Noisy Quantization: theory and practice

Camille Brunet, Sébastien Loustau

Noisy quantization
theory and practice

Camille Brunet and Sébastien Loustau*

Université d'Angers, LAREMA

Abstract

The effect of errors in variables in quantization is investigated. Given a noisy sample $Z_i = X_i + \epsilon_i, i = 1, \dots, n$, where $(X_i)_{i=1,\dots,n}$ are i.i.d. with law P, we want to find the best approximation of the probability distribution P with $k \geq 1$ points called codepoints. We prove general excess risk bounds with fast rates for an empirical minimization based on a deconvolution kernel estimator. These rates depend on the behaviour of the density of P and the asymptotic behaviour of the characteristic function of the noise ϵ. This general study can be applied to the problem of k-means clustering with noisy data. For this purpose, we introduce a deconvolution k-means stochastic minimization which reaches fast rates of convergence under standard Pollard's regularity assumptions.

We also introduce a new algorithm to deal with k-means clustering with errors in variables. Following the theoretical study, the algorithm mixes different tools from the inverse problem literature and the machine learning community. Coarsely, it is based on a two-step procedure: (1) a deconvolution step to deal with noisy inputs and (2) Newton's iterations as the popular k-means.

1 Introduction

The goal of empirical vector quantization (Graf and Luschgy [2000]) or clustering (Hartigan [1975]) is to replace multivariate data by an efficient and compact representation, which allows one to reconstruct the original observations with a certain accuracy. The problem was originated in signal processing and has many applications in cluster analysis or information theory. The basic statistical model could be described as follows. Given independent and identically distributed (i.i.d.) $\mathbb{R}^d$-random variables $X_1, \dots, X_n$, with unknown law P with density f on $\mathbb{R}^d$ with respect to the Lebesgue measure, we want to propose a quantizer $g : x \in \mathbb{R}^d \mapsto \{1, \dots, k\}$, where $k \geq 1$ is a given integer. The most investigated example of such a framework is probably cluster analysis, where the aim is to build k clusters of the set of observations $X_1, \dots, X_n$. In this framework, a quantizer g assigns cluster $g(x) \in \{1, \dots, k\}$ to an observation $x \in \mathbb{R}^d$.

However, in many real-life situations, direct data $X_1, \dots, X_n$ are not available and measurement errors may occur. Then, we observe only a corrupted sample $Z_i = X_i + \epsilon_i, i = 1, \dots n$ with noisy distribution $\tilde{P}$, where $\epsilon_1, \dots, \epsilon_n$ are i.i.d. independent of $X_1, \dots, X_n$ with density η. The problem of noisy empirical vector quantization is to represent compactly and efficiently the measure P when a contaminated empirical version $Z_1, \dots, Z_n$ is observed. This problem is a particular case of inverse statistical learning (see Loustau [2013]), and is known to be an inverse problem. To our best knowledge, it has not been yet considered in the literature. This paper tries to fill this gap by giving (1) a theoretical study of this problem and (2) an algorithm to deal with clustering from a noisy dataset.

A quiet natural habit in statistical learning is to endow clustering or empirical vector quantization into the general and extensively studied problem of empirical risk minimization (see Vapnik [2000], Bartlett and Mendelson [2006], Koltchinskii [2006]). This is exactly the guiding thread of this contribution. For this purpose, we consider a metric space $(\mathcal{G}, d)$, where $\mathcal{G}$ is a class of classifiers or quantizers (possibly infinite-dimensional space). We also introduce a loss function $\ell : \mathcal{G} \times \mathbb{R}^d$ where $\ell(g, x)$ measures the loss of g at point x. In such a framework, the measure of the accuracy of g will be evaluated thanks to a distortion or risk given by:

$$R(g) = \mathbb{E}_P \ell(g, X) = \int_{\mathbb{R}^d} \ell(g, x) f(x) dx. \tag{1.1}$$

*Corresponding author, loustau@math.univ-angers.fr

To minimize the risk (1.1) given data $X_1, \dots, X_n$, it is extremely standard to consider an empirical risk minimizer (ERM) defined as:

$$\hat{g}_n \in \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \ell(g, X_i). \tag{1.2}$$

Since the pioneer's work of Vapnik, many authors have investigated the statistical performances of (1.2) in such a generality. In clustering, we want to construct a vector of codepoints $\mathbf{c} = (c_1, \dots, c_k) \in \mathbb{R}^{dk}$ to represent efficiently with $k \geq 1$ centers a set of observations $X_1, \dots, X_n \in \mathbb{R}^d$. In this framework, it is standard to consider the loss function $\ell : \mathbb{R}^{dk} \times \mathbb{R}^d$ defined as:

$$\ell(\mathbf{c}, x) := \min_{j=1,\dots,k} \|x - c_j\|^2,$$

where $\|\cdot\|$ stands for the Euclidean norm in $\mathbb{R}^d$. In this case, the empirical risk minimizer is given by:

$$\hat{c}_n = \arg\min_{\mathbf{c} \in \mathbb{R}^{dk}} \sum_{i=1}^{n} \min_{j=1,\dots k} \|X_i - c_j\|^2, \tag{1.3}$$

and is known as the popular k-means (Pollard [1981],Pollard [1982]).

In this paper, we propose to adopt a comparable strategy in the presence of noisy measurements. Since we observe a corrupted sample $Z_i = X_i + \epsilon_i$, $i = 1, \dots, n$, the empirical risk minimization (1.2) is not available. However, we can introduce a deconvolution step in the estimation procedure by constructing a kernel deconvolution estimator of the density f of the form:

$$\hat{f}_\lambda(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{Z_i - x}{\lambda} \right), \tag{1.4}$$

where $\mathcal{K}_\eta$ is a deconvolution kernel [1] and $\lambda = (\lambda_1, \dots, \lambda_d) \in \mathbb{R}_d^+$ is a multivariate bandwidth (see Section 2 for details). Given this estimator, we construct an empirical risk $R_n^\lambda(\cdot)$ by plugging (1.4) into the true risk (1.1) (see Section 2 for a precise definition). The idea was originated in Loustau and Marteau [2012] for discriminant analysis. In the sequel, a solution of this stochastic minimization is written:

$$\hat{g}_n^\lambda \in \arg\min_{g \in \mathcal{G}} R_n^\lambda(g). \tag{1.5}$$

In the first theoretical part of this work, we study the statistical performances of $\hat{g}_n^\lambda$ in (1.5) in terms of excess risk bounds. In Section 3 we state that with high probability:

$$R(\hat{g}_n^\lambda) - \inf_{g \in \mathcal{G}} R(g) \leq \psi(n), \tag{1.6}$$

where $\psi(n) \to 0$ as $n \to \infty$ is called the rate of convergence. It is a function of the complexity of $\mathcal{G}$, the behaviour of the density f, and the density of the noise η. In this paper, the behaviour of f depends on two different assumptions : a margin assumption and a regularity assumption. The margin assumption is related to the difficulty of the problem whereas the regularity assumption will be expressed in terms of anisotropic Hölder spaces. The stochastic minimization (1.5), as well as statement (1.6), are applied to the framework of clustering. Eventually, we propose to design a new algorithm to deal with finite dimensional clustering with errors in variables.

The paper is organized as follows. In Section 2, we present the general method and the main assumptions on the density η (noise assumption), the kernel in (1.4) and the density f (regularity and margin assumptions). We state the main theoretical results in Section 3, which consists in excess risk bounds as in (1.6) with fast rates of convergence (i.e. with $\psi(n) = o(1/\sqrt{n})$). These results are applied in Section 4 for the problem of finite dimensional clustering with k-means. The proposed algorithm, called noisy k-means, is based on a two-step procedure : the construction of a kernel deconvolution estimator of the density f and Newton's type iterations as the popular k-means. A complete simulation study is proposed in Section 5 to illustrate the efficiency of the method in comparison with standard k-means in the presence of errors-in-variables. Section 6 concludes the paper with a discussion whereas Section 7-8 give detailled proofs of the main results.

1. With a slight abuse of notations, we write in (1.4), for any $x = (x_1, \dots, x_d)$, $Z_i = (Z_{1,i}, \dots, Z_{d,i}) \in \mathbb{R}^d$:

$$\frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{Z_i - x}{\lambda} \right) = \frac{1}{\Pi_{i=1}^d \lambda_i} \mathcal{K}_\eta \left(\frac{Z_{1,i} - x_1}{\lambda_1}, \dots, \frac{Z_{d,i} - x_d}{\lambda_d} \right).$$

2 Deconvolution ERM

2.1 Construction of the estimator

The deconvolution ERM introduced in this paper is originally due to Loustau and Marteau [2012] in discriminant analysis (see also Loustau [2013] for such a generality in supervised classification). The main idea of the construction is to estimate the true risk (1.1) thanks to a deconvolution kernel as follows.

Let us introduce $\mathcal{K} = \prod_{i=1}^{d} \mathcal{K}_j : \mathbb{R}^d \to \mathbb{R}$ a d-dimensional function defined as the product of d unidimensional function $\mathcal{K}_j$. Besides, $\mathcal{K}$ (and also η) belongs to $L_2(\mathbb{R}^d)$ and admits a Fourier transform. Then, if we denote by $\lambda = (\lambda_1, \dots, \lambda_d)$ a set of (positive) bandwidths and by $\mathcal{F}[\cdot]$ the Fourier transform, we define $\mathcal{K}_\eta$ as:

$$\begin{aligned} \mathcal{K}_\eta \quad &: \quad \mathbb{R}^d \to \mathbb{R} \\ &\quad t \mapsto \mathcal{K}_\eta(t) = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\mathcal{K}](\cdot)}{\mathcal{F}[\eta](\cdot/\lambda)}\right](t). \end{aligned} \tag{2.1}$$

Given this deconvolution kernel, we construct an empirical risk by plugging (1.4) into the true risk $R(g)$ to get a so-called deconvolution empirical risk given by:

$$R_n^\lambda(g) = \frac{1}{n}\sum_{i=1}^{n} \ell_\lambda(g, Z_i) \text{ where } \ell_\lambda(g, Z_i) = \int_K \ell(g, x)\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_i - x}{\lambda}\right)dx. \tag{2.2}$$

Note that for technicalities, we restrict ourselves to a compact set $K \subset \mathbb{R}^d$ and study the risk minimization (1.1) only in K. Consequently, in this paper, we only provide a control of the true risk (1.1) restricted to K, namely the truncated risk:

$$R_K(g) = \int_K \ell(g, x) f(x) dx.$$

This restriction has been considered in Mammen and Tsybakov [1999] (or more recently in Loustau and Marteau [2012]). It is important to note that when f has compact support, we can see coarsely that $R_K(g) = R(g)$ for a great enough K. In the sequel, for simplicity, we write $R(\cdot)$ for the truncated risk defined above.

2.2 Assumptions

For the sake of simplicity, we restrict ourselves to moderately or mildly ill-posed inverse problem as follows. We introduce the following noise assumption **(NA)**:

(NA): There exist $(\beta_1, \dots, \beta_d)' \in \mathbb{R}_+^d$ such that:

$$|\mathcal{F}[\eta](t)| \sim \Pi_{i=1}^{d} |t_i|^{-\beta_i}, \text{as } |t_i| \to +\infty, \forall i \in \{1, \dots, d\}.$$

Moreover, we assume that $\mathcal{F}[\eta](t) \neq 0$ for all $t = (t_1, \dots, t_d) \in \mathbb{R}^d$.

Assumption **(NA)** deals with the asymptotic behaviour of the characteristic function of the noise distribution. These kind of restrictions are standard in deconvolution problems for $d = 1$ (see Fan, Meister, Butucea [1991, 2009, 2007]). In this contribution, we only deal with d-dimensional mildly ill-posed deconvolution problems, which corresponds to a polynomial decreasing of $\mathcal{F}[\eta]$ in each direction. For the sake of brevity, we do not consider severely ill-posed inverse problems (exponential decreasing)or possible intermediates (e.g. a combination of polynomial and exponential decreasing functions). Recently, Comte and Lacour [2012] propose such a study in the context of multivariate deconvolution. In our framework, the rates in these cases could be obtained through the same steps.

We also require the following assumptions on the kernel $\mathcal{K}$.

(K1) There exists $S = (S_1, \dots, S_d) \in \mathbb{R}_d^+$, $K_1 > 0$ such that kernel $\mathcal{K}$ satisfies

$$\text{supp}\mathcal{F}[\mathcal{K}] \subset [-S, S] \text{ and } \sup_{t \in \mathbb{R}^d} |\mathcal{F}[\mathcal{K}](t)| \leq K_1,$$

where $\operatorname{supp} g = \{x : g(x) \neq 0\}$ and $[-S, S] = \bigotimes_{i=1}^{d} [-S_i, S_i]$.

This assumption is trivially satisfied for different standard kernels, such as the *sinc* kernel. It arises for technicalities in the proofs and can be relaxed using a finer algebra. Moreover, in the sequel, we consider a kernel of order m, for a particular $m \in \mathbb{N}^d$.

K(m) The kernel $\mathcal{K}$ is of order $m = (m_1, \dots, m_d) \in \mathbb{N}^d$, i.e.
— $\int_{\mathbb{R}^d} \mathcal{K}(x)dx = 1$
— $\int_{\mathbb{R}^d} \mathcal{K}(x)x_j^k dx = 0$, $\forall k \leq m_j$, $\forall j \in \{1, \dots, d\}$.
— $\int_{\mathbb{R}^d} |\mathcal{K}(x)||x_j|^{m_j} dx < K_2$, $\forall j \in \{1, \dots, d\}$.

The construction of kernel of order m satisfying **(K1)** could be done by using for instance compactly supported wavelets, such as Meyer's wavelets (see Mallat [2009]). The condition **K(m)** is standard in nonparametric kernel estimation and allows to get satisfying approximations using the following assumption over the regularity of the density f.

Definition 2.1. *For some $s = (s_1, \dots, s_d) \in \mathbb{R}_d^+$, $L > 0$, we say that f belongs to the anisotropic Hölder space $\mathcal{H}(s, L)$ if the following holds:*
— *the function f admits derivatives with respect to x_j up to order $\lfloor s_j \rfloor$, where $\lfloor s_j \rfloor$ denotes the largest integer less than s_j.*
— *$\forall j = 1, \dots, d$, $\forall x \in \mathbb{R}^d$, $\forall x'_j \in \mathbb{R}$, the following Lipschitz condition holds:*

$$\left| \frac{\partial^{\lfloor s_j \rfloor}}{(\partial x_j)^{\lfloor s_j \rfloor}} f(x_1, \dots, x_{j-1}, x'_j, x_{j+1}, \dots, x_d) - \frac{\partial^{\lfloor s_j \rfloor}}{(\partial x_j)^{\lfloor s_j \rfloor}} f(x) \right| \leq L|x'_j - x_j|^{s_j - \lfloor s_j \rfloor}.$$

If a function f belongs to the anisotropic Hölder space $\mathcal{H}(s, L)$, f has an Hölder regularity s_j in each direction $j = 1, \dots, d$. As a result, it can be well-approximated pointwise using a d-dimensional Taylor formula.

3 Main results

It is well-known that the behaviour of the rates of convergence $\psi(n)$ in (1.6) is governed by the size of $\mathcal{G}$. In this paper, the size of the hypothesis space will be quantified in terms of ϵ-entropy with bracketing of the metric space $(\{\ell(g), g \in \mathcal{G}\}, L_2)$ as follows.

Definition 3.1. *Given a metric space $(\mathcal{F}, d)$ and a real number $\epsilon > 0$, the ϵ-entropy with bracketing of $(\mathcal{F}, d)$ is the quantity $\mathcal{H}_B(\mathcal{F}, \epsilon, d)$ defined as the logarithm of the minimal integer $N_B(\epsilon)$ such that there exist pairs $(f_j, g_j) \in \mathcal{F} \times \mathcal{F}$, $j = 1, \dots, N_B(\epsilon)$ such that $f_j \leq g_j$, $d(f_j, g_j) \leq \epsilon$, and where for any $f \in \mathcal{F}$, there exists a pair (f_j, g_j) with $f_j < f < g_j$.*

This notion of complexity allows to obtain uniform concentration inequalities (see Van De Geer [2000] or van der Vaart and Weelner [1996]). Indeed, to reach fast rates of convergence (i.e. faster than $n^{-1/2}$), what really matters is not the total size of the hypothesis space but rather the size of a subclass of $\mathcal{G}$, made of functions with small errors. In this paper, we use an iterative localization principle originally introduced in Koltchinskii and Panchenko [2000] (see also Koltchinskii [2006] for such a generality). More precisely, we consider functions in $\mathcal{G}$ with small excess risk as follows:

$$\mathcal{G}(\delta) = \{g \in \mathcal{G} : R(g) - \inf_{g \in \mathcal{G}} R(g) \leq \delta\}.$$

Originally, Mammen and Tsybakov [1999] (see also Tsybakov [2004]) formulated an useful condition to get fast rates of convergence in classification. This assumption is known as the margin assumption and has been generalized by Bartlett and Mendelson [2006]. coarsely speaking, a margin assumption guarantees a nice relationship between the variance and the expectation of any function of the excess loss class. In this section, we assume the following margin assumption:

Margin Assumption MA(κ) There exists some $\kappa \geq 1$ such that:

$$\forall g \in \mathcal{G}, \|\ell(g, \cdot) - \ell(g^*(g), \cdot)\|_{L_2(K)}^2 \leq \kappa_0 \left[R(g) - \inf_{g \in \mathcal{G}} R(g) \right]^{1/\kappa},$$

for some $\kappa_0 > 0$, where $g^*(g) \in \arg\min_{g^* \in \mathcal{M}} d(g, g^*)$ and $\mathcal{M}$ is the set of oracles.

Gathering with a suitable concentration inequality applied to the class $\mathcal{G}(\delta)$, this margin assumption is used to get fast rates. Note that provided that $\ell(g,\cdot)$ is bounded, $\mathbf{MA}(\kappa)$ implies $\mathbf{MA}(\kappa')$ for any $\kappa' \geq \kappa$. Interestingly, in the framework of finite dimensional clustering with k-means, Levrard [2012] proposes to give a sufficient condition to have $\mathbf{MA}(\kappa)$ with $\kappa = 1$. This condition is related to the geometry of f with respect to the optimal clusters and gives well-separated classes. It allows to interpret $\mathbf{MA}(\kappa)$ exactly as the popular margin assumption in supervised classification (see Tsybakov [2004]). In the sequel, we call the parameter κ in $\mathbf{MA}(\kappa)$ the margin parameter.

Recently, Lecué and Mendelson [2012] points out that one could wish non-exact oracle inequalities with fast rates under a weaker assumption. The idea is to relax significantly the margin assumption and use the loss class $\{\ell(g), g \in \mathcal{G}\}$ in $\mathbf{MA}(\kappa)$ instead of the excess loss class $\{\ell(g) - \ell(g^*), g \in \mathcal{G}\}$. This framework is not considered in this paper for concision but we refer the interested reader to Loustau [2013] for completeness.

3.1 A general excess risk bound

We are now on time to state the main risk bound.

Theorem 3.1. *Suppose* ***(NA)***, ***(K1)***, *and* ***MA*** (κ) *holds for some margin parameter* $\kappa \geq 1$. *Suppose* $f \in \mathcal{H}(s, L)$ *for some* $s \in \mathbb{R}_d^+$ *and* ***K*** (m) *holds with* $m = \lfloor s \rfloor$. *Suppose there exists* $0 < \rho < 1$, $c > 0$ *such that for every* $\epsilon > 0$*:*

$$\mathcal{H}_B(\{\ell(g), g \in \mathcal{G}\}, \epsilon, L_2) \leq c\epsilon^{-2\rho}. \tag{3.1}$$

Then, for any $t > 0$, *there exists some* $n_0(t) \in \mathbb{N}^*$ *such that for any* $n \geq n_0(t)$, *with probability greater than* $1 - e^{-t}$, *the deconvolution ERM* $\hat{g}_n^\lambda$ *is such that:*

$$R(\hat{g}_n^\lambda) - \inf_{g \in \mathcal{G}} R(g) \leq C n^{-\tau_d(\kappa,\rho,\beta,s)},$$

where $C > 0$ *is independent of* n *and* $\tau_d(\kappa, \rho, \beta, s)$ *is given by:*

$$\tau_d(\kappa, \rho, \beta, s) = \frac{\kappa}{2\kappa + \rho - 1 + (2\kappa - 1)\sum_{j=1}^d \beta_j / s_j},$$

and $\lambda = (\lambda_1, \ldots, \lambda_d)$ *is chosen as:*

$$\lambda_j \approx n^{-\frac{2\kappa - 1}{2\kappa s_j}\tau_d(\kappa,\rho,\beta,s)}, \forall j = 1, \ldots d.$$

The proof of this result is postponed to Section 7. We list some remarks below.

Remark 3.1 (Comparison with Koltchinskii [2006] or Mammen and Tsybakov [1999])**.** *This result gives the order of the rate of convergence in the presence of errors in variables. The risk of the estimator* $\hat{g}_n^\lambda$ *mimics the risk of the best candidate in* $\mathcal{G}$, *up to this rate. The price to pay for the error-in-variables model depends on the asymptotic behaviour of the characteristic function of the noise distribution. If* $\beta = 0 \in \mathbb{R}^d$ *in the noise assumption* ***(NA)***, *the residual term in Theorem 3.1 satisfies:*

$$\psi(n) = n^{-\frac{\kappa}{2\kappa + \rho - 1}}.$$

It corresponds to the standard fast rates in the noise-free case stated (see Koltchinskii [2006] for such a generality or Mammen and Tsybakov [1999] in discriminant analysis).

Remark 3.2 (Comparison with Loustau [2013])**.** *In comparison with Loustau [2013], these rates deal with an anisotropic behaviour of the density* f. *If* $s_j = s$ *for any direction, we obtain the same asymptotics as in Loustau [2013] for supervised classification, namely:*

$$\psi(n) = n^{-\frac{\kappa s}{s(2\kappa + \rho - 1) + (2\kappa - 1)\sum_{j=1}^d \beta_j}}.$$

The result of Theorem 3.1 gives a generalization of Loustau [2013] to the anisotropic case, in an unsupervised framework. Moreover, it allows to deal with a non unique oracle g^* *by introducing a more complicated geometry in the proofs.*

Remark 3.3 (The anisotropic case is of practical interest)**.** *The result of Theorem 3.1 gives some insights into the noisy quantization problem with an anisotropic density* f. *In this problem, due to the anisotropic behaviour of the density, the choice of the multidimensional bandwidths* λ_j, $j = 1, \ldots, d$ *are more complicated. This result is of practical interest since it allows to consider different bandwidth coordinates for the deconvolution ERM. In finite dimensional noisy clustering with* $k \geq 2$, *this situation arises when the optimal centers are not uniformly distributed. This problem has not been treated in Loustau [2013] or Loustau and Marteau [2012].*

Remark 3.4 (Fast rates). *The most favorable case arises when $\rho \to 0$ and β is small, whereas at the same time density f has sufficiently high Hölder exponents s_j. Indeed, fast rates occur when $\tau_d(\kappa, \rho, \beta, s) \geq 1/2$, or equivalently, $(2\kappa - 1)\sum \beta_j/s_j < 1 - \rho$. If $\rho = 0$ and $\kappa = 1$ (see the particular case of Section 3.2), we have the following condition to get fast rates:*

$$\sum_{j=1}^{d} \frac{\beta_j}{s_j} < 1.$$

Remark 3.5 (Choice of λ). *The optimal choice of λ in Theorem 3.1 optimizes a bias variance decomposition as in Loustau [2013]. This choice depends on unknown parameters such as the margin parameter κ, the Hölder exponents $(s_1, \ldots, s_d)$ of the density f and the degree of illposedness β. A challenging open problem is to derive adaptive choice of λ to lead to adaptive fast rates of convergence. This is the purpose of future works.*

Remark 3.6 (Comparison with Comte and Lacour [2012]). *It is also important to note that the optimal choice of the multivariate bandwidth λ does not coincide with the optimal choice of the bandwidth in standard nonparametric anisotropic density deconvolution. Indeed, it is stated in Comte and Lacour [2012] that under the same regularity and ill-posedness assumptions, the optimal choice of the bandwidth $\lambda = (\lambda_1, \ldots, \lambda_d)$ has the following form:*

$$\lambda_u \approx n^{-\frac{1}{s_u\left(2+\sum_{j=1}^{d}\frac{2\beta_j+1}{s_j}\right)}}.$$

The proposed asymptotic optimal calibration of Theorem 3.1 is rather different. It depends explicitly on parameter ρ, which measures the complexity of the decision set $\mathcal{G}$, and the margin parameter $\kappa \geq 1$. It shows rather well that our bandwidth selection problem is not equivalent to standard nonparametric estimation problems. It illustrates one more time that our procedure is not a plug-in procedure.

3.2 Application to noisy clustering

One of the most popular issue in data mining or machine learning is to learn clusters from a big cloud of data. This problem is known as clustering. It has received many attention in the last decades. In this paragraph, we apply the general upper bound of Theorem 3.1 to the framework of noisy clustering. To frame the problem of noisy clustering into the general study of this paper, we first introduce the following notation. Given some known integer $k \geq 2$, let us consider $\mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{C}$ the set of possible codebooks, where $\mathcal{C} \subseteq \mathbb{R}^{dk}$ is compact. The loss function $\ell : \mathbb{R}^{dk} \times \mathbb{R}^d$ is defined as:

$$\ell(\mathbf{c}, x) = \min_{j=1,\ldots k} \|x - c_j\|^2,$$

where $\|\cdot\|$ stands for the standard euclidean norm on $\mathbb{R}^d$. The corresponding true risk or clustering risk is given by $R(\mathbf{c}) = \mathbb{E}_P \ell(\mathbf{c}, X)$. In the sequel, we introduce a constant $M \geq 0$ such that $\|X\|_\infty \leq M$. This boundedness assumption ensures $\ell(\mathbf{c}, X)$ to be bounded. The performances of the empirical minimizer defined in (1.3) have been widely studied in the literature. Consistency was shown by Pollard [1981] when $\mathbb{E}\|X\|^2 < \infty$ whereas Linder, Lugosi, and Zeger [1994] or Biau, Devroye, and Lugosi [2008] gives rates of convergence of the form $\mathcal{O}(1/\sqrt{n})$ for the excess clustering risk defined as $R(\hat{\mathbf{c}}_n) - R(c^*)$, where $c^* \in \mathcal{M}$ the set of all possible optimal clusters. More recently, Levrard [2012] proposes fast rates of the form $\mathcal{O}(1/n)$ under Pollard's regularity assumptions. It improves a previous result of Antos, Györfi, and György [2005]. The main ingredient of the proof is a localization argument in the spirit of Blanchard, Bousquet, and Massart [2008].

In this section, we study the problem of clustering where we have at our disposal a corrupted sample $Z_i = X_i + \epsilon_i$, $i = 1, \ldots, n$ where the ϵ_i's are i.i.d. with density η satisfying **(NA)** of Section 2. For this purpose, we introduce the following deconvolution empirical risk minimization:

$$\arg\min_{\mathbf{c}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} \ell_\lambda(\mathbf{c}, Z_i), \tag{3.2}$$

where $\ell_\lambda(\mathbf{c}, z)$ is a deconvolution k-means loss defined as:

$$\ell_\lambda(\mathbf{c}, z) = \int_K \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right) \min_{j=1,\ldots k} \|x - c_j\|^2 dx.$$

The kernel $\mathcal{K}_\eta$ is the deconvolution kernel introduced in Section 2 with $\lambda = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}_+^d$ a set of positive bandwidths chosen later on. We investigate the generalization ability of the solution of

(3.2) in the context of Pollard's regularity assumptions. For this purpose, we will use the following regularity assumptions on the source distribution P.

Pollard's Regularity Condition (PRC): The distribution P satisfies the following two conditions:

1. P has a continuous density f with respect to Lebesgue measure on $\mathbb{R}^d$,
2. The Hessian matrix of $\mathbf{c} \longmapsto P\ell(\mathbf{c}, .)$ is positive definite for all optimal vector of clusters $\mathbf{c}^*$.

It is easy to see that using the compactness of $\mathcal{B}(0, M)$, $\|X\|_\infty \leq M$ and **(PRC)** ensures that there exists only a finite number of optimal clusters $\mathbf{c}^* \in \mathcal{M}$. This number is denoted as $|\mathcal{M}|$ in the rest of this section. Moreover, Pollard's conditions can be related to the margin assumption **MA(κ)** of Section 3 thanks to the following lemma due to Antos, Györfi, and György [2005].

Lemma 3.1 (Antos, Györfi, and György [2005])**.** *Suppose $\|X\|_\infty \leq M$ and **(PRC)** holds. Then, for any $\mathbf{c} \in \mathcal{B}(0, M)$:*

$$\|\ell(\mathbf{c}, \cdot) - \ell(\mathbf{c}^*(\mathbf{c}), \cdot)\|_{L_2} \leq C_1 d(\mathbf{c}, \mathbf{c}^*(\mathbf{c}))^2 \leq C_1 C_2 \left(R(\mathbf{c}) - R(\mathbf{c}^*(\mathbf{c}))\right),$$

where $c^*(\mathbf{c}) \in \arg\min_{\mathbf{c}^*} d(\mathbf{c}, \mathbf{c}^*)$ and $d(\cdot, \cdot)$ stands for the Euclidean distance in the space of codebooks $\mathbb{R}^{dk}$.

Lemma 3.1 ensures a margin assumption **MA(κ)** with $\kappa = 1$ (see Section 3). It is useful to derive fast rates of convergence. Recently, Levrard [2012] has pointed out sufficient conditions to have **(PRC)** as follows. Denote ∂V_i the boundary of the Voronoi cell V_i associated with c_i, for $i = 1, \ldots, k$. Then, a sufficient condition to have **(PRC)** is to control the sup-norm of f on the union of all possible $|\mathcal{M}|$ boundaries $\partial V^{*,m} = \cup_{i=1}^k \partial V_i^{*,m}$, associated with $c_m^* \in \mathcal{M}$ as follows:

$$\|f_{|\cup_{m=1}^{\mathcal{M}} \partial V^{*,m}}\|_\infty \leq c(d) M^{d+1} \inf_{m=1,\ldots,|\mathcal{M}|, i=1,\ldots k} P(V_i^{*,m}),$$

where $c(d)$ is a constant depending on the dimension d. As a result, the margin assumption is guaranteed when the source distribution P is well concentrated around its optimal clusters, which is related to well-separated classes. From this point of view, the margin assumption **MA(κ)** can be related to the margin assumption in binary classification.
We are now ready to state the main result of this paragraph.

Theorem 3.2. *Assume **(NA)** holds, P satisfies **(PRC)** with density $f \in \mathcal{H}(s, L)$ and $\mathbb{E}\|\epsilon\|^2 < \infty$. Then, for any $t > 0$, for any $n \geq n_0(t)$, denoting by $\hat{\mathbf{c}}_n^\lambda$ a solution of (3.2), we have with probability higher than $1 - e^{-t}$:*

$$R(\hat{\mathbf{c}}_n^\lambda) - \inf_{\mathbf{c} \in \mathcal{C}} R(\mathbf{c}) \leq C \sqrt{\log\log(n)} n^{-\frac{1}{1+\sum_{j=1}^d \beta_j / s_j}},$$

where $C > 0$ is independent of n and $\lambda = (\lambda_1, \ldots, \lambda_d)$ is chosen as:

$$\lambda_j \approx n^{-\frac{1}{2 s_j (1+\sum_{j=1}^d \beta_j / s_j)}}, \forall j = 1, \ldots d.$$

.

The proof is postponed to Section 7.

Remark 3.7 (Fast rates of convergence)**.** *Theorem 3.2 is a direct application of Theorem 3.1 in Section 3. The order of the residual term in Theorem 3.2 is comparable to Theorem 3.1. Due to the finite dimensional hypothesis space $\mathcal{C} \subset \mathbb{R}^{dk}$, we apply the previous study to the case $\rho = 0$. It leads to the fast rates $O\left(n^{-\frac{1}{1+\sum_{i=1}^d \beta_j / s_j}}\right)$, up to an extra $\sqrt{\log\log n}$ term. This term is due to the localization principle of the proof, which consists in applying iteratively a concentration inequality due to Bousquet [2002]. In the finite dimensional case, when $\rho = 0$, we pay an extra $\sqrt{\log\log n}$ term in the rate by solving the fixed point equation. Note that using for instance Levrard [2012], this term can be avoided. It is out of the scope of the present paper.*

Remark 3.8 (Optimality)**.** *Lower bounds of the form $\mathcal{O}(1/\sqrt{n})$ have been stated in the direct case by Bartlett, Linder, and Lugosi [1998] for general distribution. An open problem is to derive lower bounds in the context of Theorem 3.2. For this purpose, we need to construct configurations where both Pollard's regularity assumption and noise assumption **(NA)** could be used in a careful way. In this direction, Loustau and Marteau [2012] suggests lower bounds in a supervised framework under both margin assumption and **(NA)**.*

4 Noisy k-means algorithm

When we consider direct data $X_1, \dots, X_n$, we want to minimize the empirical risk defined in (1.2), over $\mathbf{c} = (c_1, \dots, c_k) \in \mathbb{R}^{dk}$ the set of k possible centers. In this direction, the basic iterative procedure of k-means was proposed by Lloyd in a seminal work (Lloyd [1982], first published in 1957 in a Technical Note of Bell Laboratories). The procedure calculates, from an initialization of k centers, the associated Voronoï cells and updates the centers with the means of the data on each Voronoï cell. The k-means with Lloyd algorithm is considered as a staple in the study of clustering methods. The time complexity is approximately linear, and appears as a good algorithm for clustering spherical well-separated classes, such as a mixture of gaussian vectors. However, in many real-life situations, direct data are not available and measurement errors may occur. In social science, many data are collected by human pollster, with a possible contamination in the survey process. In medical trials, where chemical or physical measurements are treated, the diagnostic is affected by many nuisance parameters, such as the measuring accuracy of the considered machine, gathering with a possible operator bias due to the human practitionner. Same kinds of phenomenon occur in astronomy or econometrics (see Meister [2009]). However, to the best of our knowledge, these considerations are not taken into account in the clustering task. The main implicit argument is that these errors are zero mean and could be neglected at the first glance. The aim of this section is to design a new algorithm to perform clustering over contaminated datasets and to show that it can significantly improve the expected performances of a standard clustering algorithm which neglect this additional source of randomness.

When considering indirect data $Z_1, \dots, Z_n$, a deconvolution empirical risk is defined as:

$$\frac{1}{n}\sum_{i=1}^{n} \ell_\lambda(\mathbf{c}, Z_i) = \int \min_{j=1,\dots,k} \|x - c_j\|^2 \, \hat{f}_\lambda(x) dx. \tag{4.1}$$

Reasonably, a noisy clustering algorithm could be adapted, following the direct case and the construction of the standard k-means. In this section, the purpose is two-fold: on the one hand, a clustering algorithm for indirect data derived from first order conditions is proposed. On the second hand, practical and computational considerations of such an algorithm are discussed.

4.1 First order conditions

Let us consider a corrupted data sample $Z_i = X_i + \epsilon_i$, $i = 1, \dots, n$. The following theorem gives the first order conditions to minimize the deconvolution empirical risk (4.1). In the sequel, $\nabla F(x)$ denotes the gradient of a function $F : \mathbb{R}^{dk} \to \mathbb{R}$ at point $x \in \mathbb{R}^{dk}$.

Theorem 4.1. *Suppose assumptions of Theorem 3.2 are satisfied. Then, for any $\lambda > 0$:*

$$\mathbf{c}_{\ell,j} = \frac{\sum_{i=1}^n \int_{V_j} x_\ell \mathcal{K}_\eta\left(\frac{Z_i - x}{\lambda}\right) dx}{\sum_{i=1}^n \int_{V_j} \mathcal{K}_\eta\left(\frac{Z_i - x}{\lambda}\right) dx}, \, \forall \ell \in \{1, \dots, d\}, \forall j \in \{1, \dots, k\} \Rightarrow \nabla \sum_{i=1}^{n} \ell_\lambda(\mathbf{c}, Z_i) = 0_{\mathbb{R}^{dk}}, \tag{4.2}$$

where $\mathbf{c}_{\ell,j}$ stands for the ℓ-th coordinates of the j-th centers, whereas V_j is the Voronoï cell associated with center j of $\mathbf{c}$:

$$V_j = \{x \in \mathbb{R}^d : \min_{u=1,\dots,k} \|x - c_u\| = \|x - c_j\|\}.$$

The proof is based on the calculation of the directional derivatives of the deconvolution empirical risk (4.1). It is postponed to Section 7.

Remark 4.1 (Comparison with k-means)**.** *It is easy to see that a similar result can be shown with the k-means. Indeed, a necessary condition to minimize the standard empirical risk (1.3) is as follows:*

$$\mathbf{c}_{\ell,j} = \frac{\sum_{i=1}^n \int_{V_j} x_\ell \delta_{X_i} dx}{\sum_{i=1}^n \int_{V_j} \delta_{X_i} dx}, \, \forall \ell \in \{1, \dots, d\}, \forall j \in \{1, \dots, k\},$$

where δ_{X_i} is the Dirac function at point X_i. Theorem 4.1 proposes a same kind of condition in the errors-in-variable case replacing the Dirac function by a deconvolution kernel.

Remark 4.2 (A kernelized k-means)**.** *The previous representation of k-means can lead to a kernelized version of the algorithm in the noiseless case. Indeed, we can replace the Dirac function by a standard kernel function (such as the indicator function) with a sufficiently small bandwidth. This idea has been already presented in discriminant analysis (see Loustau and Marteau [2012]) where optimality in the minimax sense is proved.*

1. Initialize the centers $\mathbf{c}^{(0)} = (c_1^{(0)}, \ldots, c_k^{(0)}) \in \mathbb{R}^{dk}$
2. Estimation step:
   (a) Compute the deconvoluting Kernel $\mathcal{K}_\eta$ and its FFT $\mathcal{F}(\mathcal{K}_\eta)$.
   (b) Build a histogram of 2-d grid using linear binning rule and compute its FFT: $\mathcal{F}(\hat{f}_Z)$.
   (c) Compute: $\mathcal{F}(\hat{f}) = \mathcal{F}(\mathcal{K}_\eta)\mathcal{F}(\hat{f}_Z)$.
   (d) Compute the Inverse FFT of $\mathcal{F}(\hat{f})$ to obtain the density estimated of X: $\hat{f} = \mathcal{F}^{-1}(\mathcal{F}(\hat{f}))$.
3. Repeat until convergence:
   (a) Assign data points to closest clusters in order to compute the Voronoi diagram.
   (b) Re-adjust the center of clusters with equation (4.3).
4. Compute the final partition by assigning data points to the final closest clusters $\hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_k)$.

Figure 1: The algorithm of Noisy k-means.

Remark 4.3 (A simpler representation)**.** *We can note that by switching the integral with the sum in equation (4.2), the first order conditions on* $\mathbf{c}$ *can be rewritten as follows :*

$$\mathbf{c}_{\ell,j} = \frac{\int_{V_j} x_\ell \hat{f}_\lambda(x)dx}{\int_{V_j} \hat{f}_\lambda(x)dx}, \forall \ell \in \{1, \ldots, d\}, \forall j \in \{1, \ldots, k\}, \tag{4.3}$$

where $\hat{f}_\lambda(x) = 1/n \sum_{i=1}^n \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_i - x}{\lambda}\right)$ *is the kernel deconvolution estimator of the density* f*. This property is at the core of the algorithm presented in Section 4.2.*

4.2 The noisy k-means algorithm

In the same spirit of the k-means algorithm of Lloyd (Lloyd [1982]), we derive therefore an iterative algorithm, named noisy k-means, which enables to find a reasonable partition of the direct data from a corrupted sample. The noisy k-means algorithm consists in two steps (see Figure 1) : (1) a deconvolution estimation step in order to estimate the density f from the corrupted data and (2) an iterative Newton's procedure according to (4.3). This second step can be repeated several times until a stable solution is available.

4.2.1 Estimation step

In this step, the purpose is to estimate the density f from indirect observations $Z_1, \ldots, Z_n$. Let us denote by f_Z the density of corrupted data Z. Then f_Z is the convolution product of the densities f and η denoted by $f_Z = f * \eta$. Consequently, the following relation holds : $\mathcal{F}[f] = \mathcal{F}[f_Z]/\mathcal{F}[\eta]$. A natural property for the Fourier transform of an estimator $\hat{f}$ can be deduced:

$$\mathcal{F}[\hat{f}] = \widehat{\mathcal{F}}[f_Z]/\mathcal{F}[\eta], \tag{4.4}$$

where $\widehat{\mathcal{F}}[f_Z](t) = 1/n \sum_{i=1}^n e^{i\langle t, Z_i\rangle}$ is the Fourier transform of the data. These considerations explain the introduction of the deconvolution kernel estimator (1.4) presented in Section 2. In practice, deconvolution estimation involves n numerical integrations for each grid where the density needs to be estimated. Consequently, a direct programming of such a problem is time consuming when the dimension d of the problem increases. In order to speed the procedure, we have used the Fast Fourier Transform (FFT). In particular, we have adapted the FFT algorithm for the computation of multivariate kernel estimators proposed by [30] to the deconvolution problem. Therefore, the FFT of the deconvoluting kernel is first computed. Then, the Fourier transform of data $\widehat{\mathcal{F}}[f_Z]$ is computed via a discrete approximation: an histogram on a grid of 2 dimensional cells is built before applying the FFT as it was proposed in [30]. Then, the discrete Fourier transform of f is obtained from equation (4.4) and an estimation of f is found by an inverse Fourier transform.

4.2.2 Newton's iterations step

The center of the jth group on the ℓth component can therefore be computed from (4.3) as follows :

$$\mathbf{c}_{\ell,j} = \frac{\int_{V_j} x_\ell \hat{f}_\lambda(x)dx}{\int_{V_j} \hat{f}_\lambda(x)dx},$$

where V_j stands for the Voronoi cell of the group j.

5 Experimental validation

Evaluation of clustering algorithms is not an easy task (see von Luxburg, Williamson, and Guyon [2009]). In supervised classification, cross-validation techniques are standard to evaluate learning algorithms such as classifiers. The principle is to divide the sample into V subsets, the first $V-1$ are used for training the considered classifiers whereas the last one is used for testing these classifiers. Unfortunately, in an unsupervised framework - such as clustering - the performances of new algorithms depend on what one is trying to do. Many often, a natural grouping of a set of points is not necessarily unique. In this section, we propose two experimental settings to illustrate the efficiency of noisy k-means with different criteria.

These experimental settings are based on simulations of gaussian mixtures with additive random noise. We want to emphasize that this additional source of randomness does not have to be neglected for both clustering, or quantization. For this purpose, we compare noisy k-means algorithm (based on a deconvolution step) with standard k-means (a direct algorithm) using Lloyd algorithm, where the random initialization is common for both method. This allows us to reduce the dependence to the initialization of the measure of performances, due to the non-convexity of the considered problem (see Bubeck [2002]).

Dealing with these noisy Gaussian mixtures, we investigate two different problems, called the noisy clustering problem and the noisy quantization problem. The problem of noisy clustering could be summarized as follows:

— Can we separate several Gaussian mixtures from a set of noisy Gaussian mixtures ?

The problem of noisy quantization could be stated as follows :

— Are we abble to summarize the unobserved sample X_i, $i = 1, \dots, n$ from a sequence of i.i.d. $Z_i = X_i + \epsilon_i$, $i = 1, \dots, n$?

Equivalently, one also could adress the problem of estimation of the mean of each Gaussian mixture when a contaminated sample is available. The answer to these questions is proposed in the sequel and depends on several parameters in our models, such as the level of noise ϵ, the type of noise (Laplace or Gaussian) and the number $k = 2$ or $k = 4$ of Gaussian mixtures.

5.1 Experimental setting

We consider two different spherical gaussian mixtures for the unobserved sample X_i, $i = 1, \dots, n$ and an additive Gaussian or Laplace noise for the ϵ_i, $i = 1, \dots, n$ with increasing vertical variance $u \in \{1, \dots, 10\}$.

5.1.1 The first experiment

The first model is a mixture of 2 Gaussian vectors in $\mathbb{R}^2$. For $\mathbb{L} \in \{\mathcal{L}, \mathcal{N}\}$ and $u \in \{1, \dots, 10\}$, the model **Mod1**$(\mathbb{L}, u)$ is generated as:

$$Z_i = X_i + \epsilon_i(u),\ i = 1, \dots, n, \quad \textbf{Mod1}(\mathbb{L}, u)$$

where $(X_i)_{i=1}^n$ are i.i.d. with density:

$$f = 1/2 f_{\mathcal{N}(0_2, I_2)} + 1/2 f_{\mathcal{N}\left((5,0)^T, I_2\right)},$$

$(\epsilon_i(u))_{i=1}^n$ are i.i.d. with law $\mathbb{L}$ with zero mean $(0,0)^T$ and covariance matrix $\Sigma(u) = \begin{pmatrix} 1 & 0 \\ 0 & u \end{pmatrix}$ for $u \in \{1, \dots, 10\}$. We also consider two cases for $\mathbb{L}$, namely a two-dimensional Laplace ($\mathcal{L}$) or Gaussian ($\mathcal{N}$) noise.

5.1.2 The second experiment

The second model is a mixture of 4 Gaussian vectors in $\mathbb{R}^2$. For $\mathbb{L} \in \{\mathcal{L}, \mathcal{N}\}$ and $u \in \{1, \dots, 10\}$, the model **Mod2**$(\mathbb{L}, u)$ is generated as:

$$Z_i = X_i + \epsilon_i(u),\ i = 1, \dots, n, \quad \textbf{Mod2}(\mathbb{L}, u)$$

where: $(X_i)_{i=1}^n$ are i.i.d. with density

$$f = 1/4 f_{\mathcal{N}(0_2, I_2)} + 1/4 f_{\mathcal{N}\left((5,0)^T, I_2\right)} + 1/4 f_{\mathcal{N}\left((0,5)^T, I_2\right)} + 1/4 f_{\mathcal{N}\left((5,5)^T, I_2\right)}.$$

The noise variables $(\epsilon_i(u))_{i=1}^n$ are i.i.d. with law $\mathbb{L} \in \{\mathcal{L}, \mathcal{N}\}$ with zero mean $(0,0)^T$ and covariance matrix $\Sigma(u) = \begin{pmatrix} 1 & 0 \\ 0 & u \end{pmatrix}$ for $u \in \{1, \dots, 10\}$, where $\mathcal{L}$ is a two-dimensional Laplace and $\mathcal{N}$ is a Gaussian noise.

5.1.3 Performance measurements

For each experiment, we propose to compare the performances of Noisy k-means with respect to k-means by computing three different criteria, which corresponds to different problems evocated in the beginning of Section 5 :

— Given noisy sample $Z_i = X_i + \epsilon_i$, $i = 1, \dots, n$, we can argue that we want to cluster the direct data by computing :

$$\mathcal{I}_n(\hat{\mathbf{c}}) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(Y_i \neq f_{\hat{\mathbf{c}}}(X_i)),\ \forall \hat{\mathbf{c}} = (\hat{c}_1, \dots, \hat{c}_k) \in \mathbb{R}^{dk}, \tag{5.1}$$

where $f_{\hat{\mathbf{c}}}(x) = \arg\min_{j=1,\dots,k} \|x - \hat{\mathbf{c}}_j\|_2^2$ and $Y_i \in \{1, 2\}$ for **Mod1**$(\mathbb{L}, u)$ (resp. $Y_i \in \{1, 2, 3, 4\}$ for **Mod2**$(\mathbb{L}, u)$) corresponds to the mixture of the point X_i.

— Given noisy sample $Z_i = X_i + \epsilon_i$, $i = 1, \dots, n$, we can argue that we want to summarize the information of the unobserved sample X_i, $i = 1, \dots, n$. In this case, we compute the following quantization error $\mathcal{Q}_n(\hat{\mathbf{c}})$ defined as :

$$\mathcal{Q}_n(\hat{\mathbf{c}}) := \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,k} \|X_i - \hat{c}_j\|_2^2,\ \forall \hat{\mathbf{c}} = (\hat{c}_1, \dots, \hat{c}_k) \in \mathbb{R}^{dk}. \tag{5.2}$$

— From an estimation point of view, we can also compute the ℓ_2-estimation error of $\hat{\mathbf{c}}$ given by:

$$\|\hat{\mathbf{c}} - \mathbf{c}^*\| := \sqrt{\sum_{j=1}^{k} \|\hat{c}_j - c_j^*\|_2^2},\ \forall \hat{\mathbf{c}} = (\hat{c}_1, \dots, \hat{c}_k) \in \mathbb{R}^{dk}, \tag{5.3}$$

where $(\mathbf{c}_1^*, \mathbf{c}_2^*) = (0, 0, 5, 0)$ for **Mod1**$(\mathbb{L}, u)$ (resp. $(\mathbf{c}_1^*, \mathbf{c}_2^*, \mathbf{c}_3^*, \mathbf{c}_4^*) = (0, 0, 5, 0, 0, 5, 5, 5)$ for **Mod2**$(\mathbb{L}, u)$).

For each criterion, we study the behaviour of the Lloyd algorithm (standard k-means) with two different noisy k-means, corresponding to two different choice of bandwidths λ in the estimation step (see Figure 1). For a grid $\Lambda \subseteq [0.1, 5]^2$ of 10×10 parameters, we compute $\lambda_{\mathcal{I}}$ defined as the minimizer of (5.1) over the grid Λ whereas $\lambda_{\mathcal{Q}}$ is the minimizer of (5.2). Then, we have three clustering algorithms denoted as $\hat{\mathbf{c}}$ for standard k-means using Lloyd algorithm, and $\{\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2\}$ for noisy k-means algorithms with the same initialization and with associated bandwidth $\lambda_{\mathcal{I}}$ and $\lambda_{\mathcal{Q}}$ defined above. It is important to stress that choice of bandwidth $\lambda_{\mathcal{I}}$ and $\lambda_{\mathcal{Q}}$ are not possible in practice. Hence, an adaptive procedure to choose the bandwidth has to be performed, as in standard nonparameteric problem. This is out of the scope of the present paper where we propose to compare k-means with Noisy k-means with fixed bandwidths $\lambda_{\mathcal{Q}}$ and $\lambda_{\mathcal{I}}$. In the sequel, we illustrate the behaviour of these methods for each criterion and each experiment.

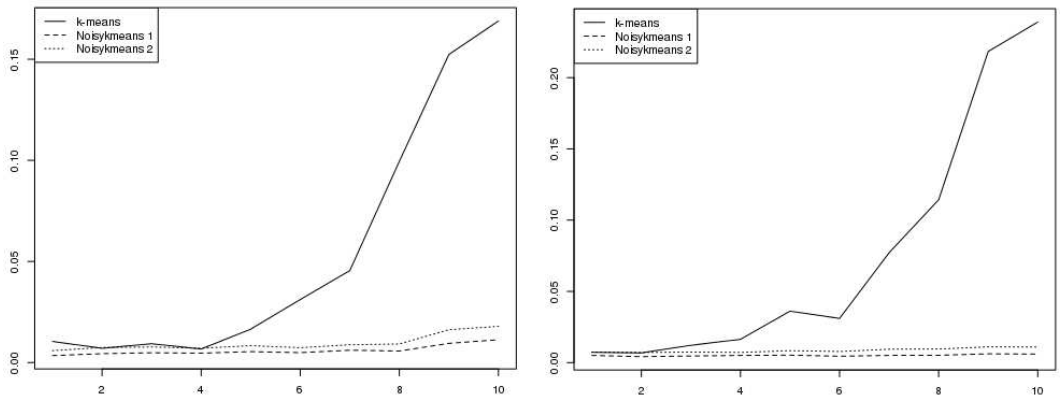
5.2 Results of the first experiment

In the first experiment, we run 100 realizations of training set $\{Z_1, \dots, Z_n\}$ from **Mod1**$(\mathbb{L}, u)$ with $n = 200$. At each realization, we run Lloyd algorithm and noisy k-means with the same random initialization.

		$\mathcal{I}_n$		$\mathcal{Q}_n$		ℓ_2	
		Lap.	Gaus.	Lap.	Gaus.	Lap.	Gaus.
	$\hat{\mathbf{c}}$	1.1	0.7	1.96	1.98	0.29	0.30
$\sigma = 1$	$\hat{\mathbf{c}}_1$	0.3	0.5	2.28	3.39	0.62	1.02
	$\hat{\mathbf{c}}_2$	0.6	0.7	1.97	1.99	0.30	0.33
	$\hat{\mathbf{c}}$	0.7	0.7	2.01	1.99	0.35	0.36
$\sigma = 2$	$\hat{\mathbf{c}}_1$	0.4	0.4	2.42	2.86	0.77	0.94
	$\hat{\mathbf{c}}_2$	0.7	0.7	2.01	2	0.36	0.38
	$\hat{\mathbf{c}}$	0.9	1.2	2.06	2.01	0.40	0.35
$\sigma = 3$	$\hat{\mathbf{c}}_1$	0.5	0.5	2.35	2.83	0.71	0.90
	$\hat{\mathbf{c}}_2$	0.8	0.7	2.02	2.05	0.38	0.43
	$\hat{\mathbf{c}}$	0.7	1.6	2.04	2.13	0.44	0.50
$\sigma = 4$	$\hat{\mathbf{c}}_1$	0.5	0.5	2.35	3.65	0.79	1.28
	$\hat{\mathbf{c}}_2$	0.7	0.7	2.04	2.09	0.43	0.56
	$\hat{\mathbf{c}}$	1.7	3.6	2.26	2.64	0.76	0.81
$\sigma = 5$	$\hat{\mathbf{c}}_1$	0.5	0.5	2.72	3.90	1.05	1.45
	$\hat{\mathbf{c}}_2$	0.8	0.8	2.15	2.30	0.55	0.74
	$\hat{\mathbf{c}}$	3.1	3.1	2.57	2.82	0.82	0.94
$\sigma = 6$	$\hat{\mathbf{c}}_1$	0.5	0.5	2.70	3.87	1.08	1.62
	$\hat{\mathbf{c}}_2$	0.7	0.8	2.12	2.33	0.55	0.78
	$\hat{\mathbf{c}}$	4.5	7.7	3.35	4.20	1.49	1.72
$\sigma = 7$	$\hat{\mathbf{c}}_1$	0.6	0.5	2.96	3.93	1.30	1.61
	$\hat{\mathbf{c}}_2$	0.9	0.9	2.21	2.50	0.68	0.94
	$\hat{\mathbf{c}}$	10.0	11.4	4.33	5.34	2.16	2.46
$\sigma = 8$	$\hat{\mathbf{c}}_1$	0.6	0.5	3.29	4.51	1.46	1.82
	$\hat{\mathbf{c}}_2$	0.9	1	2.32	2.65	0.73	1.07
	$\hat{\mathbf{c}}$	15.2	21.8	5.9	7.62	3.02	3.41
$\sigma = 9$	$\hat{\mathbf{c}}_1$	1.0	0.6	3.69	5.29	1.67	2.14
	$\hat{\mathbf{c}}_2$	1.6	1.1	2.48	2.89	0.97	1.27
	$\hat{\mathbf{c}}$	16.9	23.9	6.22	8.11	3.47	3.66
$\sigma = 10$	$\hat{\mathbf{c}}_1$	1.1	0.6	3.85	5.27	1.84	2.21
	$\hat{\mathbf{c}}_2$	1.8	1.1	2.68	3.09	1.27	1.37

Figure 2: Results of the first experiments averaged over 100 replications. Quantities $\mathcal{I}_n$, $\mathcal{Q}_n$, ℓ_2 are defined in equations (5.1)-(5.3) whereas estimators $\hat{\mathbf{c}}$ (k-means with Lloyd), $\hat{\mathbf{c}}_1$ and $\hat{\mathbf{c}}_2$ (noisy k-means with two particular bandwidths) are defined in Section 5.1. The values of σ corresponds to the variance of the vertical direction of the additive noise ϵ, which is distributed as a Laplace or a Gaussian distribution).

Clustering risk Figure 3 (a)-(b) illustrates the evolution of the clustering risk (5.1) of $\{\hat{\mathbf{c}}, \hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2\}$ when $u \in \{1, \ldots, 10\}$ (horizontal axe) in **Mod(1,$\mathbb{L}$)**.

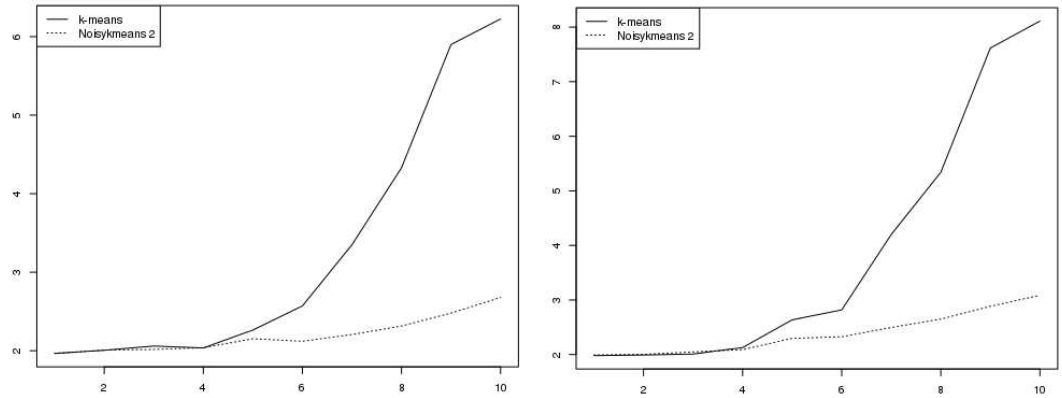


(a) Laplace error (b) Gaussian error

Figure 3: Clustering risk averaged over 100 replications from **Mod1($\mathbb{L}, u$)** with $n = 200$.

When $u \leq 4$, the results are comparable and Noisy k-means seems to slightly outperform standard k-means. However, when the level of noise in the vertical axe becomes higher (i.e. $u \geq 5$), k-means with Lloyd algorithm shows a very bad behaviour. On the contrary, noisy k-means seems robust in these situations, for both Laplace and Gaussian noise.

Quantization risk Figure 4 (a)-(b) shows the behaviour of the quantization risk (5.2) of $\hat{\mathbf{c}}$ and $\hat{\mathbf{c}}_{\mathcal{Q}}$ when u increases.

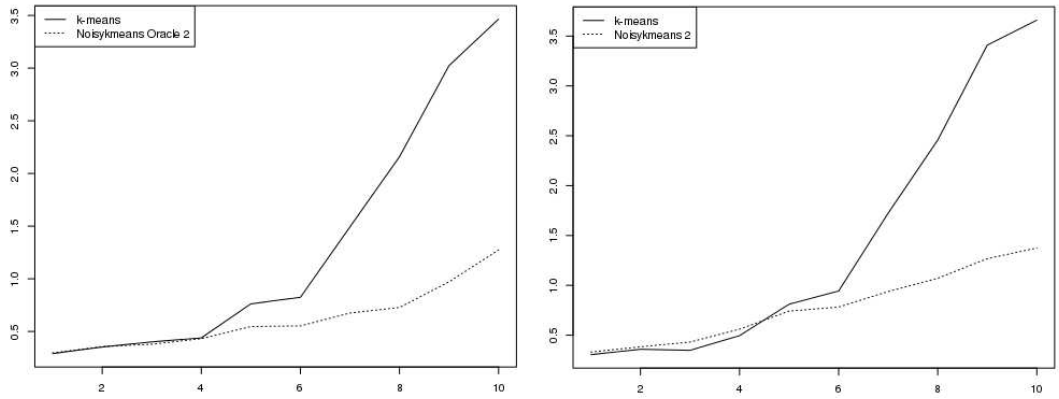


(a) Laplace error (b) Gaussian error

Figure 4: Quantization risk averaged over 100 replications from **Mod1($\mathbb{L}, u$)** with $n = 200$.

We omit $\hat{\mathbf{c}}_1$ because it shows bad performances when the variance u in **Mod(1,$\mathbb{L}$)** increases (see Figure 2). This phenomenon can be explained as follows : $\hat{\mathbf{c}}_1$ is chosen to minimize the clustering risk (5.1). As a result, the proposed codebook $\hat{\mathbf{c}}_1$ is not necessarily a good quantizer, even if it gives good Voronoï cells for clustering the set of data. On the contrary, $\hat{\mathbf{c}}_2$ outperform standard k-means when the vertical variance increases. The quantization error behaves like the clustering risk above. Laplace and Gaussian noise highlights comparable results.

L2 risk In Figure 5 (a)-(b), the ℓ_2 risk (5.3) of $\hat{\mathbf{c}}$ and $\hat{\mathbf{c}}_2$ is proposed. In this case, we can see a more efficient robustness to the noise for Noisy k-means in comparison with standard k-means. However, in comparison with the two other criteria, the ℓ_2 risk of noisy k-means increases when the variance increases. This phenomenon is comparable for Laplace and Gaussian noise, with a slightly better robustness of noisy k-means in the Laplace case.



(a) Laplace error (b) Gaussian error

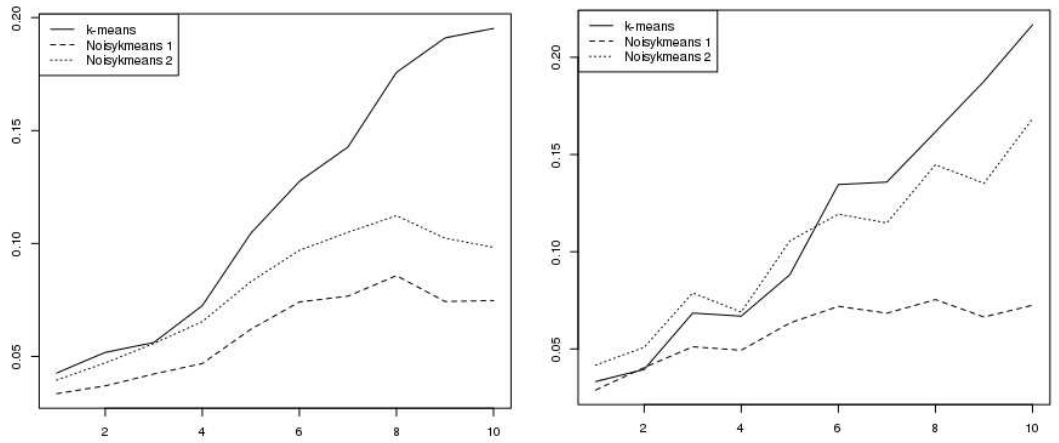
Figure 5: ℓ_2-risk averaged over 100 replications from **Mod1**$(\mathbb{L}, u)$ with $n = 200$.

Conclusion to the first experiment The first experiment shows very well the lack of efficiency of the standard k-means when we deal with errors in variables. When the variance of the noise ϵ increases, the performances of the k-means are deteriorated. On the contrary, the noisy k-means shows a good robustness to this additional source of noise for the considered criteria.

5.3 Result of the second experiment

In the second experiment, we run 100 realizations of training set $\{Z_1, \ldots, Z_n\}$ from **Mod2**$(\mathbb{L}, u)$ with $n = 200$. At each realization, we run Lloyd algorithm and Noisy k-means with the same random initialization.

Clustering risk Figure 6 (a)-(b) shows the evolution of the clustering risk (5.1) of $\{\hat{\mathbf{c}}, \hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2\}$ when $u \in \{1, \ldots, 10\}$ in **Mod**$(2,\mathbb{L})$ is proposed.



(a) Laplace error (b) Gaussian error

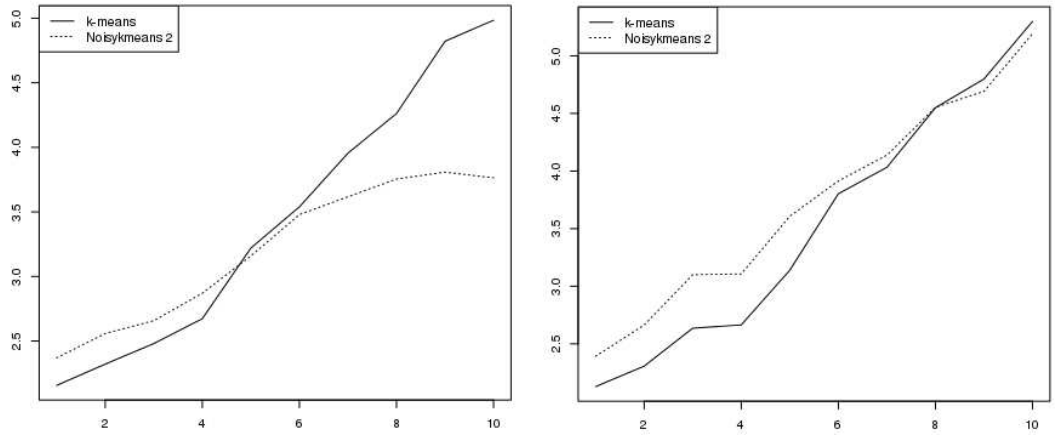
Figure 6: Clustering risk averaged over 100 replications from **Mod2**$(\mathbb{L}, u)$ with $n = 200$.

Figure 5 shows a good resistance of noisy k-means $\hat{\mathbf{c}}_1$ in the presence of a mixture of four Gaussian with errors. When the level of noise is small, $\hat{\mathbf{c}}_1$ slightly outperforms k-means $\hat{\mathbf{c}}$ and when the level of noise becomes higher (i.e. $u \geq 5$), k-means with Lloyd algorithm shows a very bad behaviour. On the contrary, noisy k-means seems more robust in these situations. However, in the presence of a Gaussian noise, $\hat{\mathbf{c}}_2$ is comparable with $\hat{\mathbf{c}}$.

		$\mathcal{I}_n$		$\mathcal{Q}_n$		ℓ_2	
		Lap.	Gaus.	Lap.	Gaus.	Lap.	Gaus.
	$\hat{\mathbf{c}}$	4.3	3.3	2.16	2.13	0.83	0.86
$\sigma = 1$	$\hat{\mathbf{c}}_1$	3.4	2.9	2.57	4.24	1.55	2.14
	$\hat{\mathbf{c}}_2$	4.0	4.2	2.37	2.39	1.28	1.29
	$\hat{\mathbf{c}}$	5.2	3.9	2.32	2.31	1.21	1.21
$\sigma = 2$	$\hat{\mathbf{c}}_1$	3.7	4.0	2.88	7.00	1.87	3.40
	$\hat{\mathbf{c}}_2$	4.7	5.1	2.56	2.66	1.67	1.70
	$\hat{\mathbf{c}}$	5.6	6.8	2.48	2.64	1.48	1.65
$\sigma = 3$	$\hat{\mathbf{c}}_1$	4.2	5.1	3.03	10.15	2.12	4.58
	$\hat{\mathbf{c}}_2$	5.6	7.9	2.66	3.10	1.79	2.21
	$\hat{\mathbf{c}}$	7.3	6.7	2.67	2.66	1.85	1.72
$\sigma = 4$	$\hat{\mathbf{c}}_1$	4.7	4.9	3.59	8.79	2.56	4.29
	$\hat{\mathbf{c}}_2$	6.5	6.9	2.87	3.11	2.21	2.23
	$\hat{\mathbf{c}}$	10.5	8.8	3.22	3.14	2.85	2.30
$\sigma = 5$	$\hat{\mathbf{c}}_1$	6.2	6.3	4.03	11.17	3.11	5.28
	$\hat{\mathbf{c}}_2$	8.3	10.6	3.16	3.61	2.82	2.80
	$\hat{\mathbf{c}}$	12.8	13.5	3.54	3.80	3.07	3.07
$\sigma = 6$	$\hat{\mathbf{c}}_1$	7.4	7.2	4.34	12.88	3.43	5.97
	$\hat{\mathbf{c}}_2$	9.7	11.9	3.48	3.91	3.37	3.17
	$\hat{\mathbf{c}}$	14.3	13.6	3.95	4.03	3.62	3.28
$\sigma = 3$	$\hat{\mathbf{c}}_1$	7.7	6.8	4.72	12.84	3.83	6.02
	$\hat{\mathbf{c}}_2$	10.5	11.5	3.62	4.14	3.69	3.30
	$\hat{\mathbf{c}}$	17.6	16.2	4.26	4.55	4.45	3.77
$\sigma = 4$	$\hat{\mathbf{c}}_1$	8.6	7.5	4.75	14.57	4.28	6.76
	$\hat{\mathbf{c}}_2$	11.2	14.5	3.75	4.55	4.12	3.76
	$\hat{\mathbf{c}}$	19.1	18.8	4.82	4.80	4.95	4.10
$\sigma = 5$	$\hat{\mathbf{c}}_1$	7.4	6.6	5.12	14.13	3.98	6.61
	$\hat{\mathbf{c}}_2$	10.2	13.5	3.81	4.69	4.11	3.91
	$\hat{\mathbf{c}}$	19.5	21.7	4.98	5.30	5.39	4.60
$\sigma = 6$	$\hat{\mathbf{c}}_1$	7.5	7.3	5.19	14.56	4.23	6.88
	$\hat{\mathbf{c}}_2$	9.8	16.8	3.76	5.19	4.33	4.40

Figure 7: Results of the second experiment averaged over 100 replications. Quantities $\mathcal{I}_n$, $\mathcal{Q}_n$, ℓ_2 are defined in equations (5.1)-(5.3) whereas estimators $\hat{\mathbf{c}}$ (k-means with Lloyd), $\hat{\mathbf{c}}_1$ and $\hat{\mathbf{c}}_2$ (noisy k-means with different bandwidths) are defined in Section 5.1. The values of σ corresponds to the variance of the vertical direction of the additive noise ϵ, which is distributed as a Laplace or a Gaussian distribution).

Quantization risk Figure 8 (a)-(b) shows the evolution of the quantization risk (5.1) of $\hat{\mathbf{c}}$ and $\hat{\mathbf{c}}_2$ when $u \in \{1, \dots, 10\}$ in **Mod(2,$\mathbb{L}$)**. We omit $\hat{\mathbf{c}}_1$ for the same reason as in **Mod(1,$\mathbb{L}$)**.

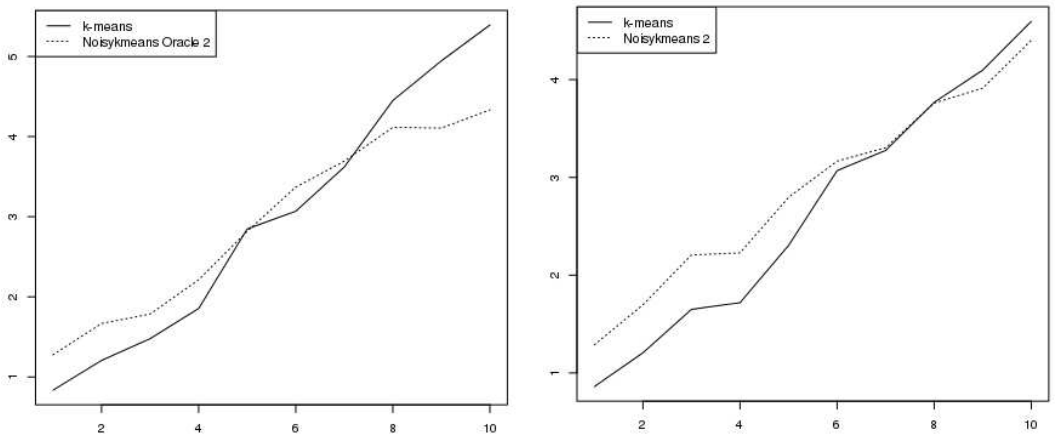


(a) Laplace error (b) Gaussian error

Figure 8: Quantization risk averaged over 100 replications from **Mod2($\mathbb{L}, u$)** with $n = 200$.

Here the evolution of the quantization risk depends strongly on the type of noise in **Mod(2,$\mathbb{L}$)**. When the noise is Laplace, $\hat{\mathbf{c}}_2$ outperforms standard k-means when the vertical variance $u \geq 5$, whereas for small variance, the results are comparable. On the contrary, when the additive noise is Gaussian, the problem seems intractable and Noisy k-means with $n = 200$ does not provide interesting results.

L2 risk Figure 9 (a)-(b) proposes the ℓ_2 risk (5.3) of $\hat{\mathbf{c}}$ and $\hat{\mathbf{c}}_2$ in **Mod2($\mathbb{L}, u$)**.



(a) Laplace error (b) Gaussian error

Figure 9: ℓ_2-risk averaged over 100 replications from **Mod1($\mathbb{L}, u$)** with $n = 200$.

The results are comparable with the Quantization risk and even worst : the Noisy k-means outperforms standard k-means for higher variance ($u \geq 8$).

Conclusion of the second experiment The performances of the k-means are deteriorated when the variance of ϵ increases in the second experiment. However, in this experiment, the problem of noisy clustering -or noisy quantization - seems more difficult. Indeed, Noisy k-means algorithms are not always significantly better than a standard k-means. In this experiment, the difficulty of the problem strongly depends on the type of noise (Gaussian or Laplace), which coincides with standard results in errors-in-variables models.

5.4 Conclusion of the experimental study

The results of this section show rather well the importance of the deconvolution step in the problem of clustering with errors-in-variables. In the presence of well-separated Gaussian mixtures with additive noise, standard k-means gives very bad performances when the variance of the noise increases. On the contrary, Noisy k-means is more robust to this additional source of randomness.

In the particular case of the first experiment, noisy k-means significantly outperforms standard k-means. Unfortunately, when the mixture is more complicated (4 modes in the second experiment), the problem of noisy clustering seems more difficult. The performances of Noisy k-means are not as good as in the first experiment.

6 Conclusion

This paper can be seen as a first attempt into the study of theoretical and practical quantization with errors-in-variables. Many problems could be considered in future works, from theoretical or practical point of view.

In the problem of risk minimization with noisy data, we provide excess risk bounds with fast rates for an empirical risk minimization based on a deconvolution kernel. The risk of the deconvolution ERM mimics the risk of the oracle, up to some residual term, called the rate of convergence. The order of these rates depends on the complexity of the hypothesis space in terms of entropy, the behaviour of the density f and the degree of ill-posedness. From the theoretical point of view, these results extend the previous study of Loustau [2013] to the unsupervised framework and to an anisotropic behaviour of the density f. These extensions could be the core of many applications in unsupervised learning with a corrupted sample, such as anomaly detection, learning principal curves, level-set estimation or quantile estimation.

A seminal example of unsupervised learning is the problem of clustering with k-means. We introduce a deconvolution kernel estimator in the standard k-means distortion, which gives rise to a new stochastic minimization. It allows us to design a new algorithm to deal with clustering with noisy observations. The construction of a noisy version of the well-known k-means is proposed in Section 4.2. The algorithm called Noisy k-means mimics the Newton's iterations of the standard k-means, after a deconvolution estimation step.

Eventually, we illustrate with a rigorous simulation study the behaviour of noisy k-means in two different Gaussian mixture framework. We investigate the ability of the algorithm to separate, quantize or estimate Gaussian mixtures when we observe a corrupted sample with additive Laplace - or Gaussian - noise. The message of this simulation study is the following : when the variance of the noise increases, a deconvolution step is necessary to deal with the inverse problem.

Based on these considerations, a natural direction of research is to look at adaptive noisy k-means. The choice of the bandwidth parameter in the algorithm is a cornerstone to have good results in practice. Then, the design of an automatic bandwidth selection is the next step to investigate. This is the purpose of a future work. Another related issue is to propose an algorithm which doesn't need the a priori knowledge of the distribution error. This problem could be adressed in the presence of repeated measurements, which is also the purpose of a future work.

7 Proofs

The main probabilistic tool for our needs is the localization principle presented in Koltchinskii [2006], which consists in using A Talagrand concentration inequality to functions in $\mathcal{G}$ with small error.

Let us first introduce the following notations. For any fixed $g \in \mathcal{G}$, we write:

$$R^{\lambda}(g) = \int_K \ell(g,x)\mathbb{E}_P \frac{1}{\lambda}\mathcal{K}\left(\frac{X-x}{\lambda}\right) dx \text{ and } R_n^{\lambda}(g) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\lambda}(g, Z_i).$$

As a result, for any fixed $g \in \mathcal{G}$, we have the following equality:

$$R_n^{\lambda}(g) - R^{\lambda}(g) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\lambda}(g, Z_i) - \mathbb{E}_{\tilde{P}}\ell_{\lambda}(g, Z).$$

With a slight abuse of notations, we also denote:

$$(R_n^{\lambda} - R^{\lambda})(g - g') = R_n^{\lambda}(g) - R^{\lambda}(g) - R_n^{\lambda}(g') + R^{\lambda}(g').$$

The same notation is used for $R^{\lambda}(\cdot)$ and $R(\cdot)$ with the quantity $(R - R^{\lambda})(g - g')$.

For a function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, the following transformations are needed:

$$\breve{\psi}_q(\delta) = \sup_{\delta_j \geq \delta} \frac{\psi(\delta_j)}{\delta_j} \text{ and } \psi_q^{\dagger}(\epsilon) = \inf\{\delta > 0 : \breve{\psi}_q(\delta) \leq \epsilon\},$$

where for some $q > 1$, $\delta_j = q^{-j}$ for $j \in \mathbb{N}^*$. Moreover, in the sequel, constant $C, C' > 0$ denote generic constants that may vary from line to line.

We are now ready to state the main ingredient of the proof of Theorem 3.1. The following lemma extends Lemma 2 in Loustau [2013].

Lemma 7.1. *Suppose there exists some function $a : \lambda \mapsto a(\lambda)$ and a constant $0 < r < 1$ such that:*

$$\forall g \in \mathcal{G}, \left|(R - R^\lambda)(g - g^*(g))\right| \leq a(\lambda) + r(R(g) - R(g^*(g))), \tag{7.1}$$

where $g^*(g) \in \arg\min_h R(h)$ can depend on g.
Then, for any $q > 1$, $\forall \delta \geq \bar{\delta}_\lambda(t)$, if $a(\lambda) \leq \delta(1-r)/4q$, we have:

$$\mathbb{P}(R(\hat{g}_n^\lambda) \geq \inf_{g \in \mathcal{G}} R(g) + \delta) \leq \log_q\left(\frac{1}{\delta}\right) e^{-t},$$

where:

$$\bar{\delta}_\lambda(t) = \max\left(\delta_\lambda(t), \frac{8q}{1-r} a(\lambda)\right),$$

for $\delta_\lambda(t) = (U_\lambda(\cdot, t))^\dagger ((1-r)/4q)$ and where we define, for some constant $K > 0$:

$$U_\lambda(\delta, t) := K\left[\mathbb{E}Z_\lambda(\delta) + \sqrt{\frac{t}{n}}\sigma_\lambda(\delta) + \sqrt{\frac{t}{n}\left(1 + 2b_\lambda(\delta)\right)\mathbb{E}Z_\lambda(\delta)} + \frac{t}{3n}\right],$$

where

$$Z_\lambda(\delta) := \sup_{g, g' \in \mathcal{G}(\delta)} \left|(R_n^\lambda - R^\lambda)(g - g')\right|,$$

$$\sigma_\lambda(\delta) := \sup_{g, g' \in \mathcal{G}(\delta)} \sqrt{\mathbb{E}_{\tilde{P}}(\ell_\lambda(g, Z) - \ell_\lambda(g', Z))^2},$$

$$b_\lambda(\delta) := \sup_{g \in \mathcal{G}(\delta)} \|\ell_\lambda(g, \cdot)\|_\infty.$$

7.1 Proof of Theorem 3.1 and 3.2

7.1.1 Proof of Theorem 3.1

The proof of Theorem 3.1 is divided into two steps. Using Lemma 7.1, we obtain the main risk bound when $|\mathcal{M}| = 1$. For the general case, we will introduce a more sophisticated localization explain in Section 4 of Koltchinskii [2006]. Moreover, we begin the proof in dimension $d = 1$ for simplicity. A slightly different algebra is precised at the end of the proof to lead to the general case.
Case 1: $|\mathcal{M}| = 1$.
When $|\mathcal{M}| = 1$, it is important to note that **MA(κ)** holds with a minimizer $g^* \in \mathcal{G}$ which does not depend on g. Then, we can write, for any $g, g' \in \mathcal{G}(\delta)$:

$$\|\ell(g) - l(g')\|_{L_2(K)} \leq \|\ell(g) - \ell(g^*)\|_{L_2(K)} + \|\ell(g') - \ell(g^*)\|_{L_2(K)} \leq 2\sqrt{\kappa_0}\delta^{1/2\kappa}.$$

Gathering with the entropy condition (3.1), we obtain:

$$\begin{aligned}
\mathbb{E} \sup_{g, g' \in \mathcal{G}(\delta)} \left|(R_n^\lambda - R^\lambda)(g - g')\right| &\leq \mathbb{E} \sup_{\|\ell(g) - \ell(g')\|_{L_2(K)} \leq 2\sqrt{\kappa_0}\delta^{1/2\kappa}} \left|(R_n^\lambda - R^\lambda)(g - g')\right| \\
&\leq C\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2\kappa}},
\end{aligned}$$

where we use in last line Lemma 1 in Loustau [2013]. Then, using the notations of Lemma 7.1:

$$\begin{aligned}
U_\lambda(\delta, t) &= K\left[\mathbb{E}Z_\lambda(\delta) + \sqrt{\frac{t}{n}}\sigma_\lambda(\delta) + \sqrt{\frac{t}{n}\left(1 + 2b_\lambda(\delta)\right)\mathbb{E}Z_\lambda(\delta)} + \frac{t}{3n}\right] \\
&\leq K\left[\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2\kappa}} + \sqrt{\frac{t}{n}}\sigma_\lambda(\delta) + \sqrt{\frac{t}{n}\left(1 + 2b_\lambda(\delta)\right)\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2\kappa}}} + \frac{t}{3n}\right].
\end{aligned}$$

It remains to control the $L^2(\tilde{P})$-diameter $\sigma_\lambda(\delta)$ and the term $b_\lambda(\delta)$ thanks to Lemma 8.1. Using again assumption $\mathbf{MA}(\kappa)$, and the unicity of the minimizer g^*, gathering with the first assertion of Lemma 8.1, we can write:

$$\sigma_\lambda(\delta) = \sup_{g,g' \in \mathcal{G}(\delta)} \sqrt{\mathbb{E}_{\tilde{P}}(l_\lambda(g,Z) - l_\lambda(g',Z))^2} \leq C\lambda^{-\beta}\sqrt{\kappa_0}\delta^{\frac{1}{2\kappa}}.$$

Now, by the second assertion of Lemma 8.1:

$$b_\lambda(\delta) = \sup_{g \in \mathcal{G}(\delta)} \|l_\lambda(g,\cdot)\|_\infty \leq C\lambda^{-\beta-1/2}.$$

It follows that:

$$U_\lambda(\delta,t) \leq K\left[\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2\kappa}} + \sqrt{t}\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1}{2\kappa}} + \sqrt{\frac{t}{n}\left(1+\lambda^{-\beta-1/2}\right)\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2\kappa}}} + \frac{t}{3n}\right]. \tag{7.2}$$

We hence have the following assertion:

$$t \leq \delta^{-\frac{\rho}{\kappa}} \wedge \sqrt{n}\lambda^{-\beta}\delta^{\frac{1-\rho}{2\kappa}} \Rightarrow U'_\lambda(\delta,t) \leq K\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2\kappa}}.$$

From an easy calculation, we hence get in this case:

$$\delta_\lambda(t) \leq K\left(\frac{\lambda^{-\beta}}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}},$$

where $K > 0$ is a generic constant. We are now on time to apply Lemma 7.1 with:

$$\delta = K\left(\frac{\lambda^{-\beta}}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}} \text{ and } t' = t + \log\log_q n.$$

In this case, note that for any $t > 0$ independent on n, the choice of λ in Theorem 3.1 warrants that, for any $n \geq n_0(t)$:

$$t + \log\log_q n \leq \delta^{-\frac{\rho}{\kappa}} \wedge \sqrt{n}\lambda^{-\beta}\delta^{\frac{1-\rho}{2\kappa}}.$$

Moreover, using Lemma 8.2, we have in dimension $d = 1$:

$$\forall g \in \mathcal{G},\ \left|(R - R^\lambda)(g - g^*)\right| \leq C\lambda^{2\kappa s/(2\kappa-1)} + \frac{1}{2}(R(g) - R(g^*)).$$

As a result condition (7.1) of Lemma 7.1 is satisfied with $r = 1/2$ and $a(\lambda) = \lambda^{2\kappa s/(2\kappa-1)}$. We can also check that for n great enough, the choice of λ in Theorem 3.1 guarantees:

$$\lambda^{2s} \leq K\left(\frac{\lambda^{-\beta}}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}}.$$

Finally, we get the result since:

$$\log_q \frac{1}{\delta}e^{-t'} \leq \left(\frac{2\kappa}{2\kappa+\rho-1}\right)\log\left(\frac{\sqrt{n}}{\lambda^{-\beta}}\right)\frac{e^{-t}}{\log_q(n)} \leq e^{-t}.$$

For the d-dimensional case, we have the same algebra by replacing $\lambda^{-\beta}$ by $\Pi_{j=1}^d \lambda_j^{-\beta_j}$ in the previous calculus and $\lambda^{2\kappa s/(2\kappa-1)}$ by $\sum_{j=1}^d \lambda_j^{2\kappa s_j/(2\kappa-1)}$ thanks to Lemma 8.2. The choice of λ_j, for $j = 1,\ldots,d$ in Theorem 3.1 allows to conclude.

Case 2: $|\mathcal{M}| \geq 2$.

When the infimum is not unique, the diameter $\sigma_\lambda^2(\delta)$ does not necessary tend to zero when $\delta \to 0$. We hence introduce the more sophisticated geometric parameter:

$$r(\sigma,\delta) = \sup_{g \in \mathcal{G}(\delta)} \inf_{g' \in \mathcal{G}(\sigma)} \sqrt{\mathbb{E}_{\tilde{P}}(\ell_\lambda(g,Z) - \ell_\lambda(g',Z))^2},\ \text{for } 0 < \sigma \leq \delta.$$

It is clear that $r(\sigma,\delta) \leq \sqrt{\sigma_\lambda^2(\delta)}$ and for $\delta \to 0$, we have $r(\sigma,\delta) \to 0$. The idea of the proof is to use a slightly modified version of Lemma 7.1 following Koltchinskii [2006]. More precisely, we have to apply a Talagrand concentration inequality to the random variable:

$$W_\lambda(\delta) = \sup_{g \in \mathcal{G}(\sigma)} \sup_{g' \in \mathcal{G}(\delta):\sqrt{\mathbb{E}_{\tilde{P}}(\ell_\lambda(g,Z)-\ell_\lambda(g',Z))^2} \leq r(\sigma,\delta)+\epsilon} \left|(R_n^\lambda - R^\lambda)(g - g')\right|.$$

This localization guarantees the upper bounds of Theorem 3.1 when $|\mathcal{M}| \geq 2$. However, to this end, we have to check (for $d = 1$ for simplicity):

$$\lim_{\epsilon \to 0} \mathbb{E} \sup_{g \in \mathcal{G}(\sigma)} \sup_{g' \in \mathcal{G}(\delta): \sqrt{\mathbb{E}_{\tilde{P}}(\ell_\lambda(g,Z) - \ell_\lambda(g',Z))^2} \leq r(\sigma,\delta)+\epsilon} \left| (R_n^\lambda - R^\lambda)(g - g') \right| \leq C \frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{1/2\kappa}, \tag{7.3}$$

and for $0 < \sigma \leq \delta$:

$$r(\sigma, \delta) \leq C \lambda^{-\beta} \delta^{1/2\kappa}. \tag{7.4}$$

Using $\mathbf{MA}(\kappa)$ and Lemma 1 in Loustau [2013], it is clear that (7.3) holds since:

$$\begin{aligned}
&\mathbb{E} \sup_{g \in \mathcal{G}(\sigma)} \sup_{g' \in \mathcal{G}(\delta): \sqrt{\mathbb{E}_{\tilde{P}}(\ell_\lambda(g,Z) - \ell_\lambda(g',Z))^2} \leq r(\sigma,\delta)+\epsilon} \left| (R_n^\lambda - R^\lambda)(g - g') \right| \\
&\leq \mathbb{E} \sup_{g \in \mathcal{G}(\sigma), g^* \in \mathcal{M}} \left| (R_n^\lambda - R^\lambda)(g - g^*) \right| + \mathbb{E} \sup_{g' \in \mathcal{G}(\delta)} \left| (R_n^\lambda - R^\lambda)(g' - g^*(g')) \right| \\
&\leq 2\mathbb{E} \sup_{(g,g^*) \in \mathcal{G}(\delta) \times \mathcal{M}} \left| (R_n^\lambda - R^\lambda)(g^* - g) \right| \\
&\leq C \frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{1/2\kappa}.
\end{aligned}$$

To check (7.4), note that with $\mathbf{MA}(\kappa)$ and the first assertion of Lemma 8.1, we have $\forall g \in \mathcal{G}(\delta), g' \in \mathcal{G}(\sigma)$:

$$\begin{aligned}
\sqrt{\mathbb{E}_{\tilde{P}}(\ell_\lambda(g,Z) - \ell_\lambda(g',Z))^2} &\leq C\lambda^{-\beta} \|\ell(g) - \ell(g')\|_{L_2(K)} \\
&\leq C\lambda^{-\beta} \delta^{1/2\kappa} + C\lambda^{-\beta} \|\ell(g^*(g)) - \ell(g^*(g'))\|_{L_2(K)},
\end{aligned}$$

for $0 < \sigma \leq \delta$. Taking the infimum with respect to $g' \in \mathcal{G}(\sigma)$, we get:

$$\|\ell(g^*(g)) - \ell(g^*(g'))\|_{L_2(K)} = 0.$$

7.1.2 Proof of Theorem 3.2

The proof of Theorem 3.2 uses a slightly different version of Theorem 3.1. First of all, an inspection of the proof of Theorem 3.1 shows that condition (3.1) in Theorem 3.1 can be replaced by the following control of the local complexity of the noisy empirical process:

$$\mathbb{E} \sup_{g,g' \in \mathcal{G}(\delta)} \left| (R_n^\lambda - R^\lambda)(g - g') \right| \leq C \frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{\frac{1-\rho}{2\kappa}}. \tag{7.5}$$

Hence, using Lemma 8.3 in the Appendix, gathering with condition **(PRC)**, we can have (7.5) with $\rho = 0$ and $\kappa = 1$.
However, the case $\rho = 0$ is not treated in Theorem 3.1 where $\rho \in (0, 1)$. From (7.5), and using the notations of Lemma 7.1, (7.2) in the proof of Theorem 3.1 becomes:

$$U_\lambda(\delta, t) \leq K \left[\frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{\frac{1}{2}} + \sqrt{t} \frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{\frac{1}{2}} + \sqrt{\frac{t}{n} \left(1 + \lambda^{-\beta - 1/2}\right) \frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{\frac{1}{2}}} + \frac{t}{3n} \right].$$

We hence have the following assertion:

$$t \leq \sqrt{n} \lambda^{-\beta} \delta^{\frac{1}{2}} \Rightarrow U_\lambda(\delta, t) \leq K \left(1 + \sqrt{t}\right) \frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{\frac{1}{2}}.$$

Using the same algebra as above, we can use Lemma 7.1 with:

$$\delta = K \left(1 + \sqrt{t'}\right) \left(\frac{\lambda^{-\beta}}{\sqrt{n}}\right)^{\frac{2}{1+\rho}} \text{ and } t' = t + \log\log_q n.$$

In this case, note that the choice of $t' = t + \log\log_q n$ gives rise to the following asymptotic:

$$\delta \approx \sqrt{\log\log n} \frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{\frac{1}{2}},$$

and leads to an extra $\sqrt{\log\log n}$ term in the rates of convergence.

7.2 Proof of Theorem 4.1

To give the first order conditions for the deconvolution empirical risk defined in (4.1) as:

$$\tilde{W}_n(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n}\int_K \min_{j=1,\dots,k}\|x-c_j\|^2\,\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_i-x}{\lambda}\right)dx,$$

let us introduce the quantity $J(\mathbf{c},z)$ defined as:

$$J(\mathbf{c},z) = \int_K \min_{j=1,\dots,k}\|x-c_j\|^2\,\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)dx.$$

For a fixed $z\in\mathbb{R}$, and for any $\mathbf{c},\mathbf{c}'\in\mathbb{R}^{dk}$, let us consider the directional derivative of the function $J(\cdot,z):\mathbb{R}^{dk}\to\mathbb{R}$, at $\mathbf{c}$ along the direction $\mathbf{c}'$ defined as:

$$\nabla_{\mathbf{c}'}J(\mathbf{c},z) = \lim_{h\to 0}\frac{J(\mathbf{c}+\mathbf{c}'h,z)-J(\mathbf{c},z)}{h}.$$

Using simple algebra, we have, denoting V_j the Voronoï cell associated to c_j and $V_j(h)$ the Voronoï cell associated with $(\mathbf{c}+h\mathbf{c}')_j$:

$$\begin{aligned}J(\mathbf{c}+\mathbf{c}'h,z)-J(\mathbf{c},z) &= \int_K\left[\min_{j=1,\dots,k}\left\|x-(\mathbf{c}+\mathbf{c}'h)_j\right\|^2-\min_{j=1,\dots,k}\|x-c_j\|^2\right]\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)dx\\ &= \sum_{j=1}^{k}\left[\int_{V_j\cap V_j(h)}\left(h^2\|c_j'\|^2-2h\langle x-c_j,c_j'\rangle\right)\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)dx\right]+\int_{V(h)^C}r(\mathbf{c},\mathbf{c}',x,h,\lambda)dx,\end{aligned}$$

where:

$$V(h) = \bigcup_{j=1}^{k}\left(V_j\cap V_j(h)\right),$$

and $x\mapsto r(\mathbf{c},\mathbf{c}',x,h,\lambda)$ is a bounded function whose precise expression is not useful. Indeed, using dominated convergence and the fact that for any $x\in K$, there exists some $h(x)>0$ such that for any $h\le h(x)$, $\mathbb{1}_{V(h)^C}(x)=0$, we arrive at:

$$\nabla_{\mathbf{c}'}J(\mathbf{c},z) = \sum_{j=1}^{k}\int_{V_j}-2\langle x-c_j,c_j'\rangle\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)dx.$$

For $\mathbf{c}'\in\{e_{ij}=(0,\dots,0,1,\dots,0)|i=1\dots d,j=1\dots k\}$ the canonical basis of $\mathbb{R}^{dk}$, one has:

$$\nabla_{e_{ij}}J(\mathbf{c},z) = -2\int_{V_j}(x_i-c_{ij})\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)dx.$$

Then a sufficient condition on $\mathbf{c}$ to have $\nabla_{e_{\ell,j}}\sum_{i=1}^{n}J(\mathbf{c},Z_i)=0$ is:

$$c_{\ell,j} = \frac{1/n\sum_{i=1}^{n}\int_{V_j}x_\ell\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_i-x}{\lambda}\right)dx}{1/n\sum_{i=1}^{n}\int_{V_j}\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_i-x}{\lambda}\right)dx}. \tag{7.6}$$

8 Appendix

8.1 Technical lemmas

Lemma 8.1. *Suppose **(NA)** holds, and $\mathcal{K}$ satisfies assumption **(K1)**. Suppose $\|f*\eta\|_\infty\le\tilde{c}_\infty$ and $\sup_{g\in\mathcal{G}}\|\ell(g,\cdot)\|_{L_2(K)}<\infty$. Then, the two following assertions hold:*

(i) *$\ell(g)\mapsto\ell_\lambda(g)$ is Lipschitz with respect to λ:*

$$\forall g,g'\in\mathcal{G},\ \|\ell_\lambda(g,\cdot)-\ell_\lambda(g',\cdot)\|_{L_2(\tilde{P})}\le C_1\Pi_{i=1}^{d}\lambda_i^{-\beta_i}\|\ell(g,\cdot)-\ell(g',\cdot)\|_{L_2(K)},$$

*where $C>0$ is a generic constant which depends on $\tilde{c}_\infty$ and constants in **(K1)**.*

(ii) *$\{\ell_\lambda(g),g\in\mathcal{G}\}$ is uniformly bounded:*

$$\sup_{g\in\mathcal{G}}\|\ell_\lambda(g,\cdot)\|_\infty\le C_2\Pi_{i=1}^{d}\lambda_i^{-(\beta_i+1/2)},$$

*where $C_2>0$ is a generic constant which depends on constants in **(K1)**.*

Proof. Using Plancherel and the boundedness assumption over $f * \eta$, we have:

$$\begin{aligned}\mathbb{E}_{\tilde{P}}(\ell_\lambda(g,Z)-\ell_\lambda(g',Z))^2 &= \int \left[\frac{1}{\lambda}\mathcal{K}_\eta(\frac{\cdot}{\lambda}) * (\mathbb{1}_K \times (\ell(g,\cdot)-\ell(g',\cdot))(z)\right]^2 f*\eta(z)dz \\ &\leq C\int \frac{1}{\lambda^2}|\mathcal{F}[\mathcal{K}_\eta(\frac{\cdot}{\lambda})](t)|^2|\mathcal{F}[\mathbb{1}_K \times (\ell(g,\cdot)-\ell(g',\cdot))](t)|^2dt \\ &\leq C\lambda^{-2\beta}\|\ell(g)-\ell(g')\|^2_{L_2(K)},\end{aligned}$$

where we use in last line the following inequalities:

$$\frac{1}{\lambda^2}\left|\mathcal{F}[\mathcal{K}_\eta(./\lambda)](s)\right|^2 = \left|\mathcal{F}[\mathcal{K}_\eta](s\lambda)\right|^2 \leq C\sup_{t\in\mathbb{R}}\left|\frac{\mathcal{F}[\mathcal{K}](t\lambda)}{\mathcal{F}[\eta](t)}\right|^2 \leq C\sup_{t\in[-\frac{L}{\lambda},\frac{L}{\lambda}]}\left|\frac{1}{\mathcal{F}[\eta](t)}\right|^2 \leq C\lambda^{-2\beta},$$

provided that **(K1)** holds.

By the same way, the second assertion holds since if $\ell(g,\cdot)\in L^2(K)$:

$$\begin{aligned}|\ell_\lambda(g,z)| &\leq \int_K \left|\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)\ell(g,x)\right|dx \\ &\leq C\sqrt{\int_K \left|\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)\right|^2 dx} \\ &\leq C\lambda^{-\beta-1/2}.\end{aligned}$$

A straightforward generalization leads to the d-dimensional case. □

Lemma 8.2. *Suppose f belongs to the anisotropic Hölder spaces $\mathcal{H}(s,L)$ with $s=(s_1,\ldots,s_d)$. Let $\mathcal{K}$ a kernel satisfying assumption **K(m)** with $m=\lfloor s\rfloor\in\mathbb{N}^d$. Suppose **MA(κ)** holds with parameter $\kappa\geq 1$. Then, we have:*

$$\forall g\in\mathcal{G},\ \left|(R-R^\lambda)(g-g^*(g))\right| \leq C\sum_{j=1}^d \lambda_j^{2\kappa s_j/(2\kappa-1)} + \frac{1}{2\kappa}(R(g)-\inf_{g\in\mathcal{G}}R(g)),$$

where $C>O$ is a generic constant.

Proof : Note that we can write:

$$(R^\lambda - R)(g-g^*) = \int_K (\ell(g,x)-\ell(g^*,x))\left(\mathbb{E}\hat{f}_\lambda(x)-f(x)\right)dx,$$

where we omit the notation $g^*=g^*(g)$ for simplicity. The first part of the proof uses Proposition 1 stated in Comte and Lacour [2012].

Proposition 8.1 (Comte and Lacour [2012])**.** *Let $B_0(\lambda)=\sup_{x_0\in\mathbb{R}^d}|f(x_0)-\mathbb{E}\hat{f}_\lambda(x_0)|$. Then, if f belongs to the anisotropic Hölder space $\mathcal{H}(s,L)$, and $\mathcal{K}$ is a kernel of order $\lfloor s\rfloor$, we have:*

$$B_0(\lambda)\leq C\sum_{j=1}^d \lambda_j^{s_j},$$

where $C>0$ denotes some generic constant.

The rest of the proof uses the margin assumption **MA(κ)** as follows:

$$\begin{aligned}\left|(R^\lambda-R)(g-g^*)\right| &\leq C\sum_{j=1}^d \lambda_j^{s_j}\int_K |\ell(g,x)-\ell(g^*,x)|dx. \\ &\leq C\sum_{j=1}^d \lambda_j^{s_j}\sqrt{\int_K |\ell(g,x)-\ell(g^*,x)|^2dx} \\ &\leq C\sum_{j=1}^d \lambda_j^{s_j}\left(R(g)-R(g^*)\right)^{\frac{1}{2\kappa}} \\ &\leq C\sum_{j=1}^d \lambda_j^{2\kappa s_j/(2\kappa-1)} + \frac{1}{2\kappa}(R(g)-\inf_{g\in\mathcal{G}}R(g)),\end{aligned}$$

where we use in last line Young's inequality:

$$xy^r \leq ry + x^{1/1-r}, \forall r<1,$$

with $r=\frac{1}{2\kappa}$.

Lemma 8.3. *Suppose* ***(PRC)***, ***(NA)*** *and the kernel assumption* ***(K1)*** *are satisfied and* $\|X\|_\infty \leq M$. *Suppose* $\mathbb{E}\|\epsilon\|^2 < \infty$. *Then:*

$$\mathbb{E} \sup_{(\mathbf{c},\mathbf{c}^*)\in\mathcal{C}\times\mathcal{M},\|\mathbf{c}-\mathbf{c}^*\|^2\leq\delta} \left|(R_n^\lambda - R^\lambda)(\mathbf{c}^* - \mathbf{c})\right| \leq C\Pi_{i=1}^d \lambda_i^{-\beta_i} \frac{\sqrt{\delta}}{\sqrt{n}},$$

where $C > 0$ *is a positive constant.*

Proof. The proof follows Levrard [2012] applied to the noisy setting. First note that in the sequel, we need to introduce the following notation:

$$(\tilde{P}_n - \tilde{P})(\ell_\lambda(\mathbf{c}, Z) - \ell_\lambda(\mathbf{c}', Z) := \frac{1}{n}\sum_{i=1}^n \left[\ell_\lambda(\mathbf{c}, Z_i) - \ell_\lambda(\mathbf{c}', Z_i)\right] - \mathbb{E}_{\tilde{P}}\left[\ell_\lambda(\mathbf{c}, Z) - \ell_\lambda(\mathbf{c}', Z)\right].$$

By smoothness assumptions over $\mathbf{c} \mapsto \min\|x - c_j\|$, for any $\mathbf{c} \in \mathbb{R}^{dk}$ and $\mathbf{c}^* \in \mathcal{M}$, we have:

$$\ell_\lambda(\mathbf{c}, z) - \ell_\lambda(\mathbf{c}^*, z) = \langle \mathbf{c} - \mathbf{c}^*, \nabla_{\mathbf{c}}\ell_\lambda(\mathbf{c}^*, z)\rangle + \|\mathbf{c} - \mathbf{c}^*\| R_\lambda(\mathbf{c}^*, \mathbf{c} - \mathbf{c}^*, z),$$

where, with Pollard [1982] we have:

$$\nabla_{\mathbf{c}}\ell_\lambda(\mathbf{c}^*, z) = -2\left(\int \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)(x - c_1^*)\mathbf{1}_{V_1^*}(x)dx, ..., \int \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)(x - c_k^*)\mathbf{1}_{V_k^*}(x)dx\right)$$

and $R_\lambda(\mathbf{c}^*, \mathbf{c} - \mathbf{c}^*, z)$ satisfies:

$$|R_\lambda(\mathbf{c}^*, \mathbf{c} - \mathbf{c}^*, z)| \leq \|\mathbf{c} - \mathbf{c}^*\|^{-1}\left(|\langle \mathbf{c} - \mathbf{c}^*, \nabla_{\mathbf{c}}\ell_\lambda(\mathbf{c}^*, z)\rangle| + \max_{j=1,...k}(|\|z - \mathbf{c}_j\| - \|x - \mathbf{c}_j^*\|\right).$$

Splitting the expectation in two parts, we obtain:

$$\begin{aligned}\mathbb{E}\sup_{\mathbf{c}^*\in\mathcal{M},\|\mathbf{c}-\mathbf{c}^*\|^2\leq\delta}|\tilde{P}_n - \tilde{P}|(\ell_\lambda(\mathbf{c}^*, .) - \ell_\lambda(\mathbf{c}, .)) &\leq \mathbb{E}\sup_{\mathbf{c}^*\in\mathcal{M},\|\mathbf{c}-\mathbf{c}^*\|^2\leq\delta}|\tilde{P}_n - \tilde{P}|\langle \mathbf{c}^* - \mathbf{c}, \nabla_{\mathbf{c}}\ell_\lambda(\mathbf{c}^*, .)\rangle \\ &+ \sqrt{\delta}\mathbb{E}\sup_{\mathbf{c}^*\in\mathcal{M},\|\mathbf{c}-\mathbf{c}^*\|^2\leq\delta}|\tilde{P}_n - \tilde{P}|(-R_\lambda(\mathbf{c}^*, \mathbf{c} - \mathbf{c}^*, .)) \qquad (8.1)\end{aligned}$$

To bound the first term in this decomposition, consider the random variable

$$Z_n = (\tilde{P}_n - \tilde{P})\langle \mathbf{c}^* - \mathbf{c}, \nabla_{\mathbf{c}}\ell_\lambda(\mathbf{c}^*, .)\rangle = \frac{2}{n}\sum_{u=1}^k\sum_{j=1}^d (c_{u,j} - c_{u,j}^*)\sum_{i=1}^n \int_{V_u}\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_i - x}{\lambda}\right)(x_j - c_{u,j})dx.$$

By a simple Hoeffding's inequality, Z_n is a subgaussian random variable. Its variance can be bounded as follows:

$$\begin{aligned}\mathrm{var}Z_n &= \frac{4}{n}\sum_{u=1}^k\sum_{j=1}^d (c_{u,j} - c_{u,j}^*)^2 \mathrm{var}\int_{V_u}\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z-x}{\lambda}\right)(x_j - c_{u,j})dx \\ &\leq \frac{4}{n}\delta\mathbb{E}\left(\int_{V_{u^+}}\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z-x}{\lambda}\right)(x_j - c_{u^+,j})dx\right)^2 \\ &\leq C\frac{4}{n}\delta\int\left|\mathcal{F}\left[\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{\cdot}{\lambda}\right)\right](t)\right|^2 \left|\mathcal{F}[(\pi_j - c_{u^+,j})1_{V_{u^+}}](t)\right|^2 dt \\ &\leq C\frac{4}{n}\delta\Pi_{i=1}^d\lambda_i^{-2\beta_i}\int_{V_{u^+}}(x_j - c_{u^+,j})^2 dx \\ &\leq C\Pi_{i=1}^d\lambda_i^{-2\beta_i}\frac{4}{n}\delta,\end{aligned}$$

where $u^+ = \arg\max_u \int_{V_u}\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z-x}{\lambda}\right)(x_j - c_{u,j})dx$ and $\pi_j : x \mapsto x_j$, and where we use the same argument as in Lemma 8.1 under assumption **(K1)**. We hence have using for instance a maximal inequality due to Massart Massart [34, Part 6.1]:

$$\mathbb{E}\left(\sup_{\mathbf{c}^*\in\mathcal{M},\|\mathbf{c}-\mathbf{c}^*\|^2\leq\delta}(\tilde{P}_n - \tilde{P})\langle \mathbf{c}^* - \mathbf{c}, \nabla_{\mathbf{c}}\ell_\lambda(\mathbf{c}^*, .)\rangle\right) \leq C\frac{\Pi_{i=1}^d\lambda_i^{-\beta_i}}{\sqrt{n}}\sqrt{\delta}.$$

We obtain for the first term in (8.1) the right order. To prove that the second term in (8.1) is smaller, note that from Pollard [1982], we have:

$$\begin{aligned}|R_\lambda(\mathbf{c}^*, \mathbf{c}-\mathbf{c}^*, z)| &\leq \|\mathbf{c}-\mathbf{c}^*\|^{-1}\left(\langle \mathbf{c}-\mathbf{c}^*, \nabla_{\mathbf{c}}\ell_\lambda(\mathbf{c}^*, z)\rangle + \max_{j=1,\dots k}(|\|z-\mathbf{c}_j\|^2 - \|z-\mathbf{c}_j^*\|^2|\right)\\ &\leq \|\nabla_{\mathbf{c}}\ell_\lambda(\mathbf{c}^*, z)\| + \|\mathbf{c}-\mathbf{c}^*\|^{-1}\sum_{j=1,\dots k}|\|z-\mathbf{c}_j\|^2 - \|z-\mathbf{c}_j^*\|^2|\\ &\leq C(\Pi_{i=1}^d \lambda_i^{-\beta_i} + \|z\|)\end{aligned}$$

we we use in last line:

$$\|\nabla_{\mathbf{c}}\ell_\lambda(\mathbf{c}^*, z)\|^2 = 4\sum_{j,k}\left(\int \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)(x_j - c_{u,j}^*)\mathbf{1}_{V_u^*}(x)dx\right)^2 \leq C\Pi_{i=1}^d \lambda_i^{-2\beta_i}.$$

Hence it is possible to apply a chaining argument as in Levrard [2012] to the class

$$\mathcal{F} = \{R_\lambda(\mathbf{c}^*, \mathbf{c}-\mathbf{c}^*, \cdot), \mathbf{c}^* \in \mathcal{M}, \mathbf{c} \in \mathbb{R}^{kd} : \|\mathbf{c}-\mathbf{c}^*\| \leq \sqrt{\delta}\},$$

which has an enveloppe function $F(\cdot) \leq C(\Pi_{i=1}^d \lambda_i^{-\beta_i} + \|\cdot\|) \in L_2(\tilde{P})$ provided that $\mathbb{E}\|\epsilon\|^2 < \infty$. We arrive at the conclusion. □

References

[1] S. Graf, H. Luschgy, Foundation of quantization for probability distributions, Springer-Verlag, 2000. Lecture Notes in Mathematics, volume 1730.

[2] J. Hartigan, Clustering algorithms, Wiley, 1975.

[3] S. Loustau, Inverse statistical learning, Electronic Journal of Statistics 7 (2013) 2065–2097.

[4] V. Vapnik, The Nature of Statistical Learning Theory, Statistics for Engineering and Information Science, Springer, 2000.

[5] P. Bartlett, S. Mendelson, Empirical minimization, Probability Theory and Related Fields 135 (3) (2006) 311–334.

[6] V. Koltchinskii, Local rademacher complexities and oracle inequalties in risk minimization, The Annals of Statistics 34 (6) (2006) 2593–2656.

[7] D. Pollard, Strong consistency of k-means clustering, The Annals of Statistics 9 (1) (1981).

[8] D. Pollard, A central limit theorem for k-means clustering, The Annals of Probability 10 (4) (1982).

[9] S. Loustau, C. Marteau, Minimax fast rates for discriminant analysis with errors in variables, 2012. Bernoulli, to appear.

[10] E. Mammen, A. Tsybakov, Smooth discrimination analysis, The Annals of Statistics 27 (6) (1999) 1808–1829.

[11] J. Fan, On the optimal rates of convergence for nonparametric deconvolution problems, Annals of Statistics 19 (1991) 1257–1272.

[12] A. Meister, Deconvolution problems in nonparametric statistics, Springer-Verlag, 2009.

[13] C. Butucea, goodness-of-fit testing and quadratic functionnal estimation from indirect observations, The Annals of Statistics 35 (2007) 1907–1930.

[14] F. Comte, C. Lacour, Anisotropic adaptive kernel deconvolution, 2012. To appear in Annales de l'Institut Henri Poincaré.

[15] S. Mallat, A wavelet tour of signal processing, Elsevier/Academic Press, Amsterdam, 2009.

[16] S. Van De Geer, Empirical Processes in M-estimation, Cambridge University Press, 2000.

[17] A. W. van der Vaart, J. A. Weelner, Weak convergence and Empirical Processes. With Applications to Statistics, Springer Verlag, 1996.

[18] V. Koltchinskii, D. Panchenko, Rademacher processes and bounding the risk of function learning, in: High Dimensional Probability II, E. Giné, D. Mason and J. Wellner, eds., 2000, pp. 443–459.

[19] A. Tsybakov, Optimal aggregation of classifiers in statistical learning, The Annals of Statistics 32 (1) (2004) 135–166.

[20] C. Levrard, Fast rates for empirical vector quantization, hal.inria.fr/hal-00664068 (2012).

[21] G. Lecué, S. Mendelson, General non-exact oracle inequalities for classes with a subexponential envelope, The Annals of Statistics 40 (2) (2012) 832–860.

[22] S. Loustau, Anisotropic oracle inequalities in noisy quantization, 2013. Available in H.A.L.

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

[24] G. Biau, L. Devroye, G. Lugosi, On the performance of clustering in hilbert spaces, IEEE Transactions on Information Theory 54 (2) (2008).

[25] A. Antos, L. Györfi, A. György, Individual convergence rates in empirical vector quantizer design, IEEE Trans. Inform. Theory 51 (11) (2005).

[26] G. Blanchard, O. Bousquet, P. Massart, Statistical performance of support vector machines, The Annals of Statistics 36 (2) (2008) 489–531.

[27] O. Bousquet, A bennet concentration inequality and its application to suprema of empirical processes, C.R. Acad. SCI. Paris Ser. I Math 334 (2002) 495–500.

[28] P. Bartlett, T. Linder, G. Lugosi, The minimax distortion redundancy in empirical quantizer design, IEEE Trans. Inform. Theory 44 (5) (1998).

[29] S. Lloyd, Least square quantization in pcm, IEEE Transactions on Information Theory 28 (2) (1982) 129–136.

[30] M. Wand, Fast computation of Multivariate Kernel Estimators, Journal of Computational and Graphical Statistics 3 (4) (1994) 433–445.

[31] U. von Luxburg, R. Williamson, I. Guyon, Clustering: Science or art ?, 2009. Opinion paper for the NIPS workshop Clustering: Science or Art.

[32] S. Bubeck, How the initialization affects the k means, IEEE Tansactions on Information Theory 48 (2002) 2789–2793.

[33] S. Sandilya, S. Kulkarni, Principal curves with bounded turn, IEEE Tansactions on Information Theory 48 (2002) 2789–2793.

[34] P. Massart, Concentration inequalities and model selection, 2007. Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics, Springer.