# Bandwidth selection in clustering with errors in variables

Sébastien Loustau, Simon Souchet

# Bandwidth selection in clustering with errors in variables

Sébastien Loustau and Simon Souchet*

**Abstract**

We consider the problem of clustering when we observe a corrupted sample. We test two bandwidth selection methods. It allows to deal with the isotropic and the anisotropic problem of selecting the bandwidth of a deconvolution kernel. These methods are based on Lepski's type procedure. The first method (see [5]) compares empirical risks associated with different bandwidths by using ICI (Intersection of Confidence Intervals) rule whereas the second one (see [6]) computes the gradient of the empirical risk and allows us to construct an anisotropic data-driven bandwidth. Numerical experiments are proposed to illustrate the efficiency of the methods.

Keywords : Bandwidth selection, Errors-in-variables, K-means, Lepski's method.

## 1 Introduction

The problem of clustering is focal in data mining. It has received many attention in the literature (see [11], [10], [19], [20], [16]). The purpose is to learn clusters from a big cloud of data. Probabilistic assumptions, such as the "i.i.d." paradigm, is mostly often used in clustering. The aim becomes to summarize a probability distribution $P$ thanks to an i.i.d. sequence of observations $X_1, \dots, X_n$ with law $P$. However, in many applications, direct data are not available and measurement errors may arise. The problem of noisy clustering is to represent compactly and efficiently the measure $P$ when a contaminated empirical version $Z_1, \dots, Z_n$ is observed. This framework is a particular case of inverse statistical learning (see [17]), and is known to be an inverse problem. In [3], an algorithm called noisy $k$-means is introduced to deal with this issue. The principle is to plug a deconvolution kernel in the standard $k$-means optimization, for a well-chosen bandwidth parameter optimizing a bias-variance trade-off. In this paper, we propose two bandwidth selection method in order to select automatically the bandwidth in noisy $k$-means algorithm.

### 1.1 Model and notations

We first introduce the following notations. Suppose we observe a corrupted sample $Z_i$, $i = 1, \dots, n$ of i.i.d. observations satisfying:

$$Z_i = X_i + \epsilon_i, \ i = 1, \dots, n. \tag{1.1}$$

We denote by $f$ the unknown density (with respect to the Lebesgue measure on $\mathbb{R}^d$) of the i.i.d. sequence $X_1, X_2, ..., X_n$ and $\eta$ the known density of the i.i.d. random variables $\epsilon_1, \epsilon_2, ..., \epsilon_n$, independent of the sequence $(X_i)_{i=1}^n$. Given some integer $k \geq 1$, we are looking at $k$ clusters from $f$ when a contaminated empirical version $Z_1, \dots, Z_n$ is observed. This problem is a particular case of inverse statistical learning which has deserved particular attention in [17]. For a bandwidth $h \in \mathbb{R}_+^d$ and a given kernel $\mathcal{K}_h(\cdot) = \mathcal{K}(\cdot/h)/h$, let us introduce a deconvolution kernel $\widetilde{\mathcal{K}}_h(\cdot)$ defined as:

$$\widetilde{\mathcal{K}}_h(t) = \mathcal{F}\left[\frac{\mathcal{F}[\mathcal{K}_h]}{\mathcal{F}[\eta]}\right](t), \tag{1.2}$$

*Université d'Angers, LAREMA, loustau@math.univ-angers.fr, souchet@math.univ-angers.fr

provides that $\eta$ has non-null Fourier transform, where $\mathcal{F}[\cdot]$ denotes the usual Fourier transform. With such a kernel, [3] propose a collection of noisy $k$-means minimizers:

$$\hat{\mathbf{c}}_h := \arg\min_{\mathbf{c}\in\mathcal{C}} \widehat{R}_h(\mathbf{c}), \quad h > 0, \tag{1.3}$$

where $\widehat{R}_h(\mathbf{c})$ depends on a kernel deconvolution estimator $\hat{f}_h(\cdot) = \frac{1}{n}\sum_{i=1}^n \widetilde{\mathcal{K}}_h(Z_i - \cdot)$ according to:

$$\widehat{R}_h(\mathbf{c}) = \int_{\mathcal{B}(0,1)} \ell(\mathbf{c}, x)\hat{f}_h(x)dx = \frac{1}{n}\sum_{i=1}^n \ell_h(\mathbf{c}, Z_i). \tag{1.4}$$

In (1.4), $\ell_h(\mathbf{c}, Z)$ is the following convolution product:

$$\ell_h(\mathbf{c}, Z) := \left[\widetilde{\mathcal{K}}_h * (\ell(\mathbf{c}, \cdot)\ \mathbb{1}_{\mathcal{B}(0,1)}(\cdot))\right](Z) = \int_{\mathcal{B}(0,1)} \widetilde{\mathcal{K}}_h(Z - x)\ell(\mathbf{c}, x)dx, \quad \mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{C},$$

and $\mathcal{C} := \left\{\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^{dk} \ : \ c_j \in \mathcal{B}(0,1),\ j = 1, \ldots, k\right\}$ is the set of possible codebooks and $\ell(\mathbf{c}, x)$ is the standard $k$-means loss function:

$$\ell(\mathbf{c}, x) = \min_{j=1,\ldots,k} |x - c_j|_2^2,$$

where $|\cdot|_2$ is the Euclidean norm in $\mathbb{R}^d$.

### 1.2 The Noisy $k$-means algorithm

The algorithm of noisy $k$-means is an alteration of the popular $k$-means algorithm (see [11]). It gives an approximation of the solution of the optimization problem (1.3). Following the direct case of the $k$-means, we compute in [3] an iterative procedure based on Newton optimization. The update at each ieration is performed according to the first order conditions:

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

where $V_j$ stands for the Voronoi cell of group $j$ and $\hat{f}_h(\cdot)$ is the deconvolution kernel estimator. The algorithm and its theoretical foundations are more detailled in [3], where an experimental study reveals a good robustness to the errors in (1.1) when the level of variance increases. However, a data-driven choice of $h$ has to be done to reach these performances.

### 1.3 The bandwidth selection problem

An appropriate choice of the bandwidth provides in [3] fast rates in noisy clustering thanks to the following bias-variance decomposition:

$$\begin{aligned} R(\hat{\mathbf{c}}_h, \mathbf{c}^\star) \leq (R - \widehat{R}_h)(\hat{\mathbf{c}}_h, \mathbf{c}^\star) &\leq (R - R_h)(\hat{\mathbf{c}}_h, \mathbf{c}^\star) + (R_h - \widehat{R}_h)(\hat{\mathbf{c}}_h, \mathbf{c}^\star) \\ &=: \mathrm{bias}(h) + \mathrm{var}(h), \end{aligned} \tag{1.5}$$

where $\widehat{R}_h(\cdot)$ is defined in (1.4) with associated minimizer $\hat{\mathbf{c}}_h$, whereas $R_h(\cdot) = \mathbb{E}\widehat{R}_h(\cdot)$. This bias-variance decomposition is by and large comparable to the usual bias-variance decomposition in nonparametric statistics. In particular, under a regularity assumption over the density $f$, gathering with a polynomial decreasing of the characteristic function of the noise distribution $\eta$, [3] states fast rates of convergence for $\hat{\mathbf{c}}_{\bar{h}}$. This result holds for a deterministic choice of the bandwidth $\bar{h} = (\bar{h}_1, \ldots, \bar{h}_d)^\top$ in (1.3) according to:

$$\bar{h}_j = n^{-\frac{1}{2s_j(1+\sum_{j=1}^d \beta_j/s_j)}},$$

where $s \in \mathbb{R}_+^d$ is the regularity index of $f$ in terms of Hölder space whereas $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}_+^d$ is the degree of ill-posedness. This choice depends on unknwon parameters and a data-driven choice of $h$ has to be investigated.

One of the most popular method for choosing the bandwidth is suggested by [14] in a gaussian white noise model. It is based on the *Lepski's principle* ([15]). The idea is to test several estimators (by comparison) for different values of the bandwidth. This work is at the origin of various theoretical papers dealing with adaptive minimax bounds in nonparametric estimation (see for instance [8], [18], [4]). From the practical point of view, Lepski's method has also received further development, such as the intersection of confidence intervals (ICI) rule (see [12]). This algorithm reveals computational advantages in comparison to the traditional Lepski's procedure, or even traditional cross-validation techniques since it does not require to compute all the estimators of the family. It was originally designed for a problem of gaussian filtering, which is at the core of many applications in image processing (see [13], [1] and references therein). In a deconvolution setting as well, [7] obtain adaptive optimal results (for pointwise and global risks) using an improvement of the standard Lepski's principle (see also [9]).

In this paper, we investigate the problem of bandwidth selection in the family $\{\hat{\mathbf{c}}_h,\ h \in \mathcal{H}\}$. For this purpose, the empirical risk $\widehat{R}_h(\cdot)$ is of first interest. This quantity will be evaluated for different values of $h \in \mathcal{H}$, where $\mathcal{H} \subset \mathbb{R}_+$ is a given grid of bandwidth. More precisely, the computation of (1.4) for increasing values of $h$ will be at the core of the ICI rule defined in Section 2. However, as it was shown in [6], this empirical risk is not suitable in the anisotropic case. As a result, we introduce in the sequel a second bandwidth selection based on the computations of the gradient of $\widehat{R}_h(\cdot)$. For any given $h \in \mathbb{R}_+^d$, we also defined the empirical gradient as:

$$\widehat{G}_h(\mathbf{c}) = \left( \frac{1}{n} \sum_{i=1}^{n} 2 \int_{V_j(\mathbf{c})} (x_u - c_{uj}) \widetilde{\mathcal{K}}_h(Z_i - x) dx \right)_{u=1,\ldots,d, j=1,\ldots,k} \in \mathbb{R}^{dk},\ \forall \mathbf{c} \in \mathbb{R}^{dk}, \tag{1.6}$$

where for any $j = 1, \ldots, k$, $V_j(\mathbf{c}) := \{x \in [0,1]^d : \arg\min_{a=1,\ldots,k} |x - c_a|_2 = j\}$ is the Voronoï cells associated to $\mathbf{c}$, and $x_u$ denotes the $u^{th}$ coordinate of $x \in \mathbb{R}^d$. In the sequel, we suggest to compare (1.6) at different values of $h$ in order to construct an data-driven bandwidth $\widehat{h}$ in the anisotropic framework.

Note that to construct the family of estimators $\{\widehat{\mathbf{c}}_h, h \in \mathcal{H}\}$, we use an alteration of the popular $k$-means algorithm of [11]. At each iteration, a deconvolution kernel function is involved in the Newton optimiszation. Unfortunately, the minimization problem is not convex and we can only compute a local minimizer. As a result, the solution depends strongly on the initialization step in the algorithm and affects significantly the problem of bandwidth selection. At the light of Section 2, ERC rule compares empirical risks $\widehat{R}_h(\cdot)$ at given global minimizers $\widehat{\mathbf{c}}_h$, which is not achievable in practice. In Section 3, EGC rule is introduced as a non-convex optimization problem related with the minimization of (1.6). With these considerations in mind, we expect that, up to some optimization intrinsic difficulties, computations of ERC and EGC can lead to efficient performances, at least in comparison with standard $k$-means. In this direction, some hints could be proposed, such as multiple initializations.

In our problem, we have to choose a bandwidth $h \in \mathcal{H}$. In the isotropic case, we can consider a one-dimensional grid $\mathcal{H} \subset \mathbb{R}$ made of $L$ values. We denote it as $\mathcal{H}_{\text{iso}}$ in the sequel. Equipped with this grid, we use a sequential procedure based on the ICI rule to deal with the isotropic choice of the bandwidth. Loosely speaking, for increasing values of bandwidths $h \in \mathcal{H}_{\text{iso}}$, we construct an intersection of confidence intervals and stops when this intersection is the empty set. In the anisotropic case, we we restrict the study to the two-dimensional case for computational issues ($d = 2$ in (1.1)). We consider a two-dimensional bandwidth $(h_1, h_2)$ and consider the set $\mathcal{H}_{\text{aniso}}$ of $L \times L$ values. Given this two-dimensional grid, we minimize an estimate of the bias-variance decomposition of the gradient excess risk introduced in [6]. This estimation is computed thanks to (1.6) and the introduction of an auxiliary empirical gradient defined below.

### 1.4 Outlines

The rest of this paper is organised as follows. Section 2 could be considered as a first step into the study of data-driven selection rule in the isotropic case. By considering empirical risks instead of estimators, Lepski's heuristic allows to select an isotropic bandwith in noisy clustering. In Section 3, we want to deal with a more challenging problem: the bandwidth selection with anisotropic regularity assumptions. These could be done by extending the Goldenshluger-Lepski procedure in the same flavour as we extend the Lepski's method. However, as we will see, this problem needs the introduction of a new criterion: the empirical gradient. Eventually, Section 4 is dedicated to a simulation study which shows the accuracy of these bandwidth selection methods.

## 2 Isotropic bandwidth selection

The ICI method is a now popular bandwidth selection method. It was proposed by [12] as an alteration to the theoretical Lepski's method. The implementation is very simple and does not need the computation of all the estimators in the family, in comparison to the Lepski's method. It has been successfully applied in various areas, such as image processing (see [2], [1]). In our case, we want to use an ICI-based method to implement the ERC method.

In [5], the ERC selection rule allows a theoretical well justified method to design noisy $k$-means with adaptive properties. The selected bandwidth does not depend on the regularity of the density $f$ in (1.1). The data-driven bandwidth chosen with ERC is given by:

$$\hat{h} = \max\left\{h \in h_a \ : \ \widehat{R}_{h'}(\hat{\mathbf{c}}_h) - \widehat{R}_{h'}(\hat{\mathbf{c}}_{h'}) \leq 3\delta_{h'},\ \forall h' \leq h\right\}, \tag{2.1}$$

where $\delta_h = \log n \mathrm{var}(h)$ for any $h \in \mathcal{H}$. The principal motivation to introduce ERC is to compare empirical risks instead of estimators. Then, in order to apply the ICI rule to (2.1), we choose to replace intervals centered at pointwise estimators (see [12]) by intervals centered at empirical risks $\widehat{R}_h(\widehat{\mathbf{c}}_h)$. This motivates the introduction of a sequence of intervals $(\mathcal{D}_k)_{k=1}^L$ such that:

$$\mathcal{D}_k = \left[\widehat{\mathcal{R}}_k - C\frac{h_k^{-2\bar{\beta}}\log(n)}{n}; \widehat{\mathcal{R}}_k + C\frac{h_k^{-2\bar{\beta}}\log(n)}{n}\right[, \ \forall k = 1, \ldots, L, \tag{2.2}$$

where in (2.2), $\bar{\beta} = \sum_{i=1}^d \beta_i$, $C > 0$ and for any bandwidth $h_k \in \mathcal{H}_{\mathrm{iso}}$, $\widehat{\mathcal{R}}_k := \widehat{R}_{h_k}(\widehat{c}_{h_k})$. Then, the selected bandwidth $\widehat{h}_{\mathrm{ICI}}$ according to ICI rule is selected according to:

$$\widehat{h}_{\mathrm{ICI}} := \max\left\{h_k,\ k = 1, \ldots, |\mathcal{H}_{\mathrm{iso}}| : \mathcal{I}_k \neq \emptyset\right\} \text{ where } \mathcal{I}_k = \bigcap_{j=1}^{k} \mathcal{D}_k. \tag{2.3}$$

The ICI rule (2.3) can be interpreted as follows. The first interval $\mathcal{D}_1$ is constructed thanks to (2.2) with $h_1$. Then, the second interval $\mathcal{D}_2$ is constructed with $h_2 > h_1$ and $\mathcal{I}_2 = \mathcal{D}_1 \cap \mathcal{D}_2$ is computed. If $\mathcal{I}_2 = \emptyset$, the algorithm stops and the selected bandwidth is $\widehat{h}_{\mathrm{ICI}} = h_1$. Otherwise, $\mathcal{D}_3$ is constructed and $\mathcal{I}_3 = \mathcal{I}_2 \cap \mathcal{D}_3$ is built. If $\mathcal{I}_3 = \emptyset$, the algorithm stops and $\widehat{h}_{\mathrm{ICI}} = h_2$. At each iteration $k$, a new intersection $\mathcal{I}_k$ is obtained and we stop when the result has no point. The selected bandwidth is the maximal value of $k$ such that $\mathcal{I}_k \neq \emptyset$. Figure 1 illustrates the method. It is important to notice that the chosen bandwidth made the better compromise between bias and variance of the decomposition of the excess risk. Indeed, when $k$ increases in the algorithm, the bias increases whereas the variance decreases. Then, the lengths of the $\mathcal{I}_k$'s are decreasing whereas the centers of $\mathcal{I}_k$'s have increasing variability. As a result, we propose to stop the algorithm when the intersection of intervals $\mathcal{D}_k$ becomes the empty set.

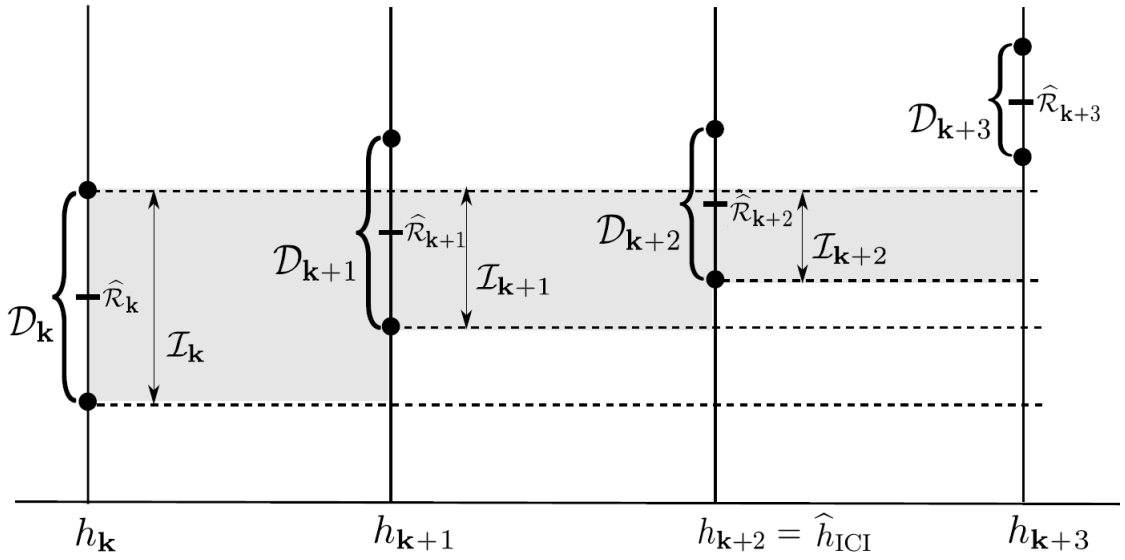



Figure 1: Illustration of ICI rule for noisy $k$-means.

It is important to stress that the proposed method depends on a threshold term $C > 0$ in (2.2). This problem was studied in [21] using the propagation method.

# 3 Anisotropic bandwidth selection

The EGC (Empirical Gradient Comparison) rule is an anisotropic bandwidth selection rule. It was motivated in [6] where general adaptive properties had been stated in kernel empirical risk minimization problems. Here, we propose to use this method in clustering with errors-in-variables to choose the bandwidth in the family of noisy $k$-means (1.3) when the regularity of $f$ depends on the direction.

The EGC rule is based on the computation of gradient empirical risk as in (1.6) instead of empirical risk as in ERC. The principal motivation to use the gradient is summarized in Section 3, where EGC rule is described precisely. In the context of noisy clustering, the data-driven bandwidth is defined as:

$$\widehat{h}_{\mathrm{EGC}} = \arg \min_{h \in \mathcal{H}_{\mathrm{aniso}}} \widehat{\mathrm{BV}}(h), \tag{3.1}$$

where $\widehat{\mathrm{BV}}(h)$ is an estimation of the bias-variance decomposition of the excess risk. This quantity is based on the introduction of an auxiliary kernel:

$$\widetilde{\mathcal{K}}_{h,h'} = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\mathcal{K}_h * \mathcal{K}_{h'}]}{\mathcal{F}[\eta]}\right](x),$$

where $\mathcal{K}_h * \mathcal{K}'_h$ stands for the convolution product between two kernel functions. This auxiliary kernel allows to compute the auxiliary gradient empirical risk $\widehat{G}_{h,h'}(\mathbf{c})$, where $\tilde{K}_{h,h'}$ is used in (1.6) instead of $\widetilde{\mathcal{K}}_h$. Then, the quantity $\widehat{\mathrm{BV}}(h)$ in (3.1) is defined as:

$$\widehat{\mathrm{BV}}(h) := \sup_{h' \in \mathcal{H}_{\mathrm{aniso}}} \left\{ |\widehat{D}_{h,h'} - \widehat{D}_{h'}|_{2,\infty} - \mathcal{M}_l(h,h') \right\} + \mathcal{M}_l^{\infty}(h), \quad \text{with } \mathcal{M}_l^{\infty}(h) := \sup_{h' \in \mathcal{H}_{\mathrm{aniso}}} \mathcal{M}_l(h',h),$$

where $|T|_{2,\infty} := \sup_{\theta} |T(\theta)|_2$ for any $T : \mathbb{R}^{dk} \to \mathbb{R}dk$ whereas $\mathcal{M}_l(h,h')$ is a majorant function (see [6] for a definition). In our framework, it is defined for the mildly ill-posed case as:

$$\mathcal{M}_l(h,h') = C\sqrt{kd}\left(\frac{\Pi_{i=1}^d h_i^{-\beta_i}}{\sqrt{n}} + \frac{\Pi_{i=1}^d (h_i \vee h'_i)^{-\beta_i}}{\sqrt{n}}\right), \tag{3.2}$$

where $C > 0$ is a positive constant. Note that in this experimental study, we also consider a Gaussian distribution for the noise $\epsilon$ in (1.1). In this case, we choose a majorant function in $\widehat{\mathrm{BV}}(h)$ as a product of exponentially decreasing functions of $h_i$, $i = 1 \ldots, d$ instead of polynomial type as in (3.2). This choice is originated in [7] where a study of the standard GL method is suggested in a deconvolution setting.

The computation of (3.1) requires many optimization steps. To overcome this computational issue, in our simulations we use simultaneously packages doParallel and foreach to provide a parallel execution of our R code on machines with multiples cores. The foreach package promotes a new looping construct for executing R code repeadtly. It is similar to the standard lapply function, but does not require the evaluation of a function. It facilitates the execution of the loop in parallel. The doParallel package registers the parallel backend with the foreach package. In our simulation study, we use a machine with 64 cores to speed up the EGC minimization (3.1).

# 4 Experiments

We generate an i.i.d. noisy sample $\mathcal{D}_n = \{Z_1, \ldots, Z_n\}$ where:

$$Z_i = X_i + \epsilon_i(u),\ i = 1, \ldots, n, \tag{4.1}$$

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

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

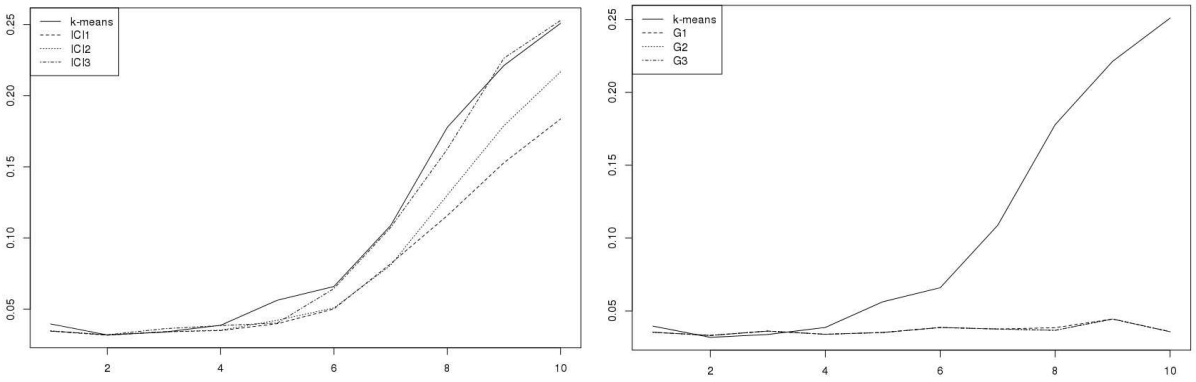
In this study, $(\epsilon_i(u))_{i=1}^n$ are i.i.d. with Gaussian noise with zero mean $(0,0)^T$ and covariance matrix $\Sigma(u) = \begin{pmatrix} 1 & 0 \\ 0 & u \end{pmatrix}$ for $u \in \{1, \ldots, 10\}$. In this setting, we propose to compare the performances of $k$-means with Noisy $k$-means by computing the empirical clustering error according to:

$$\mathcal{I}_n(\hat{\mathbf{c}}) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(Y_i \neq f_{\hat{\mathbf{c}}}(X_i)),\ \forall \hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_k) \in \mathbb{R}^{2d}, \tag{4.2}$$

where $f_{\hat{\mathbf{c}}}(x) = \arg\min_{j=1,2} |x - \hat{\mathbf{c}}_j|_2^2$ and $Y_i \in \{1, 2\}$ corresponds to the mixture of the point $X_i$.

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 $h$, with ERC or EGC. Thanks to the theoretical results, we know that each bandwidth selection method depends on some constant $C > 0$. For ERC with ICI implementation, the constant $C > 0$ is defined in (2.2) whereas for the gradient, the constant is defined in (3.2). In the sequel, we illustrate the behaviour of these methods with respect to the fluctuation of the constant $C > 0$.

Figure 3.2 (a)-(b) illustrates the evolution of the clustering risk (4.2) when $u \in \{1, \ldots, 10\}$ in the model for $k$-means and the two selection rules. For each rule, we bring into play 3 constants $C > 0$ which give rise to 3 different performances.




(a) ERC method (b) Gradient method

Figure 3.2: Clustering risk averaged over 100 replications with $n = 200$ for $k$-means (against ICI (a) and the gradient (b).

The performances of ERC method with ICI implementation depends on the constant $C > 0$ which appears in (2.2). A good calibration of this constant gives slightly better results than $k$-means (Figure 3.2 (a)). In comparison, the noisy $k$-means algorithm with EGC method significantly outperforms $k$-means or ERC (Figure 3.2 (b)). That highlights the importance in practice to choose two different bandwidths in each direction in this model, i.e. an anisotropic bandwidth. Moreover, the dependence on the constant $C > 0$ is higher for ERC than for EGC, which confirms the theoretical study stated in [6].

# 5 Conclusion

This note investigates the bandwidth selection problem in noisy $k$-means. By using theoretical results of [5] and [6], we present two data-driven bandwidth selection for both the isotropic and anisotropic case. A first simulation study reveals a good behaviour of EGC in terms of clustering. Many other problems could be adressed in the future. One can use these bandwidth selection methods in other kernel empirical risk minimization problems, such as in image denoising or local fitted likelihood. In particular, it could be a way of calibrating a local constant approximation method in image denoising with non gaussian noise by using robust loss, such as the Huber loss.

# References

[1] J. Astola, K. Egiazarian, A. Foi, and V. Katkovnik. From Local Kernel to Nonlocal Multiple-Model Image Denoising. *Int. J. Comput. Vision*, 86(1):1–32, 2010.

[2] J. Astola, K. Egiazarian, and V. Katkovnik. Adaptive Window Size Image De-noising Based on Intersection of Confidence Intervals (ICI) Rule. *J. Math. Imaging Vis.*, 16(3):223–235, 2002.

[3] C. Brunet and S. Loustau. Noisy quantization: theory and practise. Submitted, 2014.

[4] M. Chichignoud. Minimax and minimax adaptive estimation in multiplicative regression: locally Bayesian approach. *Probab. Theory Related Fields*, 153(3-4):543–586, 2012.

[5] M. Chichignoud and S. Loustau. Adaptive noisy clustering. 60 (11):1–14, 2014. IEEE Transaction on Information Theory.

[6] M. Chichignoud and S. Loustau. Bandwidth selection in ERM using the gradient. Submitted, 2014.

[7] F. Comte and C. Lacour. Anisotropic adaptive kernel deconvolution. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 49(2):569–609, 2013.

[8] A. Goldenshluger and A. Nemirovski. On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.*, 6(2):135–170, 1997.

[9] Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011.

[10] Siegfried Graf and Harald Luschgy. *Foundation of quantization for probability distributions.* Springer-Verlag, 2000. Lecture Notes in Mathematics, volume 1730.

[11] J.A. Hartigan. *Clustering algorithms.* Wiley, 1975.

[12] V. Katkovnik. A new method for varying adaptive bandwidth selection. *IEEE Trans. Image Process.*, 47(9):2567–2571, 1999.

[13] Ch. Kervrann and J. Boulanger. Optimal Spatial Adaptation for Patch-Based Image Denoising. *IEEE*, 15(10):2866–2878, 2006.

[14] O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3):929–947, 1997.

[15] O.V. Lepski. On a Problem of Adaptive Estimation in Gaussian White Noise. *Theory of Probability and its Applications*, 35(3):454–466, 1990.

[16] C. Levrard. Fast rates for empirical vector quantization. *hal.inria.fr/hal-00664068*, 2012.

[17] S. Loustau. Inverse Statistical Learning. *Electronic Journal of Statistics*, 7:2065–2097, 2013.

[18] P. Mathé. The Lepski's principle revisited. *Inverse Problems*, 22(3):L11–L15, 2006.

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

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

[21] V. Spokoiny and C. Vial. Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Statist.*, 37(5B):2783–2807, 2009.