Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Conditional quantile estimation through optimal quantization

Isabelle Charlier[a,b,c], Davy Paindaveine[a,b,*], Jérôme Saracco[c]

[a] Université Libre de Bruxelles, Département de Mathématique, Boulevard du Triomphe, Campus Plaine, CP210, B-1050, Bruxelles, Belgium

[b] ECARES, 50 Avenue F.D. Roosevelt, CP114/04, B-1050, Bruxelles, Belgium

[c] Université de Bordeaux, Institut de Mathématiques de Bordeaux, UMR CNRS 5251 et INRIA Bordeaux Sud-Ouest, équipe CQFD, 351 Cours de la Libération, 33405 Talence, France

ARTICLE INFO

Article history:
Received 2 May 2014
Received in revised form 5 August 2014
Accepted 17 August 2014
Available online 27 August 2014

Keywords:
Nonparametric regression
Optimal quantization
Quantile regression

ABSTRACT

In this paper, we use quantization to construct a nonparametric estimator of conditional quantiles of a scalar response Y given a d-dimensional vector of covariates X. First we focus on the population level and show how optimal quantization of X, which consists in discretizing X by projecting it on an appropriate grid of N points, allows to approximate conditional quantiles of Y given X. We show that this approximation is arbitrarily good as N goes to infinity and provide a rate of convergence for the approximation error. Then we turn to the sample case and define an estimator of conditional quantiles based on quantization ideas. We prove that this estimator is consistent for its fixed-N population counterpart. The results are illustrated on a numerical example. Dominance of our estimators over local constant/linear ones and nearest neighbor ones is demonstrated through extensive simulations in the companion paper Charlier et al. (2014).

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

In numerous applications, one considers regression modeling to assess the impact of a d-dimensional vector of covariates X on a scalar response variable Y. It is then classical to consider the conditional mean and variance functions

$$x \mapsto \mathrm{E}[Y|X=x] \quad \text{and} \quad x \mapsto \mathrm{Var}[Y|X=x], \tag{1.1}$$

respectively. A much more thorough picture, however, is obtained by considering, for various $\alpha \in (0,1)$, the conditional quantile functions

$$x \mapsto q_\alpha(x) = \inf\big\{y \in \mathbb{R} : F(y|x) \geq \alpha\big\}, \tag{1.2}$$

where $F(\cdot|x)$ denotes the conditional distribution of Y given $X = x$. These conditional quantile functions completely characterize the conditional distribution of Y given X, whereas (1.1), in contrast, only measures the impact of X on Y's

[*] Corresponding author at: Université Libre de Bruxelles, Département de Mathématique, Boulevard du Triomphe, Campus Plaine, CP210, B-1050, Bruxelles, Belgium.

E-mail addresses: ischarli@ulb.ac.be (I. Charlier), dpaindav@ulb.ac.be (D. Paindaveine), Jerome.Saracco@math.u-bordeaux1.fr (J. Saracco).

http://dx.doi.org/10.1016/j.jspi.2014.08.003

0378-3758/© 2014 Elsevier B.V. All rights reserved.

location and scale, hence may completely miss to capture a possible impact of X on the shape of Y's distribution, for instance.

An important application of conditional quantiles is that they provide reference curves or surfaces (the graphs of $x \mapsto q_\alpha(x)$ for various α) and conditional prediction intervals (intervals of the form $I_\alpha(x) = [q_\alpha(x), q_{1-\alpha}(x)]$, for fixed x) that are widely used in many different areas. In medicine, reference growth curves for children's height and weight as a function of age are considered. Reference curves are also of high interest in economics (e.g., to study discrimination effects and trends in income inequality), in ecology (to observe how some covariates can affect limiting sustainable population size), and in lifetime analysis (to assess influence of risk factors on survival curves), among many others.

Quantile regression, that concerns the estimation of conditional quantile curves, was introduced in the seminal paper Koenker and Bassett (1978), where the focus was on linear regression. Since then, there has been much research on quantile regression, in particular in the nonparametric regression framework. Kernel and nearest-neighbor estimators of conditional quantiles were investigated in Bhattacharya and Gangopadhyay (1990), while Yu and Jones (1998) focused on local linear quantile regression and double-kernel approaches. Many other estimators were also considered; see, among others, Fan et al. (1994), Gannoun et al. (2002), Heagerty and Pepe (1999), or Yu et al. (2003). In this work, we introduce a new nonparametric regression quantile method, based on *optimal quantization*.

In probability theory, optimal quantization refers to the problem of finding the best approximation of a continuous d-dimensional probability distribution P by a discrete probability distribution charging a fixed number N of points. In other words, the d-dimensional random vector X needs to be approximated by a random vector $\widetilde{X}^N$ that may assume at most N values. Quantization was extensively investigated in (numerical) probability, finance, stochastic processes, and numerical integration (see, e.g., Zador (1964), Pagès (1998), Pagès et al. (2004a,b), and Bally et al. (2005)), but it was barely used in statistics—Sliced Inverse Regression (Azaïs et al., 2012) and clustering (Fischer, 2010, 2014) are the only statistical applications we are aware of. Yet, quantization is a natural tool in nonparametric quantile regression. In this context, indeed, quantization automatically takes care of the localization-in-x required in any nonparametric regression method. The resulting quantization-based estimators inherently are based on adaptive bandwidths, hence may dominate the local constant and local linear estimators from Yu and Jones (1998), that typically involve a unique global bandwidth. Quantization-based estimators also provide a refinement over nearest-neighbor estimators (such as those from Bhattacharya and Gangopadhyay (1990)) since, unlike the latter, the number of "neighbors" the former consider depends on the point x at which $q_\alpha(x)$ is to be estimated.

The outline of the paper, that mostly focuses on theoretical aspects, is as follows. Section 2 discusses quantization and provides some results on quantization, both of a theoretical and algorithmic nature. Section 3 describes how to approximate conditional quantiles through optimal quantization, which is achieved by replacing X in the definition of conditional quantiles by its L_p-optimal quantized version $\widetilde{X}^N$ (for some fixed N). The convergence rate of this approximation to the true conditional quantiles is obtained. Section 4 defines the corresponding estimator and proves its consistency (for the fixed-N approximated conditional quantiles). The results are illustrated on a numerical example, in which a smooth variant of the proposed estimator based on the bootstrap is also introduced. Section 5 provides some final comments. Eventually, the Appendix collects technical proofs.

2. Optimal quantization

In this section, we define the concept of L_p-norm optimal quantization and state the main results that will be used in the sequel (Section 2.1). Then we describe a stochastic algorithm that allows to perform optimal quantization (Section 2.2), and provide some convergence results for this algorithm (Section 2.3).

2.1. Definition and main results

Let X be a random d-vector defined on a probability space $(\Omega, \mathcal{F}, P)$, with distribution P_X, and fix a real number $p \geq 1$ such that $\mathrm{E}[|X|^p] < \infty$ (throughout, $|\cdot|$ denotes the Euclidean norm). Quantization replaces X with an appropriate random d-vector $\pi(X)$ that assumes at most N values. In optimal L_p-norm quantization, the vector $\pi(X)$ minimizes the L_p-norm quantization error

$$\|\pi(X) - X\|_p, \quad \text{with } \|Z\|_p := \big(\mathrm{E}\big[|Z|^p\big]\big)^{1/p}.$$

This optimization problem is equivalent to finding an N-grid of $\mathbb{R}^d$ — γ^N, say — such that the projection $\widetilde{X}^{\gamma^N} = \mathrm{Proj}_{\gamma^N}(X)$ of X on the (Euclidean-)nearest point of the grid minimizes the *quantization error* $\|\widetilde{X}^{\gamma^N} - X\|_p$. This definition leads to two natural questions: does such a minimum always exist? How does this minimum behave as N goes to infinity?

Existence (but not unicity) of an optimal N-grid – that is, a grid minimizing this quantization error – has been obtained under the assumption that P_X does not charge any hyperplane; see Pagès (1998). Irrespective of the sequence of optimal grids considered, $\widetilde{X}^N$ converges to X in L_p. This is a direct corollary of the following result, which is often referred to as Zador's theorem (Zador, 1964) and provides the rate of convergence of the quantization error; see, e.g., Graf and Luschgy (2000) for a proof.

Download English Version:

https://daneshyari.com/en/article/1147717

Download Persian Version:

https://daneshyari.com/article/1147717

Daneshyari.com