Journal of Multivariate Analysis 140 (2015) 220–233

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Vector quantization and clustering in the presence of censoring

Svetlana Gribkova *

Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie Paris VI, 4 place Jussieu, 75005 Paris, France
Cancer et génôme: Bioinformatique, biostatistiques et épidémiologie d'un système complexe, INSERM U900, Mines ParisTech, Institut Curie, 26 rue d'Ulm - 75248 Paris cedex 05, France[1]

ARTICLE INFO

Article history:
Received 4 August 2014
Available online 27 May 2015

AMS subject classifications:
62N01
62N02
62H30
62P10

Keywords:
Clustering
Quantization
Random censoring
k-means
Kaplan–Meier estimator

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 the 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.

1. Introduction

Vector quantization and *k*-clustering are two closely related issues. The former corresponds to a probabilistic problem of finding the optimal way to represent a distribution of random vector by a discrete distribution with a *k*-point support. Some general references on the subject are Gersho and Gray [7], Graf and Luschgy [8], Linder [14]. 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 separated as possible (see, for example, Lloyd [15] or MacQueen [17]). 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. Due to this specificity, the multivariate survival data cannot be analyzed by means of the standard clustering methods. For a complete introduction to survival analysis, we refer the reader to Fleming and Harrington [5].

To facilitate the discussion, we now set 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 the presence of censoring, instead of observing T directly, one observes a couple

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

* Correspondence to: Cancer et génôme: Bioinformatique, biostatistiques et épidémiologie d'un système complexe, INSERM U900, Mines ParisTech, Institut Curie, 26 rue d'Ulm - 75248 Paris cedex 05, France.
E-mail address: s.gribkova@mail.ru.
[1] Present address.

http://dx.doi.org/10.1016/j.jmva.2015.05.015
0047-259X/

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

$$(Y_i, \gamma_i, X_i)_{1\le i\le n} \tag{1}$$

of the random vector (Y, γ, X).

In the present article, we introduce a new optimal quantization procedure for such random vectors with one censored component. We then apply it in order to construct a new k-clustering algorithm which is valid in the presence of censored observations i.e. which is able to detect groups among n subjects with respect to their characteristics $(T_i, X_i)_{1\le i\le n}$, having at the input only their censored versions $(Y_i, \gamma_i, X_i)_{1\le i\le n}$. 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 clinicians to make their diagnostics more precise. For each patient, the available information includes 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 well adapted to the situation of unknown cancer subtypes. In this context, some of the existing approaches are discussed in Bair and Tibshirani [1]. In particular, it is mentioned that an approach by clustering with respect to X only 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 [1] propose to select among the components of X only the variables correlated with T by using Cox model 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 homogeneous. Then insurers have a possibility to fix prices 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. The 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 consistency 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 and 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 Borel measurable mapping $q : \mathbb{R}^{d+1} \to \mathcal{C}$, where $\mathcal{C} = \{c_1, \ldots, c_k\}$ is a collection of k distinct points of $\mathbb{R}^{d+1}$ called a codebook.

Let Q_k be the set of all k-point quantizers and suppose from now on that the following assumption is satisfied.

Assumption 1. Assume that $\mathbb{E}_P\|(T, X)\|^2 < \infty$.

The error (distortion) of an arbitrary k-point quantizer representing (T, X) by $q(T, X)$ may be defined by

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

where $\|\cdot\|$ is the Euclidean norm. 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 $D(P, q^*) = D_k^*(P)$, that is if it does the minimal possible error in representing (T, X) by $q(T, X)$. It was shown (see Linder [14]) that such a quantizer always exists. Moreover, any optimal quantizer belongs to the class of nearest neighbor quantizers, i.e. q^* is a mapping which associates each point (t, x) of $\mathbb{R}^{d+1}$ with the closest to (t, x) vector from the codebook, that is,

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