left\{ \underset{1\leq k\leq d(n_1,n_2)}{\arg\max} \left(s(n_1,n_2)\widehat{T}_k - \beta_k \text{pen}(n_1,n_2)\right)\right\}, \qquad (4)$$

où $d(n_1,n_2)$ *et* $pen(n_1,n_2) \to +\infty$ quand $n_1,n_2 \to +\infty$, les β_k sont des facteurs de pénalisation, et $s(n_1,n_2)$ un terme de normalisation qui dépend de la vitesse de convergence des estimateurs de p_1 et p_2. Au final nous calculons la statistique de test au rang sélectionné, soit $T(n_1,n_2) = \widehat{T}_{S(n_1,n_2)}$.

3.2 Résultats fondamentaux

Nous évoquons ci-dessous le comportement asymptotique sous l'hypothèse nulle H_0 du rang sélectionné $S(n_1,n_2)$ défini en (4), ainsi que de la statistique de test (3). Les démonstrations, disponibles dans Milhaud et al. (2020), sont inspirées de l'article en *one-sample* réalisé par Pommeret et al. (2019).

Théorème 1: Sous H_0 et sous certaines hypothèses peu restrictives, le rang $S(n_1,n_2)$ sélectionné converge en probabilité vers 1 quand $n_1,n_2 \to +\infty$.

Corollaire: Sous H_0 et sous certaines hypothèses peu restrictives, $T(n_1,n_2)$ converge en loi vers une loi du χ^2 à un degré de liberté lorsque $n_1,n_2 \to +\infty$.

On considère maintenant la collection d'hypothèses alternatives de type H_1, définies comme suit: il existe $q \in \mathbb{N}^*$ tel que

$$H_1(q): \quad f_{1,j} = f_{2,j},\ j = 1,\dots,q-1, \quad \text{et} \quad f_{1,q} \neq f_{2,q},$$

permettant de détecter à partir de quel ordre q l'écart entre f_1 et f_2 apparait. En utilisant le même raisonnement, on peut énoncer le résultat suivant qui donne la déviation asympotique de la statistique de test sous $H_1(q)$.

Théorème 2: Sous H_1 et sous certaines hypothèses peu restrictives, le rang sélectionné $S(n_1,n_2) \to s \geq q$, et $T(n_1,n_2) \to +\infty$ quand $n_1,n_2 \to +\infty$.

Concrètement, en fixant un niveau de confiance du test à 5%, ces résultats permettent de tester H_0 en comparant la statistique obtenue avec le quantile à 95% de la loi du χ^2 à un degré de liberté. Si la statistique excède ce quantile, on rejette H_0. Le cas contraire, on ne rejette pas H_0 et on conclut à l'homogénéité.

4 Application à la mortalité

Pour comprendre l'hétérogénéité de populations assurées, nous appliquons notre procédure de test sur des données d'âge au décès. Ces jeux de données proviennent d'études conduites par l'*Institut des Actuaires* et couvrent la

Table 1 Caractéristiques de la distribution de l'âge au décès, du poids p estimé de la composante inconnue f_i dans le mélange, et statistiques de test (triangle supérieur) avec p-valeur (triangle inférieur) du test sur les trois portefeuilles.

	Taille (n)	Espérance de vie	Poids $\hat{p}$	P1	P2	P3
P1	1 251	75.42	0.4603	—	23.28	0.717
P2	7 356	74.91	0.7003	1.4e-06	—	18.48
P3	3 456	75.56	0.6281	0.397	1.7e-05	—

période 2007-2011. Nous considérons trois populations de femmes ayant souscrit des garanties décès. Les caractéristiques des portefeuilles sont donnés dans la partie gauche du Tableau 1 et les densités de probabilité de l'âge au décès de chaque population apparaissent Figure 1. Dans le Tableau 1, nous affichons aussi l'espérance de vie estimée par population. Elle reste stable dans les trois populations, ce qui pourrait suggérer des profils de mortalité comparables. En observant mieux les densités mélange en Figure 1, nous réalisons que la comparaison des composantes n'est pas triviale.

Avant de procéder au test, il est indispensable de définir les densités connues, autrement dit g_1 et g_2 dans (1). N'importe quelle sous-population admet un profil de mortalité dont une partie correspond à celui de la population nationale. Ainsi, nous fixons $g_1 = g_2$, calibrées sur les observations de mortalité de la population nationale française[1]. Nous modélisons cette mortalité par la loi de Gompertz (i.e. $g(x) = b\exp(ax)\exp(-b/a(\exp(ax)-1))$), et calibrons les paramètres a et b sur la période 2007-2011. On obtient $a = 0.125$ et $b = 2.182 \times 10^{-6}$. A titre d'illustration, la Figure 1 décrit la densité de Gompertz estimée pour notre population nationale (en haut à gauche). En

[1] Données téléchargées depuis http://www.mortality.org en Septembre 2019.

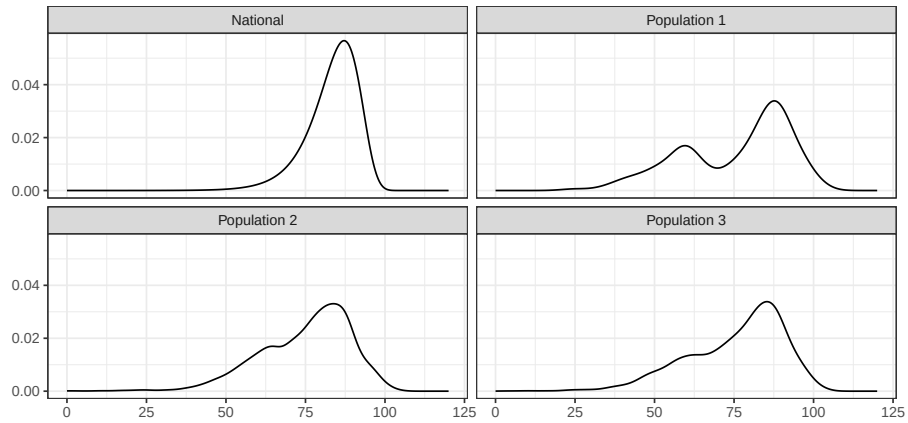

Fig. 1 Densités de probabilité de l'âge au décès pour la population féminine nationale française, puis pour chacun des trois portefeuilles étudiés.

utilisant ce calibrage, on peut calculer les poids estimés $\hat{p}_i$ du modèle (2), répertoriés eux aussi Tableau 1. Enfin, le Tableau 1 résume les résultats des tests d'hypothèse des portefeuilles pris deux-à-deux. Il y figure les valeurs de la statistique de test (en haut à droite), et les p-valeurs du test d'hypothèse (rappelons qu'une *p-valeur inférieure à 5%* conduit à rejeter H_0).

On constate que les populations P1 et P3 sont considérées homogènes en termes de profil de mortalité, tandis que les autres couples (P1,P2) et (P2,P3) se distinguent par des composantes f_i différentes. Pourtant, un premier coup d'oeil à la Figure 1 aurait sans doute conduit à ne fusionner aucune de ces trois populations, ce qui montre la finesse du modèle. Pour expliquer ces résultats, un facteur essentiel intervient ici: le niveau de souscription, qui dépend généralement de la somme assurée. Typiquement, la souscription aux garanties décès des portefeuilles P1 et P3 nécessite le passage d'un examen médical, car les sommes assurées y sont importantes. Or il est de notoriété publique que le niveau de richesse a un impact important sur la mortalité, ce qui expliquerait les similitudes entre P1 et P3. Ici, les portefeuilles P1 et P3 proviennent de deux assureurs différents, ayant des stratégies de souscription différentes. Pourtant, il apparait clair que ces populations pourraient être fusionnées pour améliorer des tâches impactant directement la gestion des risques, telles que la tarification ou la réassurance.

Remerciements Ces travaux de recherche ont été réalisés dans le cadre de la Chaire DIALog (Digital Insurance and Long-term risks), placée sous l'égide de la fondation du risque en partenariat avec CNP Assurances, et l'ISFA, Université Claude Bernard Lyon 1 (UCBL).

References

Bordes L, Vandekerkhove P (2010) Semiparametric two-component mixture model with a known component: an asymptotically normal estimator. Mathematical Methods of Statistics 19(1):22–41

Celisse A, Robin S (2010) A cross-validation based estimation of the proportion of true null hypotheses. Journal of Statistical Planning and Inference 140(11):3132–3147

Kallenberg WC, Ledwina T (1995) Consistency and monte carlo simulation of a data driven version of smooth goodness-of-fit tests. The Annals of Statistics pp 1594–1608

Ledwina T (1994) Data-driven version of neyman's smooth test of fit. Journal of the American Statistical Association 89(427):1000–1005

Milhaud X, Pommeret D, Salhi Y, Vandekerkhove P (2020) Semiparametric two-sample mixture components comparison test, URL https://hal.archives-ouvertes.fr/hal-02491127, preprint

Pommeret D, Vandekerkhove P, et al. (2019) Semiparametric density testing in the contamination model. Electronic Journal of Statistics 13(2):4743–4793