Vector quantization and clustering in presence of censoring

Svetlana Gribkova

Vector quantization and clustering in presence of censoring

Svetlana Gribkova[1]

October 19, 2014

Abstract

We consider the problem of optimal vector quantization for random vectors with one censored component and applications to clustering of censored observations. We introduce the definitions of the empirical distortion and of the empirically optimal quantizer in presence of censoring and we establish the almost sure consistency of empirical design. Moreover, we provide a non asymptotic exponential bound for the difference between the performance of the empirically optimal k-quantizer and the optimal performance over the class of all k-quantizers. As a natural application of the new quantization criterion, we propose an iterative two-step algorithm allowing for clustering of multivariate observations with one censored component. This method is investigated numerically through applications to real and simulated data.

Key words: Clustering, quantization, random censoring, k-means, Kaplan-Meier estimator.

Short title: Quantization and clustering under censoring.

[1] Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie Paris VI, 4 place Jussieu, 75005 Paris, France, E-mail: svetlana.gribkova@etu.upmc.fr

1 Introduction

Vector quantization and k-clustering are the two closely related issues. The former corresponds to a probabilistic problem of finding the optimal way to represent a continuous distribution of random vector by a discrete distribution with a k-point support. Some general references on the subject are [Gersho and Gray, 1992], [Graf and Luschgy, 1994] and [Linder, 2002]. The latter is a statistical problem of partitioning a set of i.i.d. observations of a random vector into k groups as homogeneous and as compact as possible (see, for example, [Lloyd, 1982] or [MacQueen, 1967]). The existing methodology in both settings supposes the availability of an i.i.d. complete data sample of the random vector of interest. That is commonly not the case in the context of survival analysis where observations include a lifetime variable, which may not be directly observed for the reason of censoring. For a complete introduction to survival analysis, we refer the reader to [Fleming and Harrington, 1991]. Due to this specificity, the multivariate survival data cannot be analyzed by means of the standard clustering methods.

In the present article, we first introduce a new optimal quantization procedure for random vectors with one censored component. Then, we consider its important practical application, that is a new k-clustering algorithm valid in presence of censored observations. To facilitate the discussion, we now fixe some notation. In the sequel, we will be concerned with a random vector of the form (T, X), where T is a univariate random variable subjected to right random censoring and X is a d-dimensional observed vector of quantitative covariates. In presence of censoring, instead of observing T directly, one observes a couple

$$(Y, \delta) = (\min(T, C), \mathbb{1}_{T \leq C}),$$

where C is a censoring random variable. Therefore, the available observations are composed of i.i.d. replications

$$(Y_i, \delta_i, X_i)_{1 \leq i \leq n} \tag{1}$$

of the random vector (Y, δ, X). Our first aim is to define the optimal quantization procedure for (T, X) given the incomplete sample (1). Next, we will propose a clustering algorithm detecting groups among n subjects with respect to their characteristics $(T_i, X_i)_{1 \leq i \leq n}$, having at the input only their censored versions $(Y_i, \delta_i, X_i)_{1 \leq i \leq n}$. We outline that we focus on a non supervised learning task, hence the variable T is not considered as the response.

Before explaining the difficulties of the quantization and clustering in the described setting, let us discuss some of their relevant applications. In the medical domain, finding subtypes of a disease is an important task for the personalization of the treatment. To illustrate the point, let us consider n patients suffering from the same type of cancer. This same type of the disease is often represented by several unknown subtypes which differ through biological features and clinical characteristics. Identifying such subtypes permits to clinicians to make their diagnostics more precise. For each patient, the available information include the survival time T (may be censored) and the observed vector X of biological and/or clinical characteristics. From the point of view of the statistical methodology, clustering methods are the

best adapted to the situation of unknown cancer subtypes. In this context, some of the existing approaches are discussed in [Bair and Tibshirani, 2004]. In particular, it is mentioned that an approach by clustering only with respect to X may lead to groups differing through the biological features but unrelated to patient survival, which are not of prime interest for clinicians. Therefore, it is desirable to perform clustering taking into account the censored survival time. To that aim, in the context of genetic data, [Bair and Tibshirani, 2004] propose to select among the components of X only the variables correlated with T (having a large Cox score) and to apply then a non supervised clustering method with respect to the selected covariates. The approach that we propose in this paper may be an alternative way to detect groups related to the survival time without excluding covariates. The idea is that our algorithm performs clustering with respect to the whole vector (T, X) using as the input the available set of incomplete observations. As the variable T participates in the procedure directly, the constitution of groups takes naturally into account the survival time.

We note that the field of applications of our method is not confined to the medical domain. For instance, in life insurance the population of policyholders is commonly heterogeneous. When it comes to optimize the mortality management, a relevant task consists in segmenting risks into classes that are homogenous. Insurers are then able to price taking into account the specific mortality risk of each class. The standard information available for performing such a partition composes of the residual lifetimes of policyholders and their associated geographical, socio-professional, etc. characteristics. The lifetimes of subjects are subjected to censoring (for example, in case of the cancellation of their insurance contract) and the issue of finding homogeneous risk classes brings us back to the initial mathematical problem.

The rest of the article is organized as follows. Next section summarizes some basic results from the vector quantization theory. In Section 3, we propose a generalized definition of the empirical distortion adapted to the presence of censoring and we define the empirically optimal quantizer as its minimizer. Section 4 deals with the asymptotic results for the distortion of the empirically optimal quantizer. The new clustering algorithm is presented in Section 5. Section 6 proceeds with applications on simulated an real life data sets.

2 Vector quantization

We now come back to the quantization problem and we start by giving some preliminary definitions. As we have already mentioned, the k-quantization consists in summarizing the distribution P of the random vector (T, X) of $\mathbb{R}^{d+1}$ by a discrete distribution with a k-point support. The classical way to do that consists in replacing (T, X) with $q(T, X)$, where q is a k-point quantizer, that is a mapping $q : \mathbb{R}^{d+1} \to \mathcal{C}$, where

$$\mathcal{C} = \{(c_1, \ldots, c_k), c_i \in \mathbb{R}^{d+1}, \text{ for } i = 1, \ldots, k\}$$

is a k-point subset of $\mathbb{R}^{d+1}$ called a codebook.

Let Q_k be the set of all k-point quantizers. The error (distortion) of an arbitrary k-point quantizer representing (T, X) by $q(T, X)$ may be defined by

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

The optimal performance over the class of k-point quantizers is given by

$$D_k^*(P) = \inf_{q \in Q_k} D(P, q).$$

A quantizer q^* is called optimal if $D(P, q^*) = D_k^*(P)$. It was shown (see [Linder, 2002]), that such quantizer exists. Moreover, q^* is a nearest neighbor quantizer, that is,

$$q^*(t, x) = \arg\min_{c_i \in \mathcal{C}} ||(t, x) - c_i||^2,$$

with the distortion

$$D(P, q^*) = \inf_{\mathcal{C} \in (\mathbb{R}^{d+1})^k} \mathbb{E} \min_{c_i \in \mathcal{C}} ||(T, X) - c_i||^2. \tag{3}$$

The last assertion means that the task of determining the optimal quantizer is reduced to the class of the nearest neighbor quantizers, which are entirely characterized by their codebooks.

In practice, the distribution P of (T, X) is unknown and the optimal quantizer q^* cannot be calculated. However, if one has access to an i.i.d. sample $(T_i, X_i)_{1 \leq i \leq n}$ of (T, X), it is possible to replace the distortion with respect to P with the distortion with respect to the empirical measure P_n induced by the sample. Therefore, the minimization of (2) is replaced by the minimization, with respect to $\mathcal{C}$, of the empirical distortion

$$D(P_n, q) = \frac{1}{n} \sum_{i=1}^{n} \min_{c_j \in \mathcal{C}} ||(T_i, X_i) - c_j||^2. \tag{4}$$

A quantizer $q_n^* \in Q_k$, minimizing (4) is called to be empirically optimal. We recall that the optimal quantization is closely connected to k-clustering. Indeed, given a set of i.i.d. observations of (T, X), the codevectors of the optimal codebook are the coordinates of the optimal cluster centers. Unfortunately, the task of numerical minimization of (4) over all codebooks of size k is NP-hard, therefore $\mathcal{C}$ can only be approximated using k-means type iterative algorithms.

In presence of censoring, the observed vector is no longer (T, X) but $(Y, \delta, X) = (\min(T, C), \mathbb{1}_{T \leq C}, X)$, as T is not directly observed. Therefore, the classical definition of the empirical distortion involves unobserved quantities $(T_i)_{1 \leq i \leq n}$ and can not be used in our setting. From the point of view of clustering, the specificity of censored data is that the distances between censored observations and other points are not observed and the standard iterative clustering algorithms based on the distances between the observations fail to work.

3 Quantization under censoring

In this section, we are concerned with the quantization of the random vector (T, X) taking its values in $\mathbb{R}^{d+1}$, in presence of censoring acting on the variable T. As we have already seen, in this setting, we dispose only of i.i.d. replications

$$(Y_i, \delta_i, X_i)_{1 \leq i \leq n}$$

of the observed vector (Y, δ, X). Therefore, the unknown distribution P of (T, X) can not be estimated by the empirical distribution function and the classical definition of the empirical distortion is not appropriate. The basic idea of our approach is to estimate P by another random measure arising from the available observations. This measure is associated with the estimator of the distribution function of (T, X) due to [Stute, 1993]. Its consistency requires the following identifiability assumptions which will supposed to be satisfied throughout the paper:

- T and C are independent
- $P(T \leq C|X, Y) = P(T \leq C|Y)$

This set of assumptions is standard in survival analysis. For more details, we refer to [Stute, 1993], [Stute, 1996], [Stute, 1999], [Gannoun et al., 2007], [Lopez, 2009] and [Sánchez Sellero et al., 2005].

Let $Y_{[i:n]}$ be the i-th order statistics of the sample $(Y_1, \ldots, Y_n)$. We will denote by $\delta_{[i:n]}$ and $X_{[i:n]}$ the corresponding realizations of the indicator and of the covariate. With this notation, the estimator of [Stute, 1993] takes the following form:

$$\hat{F}_n(t, x) = \sum_{i=1}^{n} W_{[i:n]} \mathbb{1}_{Y_{[i:n]} \leq t, X_{[i:n]} \leq x}, \quad t \in \mathbb{R},\ x \in \mathbb{R}^d, \tag{5}$$

where $W_{[i:n]}$ is the weight assigned to $Y_{[i:n]}$ by the univariate Kaplan-Meier estimator (see [Kaplan and Meier, 1958]) evaluated from the sample $(Y_i, \delta_i)_{1 \leq i \leq n}$. It has the following expression (see [Stute and Wang, 1993]):

$$W_{[i:n]} = \frac{\delta_{[i:n]}}{n - i + 1} \prod_{j=1}^{i-1} \left(\frac{n - j}{n - j + 1} \right)^{\delta_{[j:n]}}, \quad i = 1, \ldots, n. \tag{6}$$

We will adopt here an alternative form of the estimator (5) which was derived by [Satten and Datta, 2001]. Let W_{in} denote the weight attributed to the i-th observation and let $\hat{G}$ be a Kaplan-Meier estimator of the distribution function G of the censoring variable C. Using this notation, the estimator (5) can be written as follows:

$$\hat{F}_n(t, x) = \sum_{i=1}^{n} W_{in} \mathbb{1}_{Y_i \leq t, X_i \leq x}, \quad \text{with} \quad W_{in} = \frac{\delta_i}{n(1 - \hat{G}(Y_i-))}, \tag{7}$$

where $G(y-)$ denotes the left-hand limit of G at y.

In presence of censoring, the lack of large observations creates some difficulties for estimating tails of distributions and can make estimators inconsistent in the neighborhood of the upper bound of support. A common approach for overcoming this difficulty (see for instance [Heuchenne, 2008]) consists in truncating the distribution by a compact set $[0, \tau]$ strictly included in its support. However, the truncation τ may be chosen arbitrarily close to the upper bound of the support. This choice seems to be the best adapted to our theory. In what follows, we will deal with the truncated distribution $P^\tau := P_{(T,X)|T\leq\tau}$.

Let us consider now the corresponding distribution function $F^\tau(t,x)$ and the following estimator of it:

$$F_n^\tau(t,x) = \frac{\sum_{i=1}^n W_{in}\mathbb{1}_{Y_i\leq t, X_i\leq x}\mathbb{1}_{Y_i\leq\tau}}{\sum_{i=1}^n W_{in}\mathbb{1}_{Y_i\leq\tau}}, \quad t\in\mathbb{R},\ x\in\mathbb{R}^d. \tag{8}$$

We note that (8) is an adaptation of the estimator (5) by introducing truncation and by normalizing a sum of its weights by 1. This estimator induces a probability measure

$$\mathcal{P}_n^\tau = \sum_{i=1}^n W_{in}^\tau \delta_{(Y_i,X_i)}, \text{ with } \quad W_{in}^\tau = \frac{W_{in}\mathbb{1}_{Y_i\leq\tau}}{\sum_{i=1}^n W_{in}\mathbb{1}_{Y_i\leq\tau}}.$$

Now, it is natural to define the empirical distortion under censoring as the distortion with respect to the empirical distribution $\mathcal{P}_n^\tau$:

$$\begin{aligned}\mathcal{D}(\mathcal{P}_n^\tau, q) &= \frac{\sum_{i=1}^n W_{in}||(Y_i,X_i)-q(Y_i,X_i)||^2\mathbb{1}_{Y_i\leq\tau}}{\sum_{i=1}^n W_{in}\mathbb{1}_{Y_i\leq\tau}} \\ &= \sum_{i=1}^n W_{in}^\tau||(Y_i,X_i)-q(Y_i,X_i)||^2. \end{aligned}\tag{9}$$

In this context, a quantizer $q_n^* \in Q_k$ will be called empirically optimal when it minimizes the empirical distortion (9). This quantizer always exist due to [Pollard, 1982b]. The next section will be concerned with the asymptotic properties of q_n^*, that is with the convergence of its distortion $\mathcal{D}(\mathcal{P}^\tau, q_n^*)$ towards the optimal distortion $D_k^*(P^\tau)$ and with its rate.

4 Consistency of the empirical design

This section studies the asymptotic behavior of the empirically optimal quantizer q_n^*. Theorem 1 establishes the almost sure convergence of the distortion of q_n^* towards the minimal distortion $D_k^*(P^\tau)$. Theorem 2 provides an exponential inequality for the difference between these two distortions. Corollary 1 gives the rate of the almost sure convergence.

4.1 Almost sure convergence

The following Theorem 1 establishes the almost sure convergence of the distortion of the empirically optimal quantizer. The proof is based on the fact that the absolute value of the difference between two distortions is bounded by a Wasserstein distance

between the probability measure $\mathcal{P}_n^\tau = \sum_{i=1}^n W_{in}^\tau \delta_{(T_i,X_i)}$ and the conditional distribution P^τ of (T,X), given $T \leq \tau$. We show that the indicated distance converges almost surely to zero.

Theorem 1. *For all $k \geq 1$, the empirically optimal k-quantizer satisfies*

$$D(P^\tau, q_n^*) \xrightarrow[n\to\infty]{a.s.} D_k^*(P^\tau). \tag{10}$$

Proof. We recall that the Wasserstein distance between two probability measures μ and ν is defined by

$$\rho(\mu,\nu) = \inf_{X\sim\mu, Y\sim\nu} (\mathbb{E}||X-Y||^2)^{1/2},$$

where the infinium is taken over all random vectors (X,Y) having marginal distributions μ and ν, respectively. Any nearest neighbor quantizer satisfies (see [Linder, 2002]):

$$|D(\mu,q)^{1/2} - D(\nu,q)^{1/2}| \leq \rho(\mu,\nu),$$

and

$$|D_N^*(\mu)^{1/2} - D_N^*(\nu)^{1/2}| \leq \rho(\mu,\nu).$$

Applying these inequalities to the probability measures $\mathcal{P}_n^\tau$ and P^τ, we obtain

$$|D(\mathcal{P}_n^\tau, q_n^*)^{1/2} - D(P^\tau, q^*)^{1/2}| \leq \rho(\mathcal{P}_n^\tau, P^\tau). \tag{11}$$

By the following, we will show that the right-hand side of (11) converges to zero almost surely, which implies the assertion of the theorem. To this aim, we recall that $\rho(\mathcal{P}_n^\tau, P^\tau) \xrightarrow[n\to\infty]{a.s.} 0$ is equivalent to

$$P\left[\mathcal{P}_n^\tau \underset{n\to\infty}{\Rightarrow} P^\tau\right] = 1 \quad \text{and} \quad P\left[\int ||(t,x)||^2 d\mathcal{P}_n^\tau(t,x) \underset{n\to\infty}{\to} \int ||(t,x)||^2 dP^\tau(t,x)\right] = 1, \tag{12}$$

where $\Rightarrow$ denotes the weak convergence (see, for example, [Rachev and Rüschendorf, 1998]). We will prove that both conditions of (12) are satisfied. At first, let us recall ([Stute, 1993]) that, for any measurable function $\phi(t,x) : \mathbb{R}^{d+1} \to \mathbb{R}$, we have

$$\int \phi(t,x) dF_n(t,x) \xrightarrow[n\to\infty]{a.s.} \mathbb{E}_P[\phi(T,X)]. \tag{13}$$

The same property is true for the modified estimator. Indeed,

$$\begin{aligned} \int \phi(t,x) dF_n^\tau(t,x) &= \frac{\sum_{i=1}^n W_{in}\phi(Y_i,X_i)\mathbb{1}_{Y_i\leq\tau}}{\sum_{i=1}^n W_{in}\mathbb{1}_{Y_i\leq\tau}} \\ &= \frac{\int \phi(t,x)\mathbb{1}_{t\leq\tau} dF_n(t,x)}{\int \mathbb{1}_{t\leq\tau} dF_n(t,x)} \xrightarrow[n\to\infty]{} \mathbb{E}_{P^\tau}(\phi(T,X)) \quad \text{a.s.} \end{aligned}$$

Now, take $\phi(t,x) = ||(t,x)||^2$. This leads to

$$P\left(\int ||(t,x)||^2 d\mathcal{P}_n^\tau(t,x) \xrightarrow[n\to\infty]{} \int ||(t,x)||^2 dP^\tau(t,x)\right) = 1. \tag{14}$$

Thus, it remains to prove that $P(\mathcal{P}_n^\tau \underset{n\to\infty}{\Rightarrow} P^\tau) = 1$. To that aim, we invoke arguments similar to [Varadarajan, 1958], who showed the weak convergence of empirical measure on the set of probability one. By Lévy criterium, all we need to obtain is

$$P\left(\forall u \in \mathbb{R}^{d+1} : \psi_n(u) \to \psi(u)\right) = 1,$$

where $\psi(u) = \int \exp(i\langle (t,x), u\rangle) dP^\tau(t,x)$ and $\psi_n(u) = \int \exp(i\langle (t,x), u\rangle) dP_n^\tau(t,x)$ are the Fourrier transforms of P^τ and P_n^τ. Remark that the event

$$\Omega(u) = \{\omega : \psi_n(u) \to \psi(u)\}$$

satisfies $P^\tau(\Omega(u)) = 1$ for all u, because of property (13) applied to $\phi(u) = \exp\left(i\langle (t,x), u\rangle\right)$. Let T be a countable dense subset of $\mathbb{R}^{d+1}$ and consider an event

$$\Omega_0 = \bigcap_{u\in T} \Omega(u) \bigcap \{\mathcal{P}_n^\tau ||(t,x)|| \to P^\tau ||(t,x)||\},$$

which is of probabilty equal to one. For any $u \in \mathbb{R}^{d+1}$ and $\omega_0 \in \Omega_0$, consider a sequence $\{u_k\}_{k=1}^\infty$, such that $u_k \in T$ and $u_k \to u$. For any fixed k, we have:

$$\begin{aligned} |\psi_n(\omega_0, u) - \psi(u)| &\leq |\psi_n(\omega_0, u) - \psi_n(\omega_0, u_k)| + |\psi_n(\omega_0, u_k) - \psi(u_k)| + \\ &\quad |\psi(u_k) - \psi(u)| \\ &\leq ||u - u_k|| \left(E_{P_n^\tau} ||(T,X)|| + E_{P^\tau} ||(T,X)||\right) \\ &\quad + |\psi_n(\omega_0, u_k) - \psi(u_k)|, \end{aligned}$$

with $E_{P_n^\tau} ||(T,X)|| \underset{n\to\infty}{\to} E_{P^\tau} ||(T,X)||$, as $\omega_0 \in \Omega_0$. Moreover, $\omega_0 \in \Omega(u_k)$ implies that, for any k,

$$\lim_{n\to\infty} \sup |\psi_n(\omega_0, u) - \psi(u)| \leq 2||u - u_k|| E_{P^\tau} ||(T,X)||.$$

Now, let k tend to infinity. This concludes the proof. □

4.2 Exponential inequality

From now on, we will assume that the support of the random variable (T,X) is bounded, i.e. there exists some constant $R > 0$ such that, $P(||(T,X)|| \leq R) = 1$. In the previous section, we established the almost sure consistency of the empirical design when $n \to \infty$. This section deals with some finite sample result. More precisely, the following theorem provides a non asymptotic exponential bound for the difference between the distortion of the empirically optimal quantizer and the minimal distortion.

Theorem 2. *There exist some positive universal constants K, K_1, K_2, L_1, L_2 such that, for any $z > 4K/F_\tau^T$, with $F_\tau^T = P(T \leq \tau)$, the following inequality holds:*

$$\begin{aligned} P(\sqrt{n}|D(P^\tau, q_n^*) - D(P^\tau, q^*)| > z) &\leq 5\exp\left(-L_1 z^2 + L_2 z\right) \\ &+ 2\left[\exp(-K_1 z^2) + \exp(-\sqrt{n} K_2 z)\right] \\ &+ O\left(e^{-\sqrt{n}}\right). \end{aligned}$$

We note that the remainder term $O\left(e^{-\sqrt{n}}\right)$ does not depend on z. It arises from the control of the difference between the distribution functions of (T, X) and C and their respective estimators on sets, which are not depending on z.

For the sake of clarity, the proof of Theorem 2 is postponed to Section 7. It is based on the empirical process theory applied to classes of functions indexed by k-point quantizers. The main idea is to bound the difference between the distortions by a deviation of the supremum of some empirical process indexed by a Donsker functional class. The exponential inequality follows then from a concentration inequality of [Talagrand, 1994]. One of the main technical difficulties is that the quantizer q_n^* is optimal with respect to the empirical measure $\mathcal{P}_n^\tau$ with random weights, depending on the Kaplan-Meier estimator $\hat{G}(y)$ of the censoring variable. In order to handle such measure, we need to replace $\hat{G}(y)$ with its deterministic limit $G(y)$. To that aim, it is necessary to control $\sup_y |\sqrt{n}(\hat{G}(y) - G(y))|$, where the supremum is taken over sets which are not depending on z. This is done through an exponential inequality of [Bitouzé et al., 1999] for the Kaplan-Meier estimator.

The following corollary provides a rate of the almost sure convergence in Theorem 1.

Corollary 1. *For every probability measure P and $\tau > 0$, as $n \to \infty$*

$$|D(P^\tau, q_n^*) - D(P^\tau, q)| = O\left(\frac{\log n}{\sqrt{n}}\right) \quad a.s. \tag{15}$$

Proof. The result is a corollary of Theorem 2 and Borel-Cantelli Lemma applied to the sequence of events $\Omega_n = \{|D(P^\tau, q_n^*) - D(P^\tau, q)| > z_n\}$ with $z_n = \log n$. □

5 Clustering algorithm under censoring

Motivated by the examples given in the introduction, we are considering the following set up. Suppose that $(T_i, X_i)_{1 \le i \le n}$ are the realizations of the lifetime and of its covariates for n subjects. In presence of censoring, we do not have access to these realizations but we do observe their censored versions $(Y_i, \delta_i, X_i)_{1 \le i \le n}$. The aim of this section is to propose a way for partitioning the n subjects into k groups with respect to their unobserved realizations $(T_i, X_i)_{1 \le i \le n}$ having at hand only the corresponding censored data set. If we come back to the example of cancer patients, this task means that we want to group them with respect to their survival times and covariates, although for some of them we only know that the survival time is greater than the observed value of the censoring variable.

In Section 3 we proposed an empirical quantification criterion (9). Similarly to the non censored case, the idea consists in evaluating the coordinates of the centers of clusters by minimization of this criterion. Then, each cluster is to be composed of the observations for which its center is the nearest. However, we have to overcome several difficulties due to the presence of censoring. The first of them is that, in our case as well as in absence of censoring, the corresponding numerical minimization problem is NP-hard. Therefore, we need to define an iterative procedure which approximates the coordinates of the unknown centers and which is compatible with the censored data. The main issue in carrying out this task is that all euclidean distances related

to censored observations are unobserved. Hence, the classical k-means algorithm breaks down. Step 1 of our algorithm provides its generalization in our framework and allows for finding the centers of clusters. In contrast to the standard clustering setting, that is not sufficient for assigning labels to all observations. Indeed, each non censored observation still can be affected to cluster with the nearest center. In contrast, for censored observations, the distances to the centers are not observed and one is not able to assign them labels. Step 2 of our algorithm is a way of overcoming this difficulty by estimating the unknown distances between censored observations and the cluster centers. Based on this estimation, we assign to each censored observation the label of the center which is nearest with respect to the estimated distance.

5.1 Finding centers of clusters

At each iteration of a standard k-means algorithm, a center of cell S is actualized by the empirical mean of the composing it observations. This quantity represents a consistent estimator of

$$\mathbb{E}\left[(T,X)|(T,X)\in S\right]. \tag{16}$$

In non censored case, all observations contribute to the estimation of this expectation with the same weight. In presence of censoring, the empirical mean of the non censored observations of S is a biased estimator of (16). Its consistent version is given by a weighted sum of uncensored observations, where the corresponding weights are defined in (7) and compensate for censoring. Following this idea, we propose to actualize at each iteration the center c of cell S by

$$c=\frac{\sum_{i=1}^{n}(Y_i,X_i)^T W_{in}\mathbb{1}_{\{(Y_i,X_i)\in S,\delta_i=1\}}}{\sum_{i=1}^{n}W_{in}\mathbb{1}_{\{(Y_i,X_i)\in S,\delta_i=1\}}}. \tag{17}$$

The rest of the iterative procedure is analogous to the classical k-means. The corresponding pseudocode is presented in the frame Step 1. At the end of this step, the centers of k clusters are evaluated and all non censored observations received their labels.

Remark. Our theoretical results need to introduce a truncation bound τ which can be chosen arbitrarily close to the upper bound of the support of distribution. In practice, τ can be chosen equal to the former without significant impact on the results.

5.2 Assigning labels to censored observations

If the i-th observation is censored, one only observes that $\delta_i=0$ and that $T_i>Y_i$. The best approximation of the unobserved euclidean distance $d\left((T_i,X_i);c\right)$ from this observation to center c is given by:

$$\mathbb{E}\left[d\left((T_i,X_i);c\right)|X_i,T_i>Y_i,\delta_i=0\right].$$

Step 1 Evaluation of k centers

• Initialize the centers by $c_1^{(0)}, \ldots, c_k^{(0)}$

• Evaluate the weights W_{in} of Kaplan-Meier estimator based on the sample $(Y_i, \delta_i)_{1 \leq i \leq n}$

• **Repeat until nothing changes:** for the iteration ℓ

- Calculate Voronoï cells $S_1^\ell, \ldots, S_k^\ell$ corresponding to centers $c_1^{(\ell)}, \ldots, c_k^{(\ell)}$ for the set of non censored observations $\{(Y_i, X_i) : \delta_i = 1, i = 1, \ldots, n\}$
- For $j = 1, \ldots, k$ calculate new centers $(c_j^{(\ell+1)})_{1 \leq j \leq k}$ as

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

• The algorithm stoppes in a finite number ℓ^* of iterations. For $j = 1, \ldots, k$ attribute to observation (Y_i, X_i) with $\delta_i = 1$ a label j if $(T_i, X_i) \in S_j^{\ell^*}$.

Therefore, for each $i = 1, \ldots, n$, such that $\delta_i = 0$ we estimate the distance between (T_i, X_i) and the center $c_j^{(\ell*)}$ by the following estimator of this conditional expectation:

$$\hat{d}_{ij} = \frac{\int_{Y_i}^\infty ||(t, X_i) - c_j^{(\ell*)}||^2 d\hat{F}(t|X_i)}{\int_{Y_i}^\infty d\hat{F}(t|X_i)}, \tag{18}$$

where $\hat{F}(t|x)$ is an estimator of $F(t|X = x) = P(T \leq t|X = x)$ given by

$$\hat{F}(t|x) = \frac{1}{n} \sum_{i=1}^n W_{in} \frac{k(\frac{x - X_i}{h})}{\sum_{j=1}^n k\left(\frac{x - X_j}{h}\right)} \mathbb{1}_{Y_i \leq t}, \tag{19}$$

where $x \in \mathbb{R}^d$ and $k(x)$ is a kernel, that is a positive integrable function such that $\int_{\mathbb{R}^d} k(x)dx = 1$. Combining (18) and (19), we obtain

$$\hat{d}_{ij} = \frac{\sum_{m=1}^n W_{mn} ||(Y_m, X_i) - c_j^{(l*)}||^2 k\left(\frac{X_i - X_m}{h}\right) \mathbb{1}_{Y_m \geq Y_i}}{\sum_{m=1}^n W_{mn} k(\frac{X_i - X_m}{h}) \mathbb{1}_{Y_m \geq Y_i}}. \tag{20}$$

We present now the second step of our algorithm.

Step 2 Assigning labels to censored observations

• For each censored observation (Y_i, X_i) evaluate the estimated distances $\hat{d}_{ij}$ according to Equation (20).

• Assign to (Y_i, X_i) a label $j^* = \arg\min_j \hat{d}_{ij}$.

5.3 Number of clusters

Similarly to the other k-means type algorithms, our procedure uses the number k of clusters as the input value. However, in practice k is unknown and is to be

chosen adaptively. This issue was extensively studied in absence of censoring. A lot of criterions were proposed in the literature, a complete review and a comparative Monte Carlo study of most of them can be found in [Milligan and Cooper, 1985]. Several of these criterions can be adapted in presence of censoring. We propose here a rule for choosing the number of clusters, which is an adaptation of the criterion of [Krzanowski and Lai, 1988] proposed for non censored data. In absence of censoring, one have to calculate, for some range of values of k, the pooled within-cluster sum of squares S_k and a quantity

$$DIFF(k) = (k-1)^{2/(d+1)} S_{k-1} - k^{2/(d+1)} S_k,$$

where d+1 is the total number of the variables. [Krzanowski and Lai, 1988] proposed to chose k maximizing

$$KL(k) = \left| \frac{DIFF(k)}{DIFF(k+1)} \right|. \tag{21}$$

In our case, we propose to replace S_k by the weighted sum of squares involving only non censored observations:

$$\mathcal{D}_k = \sum_{i=1}^{n} W_{in} \min_{c_j \in \mathcal{C}} \|(Y_i, X_i) - c_j\|^2,$$

where $\mathcal{C} = \{c_1, \ldots, c_k\}$ are the centers of k clusters resulting from our iterative algorithm. Similarly to (21), the optimal number of clusters is to be chosen as the value of k maximizing

$$\left| \frac{(k-1)^{2/(d+1)} \mathcal{D}_{k-1} - k^{2/(d+1)} \mathcal{D}_k}{k^{2/(d+1)} \mathcal{D}_k - (k+1)^{2/(d+1)} \mathcal{D}_{k+1}} \right|.$$

6 Simulations and a real data analysis

6.1 Simulation study

In this section, we evaluate the performance of our algorithm on simulated data sets. We proceed in the following way. At the first step, a complete data sample with known clusters is created and a k-means algorithm is applied, in order to obtain a partition of the data into k clusters. The accuracy of this partition is then compared to that of the partition produced by our algorithm, having as the input the censored version of the initial sample.

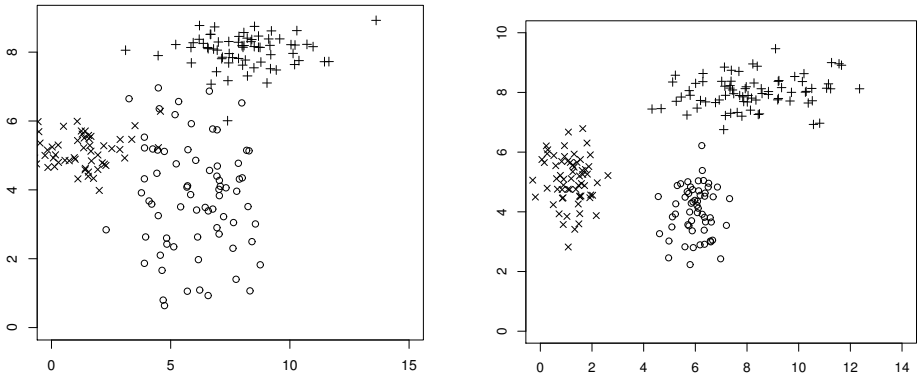
This comparison is done through the corrected Rand's statistics (see [Rand, 1971] and [Hubert and Arabie, 1985]). Rand's index permits to compare two partitions P_1 and P_2 in order to know how close they are. In the set of all possible pairs of observations let A (for "agreement") denote the number of pairs which are of one of the following types:

- Pairs of observations belonging to the same class in P_1 and P_2
- Pairs of observations belonging to a different class in P_1 and to a different class in P_2

The total number of pairs being $n(n-1)/2$, Rand's statistics is defined by $R_{P_1P_2} = 2A/(n(n-1))$. The closer $R_{P_1P_2}$ is to one, the closer are the two partitions.

Plan of the simulation study:

For each of three different levels of censoring $(15\%, 30\%, 45\%)$, we generated 1000 bivariate data sets of $n = 200$ observations. For each $j = 1, \ldots, 1000$, the j-th data set is composed of $k = 3$ clusters (clusters are supposed to be known), forming the partition denoted by $\mathcal{P}_j^0$. Data are simulated using a Gaussian mixture, in two following cases: groups are close (see Figure 1, (a)) and groups are well separated (see Figure 1, (b)). Censoring variable is chosen to have a uniform distribution.



(a) Case 1: groups are close (b) Case 2: groups are well separated

Figure 1: Examples of simulated data.

The exact scheme of simulations is the following. For each level of censoring, $k = 3$ and $j = 1, \ldots, 1000$,

- Simulate a complete data sample $(T_i^{(j)}, X_i^{(j)})_{1 \le i \le n}$. Apply a k-means algorithm, leading to a partition of these data into k clusters. Denote this partition by $\mathcal{P}_j^c$.

- Simulate a sample $(C_i^{(j)})_{1 \le i \le n}$ from censoring variable and get the censored data set as $(\min(T_i^{(j)}, C_i^{(j)}), \delta_i^{(j)}, X_i^{(j)})_{1 \le i \le n}$. Apply our algorithm described in Section 5 and denote the resulting partition into k clusters by $\mathcal{P}_j$.

- Calculate Rand's statistics $R_{\mathcal{P}_j^c \mathcal{P}_j^0}$ and $R_{\mathcal{P}_j^c \mathcal{P}_j^0}$.

The value of $R_{\mathcal{P}_j^c \mathcal{P}_j^0}$ shows how accurate is the partition created by k-means (applied to sample before censoring) with respect to the known "true" partition $\mathcal{P}_j^0$, and $R_{\mathcal{P}_j^c \mathcal{P}_j^0}$ have the same meaning for our algorithm.

Results. In Table 1 (case of close groups) and Table 2 (case of well separated groups), we present the mean values of the corrected Rand's statistics over $N = 1000$

data sets, that is

$$R^N_{\mathcal{P}\mathcal{P}^0} = 1/N \sum_{j=1}^{n} R_{\mathcal{P}_j \mathcal{P}^0_j},$$

and

$$R^N_{\mathcal{P}^c\mathcal{P}^0} = 1/N \sum_{1}^{n} R_{\mathcal{P}^c_j \mathcal{P}^0_j}.$$

Not surprisingly, both methods perform better for the well separated groups than

Level of censoring	15%	30%	45%
$R^N_{\mathcal{P}^c\mathcal{P}^0}$	0.931	0.929	0.930
$R^N_{\mathcal{P}\mathcal{P}^0}$	0.905	0.878	0.851
$R^N_{\mathcal{P}^c\mathcal{P}^0} - R^N_{\mathcal{P}\mathcal{P}^0}$	0.026	0.051	0.079

Table 1: Corrected Rand's statistics for simulated data, close groups

Level of censoring	15%	30%	45%
$R^N_{\mathcal{P}^c\mathcal{P}^0}$	0.993	0.994	0.994
$R^N_{\mathcal{P}\mathcal{P}^0}$	0.972	0.942	0.923
$R^N_{\mathcal{P}^c\mathcal{P}^0} - R^N_{\mathcal{P}\mathcal{P}^0}$	0.021	0.052	0.071

Table 2: Corrected Rand's statistics for simulated data, separated groups

for the closed ones. We remark also that the agreement between the partition by our method and the true partition decreases when the proportion of censored observations increase and the difference between the accuracy our method (having at the entry the censored sample) and that of the procedure based on its full version (which is unavailable in practice) becomes more important. However, this difference does not rise drastically and the agreement for our method remains relatively good even at 45% of censoring.

6.2 Real data analysis: PBC data

In this section, we illustrate our results by an application to a real data set. We consider the data from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver which is a rare and fatal chronic liver disease. The study had been conducted between 1974 and 1984. For a total number of 418 patients the recorded measurements are the time at risk (censored), censored indicator and 17 covariates such as a patient age, sex, clinical, biochemical and histological measurements. After excluding the participants with missing values of some covariates, we obtained a data set of 258 observations, 111 of which are non censored. The detailed description of the data set can be found in [Fleming and Harrington, 1991].

For the easier interpretability of the results we performed clustering using only variables which were shown to be important for the survival (see the study of the

same data in the regression setting conducted by [Grambsch et al.,]). These covariates are the patients' age, total serum bilirubin mentioned as one of the most important factors influencing the lifetime, serum albumin concentration and the prothrombin time. We excluded the severity of edema variable as our method permits to take into account only the quantitative covariates.

Before clustering, the observations of each variable were normalized by dividing by the corresponding range. Our algorithm has detected four clusters of patients. Table 3 presents the mean values of the survival time and of the covariates for each group and for all of the patients. The results show that the discriminative

	Survival	Age	Bilirubin	Prothrombin	Albumin
Group 1	3001.66	47.11	2.16	10.70	3.54
Group 2	2394.97	51.77	3.88	11.23	3.54
Group 3	1746.10	53.57	5.41	11.14	3.38
Group 4	1145.09	58.49	8.21	11.30	3.49
Overall means	2180.26	50.42	3.34	10.75	3.51

Table 3: Clusters found in PBC data

variables seem to be the survival time, the bilirubin level and the age. Group 1 is characterized by the most important survival time associated with the low level of bilirubin and the lowest age of patients. In contrast, Group 4 is represented by the lowest survival, a very high bilirubin level and the most important age. Groups 2 and 3 are the medium cases between 1 and 4. One can see clearly that the survival is strongly associated with the bilirubin level. This fact is in concordance with the recognized importance of the factor. The mean prothrombin and albumin levels seem to be rather close for the different groups.

In conclusion, the group of the highest risk is composed of patients with the most important level of bilirubin and great ages while the lowest risk corresponds to the youngest patients with the lowest level of bilirubin.

7 Proof of Theorem 2

In this section, we are giving the proof of the exponential inequality announced in Section 4.

Proof. We have

$$P(\sqrt{n}|D(P^\tau, q_n^*) - D(P^\tau, q^*)| > z) \quad \leq \quad P(\sqrt{n} \sup_{q \in Q_N} |D(P^\tau, q) - D(\mathcal{P}_n^\tau, q)| > z/2).$$

Using notation $f_q(y,x) = ||(y,x) - q(y,x)||^2$, $F_\tau^T = P(T \leq \tau)$, $\hat{P}_n = \sum_{i=1}^n W_{in}\delta_{(Y_i,X_i)}$

and $F^T_{\tau,n} = \hat{P}_n(T \leq \tau)$, we obtain

$$\begin{aligned}
|D(P^\tau, q) - D(P^\tau_n, q)| &= \left| \int f_q(y,x) \mathbb{1}_{y\leq\tau} \frac{dP(y,x)}{P(T\leq\tau)} - \int f_q(y,x)\mathbb{1}_{y\leq\tau} \frac{d\hat{P}_n(y,x)}{\hat{P}_n(T\leq\tau)} \right| \\
&= \left| \frac{1}{F^T_\tau} \int f_q(y,x)\mathbb{1}_{y\leq\tau} d(P-\hat{P}_n) \right. \\
&\qquad \left. + \frac{F^T_{\tau,n} - F^T_\tau}{F^T_\tau F^T_{\tau,n}} \int f_q(y,x)\mathbb{1}_{y\leq\tau} d\hat{P}_n \right|.
\end{aligned}$$

Therefore,

$$P(\sqrt{n} \sup_{q\in Q_N} |D(P^\tau, q) - D(\mathcal{P}^\tau_n, q)| > z/2) \leq T_1 + T_2,$$

where

$$\begin{aligned}
T_1 &:= P\left(\sqrt{n} \sup_{q\in Q_N} \left| \int f_q(y,x)\mathbb{1}_{y\leq\tau} d(P-\hat{P}_n) \right| > \frac{z}{4} F^T_\tau \right), \\
T_2 &:= P\left(\sqrt{n} \sup_{q\in Q_N} \left| \frac{F^T_{\tau,n} - F^T_\tau}{F^T_\tau F^T_{\tau,n}} \int f_q(y,x)\mathbb{1}_{y\leq\tau} d\hat{P}_n \right| > \frac{z}{4} \right).
\end{aligned}$$

In the following we will consider separately the terms T_1 and T_2.

1. The first term $\mathbf{T_1}$. Let us denote by $P_n(y,x,\delta) = \frac{1}{n}\sum_{i=1}^n \delta_{(Y_i,X_i,\delta_i)}$ the empirical measure of the available observations. For any function $\phi(y,x)$, we have

$$E\left[\frac{\delta}{1-G(Y-)} \phi(Y,X) \right] = E\left[\frac{E[\mathbb{1}_{Y\leq C}|Y,X]\phi(Y,X)}{1-G(Y-)} \right] = E\left[\phi(T,X)\right].$$

Therefore, the term T_1 can be written as

$$\begin{aligned}
T_1 &= P\left(\sqrt{n} \sup_{q\in Q_N} \left| \int f_q(y,x) \frac{\delta\mathbb{1}_{y\leq\tau} dP(y,x,\delta)}{1-G(y-)} \right.\right. \\
&\qquad \left.\left. - \int f_q(y,x) \frac{\delta\mathbb{1}_{y\leq\tau} dP_n(y,x,\delta)}{1-\hat{G}_n(y-)} \right| > \frac{z}{4} F^T_\tau \right) \\
&\leq T_{11} + T_{12},
\end{aligned}$$

where

$$\begin{aligned}
T_{11} &:= P\left(\sqrt{n} \sup_{q\in Q_N} \left| \int f_q(y,x) \frac{\delta\mathbb{1}_{y\leq\tau}}{1-G(y-)} d(P-P_n) \right| > \frac{z}{4} F^T_\tau \right), \\
T_{12} &:= P\left(\sqrt{n} \sup_{q\in Q_N} \left| \int f_q(y,x)\delta\mathbb{1}_{y\leq\tau} \left[\frac{\hat{G}_n(y-) - G(y-)}{(1-\hat{G}_n(y-))(1-G(y-))} \right] dP_n \right| > \frac{z}{4} F^T_\tau \right).
\end{aligned}$$

Term $\mathbf{T_{11}}$. In order to handle the term T_{11}, let us first introduce first a class of functions

$$\mathcal{F}_1 = \left\{ g_q : g_q = \frac{\delta\mathbb{1}_{t\leq\tau}}{1-G(t-)} f_q(t,x), \quad q \in Q_N \right\}, \tag{22}$$

indicated by N-quantizers. For any $u > 0$, we have the following majoration:

$$P\left(\sqrt{n}\sup_{q\in Q_N}\left|\int f_q(y,x)\frac{\delta\mathbb{1}_{y\leq\tau}}{1-G(y-)}d(P-P_n)\right|>u\right)\leq P\left(\sqrt{n}\sup_{f\in\mathcal{F}_1}|(P_n-P)f|>u\right),$$

where we used the notation $Pf = \int f dP$ and $P_n f = \int f dP_n$. The exponential inequality for T_{11} follows from a concentration inequality proposed by [Talagrand, 1994] in the form, used by [Einmahl and Mason, 2005]. Let us first remark that,

$$\mathcal{F}_1 = h_\tau(t,\delta)\times\mathcal{F}_2, \tag{23}$$

where

$$\mathcal{F}_2 := \{f_q(t,x), \quad q\in Q_N\} \quad \text{and} \quad h_\tau(t,\delta) = \frac{\delta\mathbb{1}_{t\leq\tau}}{1-G(t-)}.$$

The class $\mathcal{F}_2$ is $P-$Donsker. Indeed, as proved e.g. in [Linder, 2002], the collection of sets $\{\{(t,x) : f_q(t,x) > u\}, u > 0, q \in Q_N\}$ forms a VC-class. Therefore, the class $\mathcal{F}_2$ is VC-major by definition given in Section 2.6.4 of [van der Vaart and Wellner, 1996], and is $P-$Donsker by Theorem 2.6.14 of Section 2.6.4.

Moreover, the function $h : (t,\delta) \to \frac{\delta\mathbb{1}_{t\leq\tau}}{1-G(t-)}$ is bounded. Consequently, $\mathcal{F}_1$ is $P-$Donsker as the pointwise product of the $P-$Donsker class $\mathcal{F}_2$ and the bounded function (see the permanence property in Example 2.10.10 of [van der Vaart and Wellner, 1996]).

We are now ready to apply the inequality of [Talagrand, 1994]. It states that, for any pointwise measurable class $\mathcal{F}$, satisfying $||f||_{\mathcal{F}} := \sup_{f\in\mathcal{F}}||f|| \leq M$ for some constant $0 < M < \infty$, we have for all $u > 0$,

$$\begin{aligned}P\left(\sqrt{n}\sup_{f\in\mathcal{F}}|(P_n-P)f|\geq\frac{A_1}{\sqrt{n}}\left(E||P_n^0 f||_{\mathcal{F}}+u\right)\right) \;\leq\; & 2\Big[\exp(-A_2u^2/n\sigma_{\mathcal{F}}^2)\\ & +\exp(-A_2u/M)\Big],\end{aligned} \tag{24}$$

where $P_n^0 f = \sum_{i=1}^n \varepsilon_i f(Y_i, X_i, \delta_i)$, with i.i.d. Rademacher random variables $(\varepsilon_i)_{1\leq i\leq n}$, $\sigma_{\mathcal{F}}^2 = \sup_{f\in\mathcal{F}} Var(f(Y,X,\delta))$ and A_1, A_2 are universal constants. For any function $f \in \mathcal{F}_1$, we have $||f||_{\mathcal{F}_1} \leq 4R^2(1-G(\tau-))^{-1}$. The application of the inequality (24) to the class of functions $\mathcal{F}_1$ gives

$$\begin{aligned}P\;\left(\sqrt{n}\sup_{f\in\mathcal{F}_1}|(P_n-P)f|\geq A_1u+\frac{A_1E||P_n^0f||_{\mathcal{F}_1}}{\sqrt{n}}\right)\\ \leq 2\left[\exp(-A_2u^2/\sigma_{\mathcal{F}_1}^2)+\exp(-\sqrt{n}A_2u/M)\right],\end{aligned} \tag{25}$$

where $M := 4R^2(1-G(\tau-))^{-1}$.

Using Propostion 1 of [Einmahl and Mason, 2005], we will show that the term $A_1E||P_n^0f||_{\mathcal{F}_1}/\sqrt{n}$ is uniformly bounded by some constant B_1. Indeed, as all functions of the class $\mathcal{F}_1$ are uniformly bounded, the only condition to be verified is the inequality $N(\varepsilon,\mathcal{F}_1) \leq C\varepsilon^{-\nu}$ on covering numbers, for some constants $C, \nu \geq 1$ and every $\varepsilon \in (0,1)$. This condition is satisfied. Indeed, by [Linder, 2002], the considered class of functions is a VC class. Therefore, Theorem 2.6.7 of

[van der Vaart and Wellner, 1996] applies and gives the required bound on the covering number.

According to Proposition 1 of [Einmahl and Mason, 2005], there exists some constant B_1, such that $E||P_n^0 f||_{\mathcal{F}} \leq B_1\sqrt{n}$. The inequality (25) takes form

$$P\left(\sqrt{n}\sup_{f\in\mathcal{F}_1}|(P_n-P)f| \geq A_1u + B_1\right) \leq 2\left[\exp(-A_2u^2/\sigma_{\mathcal{F}_1}^2) + \exp(-\sqrt{n}A_2u/M)\right], \tag{26}$$

for any $u > 0$ and some universal constants A_1, B_1. For any $v > K := \min(A_1+B_1, 1)$ (26) can be rewritten in the form

$$P\left(\sqrt{n}\sup_{f\in\mathcal{F}_1}|(P_n-P)f| \geq v\right) \leq 2\left[\exp(-K_1u^2) + \exp(-\sqrt{n}K_2u)\right], \tag{27}$$

where $K_1 = A_2(\sigma_{\mathcal{F}_1}(A_1+B_1))^{-2}$ and $K_2 = A_2M^{-1}(A_1+B_1)^{-1}$. Therefore, for any $z > 4K/F_\tau^T$,

$$T_{11} \leq 2\left[\exp(-K_1z^2) + \exp(-\sqrt{n}K_2z)\right].$$

Term $\mathbf{T_{12}}$. Let us use the following decomposition,

$$\begin{aligned}
T_{12} &\leq P\left(\sqrt{n}\sup_{q\in Q_N}\left|\int f_q(y,x)\mathbb{1}_{y\leq\tau}\delta dP_n\right| \times \sup_{y\leq\tau}\left|\frac{(\hat{G}_n-G)(y-)}{1-\hat{G}_n(y-)}\right| > \frac{z}{4}F_\tau^T(1-G(\tau))\right)\\
&\leq P\left(\sqrt{n}\sup_{y\leq\tau}\left|\frac{(\hat{G}_n-G)(y-)}{1-\hat{G}_n(y-)}\right| > \frac{z}{16R^2}F_\tau^T(1-G(\tau))\right)\\
&\leq P(\mathcal{A}_n) + P(\mathcal{B}_n),
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{A}_n : &= \left\{\sqrt{n}\sup_{y\leq\tau}\left|\frac{\hat{G}_n(y-)-G(y-)}{1-\hat{G}_n(y-)}\right| > \frac{zF_\tau^T(1-G(\tau))}{16R^2}\right\}\\
&\qquad\qquad \cap\left\{\sup_{y\leq\tau}|\hat{G}_n(y-)-G(y-)| \leq \frac{1-G(\tau)}{2}\right\},
\end{aligned}$$

and

$$\mathcal{B}_n := \left\{\sup_{y\leq\tau}|\hat{G}_n(y-)-G(y-)| > \frac{1-G(\tau)}{2}\right\}.$$

For the first term we have,

$$\begin{aligned}
P(\mathcal{A}_n) &\leq P\left(\left\{\sqrt{n}\sup_{y\leq\tau}\left|\hat{G}_n(y-)-G(y-)\right| > \frac{z}{32R^2}F_\tau^T(1-G(\tau))^2\right\}\right)\\
&\leq 2.5\exp\{-2\lambda_1^2(\tau)z^2 + C\lambda_1(\tau)z\},
\end{aligned}$$

with $\lambda_1(\tau) = F_\tau^T(1-F_\tau^T)(1-G(\tau))^2/(32R^2)$, where the last inequality follows from Theorem 2 of [Bitouzé et al., 1999]. The same theorem applied to the second term gives,

$$\begin{aligned}
P(\mathcal{B}_n) &\leq P\left(\sup_{y\leq\tau}|(1-F^T(y-))(\hat{G}_n(y-)-G(y-)| > \frac{(1-G(\tau))(1-F_\tau^T)}{2}\right)\\
&\leq 2.5\exp\{-\sqrt{n}(-2\tilde{\lambda}_1^2(\tau) + C\tilde{\lambda}_1(\tau))\},
\end{aligned}$$

where $\tilde{\lambda}_1(\tau) = (1 - G(\tau))(1 - F_\tau^T))/2$.

2. The second term $\mathbf{T_2}$. The estimation of the second term T_2 is similar to that of T_{12}. Indeed,

$$\begin{aligned} T_2 &= P\left(\sqrt{n}\sup_{q\in Q_N}\left|\frac{F_{\tau,n}^T - F_\tau^T}{F_{\tau,n}^T}\int f_q(y,x)\mathbb{1}_{y\le\tau}d\hat{P}_n\right| > \frac{z}{4}F_\tau^T\right) \\ &= P\left(\sqrt{n}\sup_{q\in Q_N}\left|\frac{F_{\tau,n}^T - F_\tau^T}{F_{\tau,n}^T}\int f_q(y,x)\mathbb{1}_{y\le\tau}\frac{\delta}{1-\hat{G}_n(y-)}dP_n\right| > \frac{z}{4}F_\tau^T\right) \\ &\le P\left(\sqrt{n}\left|\frac{F_{\tau,n}^T - F_\tau^T}{F_{\tau,n}^T}\right| > \frac{z}{32R^2}F_\tau^T(1-G(\tau))\right) \\ &\quad +2.5\exp\{-\sqrt{n}(-2\tilde{\lambda}_1^2(\tau) + C\tilde{\lambda}_1(\tau))\} \\ &\quad =: T_{21} + T_{22}. \end{aligned}$$

The first term can be decomposed as $T_{21} = P(\mathcal{A}_n^{'}) + P(\mathcal{B}_n^{'})$, where

$$\mathcal{A}_n^{'} = \left\{\sqrt{n}\left|\frac{F_{\tau,n}^T - F_\tau^T}{F_{\tau,n}^T}\right| > \frac{z}{32R^2}F_\tau^T(1-G(\tau))\right\} \cap \left\{|F_{\tau,n}^T - F_\tau^T| \le F_\tau^T/2\right\},$$

and

$$\mathcal{B}_n^{'} = \left\{|F_{\tau,n}^T - F_\tau^T)| > F_\tau^T/2\right\}.$$

Using again [Bitouzé et al., 1999] we obtain,

$$\begin{aligned} P(\mathcal{A}_n^{'}) &\le P\left(\sqrt{n}\sup_{y\le\tau}\left|(1-G(y-))(\hat{F}_n(y) - F(y))\right| > \frac{z}{64R^2}(F_\tau^T)^2(1-G(\tau))^2\right) \\ &\le 2.5\exp\{-2\lambda_2^2(\tau)z^2 + C\lambda_2(\tau)z\}, \end{aligned}$$

where $\lambda_2(\tau) = (F_\tau^T)^2(1-G(\tau))^2/(64R^2)$. Moreover,

$$P(\mathcal{B}_n^{'}) \le 2.5\exp\{-\sqrt{n}(-2\tilde{\lambda}_2^2(\tau) + C\tilde{\lambda}_2(\tau))\},$$

with $\tilde{\lambda}_2(\tau) = F_\tau^T(1-G(\tau-))/2$. Bringing together all the inequalities, we obtain the assertion of the theorem. □

References

[Bair and Tibshirani, 2004] Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*, 2(4):e108.

[Bartlett et al., 1998] Bartlett, P. L., Linder, T., and Lugosi, G. (1998). The minimax distortion redundancy in empirical quantizer design. *IEEE Trans. Inform. Theory*, 44(5):1802–1813.

[Bitouzé et al., 1999] Bitouzé, D., Laurent, B., and Massart, P. (1999). A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. *Ann. Inst. H. Poincaré Probab. Statist.*, 35(6):735–763.

[Einmahl and Mason, 2005] Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.*, 33(3):1380–1403.

[Fleming and Harrington, 1991] Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York.

[Gannoun et al., 2007] Gannoun, A., Saracco, J., and Yu, K. (2007). Comparison of kernel estimators of conditional distribution function and quantile regression under censoring. *Stat. Model.*, 7(4):329–344.

[Gersho and Gray, 1992] Gersho, A. and Gray, R. (1992). *Vector Quantization and Signal Compression*. Kluwer international series in engineering and computer science: Communications and information theory. Springer US.

[Graf and Luschgy, 1994] Graf, S. and Luschgy, H. (1994). *Consistent Estimation in the Quantization Problem for Random Vectors*. Angewandte Mathematik und Informatik. Univ.

[Grambsch et al.,] Grambsch, P. M., Dickson, E. R., Wiesner, R. H., and Langworthy, A. Application of the mayo primary biliary cirrhosis survival model to mayo liver transplant patients. *Mayo Clinic Proceedings*, 64(6):699–704.

[Heuchenne, 2008] Heuchenne, C. (2008). Strong uniform consistency results of the weighted average of conditional artificial data points. *J. Statist. Plann. Inference*, 138(5):1496–1515.

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

[Kaplan and Meier, 1958] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53:457–481.

[Krzanowski and Lai, 1988] Krzanowski, W. J. and Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 44(1):23–34.

[Linder, 2002] Linder, T. (2002). *Learning-theoretic methods in vector quantization*, volume 434 of *CISM Courses and Lectures*, pages 163–210. Springer-Verlag, Vienna.

[Linder et al., 1994] Linder, T., Lugosi, G., and Zeger, K. (1994). Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Trans. Inform. Theory*, 40(6):1728–1740.

[Lloyd, 1982] Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137.

[Lopez, 2009] Lopez, O. (2009). Single-index regression models with right-censored responses. *J. Statist. Plann. Inference*, 139(3):1082–1097.

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*, pages Vol. I: Statistics, pp. 281–297. Univ. California Press, Berkeley, Calif.

[Milligan and Cooper, 1985] Milligan, G. and Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

[Pollard, 1982a] Pollard, D. (1982a). A central limit theorem for k-means clustering. *Ann. Probab.*, 10(4):919–926.

[Pollard, 1982b] Pollard, D. (1982b). Quantization and the method of k-means. *IEEE Trans. Inform. Theory*, 28(2):199–205.

[Rachev and Rüschendorf, 1998] Rachev, S. T. and Rüschendorf, L. (1998). *Mass transportation problems. Vol. II.* Probability and its Applications (New York). Springer-Verlag, New York. Applications.

[Rand, 1971] Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

[Sánchez Sellero et al., 2005] Sánchez Sellero, C., González Manteiga, W., and Van Keilegom, I. (2005). Uniform representation of product-limit integrals with applications. *Scand. J. Statist.*, 32(4):563–581.

[Satten and Datta, 2001] Satten, G. A. and Datta, S. (2001). The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *Amer. Statist.*, 55(3):207–210.

[Stute, 1993] Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *J. Multivariate Anal.*, 45(1):89–103.

[Stute, 1996] Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. Statist.*, 23(4):461–471.

[Stute, 1999] Stute, W. (1999). Nonlinear censored regression. *Statist. Sinica*, 9(4):1089–1102.

[Stute and Wang, 1993] Stute, W. and Wang, J.-L. (1993). The strong law under random censorship. *Ann. Statist.*, 21(3):1591–1607.

[Talagrand, 1994] Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22(1):28–76.

[van der Vaart and Wellner, 1996] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

[Varadarajan, 1958] Varadarajan, V. S. (1958). Weak convergence of measures on separable metric spaces. *Sankhyā*, 19:15–22.