Clustering par quantification en présence de censure

Svetlana Gribkova

[1] *Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie Paris 6, 4 place Jussieu, 75005 Paris*

Résumé. Les méthodes de clustering ont pour but de déterminer au sein d'une population hétérogène plusieurs groupes d'objets homogènes avec une séparation maximale entre les groupes. Cette problématique apparaît dans de nombreux domaines comme la biologie, la médecine ou encore l'économie et la finance. Dans cet exposé, on s'intéressera plus particulièrement au problème de clustering pour les données de survie où l'éventuelle présence de censure rend les méthodes existantes inapplicables. Le but de ce travail est de proposer une nouvelle méthode de clustering pour une configuration où les observations sont composées d'une variable de durée (éventuellement censurée) et d'un vecteur de covariables qui est observé pour tous les individus. Cette méthode utilise la notion de quantification et généralise l'algorithme classique des "$k-$means". On étudiera théoriquement la consistance de la méthode et on présentera une inégalité exponentielle qui permet de contrôler la différence entre la distorsion minimale et celle d'un quantificateur empirique optimal. En conclusion on présentera des applications numériques aux données simulées et réelles.

Mots-clés. Clustering, quantification, censure, k-means.

Abstract. Clustering methods in statistics permit do distinguish in heterogeneous population several separated homogeneous groups. This type of problem appears in numerous areas such as biology, medicine, economy or finance. In this paper, we are interested by clustering algorithms for survival data where an eventual presence of censoring makes impossible the using of existing methods. We aim to propose a new method of clustering for the configuration where observations are composed of a lifetime variable, subjected to censoring and a vector of covariates which is observed for all subjects. This method is based on the notion of quantization and generalizes the classical "$k-$means" algorithm. The consistency of the empirical design will be studied in terms of exponential inequality which permits to control the difference between the distortion of the empirically optimal quantizer and the minimal distortion. In conclusion we will present some applications to simulated and real data.

Keywords. Clustering, quantization, censoring, k-means.

1 Introduction

En analyse de survie ou en fiabilité l'objet principal de l'étude est la durée T écoulée entre le début de la période d'observation et l'occurrence d'un certain événement d'intérêt. Un exemple important dans la médecine est une étude de survie de patients atteints d'une certaine maladie où l'événement d'intérêt est le temps de décès. En assurance on s'intéresse souvent au temps écoulé jusqu'à ce qu'un accident survienne. Généralement, en plus de la durée T on observe un vecteur $X \in \mathbb{R}^d$ de variables explicatives. En médecine par exemple ce vecteur peut être composé des caractéristiques individuelles des patients comme l'âge, le sexe, les résultats de prélèvements sanguins, etc. En assurance ce sont les informations diverses sur les souscripteurs de contrat. Habituellement, le vecteur de covariables est observé pour tous les individus de la population, alors que la variable de durée peut ne pas toujours l'être. Cela est dû au phénomène de censure qui est souvent présent dans les données de survie. Dans ce contexte, on n'observe pas la variable d'intérêt T, mais le minimum entre cette variable et une autre variable aléatoire "parasite" C, appelée la censure. Dans le cas d'une étude médicale, C peut représenter la date de la fin d'étude ou celle à laquelle le patient a été perdu de vue. Pour résumer, en présence de censure, au lieu d'observer un vecteur aléatoire $(T, X) \in \mathbb{R}^{d+1}$, on observe $(Y, X, \delta) = (\min(T, C), X, \mathbb{1}_{T \leq C})$.

La population étudiée en analyse de survie est souvent hétérogène et plusieurs clusters peuvent être naturellement présents dans les données. Parmi les patients atteints d'une maladie, il peut y avoir plusieurs sous groupes de survie et des caractéristiques histologiques similaires. En assurance vie la population est forcement hétérogène, lorsque les personnes issues de plusieurs générations sont présentes. La détection des groupes homogènes apporte une information sur la structure de données et permet de modéliser séparément les données liées à chaque sous groupe, ce qui peut améliorer significativement la qualité de prévisions.

Le problème de détection de groupes d'objets similaires au sein d'une population hétérogène est bien connu en statistique. Dans le cadre classique, où l'on dispose des observations indépendantes identiquement distribuées, il existe plusieurs méthodes de partitionnement. En présence d'observations censurées, une application directe de ces méthodes fournit des résultats incorrects, car elle amène à constituer des groupes basés sur les valeurs observées de la variable Y et non pas sur celles de la durée T qui nous intéresse. Le but de ce travail est de proposer une nouvelle méthode de clustering par quantification permettant de traiter les données de survie multivariées dont une composante est une variable de durée éventuellement censurée et les autres composantes sont des variables explicatives observées pour tous les sujets.

2 Quantification en présence de censure

On considère un vecteur aléatoire (T, X) à valeurs dans R^{d+1}, de la loi P où $T \in \mathbb{R}$ est une variable de durée et $X \in \mathbb{R}^d$ est un vecteur de covariables. Au lieu d'observer (T, X) on observe les réalisations i.i.d. du triplet $(Y, X, \delta) = (\min(T, C), X, \mathbb{1}_{T \leq C})$ où C est une variable de censure.

Un N-quantificateur dans $\mathbb{R}^{d+1}$ est une application de $\mathbb{R}^{d+1}$ dans un dictionnaire $\mathbf{c}$, où $\mathbf{c} = (c_1, \ldots, c_N) \in \left(\mathbb{R}^{d+1}\right)^N$. On notera par Q_N l'ensemble des $N-$quantificateurs. Une erreur commise par un $N-$quantificateur $q \in Q_N$ (sa distorsion) est définie par

$$D(P, q) = \mathbb{E}_P ||(T, X) - q(T, X)||^2. \tag{1}$$

La distorsion minimale pour la classe de $N-$quantificateurs est donnée par

$$D(P, q^*) := D_N^*(P) = \inf_{q \in Q_N} D(P, q).$$

Un quantificateur q^* est dit optimal si $D(P, q^*) = D_N^*(P)$. On montre qu'un tel quantificateur existe et c'est un quantificateur des plus proches voisins dont la distorsion est donnée par

$$D(P, q^*) = \inf_{\mathbf{c} \in (\mathbb{R}^{d+1})^N} \mathbb{E} \min_{y_i \in \mathbf{c}} ||(T, X) - y_i||^2. \tag{2}$$

On cherchera donc par la suite le meilleur quantificateur parmi ceux des plus proches voisins. En réalité dans la plupart des cas la loi P du vecteur aléatoire (T, X) n'est pas connue et on ne peut pas obtenir un quantificateur optimal q^* par la minimisation du critère (2). Ce critère doit alors être remplacé par sa version empirique. Si on disposait des observations i.i.d. $(T_i, X_i)_{1 \leq i \leq n}$ du vecteur (T, X), on aurait remplacé la loi inconnue P par la loi empirique engendrée par les observations. Dans notre cas, les observations i.i.d. de (T, X) ne sont pas disponibles. Pour pallier cette difficulté, on remplace la fonction de répartition inconnue de (T, X) par son estimateur adapté à la présence de censure, proposé par Stute (1993). Cet estimateur est donné par

$$\hat{F}_n(t, x) = \sum_{i=1}^{n} W_{in} \mathbb{1}_{Y_i \leq t, X_i \leq x}, \quad t \in \mathbb{R},\ x \in \mathbb{R}^d, \tag{3}$$

où

$$W_{in} = \frac{\delta_i}{n(1 - \hat{G}(Y_i-))},$$

et $\hat{G}(y)$ est un estimateur de Kaplan-Meier de la fonction de distribution $G(y)$ de la variable de censure. Cet estimateur engendre une mesure $P_n = \sum_{i=1}^n W_{in} \delta_{(Y_i, X_i)}$ sur $\mathbb{R}^{d+1}$ ce qui nous amène à une définition suivante de distorsion empirique:

$$D(P_n, q) = \frac{1}{n} \sum_{i=1}^{n} W_{in} ||(Y_i, X_i) - q(Y_i, X_i)||^2. \tag{4}$$

On appellera $q_n^* \in Q_N$ un quantificateur empirique optimal s'il minimise la distorsion empirique, c'est-à-dire $D(P_n, q_n^*) = \inf_{q \in Q_N} D(P_n, q)$. Le résultat suivant montre la consistance du critère empirique.

Théorème 1. *Pour tout $N \geq 1$, la suite $(q_n^*)_{n=1,2,\ldots}$ de quantificateurs empiriques optimales $n = 1, 2, \ldots$, vérifie*

$$\lim_{n \to \infty} D(P_n, q_n^*) \underset{p.s.}{=} D_N^*(P_n).$$

3 Algorithm de clustering par quantification

Dans cette partie on propose un algorithme de clustering par quantification qui est une généralisation de méthode des "$k-$means"au cas des données censureées. On suppose que l'on a disposition des observations i.i.d. $(Y_i, X_i, \delta_i)_{1 \leq i \leq n}$ de $(Y, X, \delta) = (\min(T, C), X, \delta)$. On cherche à déterminer k groupes homogènes du point de vue de la variable de durée T qu'on n'observe pas et de la covariable X. L'algorithme se dcompose en deux étapes. En première étape on détermine les centres des k clusters, de faon suivante,

Étape 1:

- Initialiser les coordonnées de k centres $c_1^{(0)}, \ldots, c_k^{(0)}$.
- Calculer les valeurs des poids W_{in} de l'estimateur de Kaplan-Meier à partir de l'échantillon $(Y_i, X_i, \delta_i)_{1 \leq i \leq n}$
- Pour une itération ℓ avec $\ell = 0, 1, 2, \ldots$
  - Calculer les cellules de Voronoi $S_1^l, \ldots, S_k^l$ qui correspondent au centres $c_1^{(l)}, \ldots, c_k^{(\ell)}$ pour l'ensemble des observations non censurées $\{(Y_i, X_i) : \delta_i = 1\}$
  - Pour $j = 1, \ldots, k$ calculer les nouveaux centres $(c_j^{(\ell+1)})_{1 \leq j \leq k}$ comme

$$c_j^{(\ell+1)} = \frac{\sum_{i=1}^n (Y_i, X_i)^T W_{in} \mathbb{1}_{\{(Y_i, X_i) \in S_j, \delta_i = 1\}}}{\sum_{i=1}^n W_{in} \mathbb{1}_{\{(Y_i, X_i) \in S_j, \delta_i = 1\}}}.$$

    $c_j^{(\ell+1)}$ représente un estimateur de l'espérance conditionnelle $\mathbb{E}[(T, X) | (T, X) \in S_j^l]$ et dans le cadre classique est évaluée comme la moyenne empirique des observations qui se trouvent dans la cellule. Dans notre cas, ce calcul est remplacé par la moyenne prise uniquement sur les observations non censurées qui appartiennent à la cellule, chaque observation étant pondérée par son poids W_{in}.
  - L'algorithme s'arrête à une certaine itération ℓ^*. Pour $j = 1, \ldots, k$ on attribue à chaque observation (Y_i, X_i) avec $\delta_i = 1$ le label j si $(Y_i, X_i) \in S_j^{l^*}$.

A la fin de la première étape les centres de clusters sont déterminés et les observations non censurées ont obtenu leur labels. Le but de la deuxième étape consiste à attribuer les labels aux observations censurées, pour lesquelles les distances euclidiennes aux centres de clusters ne sont pas observées.

Étape 2 [Attribution de labels aux observations censurées]. Les distances entre les observations non censurées $\{(Y_i, X_i) : \delta_i = 0\}$ et les centres $c_1^{(\ell*)}, \ldots, c_k^{(\ell*)}$ ne sont pas observées, par conséquence on ne peut pas déterminer pour eux le centre le plus proche. Pour résoudre ce problème on estime la distance de chaque observation censurée (Y_i, X_i) avec $\delta_i = 0$ au centre $c_j^{(l*)}$ par

$$\hat{d}_{ij} = \frac{\int_{Y_i}^{\infty} ||(t, X_i) - c_j^{(l*)}||^2 d\hat{F}(t|X_i)}{\int_{Y_i}^{\infty} d\hat{F}(t|X_i)}, \tag{5}$$

où $\hat{F}(t|x)$ est un estimateur de $F(t|X = x) = P(T \leq t|X = x)$, défini dans Beran (1981) et donné par

$$\hat{F}(t|x) = \sum_{i=1}^{n} W_{in} \frac{K(\frac{x-X_i}{h})}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)} \mathbb{1}_{Y_i \leq t}, \tag{6}$$

avec un noyau $K(x)$, c'est-à-dire une fonction non négative intégrable avec $\int_{-\infty}^{\infty} K(x)dx = 1$. Avec (5) et (6) on obtient,

$$\hat{d}_{ij} = \frac{\sum_{m=1}^{n} W_{mn} ||(Y_m, X_i) - c^{(l*)}||^2 K\left(\frac{X_i - X_m}{h}\right) \mathbb{1}_{Y_m \geq Y_i}}{\sum_{m=1}^{n} W_{mn} K(\frac{X_i - X_m}{h}) \mathbb{1}_{Y_m \geq Y_i}}. \tag{7}$$

Pour attribuer les labels aux observations censureées,

- Pour chaque observation (Y_i, X_i) avec $\delta_i = 0$ calculer l'estimateur de distance $\hat{d}_{ij}$ en utilisant (7).
- Attribuer à (Y_i, X_i) le label $j^* = \arg\min_j \hat{d}_{ij}$.

Bibliographie

[1] Beran, R. (1981), Nonparametric regression with randomly censored survival data, *Technical report*, Univ. California, Berkeley.
[2] Stute, W. (1993), Consistent estimation under random censorship when covariables are present, *Journal of Multivariate Analysis*.