Prediction by quantization of a conditional distribution

Jean-Michel Loubes, Bruno Pelletier

Prediction by quantization of a conditional distribution

Jean-Michel Loubes* Bruno Pelletier†

February 17, 2017

Abstract. Given a pair of random vectors (X, Y), we consider the problem of approximating Y by $\mathbf{c}(X) = \{\mathbf{c}_1(X), \ldots, \mathbf{c}_M(X)\}$ where $\mathbf{c}$ is a measurable set-valued function. We give meaning to the approximation by using the principles of vector quantization which leads to the definition of a multifunction regression problem. The formulated problem amounts at quantizing the conditional distributions of Y given X. We propose a nonparametric estimate of the solutions of the multifunction regression problem by combining the method of M-means clustering with the nonparametric smoothing technique of k-nearest neighbors. We provide an asymptotic analysis of the estimate and we derive a convergence rate for the excess risk of the estimate. The proposed methodology is illustrated on simulated examples and on a speed-flow traffic data set emanating from the context of road traffic forecasting.

Index Terms: Regression analysis, vector quantization, nonparametric statistics, clustering, k-means, set-valued function, multifunction.

1 Introduction

Regression analysis encompasses important statistical methods for exploring the relationship between a response variable Y and a predictor X. Most commonly, the focus is on estimating (or modeling) the regression function $\eta(x) := \mathbb{E}[Y|X = x]$ by methods of various sorts (see e.g. Györfi et al., 2002; Ruppert et al., 2003). Over the years, alternatives to mean regression (that is, estimation of the regression function η) have been proposed and analyzed in the literature. Among these, median regression (as a special case of quantile regression Koenker, 2005) exhibits properties of robustness to outliers (Huber and Ronchetti, 2009). Another alternative is modal regression. In Lee (1989, 1993) (see also Kemp and Santos Silva, 2012; Yao and Li, 2014) the mode of the conditional distribution of Y given X is modeled as a linear function of x. A related setting is that considered in Sager and Thisted (1982) where the dependence of the conditional mode on the predictor x is monotone. Typical nonparametric approaches to conditional mode estimation resort to first estimating the conditional densities using a nonparametric method, and then to infer the mode by maximization, as in Collomb et al. (1987) for instance, and Yao et al. (2012) for a generalization of this approach using local polynomials.

Yet in the situation where the data is heterogeneous, summarizing the conditional distribution of Y given X by a single measure of location (mean, median, or mode) may be inadequate. As an illustration, consider the scatterplot represented in Figure 1. The distribution of Y given X is a mixture of two Normal distributions with equal proportions, equal variances, and means $\eta_1(x) < \eta_2(x)$, and X follows a uniform distribution over the unit interval. The difference in means $\eta_2(x) -$

*Institut de Mathématiques de Toulouse, Université Toulouse III, France
†Département de Mathématiques, IRMAR – UMR CNRS 6625, Université Rennes II, France

$\eta_1(x)$ increases with x so that the conditional distribution of Y given X is clustered into two distinct groups, all the more separate as x is large. By construction, $\eta(x) = \frac{1}{2}[\eta_1(x) + \eta_2(x)]$ and η is an increasing function of x. Thus the regression function is well representative of the average trend in the data but provides a limited summary of the distribution of Y given X since it is bimodal. Instead, the *set-valued map*, also referred to as a *multi-valued function* or a *multifunction*, defined by $x \mapsto \{\eta_1(x), \eta_2(x)\}$ would better capture the structure of the data than a real-valued map such as the conditional mean (or mode or median) function.

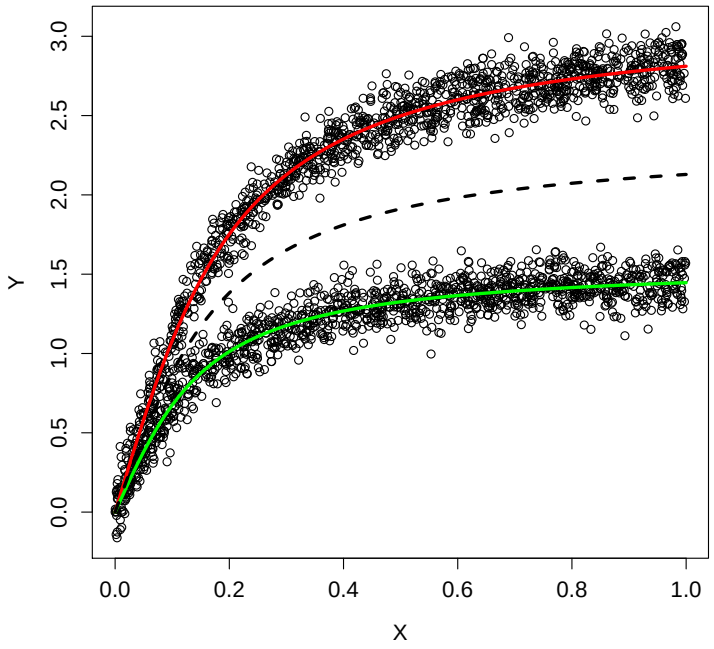

Figure 1: Scatterplot of 400 realizations of the pair (X, Y) where X follows a uniform distribution over $[0, 1]$, and where the distribution of Y given X is a mixture of two Normal distribution with weights equal to $\frac{1}{2}$, variances equal to 0.01, and means $\eta_1(x) = \arctan(8x)$ (green curve) and $\eta_2(x) = 2\arctan(6x)$ (red curve). The regression function $\mathbb{E}[Y|X = x] = \frac{1}{2}[\eta_1(x) + \eta_2(x)]$ is represented by the dashed curve.

Fitting a *finite mixture model* is a popular approach for modeling such heterogeneous data. These models are typically studied in an estimation framework (see e.g. Everitt and Hand, 1981; McLachlan and Peel, 2000; Titterington et al., 1985) where an application of the maximum likelihood principle defines the estimation method of choice. For purposes of regression analysis, a *finite mixture regression model* is obtained by conditioning a finite mixture distribution on a vector of covariates, as in Khalili and Chen (2007). For instance, the data represented in the scatterplot of Figure 1 is drawn from a Gaussian mixture regression model with two components, and the interest would be primarily in estimating the mean curves $\eta_1(x)$ and $\eta_2(x)$, in addition to the mixture proportions and the variances of the components of the model. Finite mixture regression models provide a flexible way of handling heterogeneous data and are receiving a growing attention from the statistical community, with recent results giving performance bounds even in a high-dimensional setting (see e.g. Devijver, 2015; Meynet, 2013; Städler et al., 2010). These models are also known

as mixture of experts (Jacobs et al., 1991; Jiang and Tanner, 1999) in machine learning.

Another class of methods centered on *modal regression* has developed to estimate the set of all the modes of the conditional distribution, that is, the set of points of local maximum of the conditional density, called the modal set. Following Einbeck and Tutz (2006), Chen et al. (2016) propose a plug-in nonparametric estimate of the modal set based on a kernel density estimate. This contrasts with early works on modal regression (e.g. Lee, 1989, 1993; Sager and Thisted, 1982) which focused primarily on estimating the principal mode motivated by concerns of robustness to outliers. Formally, Chen et al. (2016) defines the modal set $\mathrm{Mod}(x)$ at some point x as the set of points y in $\mathbb{R}$ where the conditional density $f_{Y|X=x}(y)$ of Y given $X = x$ satisfies $f'_{Y|X=x}(y) = 0$ and $f''_{Y|X=x}(y) < 0$ and it is assumed that $\mathrm{Mod}(x)$ is finite (we note that, as defined, $\mathrm{Mod}(x)$ is a subset of the set of points of local maximum). Hence the map $x \mapsto \mathrm{Mod}(x)$ is a multivalued function. From an algorithmic standpoint, Einbeck and Tutz (2006) propose to estimate the modal set with a conditional version of the mean-shift algorithm, which is a modified version of the mean-shift algorithm used in the context of density mode clustering (Arias-Castro et al., 2016; Cheng, 1995; Comaniciu and Meer, 2002), and Chen et al. (2016) prove that the resulting modal set estimate is consistent.

Arguably, nonparametric modal regression may prove effective especially when the conditional distributions admit only a limited number of local modes, as in the speed-flow traffic data reported in Einbeck and Tutz (2006). There, the conditional distribution of the speed of vehicules on a Californian freeway given the traffic flow is found to be bimodal over a range of small flow values, and then unimodal for larger values of the flow. In this example, the modal set is composed of at most two points. However in a situation where the conditional distributions would admit a large number of local points of maximum, then the modal set might be difficult to interpret (the modal set may even be uncountable, when this latter is defined as the set of points of local maximum). Therefore, there is a need for developing a regression methodology which could extract potentially more than one feature from the data, in a manner similar to modal regression, but while keeping their number relatively small, or even the control thereof, to preserve the interpretability of these features.

In this paper, we propose to apply the principles of vector quantization (Gersho and Gray, 1992; Graf and Luschgy, 2000; Linder, 2002) to the conditional distributions of Y given X in order to define, given M an integer, a set-valued function $x \mapsto \mathbf{c}(x) := \{c_1(x), \ldots, c_M(x)\}$ meant to capture the underlying structure of the data, thereby extending the regression problem to a *multifunction regression problem*. We define optimality in terms of the *mean squared error* or *predictive risk*

$$\mathcal{E}(\mathbf{c}) = \mathbb{E}\left[\min_{1 \le j \le M} \|Y - c_j(X)\|^2\right] \tag{1}$$

and we focus on the estimation of a multifunction $\mathbf{c}^\star$ which achieves the infimum of $\mathcal{E}(\mathbf{c})$ over the set of measurable multifunctions of the form $x \mapsto \mathbf{c}(x) := \{c_1(x), \ldots, c_M(x)\}$. We remark that when $M = 1$, $\mathcal{E}(\mathbf{c})$ coincides with the L_2 risk $\mathbb{E}[|Y - f(X)|^2]$ of a real-valued function f used in regression analysis. Hence $\mathcal{E}(\mathbf{c})$ is a natural extension of the L_2 risk to multifunctions. We emphasize that, even when $M = 1$, our objective is not to quantize the regression function η, a problem studied in Györfi and Wegkamp (2008), but to estimate an optimal multifunction $\mathbf{c}^\star$, which can be represented by the M real-valued functions $x \mapsto c_j(x)$, for $j = 1, \ldots, M$.

Given IID data $(X_1, Y_1), \ldots, (X_n, Y_n)$ with the same distribution as (X, Y), we propose a nonparametric estimate $\hat{\mathbf{c}}_n(x) := \{\hat{c}_{n,1}(x), \ldots, \hat{c}_{n,M}(x)\}$ defined by combining the approach of k-means clustering (see e.g. Duda et al., 2000, Chap. 10) with the smoothing technique of k-nearest neighbors averaging (see e.g. Györfi et al., 2002). Underlying the definition of $\hat{\mathbf{c}}_n$ is the estimation of

$\mathcal{E}(\mathbf{c})$ with k-nearest neighbors smoothing, followed by minimization of this estimate over the set of measurable multifunctions $x \mapsto \mathbf{c}(x) := \{c_1(x), \ldots, c_M(x)\}$. As will be argued further in the paper, the minimization problem over $\mathbf{c}$ can be reduced to a collection of quantization problems indexed by x, which leads to a simple algorithm for evaluating the value of the estimate $\hat{\mathbf{c}}_n(x)$ at any x. We measure the performance of the estimate by the *excess risk*

$$\mathcal{R}(\hat{\mathbf{c}}_n) = \mathbb{E}\left[\min_{1\leq j\leq M} \|Y - \hat{c}_{n,j}(X)\|^2\right] - \inf_{\mathbf{c}} \mathbb{E}\left[\min_{1\leq j\leq M} \|Y - c_j(X)\|^2\right]. \tag{2}$$

Notice that when $M = 1$, $\hat{\mathbf{c}}_n$ reduces to a single valued function $\hat{c}_{n,1} : \mathbf{R} \to \mathbb{R}$ and $\mathcal{R}(\hat{\mathbf{c}}_n) = \mathbb{E}[(\hat{c}_{n,1}(X) - \eta(X)^2]$, the expectaction of the L_2 error of the estimate $\hat{c}_{n,1}$ of the regression function η. To summarize, in the present paper, we make in particular the following contributions:

- We state a multifunction regression problem and we study its solutions.
- We propose a nonparametric multifunction estimate $\hat{\mathbf{c}}_n$ defined by combining the method of M-means with the smoothing technique of k-nearest neighbors averaging. We propose a simple companion algorithm to compute the value of the estimate.
- We prove the consistency of the estimate and we derive convergence rates on the excess risk and on a pointwise version of the excess risk.
- We propose a heuristic for automatically selecting the number of neighbours of the estimate. We also study the automatic selection of the number of quantization points. We illustrate the methods on two simulated examples and on a data set of speed records versus the location along an automobile path in the city of Toulouse, France.

The paper is organized as follows. In section 2, we summarize the foundational principles of vector quantization and of the design of empirical vector quantizors by minimization of the empirical risk. In section 3, we define the multifunction regression problem, emphasizing potential measurability issues that we address. In section 4, we define our proposed estimate $\hat{\mathbf{c}}_n$, and in section 5, we provide an asymptotic analysis of the estimate. First, in Theorem 1, we obtain a bound on the pointwise version of the excess risk, under a local regularity condition on the conditional densities. Then we derive a convergence rate on the excess risk in Theorem 2 under mild regularity conditions. In Theorem 3, we prove a convergence result of the estimte towards elements of the solution set of the multifunction regression problem. Then in section 6, we report on practical implementation details on numerical experiments and we also apply the methodology to speed data along an automobile path in the city of Toulouse, France. The proofs of the theorems are exposed in section 7.

2 Vector quantization

In this section, we collect foundational materials on vector quantization. We start by formulating the quantization problem and by defining the notion of an optimal quantizer. Then we describe the application of the principle of empirical risk minimization to the design of an empirical quantizer from IID data.

2.1 The quantization problem

Vector quantization refers to the process of discretizing a random vector by a random variable that can take only a finite number of values (Gersho and Gray, 1992; Graf and Luschgy, 2000;

Linder, 2002). Known as lossy data compression in information theory and signal processing, vector quantization forms the basic principle of the method of k-means for data clustering (Pollard, 1982b) and is also used in defining numerical integration schemes (Pagès, 1997). In this section, and we collect foundational materials on vector quantization.

Let Y be a random vector in $\mathbb{R}^p$ with distribution P_Y. Given M an integer, an M-points *quantizer* is a map $q : \mathbb{R}^p \to \mathbb{R}^p$ such that its image is a finite set $\{c_1, \ldots, c_M\}$ of M points of $\mathbb{R}^p$. Using the Euclidean norm $\|.\|$ on $\mathbb{R}^p$, the performance of a quantizer q is measured by the *distortion*

$$\mathcal{D}(q; P_Y) = \mathbb{E}[\|Y - q(Y)\|^2]. \tag{3}$$

An M-points nearest-neighbor quantizer is a quantizer $q_{\mathbf{c}}$ of the form $q_{\mathbf{c}}(x) = \arg\min_{1 \le j \le M} \|x - c_j\|$, where ties are broken arbitrarily, and where $\mathbf{c} := (c_1, \ldots, c_M)$ is a configuration, or *codebook*, of M points in $\mathbb{R}^p$. Any quantizer q defines a partition of $\mathbb{R}^p$ into the sets $q^{-1}(c_j)$, for $j = 1, \ldots, M$. In the case of a nearest-neighbor quantizer $q_{\mathbf{c}}$, the partition is called a Voronoi partition and for any $j = 1, \ldots, M$, the (closed) Voronoi cell $V_j(\mathbf{c})$ associated with c_j is defined by

$$V_j(\mathbf{c}) = \{x \in \mathbb{R}^p \, : \, \|x - c_j\| \le \|x - c_\ell\| \text{ for all } 1 \le \ell \le M\}. \tag{4}$$

Notice that $\{V_1(\mathbf{c}), \ldots, V_M(\mathbf{c})\}$ does not form a partition of $\mathbb{R}^p$ because $V_i(\mathbf{c}) \cap V_j(\mathbf{c})$ is not empty for all $1 \le i \ne j \le M$, but $q_{\mathbf{c}}^{-1}(c_j) \subset V_j(\mathbf{c})$ for all $j = 1, \ldots, M$.

2.2 Optimal quantizers

The search for an optimal quantizer minimizing the distortion can be restricted to the class of nearest-neighbor quantizers (Graf and Luschgy, 2000, Lemma 3.1). In the present work, only nearest-neighbor quantizers are considered, and a nearest-neighbor quantizer $q_{\mathbf{c}}$ is referred to by the configuration $\mathbf{c} := (c_1, \ldots, c_M)$ from which it is defined. A configuration $\mathbf{c} := (c_1, \ldots, c_M)$ will be called simply a quantizer and the distortion $\mathcal{E}(\mathbf{c}; P_Y)$ of the quantizer $\mathbf{c}$ is defined by

$$\mathcal{E}(\mathbf{c}; P_Y) := \mathcal{D}(q_{\mathbf{c}}; P_Y) = \mathbb{E}\left[\min_{1 \le j \le M} \|Y - c_j\|^2\right]. \tag{5}$$

An optimal quantizer $\mathbf{c}^\star$ is any minimizer of $\mathcal{E}(\mathbf{c}; P_Y)$ over all $\mathbf{c}$ in $(\mathbb{R}^p)^M$, that is, such that $\mathcal{E}(\mathbf{c}^\star; P_Y) = \mathcal{E}^\star(P_Y)$, where

$$\mathcal{E}^\star(P_Y) = \inf_{\mathbf{c} \in (\mathbb{R}^p)^M} \mathcal{E}(\mathbf{c}; P_Y), \tag{6}$$

and its existence is guaranteed; see e.g. Theorem 1 in Linder (2002) or Theorem 4.12 in Graf and Luschgy (2000).

2.3 Approximation of measures

The connection between vector quantization and the Wasserstein distance has long been recognized; in particular, we have

$$\mathcal{E}^\star(P_Y) = \inf\{W_2(P_Y, Q) \, : \, Q \text{ probability measure with } |\mathrm{Supp}(Q)| \le M\}, \tag{7}$$

where $W_2(P_Y, Q)$ denotes the L_2 Wasserstein distance between the probability measures P_Y and Q (Graf and Luschgy, 2000, Lemma 3.4). Hence finding an optimal quantizer for P_Y is equivalent to best approximating P_Y, in the Wasserstein distance, by a discrete measure with support of cardinality at most M. Under the regularity assumption that, for any optimal quantizer $\mathbf{c}^\star$, P_Y does

not charge the boundaries common to any two adjacent Voronoi cells, that is, if $P_Y(V_i(\mathbf{c}^\star) \cap V_j(\mathbf{c}^\star)) = 0$ for all $1 \leq i \neq j \leq M$, then the set of minimizers in (7) coincides with the set of optimal quantizers minimizing (5) by (Graf and Luschgy, 2000, Lemma 3.1) and (Graf and Luschgy, 2000, Lemma 4.4). Following Graf and Luschgy (2000), any minimizer of (7) is called an M-optimal quantizing measure. By this equivalence, any M-optimal quantizing measure is of the form $P_Y \circ q_{\mathbf{c}^\star}^{-1}$, that is, the image measure (pushforward measure) of P_Y by the quantizer map $q_{\mathbf{c}^\star}$, and it can be expressed as $P_Y \circ q_{\mathbf{c}^\star}^{-1} = \sum_{i=1}^{M} \mathbb{P}(Y \in V_i(\mathbf{c}^\star))\delta_{c_i^\star}$, where $\mathbf{c}^\star = (c_1^\star, \ldots, c_M^\star)$.

2.4 Empirical vector quantization

Empirical vector quantization refers to the quantization of the empirical measure of a random sample and forms the basis for data clustering by the method of k-means (Pollard, 1982b), where the goal is to automatically partition the data into dissimilar groups of similar items. The setting is that of a sequence $(Y_i)_{i \geq 1}$ of independent random vectors with the same distribution as Y. For each sample size n, denote by $P_Y^{(n)} := \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$ the empirical measure associated with $Y_1, \ldots, Y_n$. An empirical quantizer $\mathbf{c}_n^\star$ is any minimizer of the distortion for $P_Y^{(n)}$, that is, such that $\mathcal{E}_n(\mathbf{c}_n^\star; P_Y) = \mathcal{E}_n^\star := \inf_{\mathbf{c} \in (\mathbb{R}^p)^M} \mathcal{E}_n(\mathbf{c}; P_Y)$ where

$$\mathcal{E}_n(\mathbf{c}; P_Y) := \mathcal{E}\left(\mathbf{c}; P_Y^{(n)}\right) = \frac{1}{n}\sum_{i=1}^{n} \min_{1 \leq j \leq M} \|Y_i - c_j\|^2, \tag{8}$$

with $\mathcal{E}$ as in (5). Consistency of $\mathbf{c}_n^\star$ is shown in (Pollard, 1981, 1982b). It is shown in (Antos, 2005; Bartlett et al., 1998; Linder et al., 1994) that the excess risk $\mathbb{E}[\mathcal{E}(\mathbf{c}_n^\star; P_Y)] - \mathcal{E}^\star(P_Y)$ of an empirical quantizer decreases at a rate on the order of $\mathcal{O}(1/\sqrt{n})$ under the assumption that P_Y has bounded support. This result is extended in Biau et al. (2008) for the quantization over a separable Hilbert space. Faster convergence rates have been reported in the literature under different kind of assumptions (see e.g. Antos et al., 2005; Levrard, 2015). We mention that these rates share the property of depending only on the sample size, and not on the number of quantization points nor on the space dimension. The dependence on these parameters is only through the constant factors.

3 The multifunction regression problem

Let (X, Y) be a pair of random vectors taking values in $\mathbb{R}^d \times \mathbb{R}^p$. Following Rockafellar and Wets (2009), a *set-valued mapping* or *multifunction* $\mathbf{c} : \mathbb{R}^d \rightrightarrows \mathbb{R}^p$ is a map which to each x in $\mathbb{R}^d$ associates a subset $\mathbf{c}(x)$ of $\mathbb{R}^p$. The double arrow notation is used to distinguish multifunctions from single-valued functions and the Euclidean spaces under consideration are endowed with their Borel σ-fields. A multifunction $\mathbf{c} : \mathbb{R}^d \rightrightarrows \mathbb{R}^p$ is *closed-valued* if $\mathbf{c}(x)$ is closed for each x; it is *measurable* if for every open set $\mathcal{O} \subset \mathbb{R}^p$ the set $\mathbf{c}^{-1}(\mathcal{O})$ is measurable (Rockafellar and Wets, 2009, Definition 14.1).

Given M an integer, we consider the set $\mathcal{F}_M$ of measurable multifunctions $\mathbf{c} : \mathbb{R}^d \rightrightarrows \mathbb{R}^p$ such that $\mathbf{c}(x)$ contains exactly M points of $\mathbb{R}^p$ for each x, that is,

$$\mathcal{F}_M = \left\{ \mathbf{c} : \mathbb{R}^d \rightrightarrows \mathbb{R}^p \quad : \quad \mathbf{c} \text{ is measurable and } \#\mathbf{c}(x) = M \text{ for each } x \right\}, \tag{9}$$

where $\#$ denotes the cardinality of a set.
Notice that each multifunction in $\mathcal{F}_M$ is closed-valued. By (Rockafellar and Wets, 2009, Theorem 14.5), each closed-valued measurable multifunction admits a Castaing representation, meaning in our context that, for each $\mathbf{c}$ in $\mathcal{F}_M$, there exists M measurable functions $c_1, \ldots, c_M$ from $\mathbb{R}^d$ to $\mathbb{R}^p$

such that $\mathbf{c}(x) = (c_1(x), \ldots, c_M(x))$ for all x. We define the multifunction regression problem as the problem of best approximating Y by $\mathbf{c}(X)$, for some $\mathbf{c}$ in $\mathcal{F}_M$, in the sense of the predictive risk $\mathcal{E}(\mathbf{c})$ defined in (1) as

$$\mathcal{E}(\mathbf{c}) = \mathbb{E}\left[\min_{1\leq j\leq M} \|Y - c_j(X)\|^2\right]. \tag{10}$$

Notice that $\mathcal{E}(\mathbf{c})$ does not depend on the choice of functions $(c_1, \ldots, c_M)$ used to represent $\mathbf{c}$. Then we define a solution to the multifunction regression problem as any multifunction $\mathbf{c}^\star$ in $\mathcal{F}_M$ such that

$$\mathcal{E}(\mathbf{c}^\star) = \inf_{\mathbf{c}\in\mathcal{F}_M} \mathcal{E}(\mathbf{c}).$$

A notable difference with the conventional regression setting (corresponding to $M = 1$) is that the solution set typically contains multiple solutions when $M \geq 2$, while when $M = 1$, the solution is unique and coincides with the regression function $\eta(x) = \mathbb{E}[Y|X = x]$.

As claimed in the Introduction, minimization over $\mathbf{c}$ can be reduced to a collection of quantization problems indexed by x. This is true in the following sense. Denote by P_X the distribution of X and by $\mathcal{S}_X$ its support. By conditioning on X in the definition of $\mathcal{E}(\mathbf{c})$, let $\mathcal{E}(\mathbf{c}; x)$ be the function defined for any x in $\mathcal{S}_X$ and any $\mathbf{c} = (c_1, \ldots, c_M) \in \mathbb{R}^p \times \cdots \times \mathbb{R}^p$ by

$$\mathcal{E}(\mathbf{c}; x) = \mathbb{E}\left[\min_{1\leq j\leq M} \|Y - c_j\|^2 \big| X = x\right], \tag{11}$$

which is also equal to the conditional version of (5) (we use the same notation $\mathbf{c}$ to denote either a multifunction $\mathbb{R}^d \rightrightarrows \mathbb{R}^p$ or a point in $\mathbb{R}^p \times \cdots \times \mathbb{R}^p$ when this is clear from the context and there is no risk of confusion). Thus $\mathcal{E}(\mathbf{c}; x)$ corresponds to the distortion of the conditional distribution of Y given X at x by the M-points quantizer with codebook $(c_1, \ldots, c_M)$. It is clear that $\mathcal{E}(\mathbf{c}; x)$ is measurable in x for each $\mathbf{c}$ and continuous in $\mathbf{c}$ for each x, which imply that $\mathcal{E}(\mathbf{c}; x)$ is a Carathéodory integrand (these are the defining conditions). Therefore $\mathcal{E}(\mathbf{c}; x)$ is a normal integrand in the sense of (Rockafellar and Wets, 2009, Definition 14.27) and by (Rockafellar and Wets, 2009, Theorem 14.60) on the interchange of minimization and integration, we have

$$\inf_{\mathbf{c}\in\mathcal{F}_M} \int_{\mathcal{S}_X} \mathcal{E}(\mathbf{c}(x); x) P_X(dx) = \int_{\mathcal{S}_X} \left[\inf_{\mathbf{c}\in\mathbb{R}^p\times\cdots\times\mathbb{R}^p} \mathcal{E}(\mathbf{c}; x)\right] P_X(dx), \tag{12}$$

and for any $\bar{\mathbf{c}}$ in $\mathcal{F}_M$, the following equivalence holds:

$$\bar{\mathbf{c}} \in \operatorname*{arg\,min}_{\mathbf{c}\in\mathcal{F}_M} \int_{\mathcal{S}_X} \mathcal{E}(\mathbf{c}(x); x) P_X(dx) \iff \bar{\mathbf{c}}(x) \in \operatorname*{arg\,min}_{\mathbf{c}\in\mathbb{R}^p\times\cdots\times\mathbb{R}^p} \mathcal{E}(\mathbf{c}; x) \text{ for } P_X\text{-a.e. } x.$$

Thus minimizing $\mathcal{E}(\mathbf{c})$ over $\mathcal{F}_M$ is equivalent to minimizing $\mathcal{E}(\mathbf{c}; x)$ over $\mathbf{c} \in \mathbb{R}^p \times \cdots \times \mathbb{R}^p$, up to measurability. This issue can be resolved by considering a measurable selection from the argmin sets as follows: since $\mathcal{E}(\mathbf{c}, x)$ is a normal integrand, the multifunction $\mathcal{S}_X \rightrightarrows \mathbb{R}^p$ defined by $x \mapsto \arg\min_{\mathbf{c}\in\mathbb{R}^p\times\cdots\times\mathbb{R}^p} \mathcal{E}(\mathbf{c}; x)$ is closed-valued and measurable by (Rockafellar and Wets, 2009, Theorem 14.37), and so it admits a measurable selection (Rockafellar and Wets, 2009, Corollary 14.6), that is, a measurable function $\bar{\mathbf{c}} : \mathcal{S}_X \to \mathbb{R}^p \times \cdots \times \mathbb{R}^p$ such that $\bar{\mathbf{c}}(x) \in \arg\min_{\mathbf{c}\in\mathbb{R}^p\times\cdots\times\mathbb{R}^p} \mathcal{E}(\mathbf{c}; x)$ for all x in $\mathcal{S}_X$ ($\bar{\mathbf{c}}$ is cannonically identified with a multifunction in $\mathcal{F}_M$).

We conclude that any solution $\mathbf{c}^\star$ to the multifunction regression problem can be defined either directly as a minimizer of $\mathcal{E}(\mathbf{c})$ over $\mathcal{F}_M$ or as a measurable selection (which exists) from the collection of argmin sets $\{\arg\min_{\mathbf{c}\in\mathbb{R}^p\times\cdots\times\mathbb{R}^p} \mathcal{E}(\mathbf{c}; x) \,:\, x \in \mathcal{S}_X\}$.

4 The estimate

In this section, we define our estimate $\hat{\mathbf{c}}_n(x) := (\hat{c}_{n,1}(x), \ldots, \hat{c}_{n,M}(x))$. We also describe an optimization algorithm to compute the values of $\hat{\mathbf{c}}_n(x)$ at any point x.

4.1 Definition of the estimate

Let $(X_i, Y_i)_{i\geq 1}$ be an IID sequence of random vectors with the same distribution as (X, Y). To define $\hat{\mathbf{c}}_n(x)$, we proceed by first estimating $\mathcal{E}(\mathbf{c}; x)$ and next by minimizing the estimated distortion over $\mathbf{c}$ for each x. Clearly there is ample leeway for the first step, and for computational reasons exposed in section 4.2, we consider a k-nearest neighbors local averaging estimate of $\mathcal{E}(\mathbf{c}; x)$ of the form

$$\mathcal{E}_n(\mathbf{c}; x) = \sum_{i=1}^{n} W_{n,i}(x) \min_{1\leq j\leq M} \|Y_i - c_j\|^2, \tag{13}$$

where $\{W_{n,i}(x), i = 1, \ldots, n\}$ is the set of weights depending on the observations $X_1, \ldots, X_n$ defined as

$$W_{n,i}(x) = \frac{1}{k}\mathbf{1}\left\{X_i \text{ is among the } k \text{ nearest neighbors of } x\right\}. \tag{14}$$

As for $\mathcal{E}(\mathbf{c}; x)$ defined in (10), $\mathcal{E}_n(\mathbf{c}; x)$ is a Carathéodory integrand (hence a normal integrand) and so there exists a measurable selection from its argmin sets. Then we define our estimate $\hat{\mathbf{c}}_n$ as any measurable selection from the collection $\{\arg\min_{\mathbf{c}\in\mathbb{R}^p\times\cdots\times\mathbb{R}^p} \mathcal{E}_n(\mathbf{c}; x) \,:\, x \in \mathcal{S}_X\}$, meaning that $\hat{\mathbf{c}}_n$ is measurable and satisfies

$$\mathcal{E}_n(\hat{\mathbf{c}}_n(x); x) = \inf_{\mathbf{c}\in\mathbb{R}^p\times\cdots\times\mathbb{R}^p} \mathcal{E}_n(\mathbf{c}; x), \quad \text{for all } x \text{ in } \mathcal{S}_X. \tag{15}$$

4.2 An optimization algorithm

Minimizing $\mathcal{E}_n(\mathbf{c}; x)$, or $\mathcal{E}_n(\mathbf{c}; P_Y)$ in the non conditional setting, is known for being computationally difficult (it is NP-hard). A popular and tractable optimization algorithm for this purpose is the k-means algorithm, which proceeds iteratively by constructing a sequence of quantizers converging to a local optimum.

From a practical perspective, the local averaging estimate $\mathcal{E}_n(\mathbf{c}; x)$ defined in (13) can be minimized by considering a weighted version of the k-means algorithm, as described in Algorithm 1. Naturally, weights other than the k-nearest neighbors weights could be used. But when using the k-nearest neighbor weights (14), the algorithm is equivalent to the standard M-means algorithm

---

Algorithm 1: Conditional weighted k-means algorithm.

Input: Data $(X_1, Y_1), \ldots, (X_n, Y_n)$, weights $W_{n,i}(x)$, $i = 1, \ldots, n$, and number of quantization points M.

1. Initialize a configuration $\mathbf{c}^{(0)} = (c_1^{(0)}, \ldots, c_M^{(0)})$.
2. Iterate for $t \geq 0$ over:
   (a) *Assignment step*: Set $I_j^{(t)} = \{1 \leq i \leq n \,:\, \|Y_i - c_j\| \leq \|Y_i - c_\ell\| \text{ for all } 1 \leq \ell \leq M\}$, for each $1 \leq j \leq M$,
   (b) *Update step*: Set $c_j^{(t+1)} = \frac{\sum_{i\in I_j^{(t)}} W_{n,i}(x) Y_i}{\sum_{i\in I_j^{(t)}} W_{n,i}(x)}$.

Output: Configuration $\mathbf{c} = (c_1, \ldots, c_M)$ obtained at convergence.

---

applied to the Y_i's which correspond to the k nearest neighbors of x among the X_i's. Thus, the algorithm is rather simple to implement.

5 Asymptotic analysis

In this section, we study the convergence of the estimate $\hat{\mathbf{c}}_n$ defined in (15). We assume that (X, Y) admits a probability density f_{XY} with respect to the Lebesgue measure on $\mathbb{R}^d \times \mathbb{R}^p$ and that Y is bounded, that is, that there is $R > 0$ such that $\|Y\| \leq R$ almost surely.

We shall need the following notation. The conditional density of Y given $X = x$ at y is denoted by $f_{Y|X=x}(y)$. For any m, the closed ball of $\mathbb{R}^m$ centered at x and of radius ρ is denoted by $\mathcal{B}_m(x, \rho)$. The closed ball centered at the origin and of radius ρ is denoted by $\mathcal{B}_m(\rho)$ and the volume of $\mathcal{B}_m(1)$ is denoted by ω_m.

Recall the *excess risk* defined in (2) by

$$\mathcal{R}(\hat{\mathbf{c}}_n) = \mathbb{E}\left[\min_{1\leq j\leq M} \|Y - \hat{c}_{n,j}(X)\|^2\right] - \inf_{\mathbf{c}\in\mathcal{F}_M} \mathbb{E}\left[\min_{1\leq j\leq M} \|Y - c_j(X)\|^2\right], \tag{16}$$

where the class $\mathcal{F}_M$ is defined in (9). We also consider the *pointwise excess risk* defined by

$$\mathcal{R}(\hat{\mathbf{c}}_n; x) = \mathbb{E}\left[\min_{1\leq j\leq M} \|Y - \hat{c}_{n,j}(X)\|^2 | X = x\right] - \inf_{\mathbf{c}\in\mathbb{R}^p\times\cdots\times\mathbb{R}^p} \mathbb{E}\left[\min_{1\leq j\leq M} \|Y - c_j\|^2 | X = x\right]. \tag{17}$$

We note that by (12),

$$\inf_{\mathbf{c}\in\mathcal{F}} \mathbb{E}\left[\min_{1\leq j\leq M} \|Y - c_j(X)\|^2\right] = \mathbb{E}\left[\inf_{\mathbf{c}\in\mathbb{R}^p\times\cdots\times\mathbb{R}^p} \mathbb{E}\left[\min_{1\leq j\leq M} \|Y - c_j\|^2 | X\right]\right],$$

and so $\mathcal{R}(\hat{\mathbf{c}}_n) = \mathbb{E}\left[\mathcal{R}(\hat{\mathbf{c}}_n; X)\right]$.

5.1 Bounds on the excess risk

In Theorem 1 and Theorem 2, we establish convergence rates on the pointwise excess risk (17) and on the excess risk (16) of a sequence of estimate $\hat{\mathbf{c}}_n$. We consider local and global Lipschitz regularity conditions on the conditional densities analogous to those used in (Györfi and Kohler, 2007) for the estimation of conditional distributions.

Theorem 1. *Let x be a point in $\mathcal{S}_X$. Assume that there exists $\kappa > 0$ such that*

$$\mathbb{P}\left(\|X - x\| \leq \epsilon\right) \geq \kappa\epsilon^d, \qquad \text{for all } \epsilon > 0. \tag{18}$$

Assume that there exists $\delta > 0$ and an integrable function $h : \mathbb{R}^p \to \mathbb{R}_+$ such that

$$\left|f_{Y|X=\tilde{x}}(y) - f_{Y|X=x}(y)\right| \leq h(y)\|\tilde{x} - x\|, \qquad \text{for all } y \in \mathbb{R}^p \text{ and for all } \tilde{x} \text{ with } \|\tilde{x} - x\| \leq \delta. \tag{19}$$

Then there exists a constant $C := C(\delta, \kappa, h, R) > 0$ such that, for all k and n satisfying $\frac{k}{n} \leq \frac{1}{C}\left(\left(\frac{\log k}{k}\right)^{\frac{d}{2}} \wedge \delta^d\right)$, and any sequence $\hat{\mathbf{c}}_n$ of estimate,

$$\mathcal{R}\left(\hat{\mathbf{c}}_n; x\right) \leq \sqrt{\frac{C(pM+1)}{2}}\sqrt{\frac{\log k}{k}} + 8R^2 \exp\left(-\frac{n\delta^d}{C}\right) + o\left(\sqrt{\frac{\log k}{k}}\right). \tag{20}$$

Condition (18) is a regularity condition on the support $\mathcal{S}_X$ in a neighborhood of the point x which is used for instance in set estimation to define the notion of a standard set (see e.g. Baíllo et al., 2000).

The term $\sqrt{\frac{C(pM+1)}{2}}\sqrt{\frac{\log k}{k}}$ in the right-hand side of (20) corresponds to the excess risk of an empirical quantizer defined on a random sample of size k (Linder et al., 1994). As pointed out in Bartlett et al. (1998), the $\log k$ factor can be eliminated at the price of added technical difficulties, and we speculate that the same applies here, so that the first term in the righ hand side of (20) could be sharpened to a constant multiple of $1/\sqrt{k}$.

With the choice of $k \asymp n^{\frac{2}{d+2}}$, Theorem 1 leads to the following bound on the pointwise excess risk.

Corollary 1. *In the setting of Theorem 1, with $k \asymp n^{\frac{2}{d+2}}$, then*

$$\mathcal{R}\left(\hat{\mathbf{c}}_n(x); x\right) = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{d+2}}\right).$$

We note that the rate of Corollary 1 is slower than the $O(\sqrt{\log n/n})$ rate that would be obtained in the quantization of a sample of size n without conditioning. Hence we see that a curse of dimensionality is at play here, as expected.

To bound the excess risk, we consider a global version of the Lipschitz condition (19). We also assume that the support $\mathcal{S}_X$ is compact, which is a standard condition in regression estimation with nearest neighbors.

Theorem 2. *Suppose that $\mathcal{S}_X$ is compact, and that there exists an integrable function $h : \mathbb{R}^p \to \mathbb{R}_+$ such that*

$$\left|f_{Y|X=\tilde{x}}(y) - f_{Y|X=x}(y)\right| \leq h(y)\|\tilde{x} - x\|, \tag{21}$$

for all x and $\tilde{x}$ in $\mathcal{S}_X$. Let $\hat{\mathbf{c}}_n$ be a sequence of estimate. Then with $k \asymp n^{\frac{2}{d+2}}$,

$$\mathcal{R}(\hat{\mathbf{c}}_n) = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{d+2}}\right). \tag{22}$$

We note that the rates in both Corollary 1 and Theorem 2 depend adversely on the dimension d of the predictor variable X. This results from the smoothing over x used to estimate the conditional distortions $\mathcal{E}(\mathbf{c}; x)$. Note also that these rates do not depend on the number of quantization points M, nor on the dimension p of the response Y; the dependence on p and M is only through the constants, as exhibited in the right-hand side of (20) for instance. This is consistent with known bounds on the excess risk in empirical vector quantization (Antos, 2005; Bartlett et al., 1998; Linder et al., 1994), as seen in Section 2.4.

In the real-valued regression setting (with $M = 1$ and $p = 1$), condition (21) entails the Lipschitz continuity of the regression function $\eta(x) = E[Y|X = x]$ by the Lebesgue dominated theorem combined with the assumption that Y is bounded. Therefore the rate in Theorem 2 is suboptimal for $M = 1$ since the minimax rate of convergence of the L_2 risk for a Lipschitz continuous regression function with bounded Y is $n^{-\frac{d}{d+2}}$. Yet a striking difference between the cases $M = 1$ and $M \geq 2$ is that the equality

$$\mathcal{R}(\hat{\mathbf{c}}_n) = \mathbb{E}[(Y - \hat{\mathbf{c}}_n(X))^2] - \mathbb{E}[(Y - \eta(X))^2] = \mathbb{E}[(\hat{\mathbf{c}}_n(X) - \eta(X)]^2$$

in the case $M = 1$, which allows for the decomposition into the well-known bias/variance sum, with the bias depending on the regularity of the regression function, does not extend to the case $M \geq 2$.

Therefore when $M \geq 2$, the dependence, if any, of the convergence rate of the excess risk on the regularity of the solution(s) to the multifunction regression problem is unclear. In addition when $M \geq 2$, the solution set of a quantization problem may contain multiple solutions (and as pointed out in Pollard (1982a), few conditions enforcing uniqueness exist). Thus one might expect that irregular solutions to the multifunction regression problem do exist. Therefore a deeper analysis of the convergence rate would necessitate a fine examination of the regularity of the argmin sets of $\mathcal{E}(\mathbf{c}; x)$ over x, which is something that we remain curious about. That said, by using the connection between the quantization problem and the Wasserstein distance, we can state a convergence result on the solution, as shown in the next Section.

5.2 Convergence in solution

For any x in $\mathcal{S}_X$, let $\mathcal{C}^\star(x) = \arg\min_{\mathbf{c}\in\mathbb{R}^p\times\cdots\times\mathbb{R}^p} \mathcal{E}(\mathbf{c}; x)$, the argmin set of $\mathcal{E}(\mathbf{c}; x)$. Given two closed-valued multifunctions $\mathbf{c}_1$ and $\mathbf{c}_2$, we measure the proximity between $\mathbf{c}_1(x)$ and $\mathbf{c}_2(x)$ by their Hausdorff distance $d_H(\mathbf{c}_1(x), \mathbf{c}_2(x))$, where the Hausdorff distance between two subsets A and B of $\mathbb{R}^p$ is defined by

$$d_H(A, B) = \sup_{a\in A} \inf_{b\in B} \|a - b\| \vee \sup_{b\in B} \inf_{a\in A} \|a - b\|.$$

We state the following qualitative result under the local conditions used in Theorem 1.

Theorem 3. *Let x be a point in $\mathcal{S}_X$ and suppose that (18) and (19) hold. Let $\hat{\mathbf{c}}_n$ be a sequence of estimate. Suppose that $\frac{k}{n} \to 0$ and $\frac{k}{\log n} \to \infty$. Then the set of accumulation points of $\hat{\mathbf{c}}_n(x)$ is a nonempty subset of $\mathcal{C}^\star(x)$ and*

$$\inf_{\mathbf{c}\in\mathcal{C}^\star(x)} d_H(\hat{\mathbf{c}}_n(x), \mathbf{c}) \to 0 \quad \textit{almost surely as } n \to \infty.$$

Thus, at any x in $\mathcal{S}_X$, the accumulation points of $\hat{\mathbf{c}}_n(x)$ are optimal. In particular, if the multifunction regression problem admits a unique solution $\mathbf{c}^\star$, then $\hat{\mathbf{c}}_n(x)$ converges almost surely to $\mathbf{c}^\star(x)$ for all x in $\mathcal{S}_X$ satisfying (18) and (19). We note that (19) holds for all $x \in \mathcal{S}_X$ when (21) is satisfied, and that (18) holds for all x when the support $\mathcal{S}_X$ is a standard set in the sense of Baíllo et al. (2000).

6 Numerical experiments

In this section, we report on practical aspects for the implementation of the empirical conditional quantizer with k-nearest neighbor weights, as described in Algorithm 1. In particular, through two simulated examples, we discuss the choice of the parameter k corresponding to the number of neighbors and of the parameter M corresponding to the number of quantization points. The methodology is then applied to a real-world data set of speed records as a function of location along a daily automobile path in the city of Toulouse, France. This data is provided by Mediamobile (`http://www.mediamobile.com`).

6.1 Example 1: two-conditional clusters

In this example, we apply the methodology to a sample of $n = 2,500$ simulated points for the distribution represented in Figure 1. In details, X follows a uniform distribution over $[0, 1]$, and given X, Y follows a mixture of two normal distributions with equal weights, with both variances equal to 0.01, and with mean functions $\eta_1(x) = \arctan(8x)$ and $\eta_2(x) = 2\arctan(6x)$. The number of quantization points is set to $M = 2$ for all x in $[0, 1]$.

To select the number of neighbors k, we propose a data-driven method based on the minimization of an estimate of the average prediction error $\mathbb{E}[\mathcal{E}(\hat{\mathbf{c}}_n(X); X)]$ used to define the excess risk. For this purpose, we split the data into two parts, of size $\lfloor 2n/3 \rfloor$ and $n - \lfloor 2n/3 \rfloor$. The first part is used to construct the model, while the second part is used to estimate the mean prediction error (other fractions than $2/3 - 1/3$ could be taken so long as the size of the test set remains smaller than the size of the learning set). Specifically, given an integer k, for each $\lfloor 2n/3 \rfloor + 1 \leq i \leq n$, we determine an empirical quantizer $\hat{\mathbf{c}}_n(X_i)$ by minimization of the quantization error based on the k-nearest neighbors of X_i among the (X_j, Y_j), for $1 \leq j \leq \lfloor 2n/3 \rfloor$, that is, $\hat{\mathbf{c}}_n(X_i) := (\hat{c}_{n,1}(X_i), \ldots, \hat{c}_{n,M}(X_i))$ minimizes

$$\mathbf{c} \mapsto \frac{1}{k} \sum_{j=1}^{n} \min_{1 \leq j \leq M} (Y_j - c_{n,\ell})^2 \, \mathbf{1} \left\{ X_j \text{ is a } k\text{-NN of } X_i \text{ among } X_1, \ldots, X_{\lfloor 2n/3 \rfloor} \right\}.$$

Using this, we set

$$\hat{\mathcal{E}}_{P,n}(k) = \sum_{i=\lfloor 2n/3 \rfloor + 1}^{n} \min_{1 \leq j \leq M} (Y_i - \hat{c}_{n,j}(X_i))^2 \,,$$

which is an estimate of the average prediction error. The data-driven value of the number of neighbors is then selected as a minimizer of $\hat{\mathcal{E}}_{P,n}(k)$ over k.

In this example, $\hat{\mathcal{E}}_{P,n}(k)$ over k has been evaluated for values of k ranging from 10 to 150 by steps of 5. The graph of $\hat{\mathcal{E}}_{P,n}(k)$ as a function of k is represented in the left panel of Figure 2. The minimum of the estimated prediction error is attained at $k = 75$. This value is then used to evaluate empirical conditional quantizers at 100 equally spaced x-values ranging from 0 to 1. The resulting conditional quantizers are represented as the green and red curves in the right panel of Figure 2.

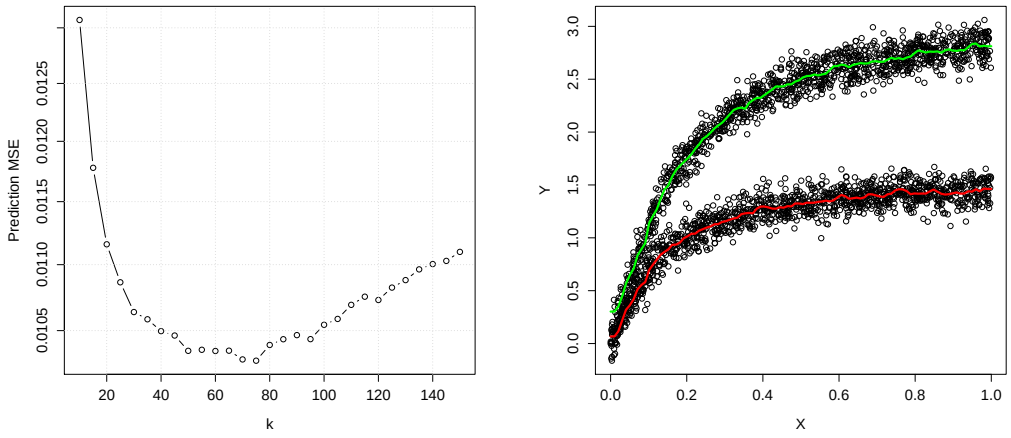

Figure 2: Left: Estimated prediction mean square error versus the number of neighbors k. The minimum is attained at $k = 75$. Right: Scatterplot of the data with the curves corresponding to the empirical quantizers with $k = 75$.

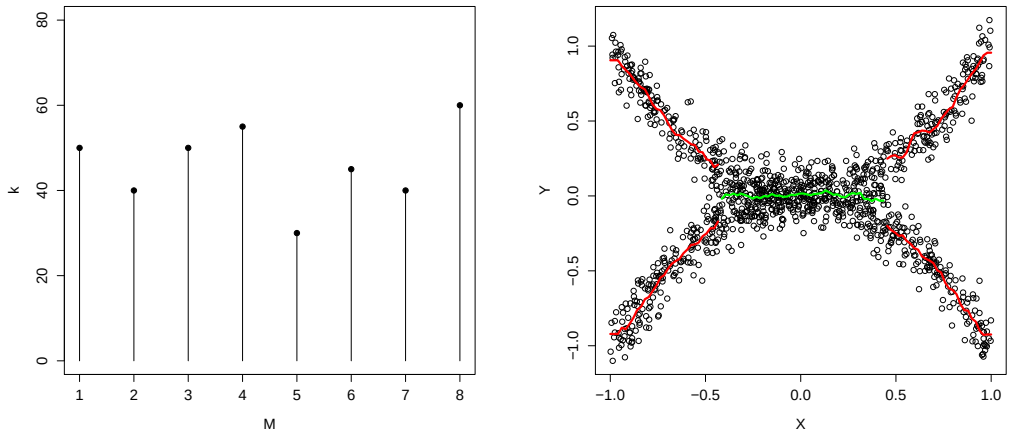

Figure 3: Left: Optimal value of k (number of neighbors) selected by minimizing the estimated mean prediction error as a function of M (number of quantization points). Right: Scatterplot of the data with the curves of the empirical quantizers. For each x, the number of quantization points is selected automatically using the gap heuristic Tibshiriani et al. (2001)

6.2 Example 2: one or two conditional clusters

In this example, we consider a pair (X, Y) where X follows a uniform distribution over $[-1, 1]$, and where given X, Y follows a mixture of normal distribution with equal proportions, with variances both equal to 0.01, and with mean functions $\eta_1(x) = x^2$ and $\eta_2(x) = -x^2$. A scatterplot of $n = 1,200$ points simulated from this distribution is represented in the right panel of Figure 3. It appears that the conditional distribution of Y given X is well concentrated around one cluster when x is approximately in the range $[-0.4, 0.4]$, while it clusters into two groups outside this interval. This calls for an automatic selection of both k (the number of neighbors) and M (the number of quantization points). Here, the goal is have k and M both depend on x.

The problematic of selecting a number a quantization points is standard in clustering analysis, where it corresponds to the selection of the number of clusters. Several heuristics have been introduced for that purpose. We shall use the gap heuristic proposed by Tibshiriani et al. (2001), whereby the number of clusters is selected by comparing the change in the within-cluster variability to that expected under a null reference distribution, which is not clustered, like a uniform or unimodal distribution. In the present setting, the difficulty in selecting both M and k lies in the lack of a global criterion to optimize. Indeed, for each k, the mean prediction error decreases with M, so this criterion cannot be used to simultaneously select k and M. Moreover, the use of an empirical heuristic, like the gap heuristic, for selecting M would require k to have been specified first.

To circumvent these issues, we propose the following method. First, for each M in a given range, a value of k is selected from the data by minimizing the mean prediction error, as described in Example 1. Denote this value by $k(M)$. In this example, we let M vary between 1 and 8; the estimated $k(M)$ are represented in Figure 3 (left) as a function of M. Next, for each x-value, and for each value of $k(M)$, we applied the gap heuristic (Tibshiriani et al., 2001) to select M. Denote this value by $M_{\text{gap}}(k(M))$. The final value of M is then selected by a majority vote. Denote this value by $\hat{M}$. At last, we select k as $\hat{k} := k(\hat{M})$. This procedure is repeated for each x-value, so the

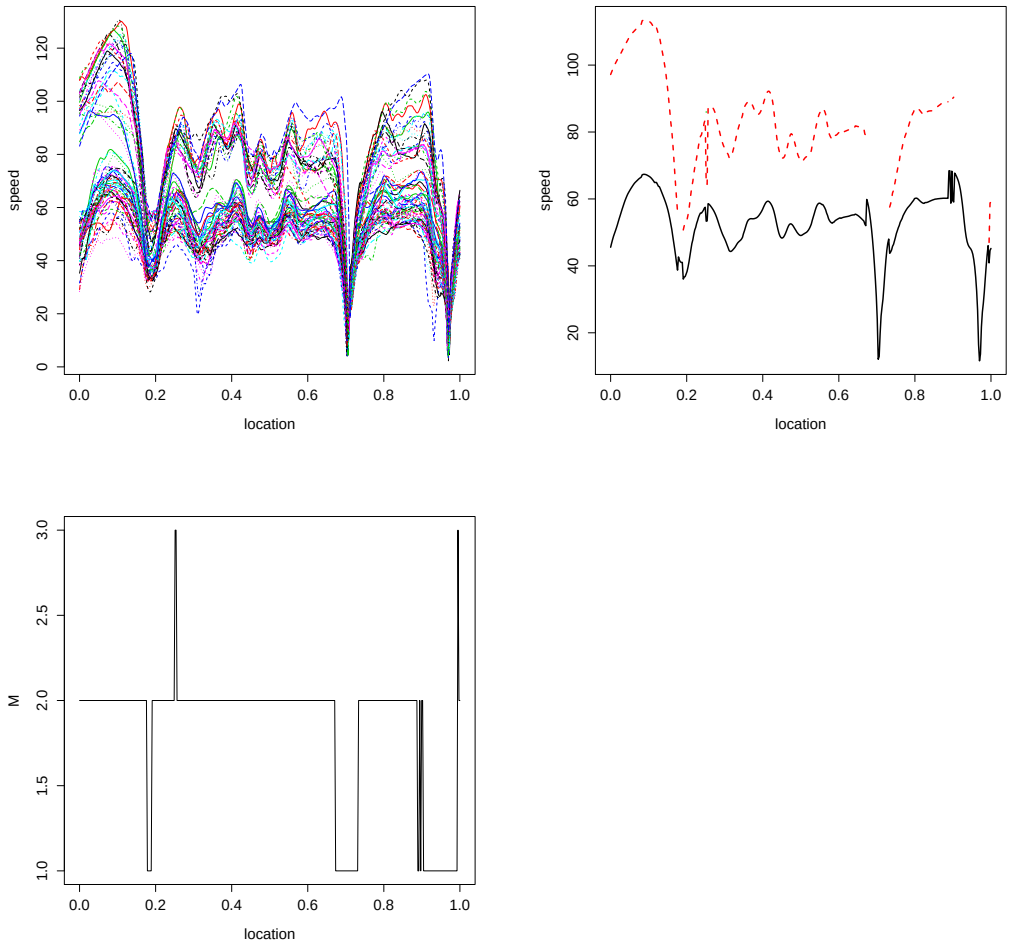

Figure 4: Top-left: Speed records as a function of location along the path. Top-right: Optimal number of quantization point selected with the gap heuristic as a function of location. Bottom-left: Quantization points (either 1, 2, or 3) as a function of location.

selected values of k and M both vary with x.

Interestingly in these simulations, for each x, the values $\{M_{\text{gap}}(k(M))\,,\, M = 1, \ldots, 8\}$ where all equal, therefore the selection of M was particularly robust to the initial value of k. It is also interesting to note that on this example the pair $(\hat{k}, \hat{M})$ selected at each x satisfies the stability relations $\hat{M} = M_{\text{gap}}(\hat{k})$ and $\hat{k} = k(\hat{M})$.

We applied this selection procedure to 100 x values equally spaced between -1 and 1. This resulted in either 1 or 2 clusters. The quantization points are represented as curves in the right panel of Figure 3.

6.3 Example: Speed data

We consider a data set of Floating Car Data (FCD) extracted from GPS devices which record the speed and location of cars at a frequency of 10 Hz. The raw data is map-matched to a network of roads. In this example, $n = 70$ vehicles have been monitored at different times and days while

moving along a given path, 10 kilometers long, and composed of sections of inner-city roads and of a freeway. The data is represented in the top-left panel of Figure 4 as 70 curves giving the speed (in kilometers per hour) as a function of the distance along the path (normalized to unit length).

As an approach to road traffic forecasting, Allain et al. (2009) propose to first cluster the speed-location curves to define prototypical speed patterns, and next to assign a new individual to the closest cluster. To implement the clustering approach, it is required that the data correspond to the same path. Yet when using FCD, vehicles may share only a small section of a trajectory, so that the number of data for a given path may be limited and this may hamper the prediction in some cases.

To cope with this issue, we propose to strongly localize the determination of the speed patterns by inferring the cluster structure of the speed conditionally on the location. It can be noticed from the top-left panel of Figure 4 that drivers have different behaviors at high speeds while vehicle speeds with small values present less variability. This difference in variabilities may be explained by the presence of traffic jams, which has a stronger effect on a freeway ride, where high speeds can no longer be attained, than on an inner-city ride, where the traffic is already constrained by speed limits, traffic signals and stop signs. The cluster structure of the traffic flow is well revealed by the conditional quantization, as represented in the top-right panel of Figure 4. The analysis yields either one or two cluster conditionally on the location which can be interpreted as corresponding to free flow and congested flow situations.

7 Proofs

We start in section 7.1 by establishing a uniform concentration inequality on the distortion. Theorem 1, Theorem 2 and Theorem 3 are proved in sections 7.2, 7.3, and 7.4 respectively.

7.1 Concentration of the distortion

Proposition 1 below gives an upper bound on the uniform deviations of $\mathcal{E}_n(\mathbf{c};x)$ to $\mathcal{E}(\mathbf{c};x)$ for each fixed point x.

Proposition 1. *In the context of Theorem 1, let x be a point in the support of the distribution of X, let $\kappa > 0$ satisfying (18) and let $\delta > 0$ and $h : \mathbb{R}^p \to \mathbb{R}_+$ integrable satisfying (19). There exists a constant $C := C(\delta, \kappa, h, R) > 0$ such that for any $\epsilon > 0$, any k and n satisfying $\frac{k}{n} \leq \frac{1}{C}\left(\epsilon^d \wedge \delta^d\right)$,*

$$\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c};x) - \mathcal{E}(\mathbf{c};x)| > \epsilon\right) \leq \frac{\omega_p^M 2^{5pM+1} R^{2pM}}{\omega_{pM}} \epsilon^{-pM} \exp\left(-\frac{k\epsilon^2}{64R^2(32R^2+\epsilon)}\right) + \exp\left(-\frac{n\epsilon^d}{C}\right) + \exp\left(-\frac{n\delta^d}{C}\right).$$

We shall need the following Lemma which is Lemma 6 in Devroye (1982).

Lemma 1 (Devroye (1982))**.** *Let $(U_i)_{i\geq 1}$ be a sequence of independent, zero mean random variables such that $|U_i| \leq c$ almost surely. For all real numbers $a_1, \ldots, a_n \geq 0$ such that $\sum_{i=1}^n a_i \leq 1$, and all $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i U_i\right| > \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2(c^2 + c\epsilon)(\sup a_i)}\right).$$

Proof. Given $\mathbf{c} := (c_1, \ldots, c_M) \in (\mathbb{R}^p)^M$. Set $Z_i = \min_{1 \le j \le M} \|Y_i - c_j\|^2$, for $i = 1, \ldots, n$. Then $\mathcal{E}_n(\mathbf{c}; x)$ can be expressed as

$$\mathcal{E}_n(\mathbf{c}; x) = \sum_{i=1}^{n} W_{n,i}(x) Z_i.$$

Let

$$\tilde{\mathcal{E}}_n(\mathbf{c}; x) = \sum_{i=1}^{n} W_{n,i}(x) \mathcal{E}(\mathbf{c}; X_i) \tag{23}$$

be a centering term. We proceed to bound the deviations of $|\mathcal{E}_n(\mathbf{c}; x) - \tilde{\mathcal{E}}_n(\mathbf{c}; x)|$ and $|\tilde{\mathcal{E}}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)|$ uniformly over $\mathbf{c}$ in $\mathcal{B}_p(R)^M$.

For any $\epsilon > 0$, we have

$$\mathbb{P}\left(\left|\mathcal{E}_n(\mathbf{c}; x) - \tilde{\mathcal{E}}_n(\mathbf{c}; x)\right| > \epsilon\right) = \mathbb{E}\left[\mathbb{P}\left(\left|\sum_{i=1}^{n} W_{n,i}(x) \left(Z_i - \mathcal{E}(\mathbf{c}; X_i)\right)\right| > \epsilon \,\middle|\, X_1, \ldots, X_n\right)\right].$$

Note that for each $i = 1, \ldots, n$, the weight $W_{n,i}(x)$ depends on the distance of x with respect to the X_i's hence it is $\sigma(X_1, \ldots, X_n)$-measurable, the random variable Z_i is almost surely bounded by $4R^2$, and $\mathbb{E}[Z_i - \mathcal{E}(c_1, \ldots, c_M; X_i)|X_1, \ldots, X_n] = 0$ almost surely. So, by applying Lemma 1 with coefficients $a_i = W_{n,i}(x)$, random variables $U_i = Z_i - \mathcal{E}(\mathbf{c}; X_i)$, conditionally on the sample $X_1, \ldots, X_n$, we obtain that for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} W_{n,i}(x) \left(Z_i - \mathcal{E}(\mathbf{c}; X_i)\right)\right| > \epsilon \,\middle|\, X_1, \ldots, X_n\right) \le 2 \exp\left(-\frac{\epsilon^2}{2[(8R^2)^2 + (8R^2)\epsilon](1/k)}\right),$$

from which it follows that

$$\mathbb{P}\left(\left|\mathcal{E}_n(\mathbf{c}; x) - \tilde{\mathcal{E}}_n(\mathbf{c}; x)\right| > \epsilon\right) \le 2 \exp\left(-\frac{k\epsilon^2}{16R^2[8R^2 + \epsilon]}\right). \tag{24}$$

To obtain a uniform bound, we consider a covering of the set $\{(c_1, \ldots, c_M) : c_i \in \mathcal{B}(R)\} = \mathcal{B}_p(R)^M$ using the distance induced by the Euclidean norm of $\mathbb{R}^{pM}$. Since $\mathcal{B}_p(R)^M$ is a compact subset of $\mathbb{R}^{pM}$, the minimal number $\mathcal{N}(\mathcal{B}_p(R)^M, \eta)$ of balls of radius η that are necessary to cover $\mathcal{B}_p(R)^M$ is of order η^{-pM}, i.e., by considering an η-packing of $\mathcal{B}_p(R)^M$, we can prove that

$$\mathcal{N}\left(\mathcal{B}_p(R)^M, \eta\right) \le \frac{\omega_p^M 2^{pM}}{\omega_{pM}} \left(\frac{R}{\eta}\right)^{pM} =: C_0 \eta^{-pM}. \tag{25}$$

Let $\mathbf{a}_1, \ldots, \mathbf{a}_{N_\eta}$ be a covering of $\{(c_1, \ldots, c_M) : c_i \in \mathcal{B}_p(R)\} = \mathcal{B}_p(R)^M$ by balls of radius $\eta > 0$ of minimal cardinality, that is, $N_\eta = \mathcal{N}(\mathcal{B}_p(R)^M, \eta)$ and for any $\mathbf{c} \in \mathcal{B}_p(R)^M$, there is at least one $\mathbf{a}_\ell$ with components $\mathbf{a}_\ell = (a_{\ell,1}, \ldots, a_{\ell,M})$ such that $\|\mathbf{c} - \mathbf{a}_i\| \le \eta$, where the norm is defined by $\|\mathbf{c} - \mathbf{a}_\ell\|^2 = \|c_1 - a_{\ell,1}\|^2 + \cdots + \|c_M - a_{\ell,M}\|^2$. By a union bound, we have

$$\mathbb{P}\left(\sup_{\mathbf{c} \in \mathcal{B}(R)^M} \left|\mathcal{E}_n(\mathbf{c}; x) - \tilde{\mathcal{E}}_n(\mathbf{c}; x)\right| > \epsilon\right) \le \sum_{\ell=1}^{N_\eta} \mathbb{P}\left(\sup_{\mathbf{c} \in \mathcal{B}(\mathbf{a}_\ell, \eta)} \left|\mathcal{E}_n(\mathbf{c}; x) - \tilde{\mathcal{E}}(\mathbf{c}; x)\right| > \epsilon\right). \tag{26}$$

Fix $1 \le \ell \le N_\eta$. For any $\mathbf{c} = (c_1, \ldots, c_M)$ in $\mathcal{B}_p(\mathbf{a}_\ell, \eta)^M$, and any $1 \le j \le M$,

$$\|Y - c_j\|^2 = \|Y - a_{\ell,j}\|^2 + \|a_{\ell,j} - c_j\|^2 + 2\langle Y - a_{\ell,j}, a_{\ell,j} - c_j\rangle \ge \|Y - a_{\ell,j}\|^2 - 4R\eta, \tag{27}$$

since $\|Y - a_{\ell,j}\| \leq 2R$ and using the fact that $\|\mathbf{c} - \mathbf{a}_\ell\| \leq \eta$ implies $\|a_{\ell,j} - c_j\| \leq \eta$. Therefore,

$$\min_{1\leq j\leq M} \|Y - c_j\|^2 \geq \min_{1\leq j\leq M} \|Y - a_{\ell,j}\|^2 - 4R\eta,$$

and by taking the expectation conditionally on X, we deduce that $\mathcal{E}(\mathbf{c}; x') \geq \mathcal{E}(\mathbf{a}_\ell; x') - 4R\eta$ for any x' in the support of the distribution of X. By exchanging c_j with $a_{\ell,j}$ in (27), the same reasoning leads to the inequality $\mathcal{E}(\mathbf{a}_\ell; x') \geq \mathcal{E}(\mathbf{c}; x') - 4R\eta$. Hence, for any $\mathbf{c}$ in $\mathcal{B}_p(\mathbf{a}_\ell, \eta)$ and for any x' in $\mathcal{S}_X$,

$$\left|\mathcal{E}(\mathbf{c}; x') - \mathcal{E}(\mathbf{a}_\ell; x')\right| \leq 4R\eta,$$

from which we deduce that

$$\left|\tilde{\mathcal{E}}_n(\mathbf{c}; x) - \tilde{\mathcal{E}}_n(\mathbf{a}_\ell; x)\right| \leq \sum_{i=1}^n W_{n,i}(x) \left|\mathcal{E}(\mathbf{c}; X_i) - \mathcal{E}(\mathbf{a}_\ell; X_i)\right| \leq 4R\eta.$$

Similarly, by considering $\mathcal{E}_n$ in place of $\mathcal{E}$ in the steps above, we also have that, for any $\mathbf{c}$ in $\mathcal{B}_p(\mathbf{a}_i, \eta)$,

$$\left|\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}_n(\mathbf{a}_i; x)\right| \leq 4R\eta.$$

Therefore, for any $1 \leq \ell \leq N_\eta$,

$$\sup_{\mathbf{c}\in\mathcal{B}_p(\mathbf{a}_\ell,\eta)} \left|\mathcal{E}_n(\mathbf{c}; x) - \tilde{\mathcal{E}}_n(\mathbf{c}; x)\right| \leq \left|\mathcal{E}_n(\mathbf{a}_\ell; x) - \tilde{\mathcal{E}}_n(\mathbf{a}_\ell, x)\right| + 8R\eta. \tag{28}$$

Then we deduce from (26) and (28) with $\eta = \epsilon/(16R)$, together with the exponential inequality in (24) and the bound on the covering number in (25), that

$$\begin{aligned}
\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} \left|\mathcal{E}_n(\mathbf{c}; x) - \tilde{\mathcal{E}}_n(\mathbf{c}; x)\right| > \epsilon\right) &\leq \mathcal{N}\left(\mathcal{B}_p(R)^M, \frac{\epsilon}{16R}\right) \max_{1\leq\ell\leq N_\epsilon} \mathbb{P}\left(\left|\mathcal{E}_n(\mathbf{a}_\ell; x) - \tilde{\mathcal{E}}_n(\mathbf{a}_\ell, x)\right| > \frac{\epsilon}{2}\right) \\
&\leq 2C_0 \left(\frac{\epsilon}{16R}\right)^{-pM} \exp\left(-\frac{k\epsilon^2}{32R^2(16R^2+\epsilon)}\right). \qquad (29)
\end{aligned}$$

Now we proceed to bound the deviations of $\left|\tilde{\mathcal{E}}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)\right|$ uniformly over $\mathbf{c}$. For any $\mathbf{c}$ in $\mathcal{B}_p(R)^M$, we have

$$\left|\tilde{\mathcal{E}}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)\right| \leq \sum_{i=1}^n W_{n,i}(x) \left|\mathcal{E}(\mathbf{c}; X_i) - \mathcal{E}(\mathbf{c}; x)\right|.$$

Let $\delta > 0$ and $h : \mathbb{R}^p \to \mathbb{R}_+$ integrable satisfying the regularity (19). Since $\|Y\|$ is bounded by R, for $\tilde{x}$ with $\|\tilde{x} - x\| \leq \delta$ and any $\mathbf{c}$ in $\mathcal{B}_p(R)^M$,

$$|\mathcal{E}(\mathbf{c}; \tilde{x}) - \mathcal{E}(\mathbf{c}; x)| \leq (4R^2) \int_{\mathbb{R}^p} \left|f_{Y|X=\tilde{x}}(y) - f_{Y|X=x}(y)\right| \mathrm{d}y \leq 4R^2 \|h\|_1 \|\tilde{x} - x\| =: L\|\tilde{x} - x\|, \tag{30}$$

where $\|h\|_1$ denotes the L^1 norm of the function h.

Denote by $X_{(k,n)}(x)$ the k^{th} nearest neighbor of x among the sample $X_1, \ldots, X_n$. Then, for any $\mathbf{c}$ in $\mathcal{B}_p(R)^M$,

$$\begin{aligned}
\left|\tilde{\mathcal{E}}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)\right| &\leq L\|X_{(k,n)}(x) - x\| \mathbf{1}\{\|X_{(k,n)}(x) - x\| \leq \delta\} \\
&\quad + \left|\tilde{\mathcal{E}}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)\right| \mathbf{1}\{\|X_{(k,n)}(x) - x\| > \delta\} \\
&\leq L\|X_{(k,n)}(x) - x\| \mathbf{1}\{\|X_{(k,n)}(x) - x\| \leq \delta\} \\
&\quad + 8R^2 \mathbf{1}\{\|X_{(k,n)}(x) - x\| > \delta\} \quad \text{almost surely,}
\end{aligned}$$

where in the last inequality we used the fact that $\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M}\left|\tilde{\mathcal{E}}_n(\mathbf{c};x)-\mathcal{E}(\mathbf{c};x)\right|\leq 8R^2$.

Hence, with probability one,

$$\begin{aligned}&\sup_{\mathbf{c}\in\mathbf{B}(R)^M}\left|\tilde{\mathcal{E}}_n(\mathbf{c};x)-\mathcal{E}(\mathbf{c};x)\right|\\&\quad\leq L\|X_{(k,n)}(x)-x\|\mathbf{1}\{\|X_{(k,n)}(x)-x\|\leq\delta\}+8R^2\mathbf{1}\{\|X_{(k,n)}(x)-x\|>\delta\}.\end{aligned}\tag{31}$$

Then,

$$\begin{aligned}&\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M}\left|\tilde{\mathcal{E}}_n(\mathbf{c};x)-\mathcal{E}(\mathbf{c};x)\right|>\epsilon\right)\\&\quad\leq\mathbb{P}\left(\left[L\|X_{(k,n)}(x)-x\|>\frac{\epsilon}{2}\right]\cap\left[\|X_{(k,n)}(x)-x\|\leq\delta\right]\right)\\&\qquad+\mathbb{P}\left(8R^2\mathbf{1}\{\|X_{(k,n)}(x)-x\|>\delta\}>\frac{\epsilon}{2}\right)\\&\quad\leq\mathbb{P}\left(\|X_{(k,n)}(x)-x\|>\frac{\epsilon}{2L}\right)+\mathbb{P}\left(\|X_{(k,n)}(x)-x\|>\delta\right).\end{aligned}\tag{32}$$

For any $0<\eta\leq\delta$, let $p_\eta=\mathbb{P}(\|X-x\|\leq\eta)$. Note that $p_\eta\geq\kappa\eta^d$ where $\kappa>0$ is defined in (18). Since,

$$\mathbb{P}\left(\|X_{(k,n)}(x)-x\|>\eta\right)=\mathbb{P}\left(\sum_{i=1}^n\mathbf{1}\left\{\|X_i-x\|\leq\eta\right\}\leq k-1\right),$$

we deduce by using Chernoff's bound that, for any $0<\eta\leq\delta$

$$\mathbb{P}\left(\|X_{(k,n)}(x)-x\|>\eta\right)\leq\exp\left(-\frac{1}{2}\left(1-\frac{k-1}{np_\eta}\right)^2np_\eta\right)\leq\exp\left(-\frac{np_\eta}{8}\right),$$

where the last inequality holds whenever $\frac{k-1}{np_\eta}\leq\frac{1}{2}$, which is implied when

$$\frac{k}{n}\leq\frac{\kappa}{2}\eta^d.\tag{33}$$

Therefore in this case

$$\mathbb{P}\left(\|X_{(k,n)}(x)-x\|>\eta\right)\leq\exp\left(-\frac{\kappa}{8}n\eta^d\right).\tag{34}$$

Hence, by reporting (34) in (32), for any $\epsilon>0$, and any k and n such that

$$\frac{k}{n}\leq\frac{\kappa}{2}\left[\left(\frac{\epsilon}{2L}\right)^d\wedge\delta^d\right],\tag{35}$$

we have

$$\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M}\left|\tilde{\mathcal{E}}_n(\mathbf{c};x)-\mathcal{E}(\mathbf{c};x)\right|>\epsilon\right)\leq\left[\exp\left(-\frac{n\epsilon^d}{C_1}\right)+\exp\left(-\frac{n\delta^d}{C_1}\right)\right]\mathbf{1}\{\epsilon\leq 8R^2\},\tag{36}$$

with $C_1=\frac{8}{\kappa}[1\vee(2L)^d]$.

Combining (29) and (36), we obtain that, for any $\epsilon>0$, and any k and n satisfying (35),

$$\begin{aligned}\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M}|\mathcal{E}_n(\mathbf{c};x)-\mathcal{E}(\mathbf{c};x)|>\epsilon\right)\quad&\leq\quad 2C_0\left(\frac{\epsilon/2}{16R}\right)^{-pM}\exp\left(-\frac{k(\epsilon/2)^2}{32R^2(16R^2+\epsilon/2)}\right)\\&\quad+\exp\left(-\frac{n(\epsilon/2)^d}{C_1}\right)+\exp\left(-\frac{n\delta^d}{C_1}\right).\end{aligned}$$

From this, and the fact that (35) is satisfied when $\frac{k}{n}\leq\frac{\kappa}{2}\left(\frac{1}{(2L)^d}\wedge 1\right)\left(\epsilon^d\wedge\delta^d\right)$, we conclude. □

7.2 Proof of Theorem 1

Let $\mathbf{c}^\star$ be an optimal quantizer, meaning that $\mathcal{E}(\mathbf{c}^\star, x) = \mathcal{E}^\star$. Denote by $\mathbf{c}_n^\star := \hat{\mathbf{c}}_n(x)$ the value of the estimate $\hat{\mathbf{c}}_n$ at the point x. Following standard arguments, we have

$$\begin{aligned}\mathcal{E}(\mathbf{c}_n^\star; x) - \mathcal{E}^\star(x) &= [\mathcal{E}(\mathbf{c}_n^\star; x) - \mathcal{E}_n(\mathbf{c}_n^\star; x)] + [\mathcal{E}_n(\mathbf{c}_n^\star; x) - \mathcal{E}_n(\mathbf{c}^\star; x)] + [\mathcal{E}_n(\mathbf{c}^\star; x) - \mathcal{E}(\mathbf{c}^\star; x)] \\ &\leq 2 \sup_{\mathbf{c} \in \mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)|, \end{aligned} \tag{37}$$

so that

$$\mathbb{E}\left[\mathcal{E}(\mathbf{c}_n^\star)\right] - \mathcal{E}^\star \leq 2\mathbb{E}\left[\sup_{\mathbf{c} \in \mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)|\right]. \tag{38}$$

Given $a > 0$, and since $\sup_{\mathbf{c} \in \mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)| \leq 8R^2$ almost surely, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{\mathbf{c} \in \mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)|\right] &= \int_0^\infty \mathbb{P}\left(\sup_{\mathbf{c} \in \mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)| > \epsilon\right) \mathrm{d}\epsilon \\ &\leq a + \int_a^\infty \mathbb{P}\left(\sup_{\mathbf{c} \in \mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)| > \epsilon\right) \mathrm{d}\epsilon. \end{aligned} \tag{39}$$

By Proposition 1 there exists constants $C_0 := C_0(R, p, M) > 0$ and $C_1 := C_1(\delta, \kappa, h, R) > 0$ such that, for any $\epsilon > 0$, and any k and n with $\frac{k}{n} \leq \frac{1}{C_1}\left(\epsilon^d \wedge \delta^d\right)$,

$$\begin{aligned}&\mathbb{P}\left(\sup_{\mathbf{c} \in \mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)| > \epsilon\right) \\ &\quad \leq C_0 \epsilon^{-pM} \exp\left(-\frac{k\epsilon^2}{C_1}\right) + \exp\left(-\frac{n\epsilon^d}{C_1}\right) + \exp\left(-\frac{n\delta^d}{C_1}\right). \end{aligned}$$

Let $a > 0$ and suppose that $\frac{k}{n} \leq \frac{1}{C_1}\left(a^d \wedge \delta^d\right)$, Then, since $\sup_{\mathbf{c} \in \mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)| \leq 8R^2$ almost surely,

$$\begin{aligned}&\int_0^\infty \mathbb{P}\left(\sup_{\mathbf{c} \in \mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)| > \epsilon\right) \mathrm{d}\epsilon \\ &\quad \leq a + C_0 \int_a^{8R^2} \epsilon^{-pM} \exp\left(-\frac{k\epsilon^2}{C_1}\right) \mathrm{d}\epsilon + \int_a^{8R^2} \exp\left(-\frac{n\epsilon^d}{C_1}\right) \mathrm{d}\epsilon + 8R^2 \exp\left(-\frac{n\delta^d}{C_1}\right) \\ &\quad \leq a + C_0 \frac{8R^2}{a^{pM}} \exp\left(-\frac{ka^2}{C_1}\right) + 8R^2 \exp\left(-\frac{na^d}{C_1}\right) + 8R^2 \exp\left(-\frac{n\delta^d}{C_1}\right). \end{aligned} \tag{40}$$

Taking $a = c\sqrt{\frac{\log k}{k}}$, with $c = \sqrt{\frac{C_1(pM+1)}{2}}$, we have

$$\frac{1}{a^{pM}} \exp\left(-\frac{ka^2}{C_1}\right) = c^{-pM} \frac{1}{\sqrt{k}(\log k)^{pM/2}} = o\left(\sqrt{\frac{\log k}{k}}\right),$$

and since $\frac{k}{n} \leq \frac{1}{C_1}\left(a^d \wedge \delta^d\right)$,

$$\exp\left(-\frac{na^d}{C_1}\right) \leq \exp\left(-k\right) = o\left(\sqrt{\frac{\log k}{k}}\right).$$

Therefore

$$\int_0^\infty \mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c};x) - \mathcal{E}(\mathbf{c};x)| > \epsilon\right) \mathrm{d}\epsilon$$
$$\leq \sqrt{\frac{C_1(pM+1)}{2}}\sqrt{\frac{\log k}{k}} + 8R^2 \exp\left(-\frac{n\delta^d}{C_1}\right) + o\left(\sqrt{\frac{\log k}{k}}\right)$$

for all k and n satisfying $\frac{k}{n} \leq \frac{1}{C_1}\left(\left(\frac{C(pM+1)}{2}\right)^{\frac{d}{2}}\left(\frac{\log k}{k}\right)^{\frac{d}{2}} \wedge \delta^d\right)$. This inequality is implied when $\frac{k}{n} \leq \frac{1}{C_1}\left(\left(\frac{C(pM+1)}{2}\right)^{\frac{d}{2}} \wedge 1\right)\left(\left(\frac{\log k}{k}\right)^{\frac{d}{2}} \wedge \delta^d\right)$ and from this we conclude with any choice of constant C (in the statement of Theorem 1) larger than $\left[\frac{1}{C_1}\left(\left(\frac{C(pM+1)}{2}\right)^{\frac{d}{2}} \wedge 1\right)\right]^{-1} \vee C_1$.

7.3 Proof of Theorem 2

For any x in $\mathcal{S}_X$ and $\mathbf{c}$ in $\mathcal{B}_p(R)^M$, let $\tilde{\mathcal{E}}_n(\mathbf{c};x)$ be defined in (23). Using (37), for any x in $\mathcal{S}_X$, we have

$$\begin{aligned}\mathcal{E}(\hat{\mathbf{c}}_n(x);x) - \mathcal{E}^\star(x) &\leq 2\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c};x) - \mathcal{E}(\mathbf{c};x)| \\ &\leq 2\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} \left|\mathcal{E}_n(\mathbf{c};x) - \tilde{\mathcal{E}}_n(\mathbf{c};x)\right| + 2\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} \left|\tilde{\mathcal{E}}_n(\mathbf{c};x) - \mathcal{E}(\mathbf{c};x)\right|.\end{aligned}$$

Hence

$$\mathbb{E}\left[\mathcal{E}(\hat{\mathbf{c}}_n(x);x) - \mathcal{E}^\star(x)\right] \leq 2\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} \left|\mathcal{E}_n(\mathbf{c};x) - \tilde{\mathcal{E}}_n(\mathbf{c};x)\right|\right] + 2\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} \left|\tilde{\mathcal{E}}_n(\mathbf{c};x) - \mathcal{E}(\mathbf{c};x)\right|\right] \tag{41}$$

and we proceed to bound the two terms on the right-hand side of (41).

To bound the first term, we use the concentration inequality (29) from the proof of Proposition 1 to deduce that, for any $\epsilon > 0$ and for any x in $\mathcal{S}_X$,

$$\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} \left|\mathcal{E}_n(\mathbf{c};x) - \tilde{\mathcal{E}}_n(\mathbf{c};x)\right| > \epsilon\right) \leq C\epsilon^{-pM}\exp\left(-\frac{k\epsilon^2}{C}\right),$$

where $C > 0$ is a constant not depending on x. Hence, by proceeding as in the first part of (40), followed by integrating over x, we deduce that

$$\mathbb{E}\left[\int_{\mathbb{R}^d} \sup_{\mathbf{c}\in B_p(R)^M} \left|\mathcal{E}_n(\mathbf{c};x) - \tilde{\mathcal{E}}_n(\mathbf{c};x)\right| P_X(\mathrm{d}x)\right] \leq C\sqrt{\frac{\log k}{k}} \tag{42}$$

for some constant $C > 0$.

To bound the second term, let $h : \mathbb{R}^p \to \mathbb{R}_+$ be an integrable function satisfying (21). Since $\|Y\|$ is bounded by R, for any $\mathbf{c} \in \mathcal{B}_p(R)^M$ and any x and $\tilde{x}$ in $\mathcal{S}_X$, we have

$$|\mathcal{E}(\mathbf{c};\tilde{x}) - \mathcal{E}(\mathbf{c};x)| \leq 4R^2\|h\|_1\|\tilde{x} - x\|.$$

Hence

$$\left|\tilde{\mathcal{E}}_n(\mathbf{c};x)-\mathcal{E}(\mathbf{c};x)\right| \leq \sum_{i=1}^n W_{n,i}(x)\left|\mathcal{E}(\mathbf{c};X_i)-\mathcal{E}(\mathbf{c};x)\right| \leq 4R^2\|h\|_1\|X_{(k,n)}(x)-x\|$$

and so

$$\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M}\left|\tilde{\mathcal{E}}_n(\mathbf{c};x)-\mathcal{E}(\mathbf{c};x)\right|\right] \leq 4R^2\|h\|_1\mathbb{E}\left[\|X_{(k,n)}(x)-x\|\right]. \tag{43}$$

Using (Liitiainen et al., 2010, Theorem 3.2), we have

$$\mathbb{E}\left[\|X_{(k,n)}(X)-X\|\right] \leq 2\sqrt{d}\,\mathrm{diam}(\mathcal{S}_X)\left(\frac{k}{n}\right)^{\frac{1}{d}}, \tag{44}$$

where X denotes a random variable with distribution P_X and independent from the sample, and where $\mathrm{diam}(\mathcal{S}_X)$ denotes the diameter of $\mathcal{S}_X$. By inserting (44) in (43) we obtain

$$\mathbb{E}\left[\int_{\mathbb{R}^d}\sup_{\mathbf{c}\in B_p(R)^M}\left|\tilde{\mathcal{E}}_n(\mathbf{c};x)-\mathcal{E}(\mathbf{c};x)\right|P_X(\mathrm{d}x)\right] \leq 8R^2\|h\|_1\sqrt{d}\,\mathrm{diam}(\mathcal{S}_X)\left(\frac{k}{n}\right)^{\frac{1}{d}}, \tag{45}$$

and we conclude by combining (42) and (45).

7.4 Proof of Theorem 3

Let $P_{n,k}=\sum_{i=1}^n W_{n,i}(x)\delta_{Y_i}$. By (Graf and Luschgy, 2000, Theorem 4.21), the result will hold if

$$W_2\left(P_{n,k},P_{Y|X=x}\right)\to 0 \qquad \text{almost surely.} \tag{46}$$

Recall that a sequence (Q_n) of probability measures converges to Q in the (L_2) Wasserstein distance if (Q_n) converges weakly to Q and if $\int\|y\|^2dQ_n(y)\to\int\|y\|^2dQ(y)$ as $n\to\infty$. Thus, since $\|Y\|\leq R$ almost surely, it suffices to show that $P_{n,k}=\sum_{i=1}^n W_{n,i}(x)\delta_{Y_i}$ converges weakly to $P_{Y|X=x}$ almost surely.

Towards proving this, let g be a continuous and bounded function over $\mathbb{R}^p$ and let $m(x)=\mathbb{E}[g(Y)|X=x]$. Proceeding as in the proof of (24), we get

$$\mathbb{P}\left(\left|\sum_{i=1}^n W_{n,i}(x)g(Y_i)-\sum_{i=1}^n W_{n,i}(x)m(X_i)\right|>\epsilon\right) \leq 2\exp\left(-\frac{k\epsilon^2}{4\|g\|_\infty\left(2\|g\|_\infty+\epsilon\right)}\right). \tag{47}$$

Let $\delta>0$ and $h:\mathbb{R}^p\to\mathbb{R}_+$ integrable satisfying the regularity condition (19). For any $\tilde{x}$ with $\|\tilde{x}-x\|\leq\delta$,

$$|m(\tilde{x})-m(x)| \leq \|g\|_\infty\|h\|_1\|\tilde{x}-x\| =: L\|\tilde{x}-x\|.$$

So

$$\left|\sum_{i=1}^n W_{n,i}(x)m(X_i)-m(x)\right| \leq L\|X_{(k,n)}(x)-x\|\mathbf{1}\left\{\|X_{(k,n)}-x\|\leq\delta\right\}+2\|g\|_\infty\mathbf{1}\left\{\|X_{(k,n)}-x\|>\delta\right\}.$$

Hence, for any $\epsilon>0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n W_{n,i}(x)m(X_i)-m(x)\right|>\epsilon\right) \leq \mathbb{P}\left(\|X_{(k,n)}(x)-x\|>\frac{\epsilon}{2L}\right)+\mathbb{P}\left(\|X_{(k,n)}(x)-x\|>\delta\right).$$

Using (34), for any k and n such that

$$\frac{k}{n} \leq \frac{\kappa}{2}\left(\left(\frac{\epsilon}{2L}\right)^d \wedge \delta^d\right), \tag{48}$$

where κ is defined in (18), we have

$$\mathbb{P}\left(\left|\sum_{i=1}^n W_{n,i}(x)m(X_i) - m(x)\right| > \epsilon\right) \leq \exp\left(-\frac{n\epsilon^d}{C_1}\right) + \exp\left(-\frac{n\delta^d}{C_1}\right), \tag{49}$$

with $C_1 = \frac{8}{\kappa}[1 \vee (2L)^d]$. Combining (47) and (49), we deduce that for any $\epsilon > 0$ and k and n satisfying (48),

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{i=1}^n W_{n,i}(x)g(Y_i) - m(x)\right| > \epsilon\right)\\ &\quad\leq 2\exp\left(-\frac{k(\epsilon/2)^2}{4\|g\|_\infty\,(2\|g\|_\infty + \epsilon/2)}\right) + \exp\left(-\frac{n(\epsilon/2)^d}{C_1}\right) + \exp\left(-\frac{n\delta^d}{C_1}\right).\end{aligned} \tag{50}$$

Since $\frac{k}{n} \to 0$, (48) is satisfied for all n large enough. Now for any $\epsilon > 0$, the last two terms in the right-hand side of (48) are summable over n, and the first term is summable if $\frac{k}{\log n} \to \infty$. So by the Borel-Cantelli Lemma, $\sum_{i=1}^n W_{n,i}(x)g(Y_i)$ converges almost surely to $m(x)$. Hence $P_{n,k}$ converges weakly to $P_{Y|X=x}$ almost surely which implies that (46) holds.

Acknowledgments

We are grateful to two anonymous referees and one Associate Editor for their helpful comments.

References

Allain, G., F. Gamboa, P. Goudal, J.-M. Loubes, and E. Maza (2009). A statistical framework for road traffic prediction. *16th ITS World Congress and Exhibition on Intelligent Transport Systems and Services*.

Antos, A. (2005). Improved minimax bounds on the test and training distortion of empirically designed vector quantizers. *IEEE Transactions on Information Theory 51*(11), 4022–4032.

Antos, A., L. Györfy, and A. György (2005). Individual convergence rates in empirical vector quantizer design. *IEEE Transactions on Information Theory 51*(11), 4013–4022.

Arias-Castro, E., D. Mason, and B. Pelletier (2016). On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research 17*(43), 1–28.

Baíllo, A., A. Cuevas, and A. Justel (2000). Set estimation and nonparametric detection. *Canadian Journal of Statistics 28*, 765–782.

Bartlett, P., T. Linder, and G. Lugosi (1998). The minimax distortion redundancy in empirical qunatizer design. *IEEE Transactions on Information Theory 44*(5), 1802–1813.

Biau, G., L. Devroye, and G. Lugosi (2008). On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory 54*(2), 781–790.

Chen, Y.-C., C. R. Genovese, R. J. Tibshirani, and L. Wasserman (2016). Nonparametric modal regression. *Annals of Statistics 44*(2), 489–514.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 17*(8), 790–799.

Collomb, G., W. Hardle, and S. Hassani (1987). A note on prediction via estimation of the conditional mode function. *Journal of Statistical Planning and Inference 15*, 227–236.

Comaniciu, D. and P. Meer (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*(5), 1–18.

Devijver, E. (2015). Finite mixture regression: a sparse variable selection by model selection for clustering. *Electron. J. Stat. 9*(2), 2642–2674.

Devroye, L. (1982). Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. *Z. Wahrsch. Verw. Gebiete 61*(4), 467–481.

Duda, R., P. Hart, and D. Stork (2000). *Pattern Classification* (Second Edition ed.). Wiley-Interscience, New-York.

Einbeck, J. and G. Tutz (2006). Modeling beyond regression functions: an application of multimodal regression to speed-flow data. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 55*(4), 461475.

Everitt, B. S. and D. J. Hand (1981). *Finite mixture distributions*. Chapman & Hall, London-New York. Monographs on Applied Probability and Statistics.

Gersho, A. and R. Gray (1992). *Vector Quantization and Signal Compression*. Kluwer Academic Press, Boston.

Graf, S. and H. Luschgy (2000). *Foundations of quantization for probability distributions*, Volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin.

Györfi, L. and M. Kohler (2007). Nonparametric estimation of conditional distributions. *IEEE Transactions on Information Theory 53*(5), 1872–1879.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New-York.

Györfi, L. and M. Wegkamp (2008). Quantization for nonparametric regression. *IEEE Transactions on Information Theory 54*(2), 867–874.

Huber, P. and E. Ronchetti (2009). *Robust Statistics* (Second ed.). Wiley Series in Probability and Statistics. Wiley.

Jacobs, R., M. Jordan, S. Nowlan, and G. Hinton (1991). Adaptive mixture of local experts. *Neural Computation 3*, 79–87.

Jiang, W. and M. Tanner (1999). Hierarchical mixture of experts for exponential family regression models: approximation and maximum likelihood estimation. *The Annals of Statistics 27*, 987–1011.

Kemp, G. and J. Santos Silva (2012). Regression towards the mode. *Journal of Econometrics 170*(1), 92–101.

Khalili, A. and J. Chen (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association 102*(409), 1025–1038.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Lee, M.-J. (1989). Mode regression. *Journal of Econometrics 42*(3), 337–349.

Lee, M.-J. (1993). Quadratic mode regression. *Journal of Econometrics 57*(1-3), 1–19.

Levrard, C. (2015). Nonasymptotic bounds for vector quantization in hilbert spaces. *The Annals of Statistics 43*(2), 592–619.

Liitiainen, E., F. Corona, and A. Lendasse (2010). Residual variance estimation using a nearest neighbor statistic. *Journal of Multivariate Analysis 101*(4), 811–823.

Linder, T. (2002). Learning-theoretic methods in vector quantization. In *Principles of Nonparametric Learning*, Volume 434 of *International Centre for Mechanical Sciences, Courses and Lectures*, pp. 163–210. Springer, Vienna.

Linder, T., G. Lugosi, and K. Zeger (1994). Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory 40*(6), 1728–1740.

McLachlan, G. and D. Peel (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

Meynet, C. (2013). An ℓ_1-oracle inequality for the Lasso in finite mixture Gaussian regression models. *ESAIM Probab. Stat. 17*, 650–671.

Pagès, G. (1997). A space quantization method for numerical integration. *Journal of Computational and Applied Mathematics 89*, 1–38.

Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics 9*, 135–140.

Pollard, D. (1982a). A central limit theorem for k-means clustering. *The Annals of Probability 10*(4), 919–926.

Pollard, D. (1982b). Quantization and the method of k-means. *IEEE Transactions on Information Theory 28*(2), 199–205.

Rockafellar, R. and R.-B. Wets (2009). *Variational Analysis*. Springer-Verlag Berlin Heidelberg.

Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge.

Sager, T. and R. Thisted (1982). Maximum likelihood estimation of isotonic modal regression. *The Annals of Statistics 10*, 690–707.

Städler, N., P. Bühlmann, and S. van de Geer (2010). ℓ_1-penalization for mixture regression models. *TEST 19*(2), 209–256.

Tibshiriani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Statist. Soc. B 63*, 411–423.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical analysis of finite mixture distributions*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Yao, W. and L. Li (2014). A new regression model: modal linear regression. *Scandinavian Journal of Statistics 41*(3), 656–671.

Yao, W., B. Lindsay, and L. Runze (2012). Local modal regression. *Journal of Nonparametric Statistics 24*(3), 647–663.