Prediction by quantization of a conditional distribution

Jean-Michel Loubes, Bruno Pelletier

Prediction by quantization of a conditional distribution

Jean-Michel Loubes* Bruno Pelletier†

April 7, 2016

Abstract. We consider the problem of quantizing the conditional distribution of a random variable Y given a random vector X. We propose an empirical quantizer defined by combining the principles of k-means clustering with the nonparametric smoothing technique of k-nearest neighbors. We provide an asymptotic analysis of the estimate and we derive a bound on the error rate of the quantizer. The proposed methodology is illustrated on simulated examples and on a speed-flow traffic data set used in the context of road traffic forecasting.

Index Terms: Regression analysis, vector quantization, nonparametric statistics, clustering, k-means.

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 $\mu(x) := \mathbb{E}[Y|X = x]$ 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 μ) 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 mode regression. In Lee (1989, 1993) (see also Kemp and Santos Silva, 2012) 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.

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 $\mu_1(x) < \mu_2(x)$, and X follows a uniform distribution over the unit interval. The difference in means $\mu_2(x) - \mu_1(x)$ 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, $\mu(x) = \frac{1}{2}[\mu_1(x) + \mu_2(x)]$ and μ 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.

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

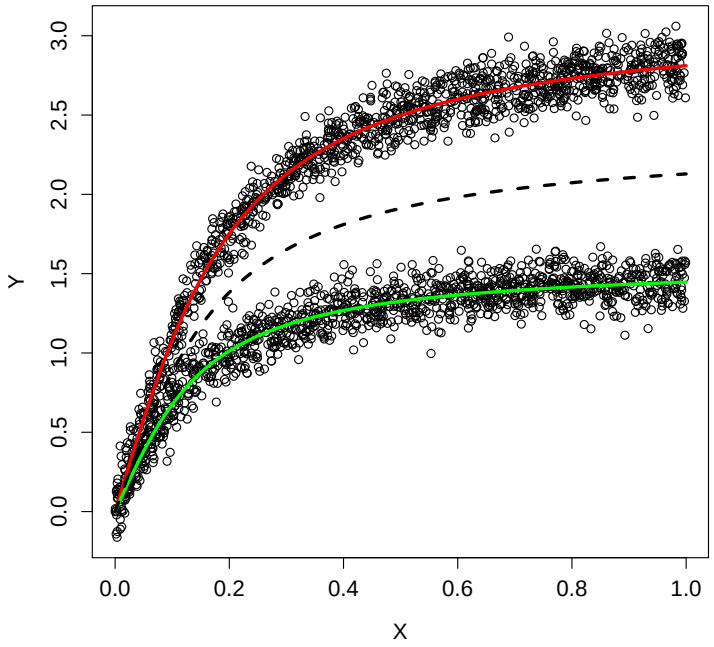

Figure 1: Scatterplot of 400 realizations of the pair (X, Y) where X follows a uniform distribution over $[0, 1]$, and where the distribution of Y given X is a mixture of two Normal distribution with weights equal to $\frac{1}{2}$, variances equal to 0.01, and means $\mu_1(x) = \arctan(8x)$ (green curve) and $\mu_2(x) = 2\arctan(6x)$ (red curve). The regression function $\mathbb{E}[Y|X = x] = \frac{1}{2}[\mu_1(x) + \mu_2(x)]$ is represented by the dashed curve.

Fitting a finite mixture model is a popular approach for modeling 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. In the context of a 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 $\mu_1(x)$ and $\mu_2(x)$, 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.

In a non parametric setting, multi-modality of the conditional distribution of Y given X is considered in Chen et al. (2014). Their proposal is to estimate all the modes of the conditional distribution, that is, the set of points of local maximum of the conditional density, called the modal set. To estimate the modal set, Einbeck and Tutz (2006) proposes 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., 2015; Cheng, 1995; Comaniciu and Meer, 2002),

and the resulting modal set estimate is consistent (Chen et al., 2014). The modal set may prove useful in a regression context 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. But in a more general setting where the conditional distribution may potentially admit a large number of local points of maxima, the modal set may be difficult to interpret (the modal set may even be uncountable).

In this paper, we consider the problem of predicting a random variable Y from a random vector X from data which may potentially be heterogeneous but where, in contrast with classical regression, the prediction from X is in the form of a finite set of cardinality potentially larger than 1. More precisely, we propose to quantize the conditional distribution of Y given X, that is, to approximate the conditional distribution of Y given X by a discrete measure with finite support. This is achieved by combining the principles of vector quantization (Gersho and Gray, 1992; Graf and Luschgy, 2000; Linder, 2002) with a smoothing technique used in nonparametric statistics (see e.g. Györfi et al., 2002) and we make the following contributions:

- We propose a method to quantize the distribution of Ygiven X 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).

- We provide an asymptotic analysis of the estimate. In particular, we prove that the proposed conditional quantizer is consistent, and we derive a bound on the quantization error.

- We propose heuristics for automatically selecting the number of neighbors and 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 start by reviewing the foundational principles of vector quantization, and we define an empirical conditional quantizer by minimization of an empirical risk. In section 3, we provide an asymptotic analysis of the empirical conditional quantizer. First, in Theorem 1, we establish the almost sure uniform convergence of the empirical distortion towards the conditional distortion. Next in Theorem 2, we obtain a bound on the distortion rate of the conditional quantizer. Then under additional technical assumptions, we prove in Theorem 3 that the accumulation points of any empirical conditional quantizer sequence are optimal conditional quantizers. In section 4, 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. In section 5, we discuss open problems and extensions. The proofs of the results are exposed in section 6 and some technical lemmas are collected in Appendix A, at the end of the paper.

2 Quantization of conditional distributions

In this section, we start by reviewing foundational materials on vector quantization. We pursue with the definition of the vector quantization of a conditional distribution and we introduce an empirical conditional quantizer constructed from a random sample. Then we present a simple algorithm to compute the conditional quantizer from the sample.

2.1 Vector quantization

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).

2.1.1 The quantization problem

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 set $\{c_1, \ldots, c_M\}$ of M points in $\mathbf{R}^p$. Using the Euclidean norm $\|.\|$ on $\mathbb{R}^p$, the performance of a quantizer q is measured by the *distortion*

$$\mathcal{D}(q; P_Y) = \mathbb{E}[\|Y - q(Y)\|^2]. \tag{1}$$

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 of M points in $\mathbb{R}^p$. Any quantizer q defines a partition of $\mathbb{R}^p$ into the sets $q^{-1}(c_i)$, for $i = 1, \ldots, M$. In the case of a nearest-neighbor quantizer $q_{\mathbf{c}}$, the partition is called a Voronoi partition and for any $i = 1, \ldots, M$, the (closed) Voronoi cell $V_i(\mathbf{c})$ associated with c_i is defined by

$$V_i(\mathbf{c}) = \{x \in \mathbb{R}^p \, : \, \|x - c_i\| \le \|x - c_j\|\}. \tag{2}$$

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 $q_{\mathbf{c}}^{-1}(c_i) \subset V_i(\mathbf{c})$ for all $i = 1, \ldots, M$.

2.1.2 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 will be considered, and a nearest-neighbor quantizer $q_{\mathbf{c}}$ will be 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

$$\mathcal{E}(\mathbf{c}; P_Y) := \mathcal{D}(q_{\mathbf{c}}; P_Y) = \mathbb{E}\left[\min_{1 \le j \le M} \|Y - c_j\|^2\right]. \tag{3}$$

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

$$\mathcal{E}^\star(P_Y) = \inf_{\mathbf{c} \in (\mathbb{R}^p)^M} \mathcal{E}(\mathbf{c}; P_Y), \tag{4}$$

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

2.1.3 Approximation of measures

The problem of finding an optimal quantizer for the probability measure P_Y is also equivalent to the one of best approximating P_Y by a discrete measure with support of cardinality at most M, in the sense that

$$\mathcal{E}^\star(P_Y) = \inf\{W_2(P_Y, Q) \, : \, Q \text{ probability measure with } |\mathrm{Supp}(Q)| \le M\}, \tag{5}$$

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). 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 (5) coincides with the set of optimal quantizers minimizing (3) (Graf and Luschgy, 2000, Lemma 4.4). Following Graf and Luschgy (2000), any minimizer of (5) 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 notice that 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 $\mathbf{c}^\star = (c_1^\star, \ldots, c_M^\star)$.

2.1.4 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

$$\mathcal{E}_n(\mathbf{c}; P_Y) := \mathcal{E}\left(\mathbf{c}; P_Y^{(n)}\right) = \frac{1}{n}\sum_{i=1}^{n} \min_{1\leq j\leq M} \|Y_i - c_j\|^2, \tag{6}$$

with $\mathcal{E}$ as in (3). 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).

2.2 Conditional vector quantization

Let (X, Y) be a pair of random vectors in $\mathbb{R}^d \times \mathbb{R}^p$. By conditioning on X in (3), given M an integer and $\mathbf{c}$ an M-points quantizer, we define the distortion of the conditional distribution of Y given X at x by

$$\mathcal{E}(\mathbf{c}; x) = \mathbb{E}\left[\min_{1\leq j\leq M} \|Y - c_j\|^2 \big| X = x\right], \tag{7}$$

and we set

$$\mathcal{E}^\star(x) = \inf\left\{\mathcal{E}(\mathbf{c}; x) \,:\, \mathbf{c} \in (\mathbb{R}^p)^M\right\}.$$

Note that $\mathcal{E}(\mathbf{c}; x)$ and $\mathcal{E}^\star(x)$ are meant to be defined P_X-almost everywhere.

Let $(X_i, Y_i)_{i\geq 1}$ be a sequence of i.i.d. random vectors with the same distribution as (X, Y). Our proposal is to construct an empirical quantizer of the conditional distribution of Y given X by first estimating $\mathcal{E}(\mathbf{c}; x)$, and next by minimizing the estimated distortion. We consider a local averaging estimate of $\mathcal{E}(\mathbf{c}; x)$ of the form

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

where $\{W_{n,i}(x) : i = 1, \ldots, n\}$ is a set of weights depending on $X_1, \ldots, X_n$ satisfying $W_{n,i}(x) \geq 0$ and $\sum_{i=1}^n W_{n,i}(x) = 1$. Using (8), we define the empirical conditional quantizer at x as any minimizer $\mathbf{c}_n^\star(x)$ of $\mathcal{E}_n(\mathbf{c}; x)$, that is, satisfying $\mathcal{E}_n(\mathbf{c}_n^\star; x) = \mathcal{E}_n^\star(x)$, where $\mathcal{E}_n^\star(x) = \inf_{\mathbf{c} \in (\mathbb{R}^p)^M} \mathcal{E}_n(\mathbf{c}; x)$. In section 3, we provide an asymptotic analysis of the empirical conditional quantizer when the $W_{n,i}(x)$'s are the weights corresponding to averaging with the method of k nearest neighbors, that is, when

$$W_{n,i}(x) = \frac{1}{k} \mathbf{1} \left\{ X_i \text{ is among the } k \text{ nearest neighbors of } x \right\}. \tag{9}$$

Minimizing $\mathcal{E}_n(\mathbf{c}; x)$, or $\mathcal{E}_n(\mathbf{c}; P_Y)$ in the non conditional setting, is 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. We emphasize that, from a practical perspective, the local averaging estimate $\mathcal{E}_n(\mathbf{c}; x)$ defined in (8) can be minimized by considering a weighted version of the k-means algorithm, as described in Algorithm 1. In particular, when using the k-nearest neighbor weights (9), the algorithm is equivalent to the standard k-means algorithm applied to the y_i's which correspond to the k nearest neighbors of x among the X_i's.

Algorithm 1: Conditional weighted k-means algorithm.

Input: Data $(X_1, Y_1), \ldots, (X_n, Y_n)$, weights $W_{n,i}(x)$, for $i = 1, \ldots, n$, 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.

3 Asymptotic analysis

In this section, we provide an asymptotic analysis of the conditional quantization scheme defined by minimizing (8) using the k nearest neighbors weights (9). We assume that (X, Y) admits a probability density f_{XY} with respect to the Lebesgue measure on $\mathbb{R}^d \times \mathbb{R}^p$ which is Lipschitz continuous over $\mathbb{R}^d \times \mathbb{R}^p$. We also assume that Y is bounded, that is, that there is $R > 0$ such that $\|Y\| \leq R$ almost surely. The marginal probability density of X is denoted by f_X. The volume of the d-dimensional unit ball is denoted by ω_d.

We start by establishing the almost sure uniform convergence of the empirical distortion towards the distortion for any fixed point x, and uniformly over all quantizers $\mathbf{c}$ in $\mathcal{B}(R)^M$.

Theorem 1. *Suppose that* $\frac{k}{n} \to 0$ *and* $\frac{k}{\log n} \to \infty$. *For any* x *with* $f_X(x) > 0$,

$$\sup_{\mathbf{c} \in \mathcal{B}(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)| \to 0, \quad \textit{almost surely},$$

We note that, from the exponential inequality (26) in the proof of Theorem 1, it follows that $r_n \sup_{\mathbf{c} \in \mathcal{B}(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)| \to 0$ almost surely for any choice of sequence $r_n \to \infty$ such that $r_n \ll \left(\frac{n}{\log n}\right)^{\frac{1}{d}} \wedge \left(\frac{k}{\log n}\right)^{\frac{1}{2}}$.

Next, we establish a result which gives the quantization error rate at x of a sequence of the empirical conditional quantizers.

Theorem 2. *Let x be a point such that $f_X(x) > 0$. There exists constants $C > 0$ and $\delta > 0$ depending only on x and f_{XY} such that, for all k and n satisfying $\frac{k^{\frac{d+2}{2}}}{(\log k)^{\frac{d}{2}}} \leq \sqrt{\frac{C(pM+1)}{2}} n$, and any sequence $\mathbf{c}_n^\star$ of empirically optimal quantizers,*

$$\mathbb{E}\left[\mathcal{E}(\mathbf{c}_n^\star; x)\right] - \mathcal{E}^\star(x) \leq \sqrt{\frac{C(pM+1)}{2}} \sqrt{\frac{\log k}{k}} + 8R^2 \exp\left(-\frac{n\delta^d}{C}\right) + o\left(\sqrt{\frac{\log k}{k}}\right). \tag{10}$$

The term $\sqrt{\frac{C(pM+1)}{2}} \sqrt{\frac{\log k}{k}}$ in the right-hand side of (10) corresponds to the quantization error rate in the case of a 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 conjecture that the same applies here, so that the first term in the bound in (10) could be sharpened to a constant multiple of $1/\sqrt{k}$. However, this quantization rate is not attained here due to the smoothing in X of the quantization error. Indeed, with the choice of $k \asymp n^{\frac{2}{d+2}}$, Theorem 2 leads to the rate

$$\mathbb{E}\left[\mathcal{E}(\mathbf{c}_n^\star; x)\right] - \mathcal{E}^\star(x) = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{d+2}}\right),$$

which 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.

With additional assumptions, we prove in Theorem 3 that the accumulation points of any sequence $(\mathbf{c}_n^\star)$ of empirical conditional quantizers are optimal conditional quantizers. As in Pollard (1982a), a regularity condition is needed to guarantee that the distortion is twice continuously differentiable with respect to the quantization points (Pollard, 1982a, Lemma C). In our context, this means that, for P_X-almost all x, the map $\mathbf{c} \mapsto \mathcal{E}(c; x)$ is twice continuously differentiable at each $\mathbf{c}$ with pairwise distinct components. More precisely, given $\mathbf{c} := (c_1, \ldots, c_M) \in \mathcal{B}_p(R)^M$, let $V_i = \{x \in \mathbb{R}^p : \|x - c_i\| \leq \|x - x_j\|\}$ be the Voronoi cell associated with c_i. Given adjacent cells V_i and V_j, denote by S_{ij} the boundary between the Voronoi cells V_i and V_j inside the ball $\mathcal{B}_p(R)$, that is

$$S_{ij} = V_i \cap V_j \cap \mathcal{B}_p(R).$$

The regularity condition used in (Pollard, 1982a, Lemma C) is the existence of the surface integrals $\int_{S_{ij}} f_{Y|X=x}(y)(y-z)'(y-z)\sigma(\mathrm{d}y)$ for each fixed point $z \in \mathbb{R}^p$ and their continuous dependence on the quantization points $\mathbf{c} = (c_1, \ldots, c_M)$, where σ is the $(p-1)$-dimensional surface measure on S_{ij}. In the present contest, by Lemma 1 applied with the function $g(y) = f_{Y|X=x}(y)(y_k - z_k)(y_\ell - z_\ell)$ for each $1 \leq k, \ell \leq M$, it follows that this regularity condition is satisfied. Therefore for P_X-almost all x, the map $\mathbf{c} \mapsto \mathcal{E}(\mathbf{c}; x)$ is twice continuously differentiable at each $\mathbf{c}$ with pairwise distinct components. Given any configuration $\mathbf{c}$ in $\mathcal{B}_p(R)^M$ with pairwise distinct components, let $\mathbf{H}(c)$ be the Hessian matrix of $\mathbf{c} \mapsto \mathcal{E}(\mathbf{c}; x)$ at $\mathbf{c}$. Then we arrive at the following result.

Theorem 3. *Suppose that $\frac{k}{n} \to 0$ and $\frac{k}{\log n} \to \infty$. Let x be such that $f_X(x) > 0$. If for each $\mathbf{c}^\star$ in $\mathcal{C}^\star(x)$ the Hessian matrix $\mathbf{H}(\mathbf{c}^\star)$ is positive definite, then for any sequence $\mathbf{c}_n^\star$ such that $\mathcal{E}_n(\mathbf{c}_n^\star; x) = \mathcal{E}_n^\star(x)$,*

$$\inf_{\mathbf{c}^\star \in \mathcal{C}^\star(x)} \|\mathbf{c}_n^\star - \mathbf{c}^\star\| \to 0 \quad \textit{almost surely.} \tag{11}$$

As pointed out in Pollard (1982a), the condition on the Hessian matrix may be difficult to remove, even in simple cases. On the other hand, it is satisfied in some families of distribution. For instance, in a finite mixture of Gaussian distributions with M "well separated" components, a quantization points of an optimal M-points quantizer are unique, up to relabeling (see e.g. Levrard, 2015).

4 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`).

4.1 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 $\mu_1(x) = \arctan(8x)$ and $\mu_2(x) = 2\arctan(6x)$. 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 the empirical mean prediction error. 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. Specifically, given an integer k, for each $\lfloor 2n/3 \rfloor + 1 \leq i \leq n$, we determine an empirical quantizer $\mathbf{c}_n^\star(X_i)$ 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, $\mathbf{c}_n^\star(X_i) := (c_{n,1}^\star, \ldots, c_{n,M}^\star)$ minimizes

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

Using this, we set

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

which is an estimate of the mean 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.

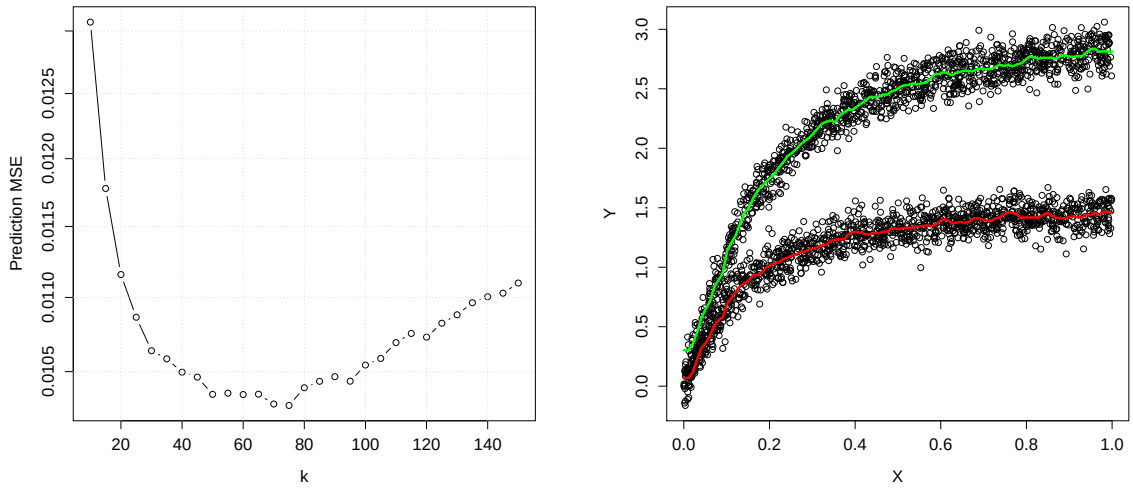

Figure 2: Left: Estimated prediction mean square error versus the number of neighbors k. The minimum is attained at $k = 75$. Right: Scatterplot of the data with the curves corresponding to the empirical quantizers with $k = 75$.

4.2 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 $\mu_1(x) = x^2$ and $\mu_2(x) = -x^2$. 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, like the gap heuristic proposed by Tibshiriani et al. (2001). 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 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.

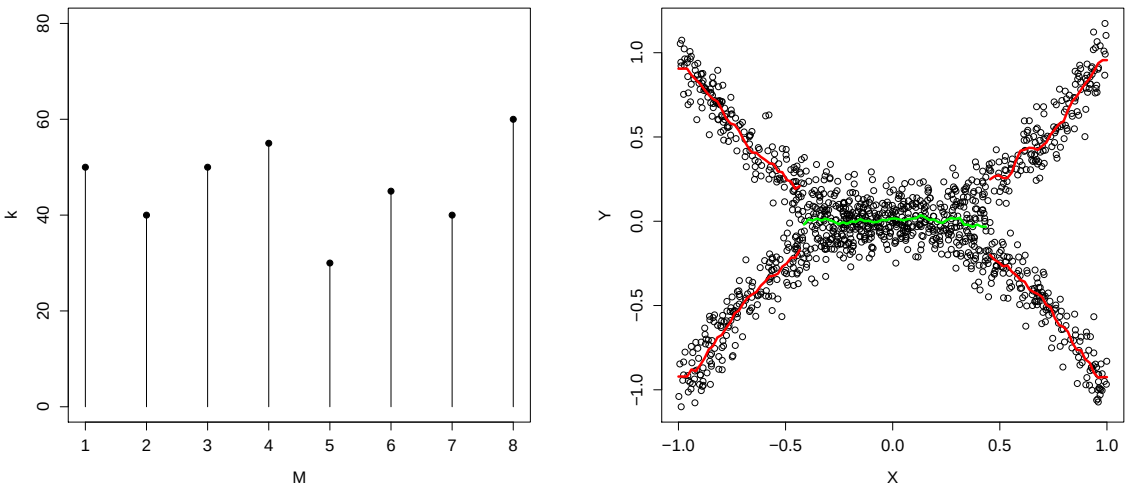

Figure 3: Left: Optimal value of k (number of neighbors) selected by minimizing the estimated mean prediction error as a function of M (number of quantization points). Right: Scatterplot of the data with the curves of the empirical quantizers. For each x, the number of quantization points is selected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

4.3 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

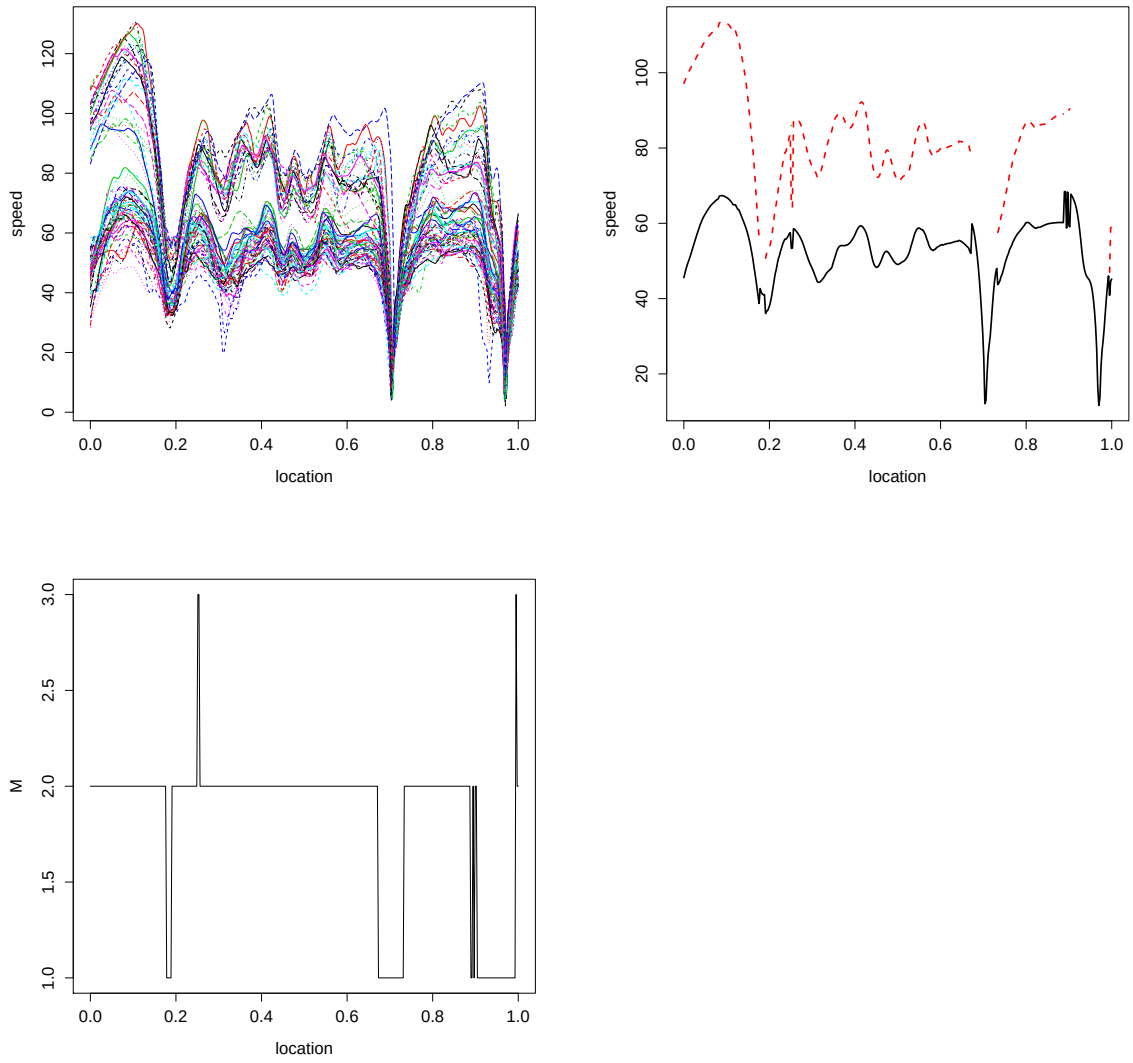

Figure 4: Top-left: Speed records as a function of location along the path. Top-right: Optimal number of quantization point selected with the gap heuristic as a function of location. Bottom-left: Quantization points (either 1, 2, or 3) as a function of location.

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.

5 Discussion

First of all, it is worth pointing out that the proposed conditional quantization scheme is completely nonparametric. This stands in contrast with a finite mixture regression model where the distribution of Y given X is a finite mixture with components belonging to some parametric family of distributions (the Gaussian family being a classical choice). Hence we do not use likelihood methods and an EM-like criterion to estimate an unobserved variable that would indicate the cluster to which each the observation belongs.

We did not either assume a regression model where the link function between Y and X belongs to a functional class defined by its smoothness or its complexity. Hence our results are pointwise, that is, they apply to any fixed x with $f_X(x) > 0$. A minimal amount of smoothness has been assumed through the condition that (X, Y) admits a Lipschitz continuous probability density function. However, for a fixed M not varying with x, this condition does not imply any regularity on the quantization points when viewed as functions of x (up to labelling), even in the case where optimal quantizers are unique, as in Theorem 3. In practice, though, it may be interesting to require that the quantization points depend smoothly on x. This question is left open for future research.

6 Proofs

We start in section 6.1 with proving a continuity for surface integrals over the boundary of two adjacent Voronoi cells. Next in section 6.2 we establish a concentration inequality on the distortion that is used in the proofs of our main results. Theorem 1, Theorem 2 and Theorem 3 are proved in sections 6.3, 6.3, and 6.3 respectively.

6.1 Regularity of surface integrals over boundaries of Voronoi cells

The following lemma states a continuity property of some surface integrals over the boundary between two adjacent cells of a Voronoi partition. Given $\mathbf{c} := (c_1, \dots, c_M) \in \mathcal{B}_p(R)^M$, denote by S_{ij} the boundary between the Voronoi cells V_i and V_j inside the ball $\mathcal{B}_p(R)^M$, that is

$$S_{ij} = V_i \cap V_j \cap \mathcal{B}_p(R),$$

where

$$V_i = \{x \in \mathbb{R}^p \,:\, \|x - c_i\| \leq \|x - x_j\|\}\,.$$

Note that S_{ij} depends on $\mathbf{c}$. The $(p-1)$-dimensional surface measure on S_{ij} is denoted by σ.

Lemma 1. *Let $g : \mathbb{R}^p \to \mathbb{R}$ be a continuous function. Given $\mathbf{c} := (c_1, \dots, c_M) \in \mathcal{B}_p(R)^M$, for any $1 \leq i \neq j \leq M$, the surface integral $\int_{S_{ij}} g(x)\sigma(\mathrm{d}x)$ depends continuously on $\mathbf{c}$ in any neighborhood of a configuration $\mathbf{c}$ with pairwise distinct components.*

Proof. Let $\tilde{\mathbf{c}} = (\tilde{c}_1, \dots, \tilde{c}_M)$ be a configuration such that $\|\tilde{c}_i - c_i\| \leq \epsilon$ for all $1 \leq i \leq M$ and with $\epsilon > 0$ small enough that the components of $\tilde{\mathbf{c}}$ are pairwise distinct. Denote by $\tilde{V}_i$, for $i = 1, \dots, M$, the Voronoi cells of $\tilde{\mathbf{c}}$, and by $\tilde{S}_{ij} = \tilde{V}_i \cap \tilde{V}_j \cap \mathcal{B}_p(R)$.

Fix $1 \leq i \neq j \leq M$. Let $\Sigma_{ij} = \{x \in \mathbb{R}^p \,:\, \|x - c_i\| = \|x - c_j\|\}$ be the hyperplane of points equidistant from c_i and c_j. Similarly, define $\tilde{\Sigma} = \{x \in \mathbb{R}^p \,:\, \|x - \tilde{c}_i\| = \|x - \tilde{c}_j\|\}$. Denote by π the orthogonal projection on Σ_{ij}. Notice that $S_{ij} \subset \Sigma_{ij}$. Notice also that whenever $(\tilde{c}_i, \tilde{c}_j)$ is close enough to (c_i, c_j): (i) the restriction of π to $\tilde{\Sigma}_{ij}$ is a diffeomorphism between $\tilde{\Sigma}_{ij}$ and Σ_{ij}, and (ii) $\pi(\tilde{S}_{ij})$ is included in $\mathcal{B}_p(2R)$.

Hence, whenever $(\tilde{c}_i, \tilde{c}_j)$ is close enough to (c_i, c_j), we have

$$\begin{aligned}
&\left| \int_{S_{ij}} g(x)\sigma(\mathrm{d}x) - \int_{\tilde{S}_{ij}} g(x)\sigma(\mathrm{d}x) \right| \\
&= \left| \int_{S_{ij}} g(x)\sigma(\mathrm{d}x) - \int_{\pi(\tilde{S}_{ij})} (g \circ \pi^{-1})(x) \left|\det(\pi^{-1})\right| \sigma(\mathrm{d}x) \right| \\
&\leq \sup_{x \in S_{ij} \cap \pi(\tilde{S}_{ij})} \left| (g \circ \pi^{-1})(x) |\det(\pi^{-1})| - g(x) \right| \sigma\left(S_{ij} \cap \pi(\tilde{S}_{ij}) \right) \\
&\quad + \left(1 + \left|\det(\pi^{-1})\right|\right) \sup_{x \in \Sigma_{ij} \cap \mathcal{B}_p(2R)} |g(x)| \sigma\left(S_{ij} \Delta \pi(\tilde{S}_{ij}) \right) \\
&\leq \omega_{p-1} R^{p-1} \sup_{x \in \Sigma_{ij} \cap \mathcal{B}_p(R)} \left| (g \circ \pi^{-1})(x) - g(x) \right| \\
&\quad + \left(1 + \left|\det(\pi^{-1})\right|\right) \sup_{x \in \Sigma_{ij} \cap \mathcal{B}_p(2R)} |g(x)| \sigma\left(S_{ij} \Delta \pi(\tilde{S}_{ij}) \right), \qquad (12)
\end{aligned}$$

where $S_{ij}\Delta\pi(\tilde{S}_{ij})$ denotes the symmetric difference between S_{ij} and $\pi(\tilde{S}_{ij})$, and where π^{-1} denotes implicitly the inverse of the restriction of π to $\tilde{\Sigma}_{ij}$. Since $g \circ \pi^{-1}$ converges uniformly to g on any compact subset of Σ_{ij} and since $\left|\det(\pi^{-1})\right| \to 1$ the first term on the right hand side of (12) converges to 0 as $\tilde{\mathbf{c}}$ tends to $\mathbf{c}$.

When $p = 1$, notice that $\pi(\tilde{S}_{ij}) = S_{ij}$ in which case the second term on the right hand side of (12) is equal to 0. When $p \geq 2$, we prove that $\sigma\left(S_{ij}\Delta\pi(\tilde{S}_{ij}) \right) \to 0$ as $\epsilon \to 0$.

By definition, $S_{ij}\Delta\pi(\tilde{S}_{ij}) = (\pi(\tilde{S}_{ij}) \cap S_{ij}^c) \cup (S_{ij} \cap \pi(\tilde{S}_{ij})^c)$. Let $y \in \pi(\tilde{S}_{ij}) \cap S_{ij}^c$. There is $\tilde{x}$ in $\tilde{S}_{ij}$ such that $y = \pi(\tilde{x})$. Since $\tilde{x} \in \tilde{S}_{ij}$, $\|\tilde{x} - \tilde{c}_i\| \leq \|\tilde{x} - \tilde{c}_k\|$ for all $1 \leq k \leq M$ which is equivalent to

$$\left\langle \tilde{x} - \tilde{c}_i, \frac{\tilde{c}_k - \tilde{c}_i}{2} \right\rangle \leq \frac{1}{4}\|\tilde{c}_k - \tilde{c}_i\|^2 \quad \text{for all } 1 \leq k \leq M.$$

We have

$$\left\langle y - c_i, \frac{c_k - c_i}{2} \right\rangle = \left\langle \pi(\tilde{x}) - \tilde{x}, \frac{c_k - c_i}{2} \right\rangle + \left\langle \tilde{x} - \tilde{c}_i, \frac{c_k - c_i}{2} \right\rangle + \left\langle \tilde{c}_i - c_i, \frac{c_k - c_i}{2} \right\rangle.$$

There is $\eta > 0$ and a constant $C_\eta > 0$ such that, whenever $\epsilon < \eta$, $\|\pi(\tilde{x}) - \tilde{x}\| \leq C_\eta \epsilon$ for all $\tilde{x}$ in $\tilde{\Sigma}_{ij} \cap \mathcal{B}_p(R)$. Hence

$$\left\langle y - c_i, \frac{c_k - c_i}{2} \right\rangle \leq RC_\eta\epsilon + \left\langle \tilde{x} - \tilde{c}_i, \frac{c_k - c_i}{2} \right\rangle + R\epsilon. \qquad (13)$$

Next,

$$\begin{aligned}
\left\langle \tilde{x} - \tilde{c}_i, \frac{c_k - c_i}{2} \right\rangle &= \left\langle \tilde{x} - \tilde{c}_i, \frac{\tilde{c}_k - \tilde{c}_i}{2} \right\rangle + \left\langle \tilde{x} - \tilde{c}_i, \frac{c_k - \tilde{c}_k - c_i + \tilde{c}_i}{2} \right\rangle \\
&\leq \frac{1}{4}\|\tilde{c}_k - \tilde{c}_i\|^2 + 2R\epsilon \\
&\leq \frac{1}{4}\|c_k - c_i\|^2 + (8R + 4)\epsilon + 2R\epsilon. \qquad (14)
\end{aligned}$$

Combining (13) and (14), we have that, for all $1 \leq k \leq M$,

$$\left\langle y - c_i, \frac{c_k - c_i}{2} \right\rangle \leq \frac{1}{4}\|c_k - c_i\|^2 + (11R + 4 + C_\eta R)\epsilon.$$

From this, and since $y \in S_{ij}^c$, we deduce that there is a constant $C > 0$, depending only on $\mathbf{c}$ and R such that $\|y - \partial S_{ij}\| \leq C\epsilon$. Hence, we have shown that $\pi(\tilde{S}_{ij}) \cap S_{ij}^c$ is included in a tubular neighborhood of ∂S_{ij} in Σ_{ij} of radius $C\epsilon$.

Since S_{ij} is convex, by Steiner's formula for the volume of parallel sets of a convex body (see e.g. Schneider (2014)), the volume of a tubular neighborhood of ∂S_{ij} in Σ_{ij} of radius $(C\epsilon)$ is a polynomial in $(C\epsilon)$ with zero constant term. Hence there is a constant $C' > 0$ such that

$$\sigma\left(\{y \in \Sigma_{ij} \,:\, \|y - \partial S_{ij}\| \leq \epsilon\}\right) = C'\epsilon + o(\epsilon).$$

From this, we conclude that $\sigma(\pi(\tilde{S}_{ij}) \cap S_{ij}^c) \to 0$ as $\epsilon \to 0$.

To prove that $\sigma(S_{ij} \cap \pi(\tilde{S}_{ij})^c) \to 0$ as $\epsilon \to 0$, we note that

$$S_{ij} \cap \pi(\tilde{S}_{ij})^c = \pi(\pi^{-1}(S_{ij})) \cap \pi(\tilde{S}_{ij}^c) = \pi\left(\pi^{-1}(S_{ij}) \cap \tilde{S}_{ij}^c\right).$$

Arguing as above, with S_{ij} exchanged with $\tilde{S}_{ij}$ and π^{-1} with π, we deduce that $\sigma\left(\pi^{-1}(S_{ij}) \cap \tilde{S}_{ij}^c\right) \to 0$ as $\epsilon \to 0$. Then by projection with π, we conclude that $\sigma(S_{ij} \cap \pi(\tilde{S}_{ij})^c) \to 0$ as $\epsilon \to 0$.

Therefore $\sigma\left(S_{ij} \Delta \pi(\tilde{S}_{ij})\right) \to 0$ as $\epsilon \to 0$ and so the second term in the right-hand side of (12) goes to 0 as $\epsilon \to 0$. □

6.2 Concentration of the distortion

Proposition 1 below gives an upper bound on the uniform deviations of $\mathcal{E}_n(\mathbf{c}; x)$ to $\mathcal{E}(\mathbf{c}; x)$. It is used in the proof of Theorem 1 for a fixed $\epsilon > 0$ in the asymptotic regime $k/n \to 0$ so that the last term in the upper bound, i.e $\mathbf{1}\left\{\frac{k}{n} > C_1\epsilon^d\right\}$, vanishes for all n large enough. Proposition 1 is also used in the proof of Theorem 2 by integrating over ϵ.

Proposition 1. *Let x be a point such that $f_X(x) > 0$. There exists positive constants $C_1 := C_1(f_{XY}, x)$, $C_2 := C_2(f_{XY}, x)$ and $\delta := \delta(f_{XY}, x)$ such that, for any $\epsilon > 0$, any k and n satisfying $\frac{k}{n} \leq \epsilon^d$,*

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathbf{c} \in \mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)| > \epsilon\right) \\
&\quad \leq \frac{\omega_p^M (4R)^{pM}}{\omega_{pM}} \epsilon^{-pM} \exp\left(-\frac{k\epsilon^2}{64R^2(32R^2 + \epsilon)}\right) + \exp\left(-\frac{n\epsilon^d}{C_2}\right) + \exp\left(-\frac{n\delta^d}{C_2}\right).
\end{aligned}$$

Proof. Given $\mathbf{c} := (c_1, \ldots, c_M) \in (\mathbf{R}^p)^M$. Set $Z_i = \min_{1 \leq j \leq M} \|Y_i - c_j\|^2$, 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

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

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 $\mathcal{B}(R)^M$.

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 2 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 $\epsilon > 0$,

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

from which it follows that

$$\mathbb{P}\left(\left|\mathcal{E}_n(\mathbf{c};x) - \tilde{\mathcal{E}}_n(\mathbf{c};x)\right| > \epsilon\right) \leq 2\exp\left(-\frac{k\epsilon^2}{16R^2[8R^2+\epsilon]}\right). \tag{15}$$

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

$$\mathcal{N}\left(\mathcal{B}(R)^M,\eta\right) \leq \frac{\omega_p^M 2^{pM}}{\omega_{pM}}\left(\frac{R}{\eta}\right)^{pM} =: C_0\eta^{-pM}. \tag{16}$$

Let $\mathbf{a}_1,\ldots,\mathbf{a}_{N_\eta}$ be a covering of $\{(c_1,\ldots,c_M) : c_i \in \mathcal{B}(R)\} = \mathcal{B}(R)^M$ by balls of radius $\eta > 0$ of minimal cardinality, that is, $N_\eta = \mathcal{N}(\mathcal{B}(R)^M,\eta)$ and for any $\mathbf{c} \in \mathcal{B}(R)^M$, there is at least one $\mathbf{a}_i$ with $\|\mathbf{c} - \mathbf{a}_i\| \leq \eta$. By a union bound, we have

$$\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}(R)^M}\left|\mathcal{E}_n(\mathbf{c};x) - \tilde{\mathcal{E}}_n(\mathbf{c};x)\right| > \epsilon\right) \leq \sum_{i=1}^{N_\eta}\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}(\mathbf{a}_i,\eta)}\left|\mathcal{E}_n(\mathbf{c};x) - \tilde{\mathcal{E}}(\mathbf{c};x)\right| > \epsilon\right). \tag{17}$$

Fix $1 \leq \ell \leq N_\eta$ and denote by $a_{\ell,j}$ the components of $\mathbf{a}_\ell$, i.e., $\mathbf{a}_\ell = (a_{\ell,1},\ldots,a_{\ell,M}) \in \mathcal{B}_p(R)^M$. For any $\mathbf{c} = (c_1,\ldots,c_M)$ in $\mathcal{B}_p(\mathbf{a}_\ell,\eta)^M$, and any $1 \leq j \leq M$,

$$\|Y - c_j\|^2 = \|Y - a_{\ell,j}\|^2 + \|a_{\ell,j} - c_j\|^2 + 2\langle Y - a_{\ell,j}, a_{\ell,j} - c_j\rangle \geq \|Y - a_{\ell,j}\|^2 - 4R\eta, \tag{18}$$

since $\|Y\| \leq R$ almost surely, and c_j and $a_{\ell,j}$ are in $\mathcal{B}(R)$. 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 $\mathcal{E}(\mathbf{c};x) \geq \mathcal{E}(\mathbf{a}_\ell;x) - 4R\eta$. By exchanging c_j with $a_{\ell,j}$ in (18), 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 $\mathcal{B}(\mathbf{a}_\ell,\eta)$,

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

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 $\mathcal{B}(\mathbf{a}_i, \eta)$,

$$|\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}_n(\mathbf{a}_i; x)| \leq 4R\eta. \tag{19}$$

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

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

Then we deduce from (17) with $\eta = \epsilon/(16R)$, together with the exponential inequality in (15) and the bound on the covering number in (16), that

$$\begin{aligned} \mathbb{P}\left(\sup_{\mathbf{c} \in \mathcal{B}(R)^M} \left| \mathcal{E}_n(\mathbf{c}; x) - \tilde{\mathcal{E}}_n(\mathbf{c}; x) \right| > \epsilon \right) &\leq \mathcal{N}\left(\mathcal{B}(R)^M, \frac{\epsilon}{16R} \right) \max_{1 \leq \ell \leq N_\epsilon} \mathbb{P}\left(\left| \mathcal{E}_n(\mathbf{a}_\ell; x) - \tilde{\mathcal{E}}_n(\mathbf{a}_\ell, x) \right| > \frac{\epsilon}{2} \right) \\ &\leq 2C_0 \epsilon^{-pM} \exp\left(-\frac{k\epsilon^2}{32R^2(16R^2 + \epsilon)} \right). \end{aligned} \tag{20}$$

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 $\mathcal{B}(R)^M$, we have

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

By Lemma 3, there exists $\delta > 0$ and $L > 0$, depending only on x and f_{XY}, such that for any $\tilde{x}$ with $\|\tilde{x} - x\| \leq \delta$

$$|\mathcal{E}(\mathbf{c}; \tilde{x}) - \mathcal{E}(\mathbf{c}; x)| \leq L\|\tilde{x} - x\|, \quad \text{for any } \mathbf{c} \text{ in } \mathcal{B}(R)^M,$$

and δ can be chosen small enough that

$$m_\delta := \inf\{ f_X(\tilde{x}) \,:\, \|\tilde{x} - x\| \leq \delta \} > 0 \tag{21}$$

by continuity of f_X. (This is actually how δ is taken in the proof of Lemma 3). 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 $\mathcal{B}(R)^M$,

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

Hence, with probability one,

$$\sup_{\mathbf{c} \in \mathbf{B}(R)^M} \left| \tilde{\mathcal{E}}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x) \right| \leq L\|X_{(k,n)}(x) - x\| \mathbf{1}\{\|X_{(k,n)}(x) - x\| \leq \delta\} + 8R^2 \mathbf{1}\{\|X_{(k,n)}(x) - x\| > \delta\}. \tag{22}$$

Then, for any $0 < \epsilon \leq 8R^2$, and since $\sup_{\mathbf{c} \in \mathbf{B}(R)^M} \left| \tilde{\mathcal{E}}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x) \right| \leq 8R^2$ almost surely,

$$\begin{aligned} &\mathbb{P}\left(\sup_{\mathbf{c} \in \mathbf{B}(R)^M} \left| \tilde{\mathcal{E}}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x) \right| > \epsilon \right) \\ &\leq \mathbb{P}\left(\left[L\|X_{(k,n)}(x) - x\| > \frac{\epsilon}{2} \right] \cap \left[\|X_{(k,n)}(x) - x\| \leq \delta \right] \right) + \mathbb{P}\left(8R^2 \mathbf{1}\{\|X_{(k,n)}(x) - x\| > \delta\} > \frac{\epsilon}{2} \right) \\ &\leq \mathbb{P}\left(\|X_{(k,n)}(x) - x\| > \frac{\epsilon}{2L} \right) \mathbf{1}\{\epsilon \leq 2L\delta \wedge 8R^2\} + \mathbb{P}\left(\|X_{(k,n)}(x) - x\| > \delta \right) \mathbf{1}\{\epsilon \leq 8R^2\}. \end{aligned} \tag{23}$$

For any $0 < \eta \leq \delta$, let $p_\eta = \mathbb{P}(\|X - x\| \leq \eta)$. Note that $p_\eta \geq m_\delta \omega_d \eta^d > 0$ by (21). 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 $0 < \eta \leq \delta$

$$\mathbb{P}\left(\|X_{(k,n)}(x) - x\| > \eta\right) \leq \exp\left(-\frac{1}{2}\left(1 - \frac{k-1}{np_\eta}\right)^2 np_\eta\right) \leq \exp\left(-\frac{np_\eta}{4}\right),$$

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

$$\frac{k}{n} \leq \frac{m_\delta \omega_d}{2}\eta^d := C_1 \eta^d.$$

Therefore in this case

$$\mathbb{P}\left(\|X_{(k,n)}(x) - x\| > \eta\right) \leq \exp\left(-\frac{n\eta^d}{4}\right). \tag{24}$$

Hence, by reporting (24) in (23), there exists a constant $C_2 > 0$, depending only on f_{XY} and x, such that, for any $\epsilon > 0$, and any k and n such that $\frac{k}{n} \leq C_1 \epsilon^d$,

$$\mathbb{P}\left(\sup_{\mathbf{c}\in\mathbf{B}(R)^M} \left|\tilde{\mathcal{E}}_n(\mathbf{c};x) - \mathcal{E}(\mathbf{c};x)\right| > \epsilon\right) \leq \left[\exp\left(-\frac{n\epsilon^d}{C_2}\right) + \exp\left(-\frac{n\delta^d}{C_2}\right)\right] \mathbf{1}\{\epsilon \leq 8R^2\}. \tag{25}$$

Combining (20) and (25), we obtain that, for any $\epsilon > 0$, and any k and n such that $\frac{k}{n} \leq C_1\epsilon^d$,

$$\begin{aligned}\mathbb{P}\left(\sup_{\mathbf{c}\in\mathbf{B}(R)^M} |\mathcal{E}_n(\mathbf{c};x) - \mathcal{E}(\mathbf{c};x)| > \epsilon\right) \quad &\leq \quad 2C_0\left(\frac{\epsilon}{2}\right)^{-pM} \exp\left(-\frac{k(\epsilon/2)^2}{32R^2(16R^2+\epsilon/2)}\right) \\ &\quad + \exp\left(-\frac{n(\epsilon/2)^d}{C_2}\right) + \exp\left(-\frac{n\delta^d}{C_2}\right),\end{aligned}$$

and from this, we conclude. □

6.3 Proof of Theorem 1

Fix $\epsilon > 0$. Since $\frac{k}{n} \to 0$ by assumption, by Proposition 1 there is a constant $C > 0$ and a real number $\delta > 0$ depending only on f_{XY} and x such that, for all n large enough,

$$\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}(R)^M} |\mathcal{E}_n(\mathbf{c};x) - \mathcal{E}(\mathbf{c};x)| > \epsilon\right) \leq C\epsilon^{-pM}\exp\left(-\frac{k\epsilon^2}{C}\right) + \exp\left(-\frac{n\epsilon^d}{C}\right) + \exp\left(-\frac{n\delta^d}{C}\right). \tag{26}$$

Under the condition that $\frac{k}{\log n} \to \infty$ as $n \to \infty$, the first term in the right-hand side of (26) is summable for each fixed ϵ, and the second and third terms are summable. Then, applying the Borel-Cantelli lemma, we conclude that $\sup_{\mathbf{c}\in\mathcal{B}(R)^M} |\mathcal{E}_n(\mathbf{c};x) - \mathcal{E}(\mathbf{c};x)| \to 0$ almost surely as n goes to infinity.

6.4 Proof of Theorem 2

Let $\mathbf{c}^\star$ be an optimal quantizer, meaning that $\mathcal{E}(\mathbf{c}^\star, x) = \mathcal{E}^\star$. Following standard arguments, we have

$$\begin{aligned}\mathcal{E}(\mathbf{c}_n^\star; x) - \mathcal{E}^\star(x) &= [\mathcal{E}(\mathbf{c}_n^\star; x) - \mathcal{E}_n(\mathbf{c}_n^\star; x)] + [\mathcal{E}_n(\mathbf{c}_n^\star; x) - \mathcal{E}_n(\mathbf{c}^\star; x)] + [\mathcal{E}_n(\mathbf{c}^\star; x) - \mathcal{E}(\mathbf{c}^\star; x)] \\ &\leq 2 \sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)|, \end{aligned} \tag{27}$$

so that

$$\mathbb{E}\left[\mathcal{E}(\mathbf{c}_n^\star)\right] - \mathcal{E}^\star \leq 2\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)|\right]. \tag{28}$$

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

$$\begin{aligned}\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)|\right] &= \int_0^\infty \mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)| > \epsilon\right) \mathrm{d}\epsilon \\ &\leq a + \int_a^\infty \mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)| > \epsilon\right) \mathrm{d}\epsilon. \end{aligned} \tag{29}$$

By Proposition 1, 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, there exist constants $C_0(R, p, M) > 0$, $C_1(f_{XY}, x) > 0$, $C_2(f_{XY}, x) > 0$ and $\delta(f_{XY}, x) > 0$ such that, for any $\epsilon > 0$, and any k and n with $\frac{k}{n} \leq C_1\epsilon^d$,

$$\begin{aligned}&\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)| > \epsilon\right) \\ &\quad\leq C_0\epsilon^{-pM} \exp\left(-\frac{k\epsilon^2}{C_2}\right) + \exp\left(-\frac{n\epsilon^d}{C_2}\right) + \exp\left(-\frac{n\delta^d}{C_2}\right).\end{aligned}$$

Let $a > 0$ and suppose that $\frac{k}{n} \leq C_1 a^d$, Then

$$\begin{aligned}&\int_0^\infty \mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c}; x) - \mathcal{E}(\mathbf{c}; x)| > \epsilon\right) \mathrm{d}\epsilon \\ &\quad\leq a + C_0 \int_a^\infty \epsilon^{-pM} \exp\left(-\frac{k\epsilon^2}{C_2}\right) \mathrm{d}\epsilon + \int_a^\infty \exp\left(-\frac{n\epsilon^d}{C_2}\right) \mathrm{d}\epsilon + 8R^2 \exp\left(-\frac{n\delta^d}{C_2}\right).\end{aligned} \tag{30}$$

Now we proceed to bound the two integrals in (30)

For the first integral in (30), we have

$$\int_a^\infty \epsilon^{-pM} \exp\left(-\frac{k\epsilon^2}{C_2}\right) \mathrm{d}\epsilon \leq \frac{1}{a^{pM+1}} \frac{C_2}{2k} \exp\left(-\frac{ka^2}{C_2}\right),$$

where we have used the inequality $\int_b^\infty e^{-u^2/2}\mathrm{d}u \leq \frac{1}{b}e^{-b^2/2}$. The upper bound in the above inequality converges to 0 whenever $a \to 0$ no faster than $1/\sqrt{k}$. Therefore with the choice of $a = c\sqrt{\frac{\log k}{k}}$ we have that

$$\int_a^\infty \epsilon^{-pM} \exp\left(-\frac{k\epsilon^2}{C_2}\right) \mathrm{d}\epsilon \leq \frac{C_2}{2kc^{pM+1}} \left(\frac{k}{\log k}\right)^{(pM+1)/2} k^{-c^2/C_2} = o\left(\sqrt{\frac{\log k}{k}}\right) \tag{31}$$

with the choice of $c = \sqrt{\frac{C_2(pM+1)}{2}}$.

For the second integral in (30), we have

$$\int_a^\infty \exp\left(-\frac{n\epsilon^d}{C_2}\right) \mathrm{d}\epsilon \leq \int_\gamma^\infty \exp\left(-\frac{n\epsilon^d}{C_2}\right) \mathrm{d}\epsilon = \int_{a(n/C_2)^{1/d}}^\infty e^{-u^d} \left(\frac{n}{C_2}\right)^{-\frac{1}{d}} \mathrm{d}u \leq \frac{C_2}{nda^{d-1}} e^{-\frac{na^d}{C_2}}.$$

Since $\frac{k}{n} \leq C_1 a^d$ and $a = c\sqrt{\frac{\log k}{k}}$,

$$\int_a^\infty \exp\left(-\frac{n\epsilon^d}{C_2}\right) \mathrm{d}\epsilon \leq \frac{cC_1C_2}{d} \frac{\sqrt{\log k}}{k^{3/2}} \exp\left(-\frac{k}{C_1C_2}\right) = o\left(\sqrt{\frac{\log k}{k}}\right). \tag{32}$$

Reporting (31) and (32) in (30), we deduce that

$$\begin{aligned}\int_0^\infty \mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c};x) - \mathcal{E}(\mathbf{c};x)| > \epsilon\right) \mathrm{d}\epsilon \\ \leq \sqrt{\frac{C_2(pM+1)}{2}}\sqrt{\frac{\log k}{k}} + 8R^2 \exp\left(-\frac{n\delta^d}{C_2}\right) + o\left(\sqrt{\frac{\log k}{k}}\right)\end{aligned}$$

for all k and n satisfying $\frac{k}{n} \leq \left(\frac{C_2(pM+1)}{2}\right)^{\frac{d}{2}} \left(\frac{\log k}{k}\right)^{\frac{d}{2}}$ and from this we conclude.

6.5 Proof of Theorem 3

Note first that, since $\mathcal{C}^\star(x)$ is included in the compact $\mathcal{B}_p(R)^M$, and since the Hessian matrix $\mathbf{H}(\mathbf{c}^\star)$ is positive definite at each $\mathbf{c}^\star$ in $\mathcal{C}^\star(x)$, $\mathcal{C}^\star(x)$ is finite. Therefore there exists $\lambda^\star > 0$ such that the eigenvalues of $\mathbf{H}(\mathbf{c}^\star)$ are included in $(\lambda^\star, \infty)$ for all $\mathbf{c}^\star$ in $\mathcal{C}^\star(x)$. Second, under the regularity assumptions of Theorem 3, the gradient of $\mathcal{E}(\mathbf{c};x)$ vanishes on $\mathcal{C}^\star$.

Given $\mathbf{c}$ in $\mathcal{B}_p(R)^M$, let $\mathbf{c}^\star_{\mathbf{c}}$ be a configuration in $\mathcal{C}^\star$ closest to $\mathbf{c}$, that is, such that $\|\mathbf{c} - \mathbf{c}^\star_{\mathbf{c}}\| = \inf_{\tilde{\mathbf{c}}\in\mathcal{C}^\star} \|\mathbf{c} - \tilde{\mathbf{c}}\|$. Using a Taylor expansion of $\mathcal{E}(\mathbf{c};x)$ at $\mathbf{c}^\star_{\mathbf{c}}$ in $\mathcal{C}^\star$ we have that $\mathcal{E}(c;x) - \mathcal{E}^\star(x) = \frac{1}{2}\mathbf{H}(\mathbf{c}^\star_{\mathbf{c}})(\mathbf{c} - \mathbf{c}^\star_{\mathbf{c}})^{\otimes 2} + o(\|\mathbf{c} - \mathbf{c}^\star_{\mathbf{c}}\|^2)$ and $\frac{1}{2}\mathbf{H}(\mathbf{c}^\star_{\mathbf{c}})(\mathbf{c} - \mathbf{c}^\star_{\mathbf{c}})^{\otimes 2} \geq \frac{\lambda^\star}{2}\|\mathbf{c} - \mathbf{c}^\star_{\mathbf{c}}\|^2$. Hence there exists $\eta > 0$ such that

$$\frac{\lambda^\star}{4}\mathrm{dist}(\mathbf{c}, \mathcal{C}^\star)^2 \leq \mathcal{E}(\mathbf{c};x) - \mathcal{E}^\star(x), \quad \text{for any } \mathbf{c} \text{ such that } \mathcal{E}(\mathbf{c};x) - \mathcal{E}^\star(x) \leq \eta. \tag{33}$$

For any $\mathbf{c}^\star_n$ in $\mathcal{C}^\star_n$, we have

$$\mathcal{E}(\mathbf{c}^\star_n) \leq \mathcal{E}_n(\mathbf{c}^\star_n) + \sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c},x) - \mathcal{E}(\mathbf{c},x)| \leq \mathcal{E}^\star + 2\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c},x) - \mathcal{E}(\mathbf{c},x)|.$$

So on the probability event that $\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c},x) - \mathcal{E}(\mathbf{c},x)| \leq \frac{\eta}{2}$, we deduce using (33) that

$$\inf_{\mathbf{c}^\star\in\mathcal{C}^\star} \|\mathbf{c}^\star_n - \mathbf{c}^\star\| \leq \frac{4}{\lambda^\star}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c},x) - \mathcal{E}(\mathbf{c},x)|\right)^{\frac{1}{2}}.$$

Hence, for any $\epsilon > 0$,

$$\begin{aligned}&\mathbb{P}\left(\inf_{\mathbf{c}^\star\in\mathcal{C}^\star} \|\mathbf{c}^\star_n - \mathbf{c}^\star\| > \epsilon\right) \\ &\quad\leq \mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c},x) - \mathcal{E}(\mathbf{c},x)| > \frac{(\lambda^\star\epsilon)^2}{16}\right) + \mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{B}_p(R)^M} |\mathcal{E}_n(\mathbf{c},x) - \mathcal{E}(\mathbf{c},x)| > \frac{\eta}{2}\right).\end{aligned} \tag{34}$$

Combining (34) with (26), we deduce that $\sum_{n\geq 1} \mathbb{P}\left(\inf_{\mathbf{c}^\star\in\mathcal{C}^\star} \|\mathbf{c}^\star_n - \mathbf{c}^\star\| > \epsilon\right) < \infty$ for any $\epsilon > 0$ and from this we conclude using the Borel-Cantelli lemma.

A Technical Lemmas

The following Lemma is Lemma 6 in Devroye (1982).

Lemma 2. *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 $\epsilon > 0$,*

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

The following lemma states that, under the conditions of the main theorem, the map $x \mapsto \mathcal{E}(\mathbf{c}; x)$ is locally Lipschitz with respect to x for each fixed configuration $\mathbf{c}$ of quantization points.

Lemma 3. *Let (X, Y) be a pair of random vectors such that $\|Y\| \leq R$ almost surely and with joint density f on $\mathbb{R}^d \times \mathbb{R}^p$. Suppose that f is Lipschitz on $\mathbb{R}^d \times \mathbb{R}^p$. Let x be a point with $f_X(x) > 0$, where f_X denotes the marginal density of X. There exists $\delta := \delta(x, f_X) > 0$ and $L := L(x, f_X) > 0$ such that $|\mathcal{E}(\mathbf{c}; \tilde{x}) - \mathcal{E}(\mathbf{c}; x)| \leq L\|\tilde{x} - x\|$ for all $\mathbf{c} := (c_1, \ldots, c_M) \in \mathcal{B}(R)^M$.*

Proof. Since $f_X(x) > 0$, there exists $\delta := \delta(x, f_X) > 0$ such that $\underline{m} := \inf\{f_X(\tilde{x}) : \tilde{x} \in \mathcal{B}(x, \delta)\} > 0$. Let $\overline{m} = \sup\{f(\tilde{x}, y) : \tilde{x} \in \mathcal{B}(x, \delta),\ y \in \mathbb{R}^p\}$ and note that $\overline{m} < \infty$. Given $\mathbf{c} \in \mathcal{B}(R)^M$, for any $\tilde{x} \in \mathcal{B}(x, \delta)$, we have

$$\mathcal{E}(\mathbf{c}; \tilde{x}) - \mathcal{E}(\mathbf{c}; x) = \int_{\mathbb{R}^p} \min_{1\leq j\leq M} \|Y - c_j\|^2 \left(\frac{f(\tilde{x}, y)}{f_X(\tilde{x})} - \frac{f(x, y)}{f_X(x)}\right) \mathrm{d}y. \tag{35}$$

Let L_f be a Lipschitz constant for f. Since $\|Y\| \leq R$ almost surely and since f is Lipschitz, f_X is Lipschitz with constant no more than $\omega_p R^p L_f =: L_{f_X}$. Then, for any $\tilde{x} \in \mathcal{B}(x, \delta)$,

$$\begin{aligned}\left|\frac{f(\tilde{x}, y)}{f_X(\tilde{x})} - \frac{f(x, y)}{f_X(x)}\right| &\leq \frac{1}{f_X(\tilde{x})}|f(\tilde{x}, y) - f(x, y)| + \frac{f(x, y)}{f_X(\tilde{x}) f_X(x)}|f_X(\tilde{x}) - f_X(x)| \\ &\leq \frac{1}{\underline{m}}\|\tilde{x} - x\| + \frac{\overline{m}}{\underline{m}^2} L_{f_X}\|\tilde{x} - x\|. \end{aligned} \tag{36}$$

Note that $\min_{1\leq j\leq M} \|Y - c_j\|^2 \leq 4R^2$ almost surely. Using this, we deduce from (35) and (36) that

$$|\mathcal{E}(\mathbf{c}; \tilde{x}) - \mathcal{E}(\mathbf{c}; x)| \leq 4R^2 \mathrm{Vol}_p(\mathcal{B}(R)) \left[\frac{1}{\underline{m}} + \frac{\overline{m} L_{f_X}}{\underline{m}^2}\right] \|\tilde{x} - x\| =: L\|\tilde{x} - x\|.$$

□

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 (2015). On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*. In press.

Bartlett, P., T. Linder, and G. Lugosi (1998). The minimax distrotion 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 (2014). Nonparametric modal regression. Preprint, arXiv:1412.1716.

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

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.

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.

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.

Schneider, R. (2014). *Convex bodies: the Brunn-Minkowski theory*, Volume 151 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge.

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.