The algorithm of noisy k-means

Camille Brunet camille.brunet@univ-angers.fr
LAREMA
Université d'Angers
2 Boulevard Lavoisier,
49045 Angers Cedex, France

Sébastien Loustau loustau@math.univ-angers.fr
LAREMA
Université d'Angers
2 Boulevard Lavoisier,
49045 Angers Cedex, France

Editor:

Abstract

In this note, we introduce a new algorithm to deal with finite dimensional clustering with errors in variables. The design of this algorithm is based on recent theoretical advances (see Loustau (2013a,b)) in statistical learning with errors in variables. As the previous mentioned papers, 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.

Keywords: Clustering, Deconvolution, Lloyd algorithm, Fast Fourier Transform, Noisy k-means.

1. Introduction

One of the most popular issue in data minning or machine learning is to learn clusters from a big cloud of data. This problem is known as clustering or empirical quantization. It has received many attention in the last decades (see Hartigan (1975) or Graf and Luschgy (2000) for introductory monographs). Moreover, 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 note 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.

```
1. Initialize the centers $\mathbf{c}^{(0)} = (c_1^{(0)}, \dots, c_k^{(0)}) \in \mathbb{R}^{dk}$
2. Repeat until convergence:
    (a) Assign data points to closest clusters.
    (b) Re-adjust the center of clusters.
3. Compute the final partition by assigning data points to the final closest clusters
    $\hat{\mathbf{c}}_{\mathbf{n}} = (\hat{c}_1, \dots, \hat{c}_k).$
```

Figure 1: *Lloyd algorithm.*

The k-means clustering problem. The k-means is one of the most popular clustering method. The principle is to give an optimal partition of the data minimizing a distortion based on the Euclidean distance. The model of k-means clustering can be written as follows. Consider a random vector $X \in \mathbb{R}^d$ with law P on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and a number of clusters $k \geq 1$. Given n i.i.d. copies $X_1, \dots, X_n$ of X, we want to build a set of centers $\mathbf{c} = (c_1, \dots, c_k) \in \mathbb{R}^{dk}$ minimizing the distortion:

$$W_P(\mathbf{c}) := \mathbb{E}_P \min_{j=1,\dots,k} \|X - c_j\|^2, \tag{1}$$

where $\| \cdot \|$ stands for the Euclidean distance in $\mathbb{R}^d$. The existence of a minimizer of (1) is guaranteed by Graf and Luschgy (2000). In the rest of the paper, a minimizer of (1) is called an oracle and is denoted by $\mathbf{c}^\star$. The problem of finite dimensional clustering becomes to estimate $\mathbf{c}^\star \in \mathbb{R}^{dk}$.

A natural way of minimizing (1) thanks to the collection $X_1, \dots, X_n$ is to consider the empirical distortion:

$$W_{P_n}(\mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,k} \|X_i - c_j\|^2, \tag{2}$$

where $P_n := (1/n) \sum \delta_{X_i}$ is the empirical measure. Thanks to an uniform law of large numbers, we can expected convergence properties for $\mathbf{c}_n^\star := \arg\min_{\mathbf{c}} W_{P_n}(\mathbf{c})$ to an oracle $\mathbf{c}^\star$. In this direction, many authors have investigated the theoretical properties of $\mathbf{c}_n^\star$. As a seminal example, Pollard (1982) proves central limit theorem and consistency result under regularity conditions.

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 algorithm is illustrated in Figure 1. The procedure calculates, from an initialization of k centers, the associated Voronoï cells and actualize 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

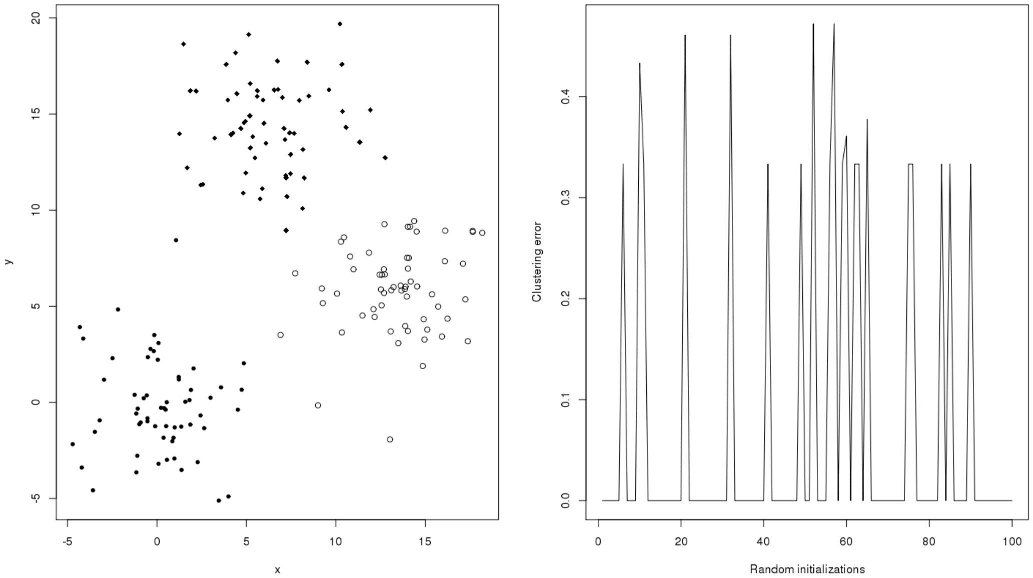

Figure 2: Left panel shows the observations $X_1, \dots, , X_n$ from a mixture of three spherical gaussian (see Section 4 for the experimental set-up). The right panel shows the classification error of 100 runs of the Lloyd algorithm with random initialization. It shows how the initialization affects the performances of the k-means.

good algorithm for clustering spherical well-separated classes, such as a mixture of gaussian vectors. However, the principal limitation in the previous algorithm is that it only reaches a local minimum of the empirical distortion $W_{P_n}(\mathbf{c})$. Indeed, Bubeck (2002) has proved that k-means does Newton iterations in the sense that step 2.(b) in Figure 1 corresponds exactly to a step of a Newton optimization. More precisely, the trajectories of two consecutives centers $\mathbf{c}^t$ and $\mathbf{c}^{t+1}$ visited by the algorithm satisfy:

$$W_{P_n}(\mathbf{c}^\alpha) \leq W_{P_n}(\mathbf{c}^t),\ \forall \mathbf{c}^\alpha = (1-\alpha)\mathbf{c}^t + \alpha \mathbf{c}^{t+1},\ \alpha \in (0,1).$$

A natural consequence of this result is that if a local minimizer of the empirical distortion is reached, then the algorithm stops to this local optimum.

This principal drawback of the k-means algorithm of Lloyd is due to the non-convexity of the empirical distortion $\mathbf{c} \mapsto W_{P_n}(\mathbf{c})$. This property appears practically on the dependence on the solution of the algorithm to the initialization. Different runs of the k-means with random initializations lead to unstable solutions. Figure 2 illustrates dramatically this phenomenon in a mixture of three spherical gaussians.

The noisy k-means clustering problem. In this paper, we are interested in the problem of noisy clustering (see Loustau (2013a)). The problem is still to minimize the distortion

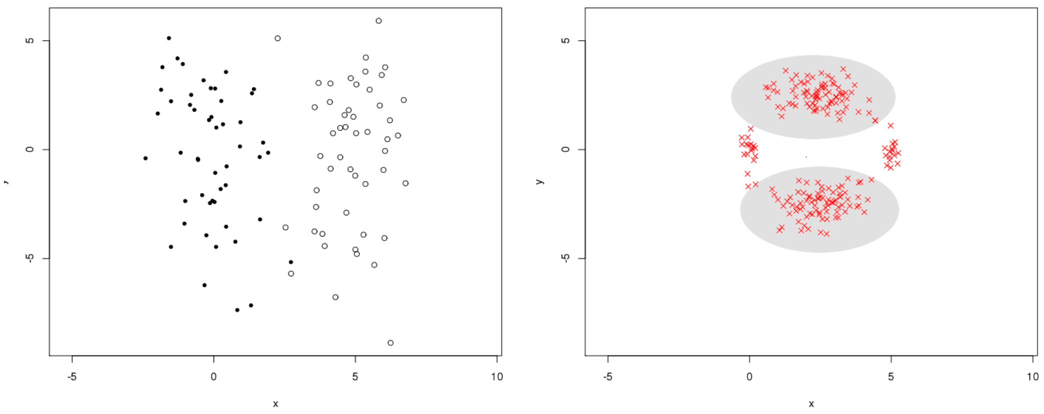

Figure 3: The experimental set-up is detailed in Section 4.1. Left panel shows an example of noisy dataset $Z_i = X_i + \epsilon_i, i = 1, \dots, n$ of two spherical gaussians (the X_i's) with additive vertical noise (the ϵ_i's) (**Mod1(9) in Section 4.1**. Right panel shows solutions $\hat{\mathbf{c}}_n$ of solutions over 100 runs of model **Mod1(9)**. In most of the runs (grey ellipsoids in the right panel), the solutions $\hat{\mathbf{c}}_n$ propose an horizontal separation, corresponding to a bad clustering of direct inputs $X_i, i = 1, \dots n$.

(1), but when we only have at our disposal a noisy sample:

$$Z_i = X_i + \epsilon_i,\ i = 1, \dots, n.$$

Here, ϵ_i, $i = 1, \dots, n$ are i.i.d. random noise with density η with respect to Lebesgue measure. As a result, the density of the observations $Z_1, \dots, Z_n$ is the convolution product $f * \eta(x) := \int f(t)\eta(x-t)dt$. For this reason, we are facing an inverse statistical learning problem (see Loustau (2013b)). The empirical measure $P_n = 1/n \sum \delta_{X_i}$ is not available and only a contaminated version $\tilde{P}_n := 1/n \sum \delta_{Z_i}$ is observable.

As a result, the empirical distortion (2) is not computable. We can only study the empirical distortion with respect to the contaminated data $Z_1, \dots, Z_n$, namely the quantity $W_{\tilde{P}_n}(\mathbf{c})$, for $\mathbf{c} \in \mathbb{R}^{dk}$. Unfortunately, a standard minimization of $W_{\tilde{P}_n}(\mathbf{c})$ seems to fail since for any fixed codebook $\mathbf{c} \in \mathbb{R}^{dk}$:

$$\mathbb{E}W_{\tilde{P}_n}(\mathbf{c}) = W_{P_Z}(\mathbf{c}) = \int \min \|x - c_j\|^2 f * \eta(x)dx \neq W_P(\mathbf{c}).$$

This phenomenon can be interpreted as follows. If we use a basic clustering algorithm based on the minimization of the empirical distortion (such as the k-means algorithm) when we deal with noisy data, the expected criterion does not coincide with the distortion. This phenomenon gives rise to two different situations, which can be summarized as follows:

- At the first glance, the inequality $W_{P_Z}(\mathbf{c}) \neq W_P(\mathbf{c})$ can be considered harmless if the following two oracle sets coincide:

$$\arg\min_{\mathbf{c}\in\mathbb{R}^{dk}} W_P(\mathbf{c}) = \arg\min_{\mathbf{c}\in\mathbb{R}^{dk}} W_{P_Z}(\mathbf{c}). \tag{3}$$

Indeed, in this case, the global minimization of $\mathbb{E}W_{\tilde{P}_n}(\mathbf{c})$ lead coarselly to the best solution $\mathbf{c}^\star$ thanks to an uniform law of large numbers. However, in practice, the global minimum is not available with standard Lloyd algorithm where only a local minimizer is guaranteed. As a result, even if the two oracle sets coincide in (3), it will be more interesting to perform a noisy version of the well-known k-means (see Section 4.2 for an illustration).

- In the general case, there is no reason that (3) holds. Indeed, if we consider an arbitrary mixture of random vectors for the distribution of X, a random additive noise can lead to different oracle $\mathbf{c}^\star$ for the distortion (1). An illustration of such a framework is proposed in Section 4.1, where the k-means is not consistent (see also Figure 3).

These considerations motivate the introduction of a clustering method which takes into account the law of the measurement errors. The noisy k-means algorithm will tackle this issue by using a deconvolution method.

Outlines. The paper is organized as follows. In Section 2, we present the theoretical foundations of the noisy k-means algorithm. This method is originated from the study of risk bounds in statistical learning with errors in variables. We present the construction of a noisy version of the standard k-means algorithm in Section 3. This algorithm mixes a multivariate kernel deconvolution strategy based on Fast Fourier Transform (FFT) with the standard iterative Lloyd algorithm of the k-means. In Section 4, we finally illustrate numerically the good efficiency of the noisy k-means algorithm to deal with noisy inputs. Section 5 concludes the paper with a discussion about the main challenging open problems.

2. Theoretical foundations of noisy k-means

The problem we have at hand is to minimize the distortion (1) thanks to an indirect set of observations $Z_i = X_i + \epsilon_i$, $i = 1, \ldots, n$. This problem is a particular case of inverse statistical learning (see Loustau (2013b)) and is known to be an inverse (deconvolution) problem. As a result, we suggest to use a deconvolution estimator of the density f in the standard k-means algorithm of Figure 1. For this purpose, consider a kernel $\mathcal{K} \in L_2(\mathbb{R}^d)$ such that $\mathcal{F}[\mathcal{K}]$ exists, where $\mathcal{F}$ denotes the standard Fourier transform with inverse $\mathcal{F}^{-1}$. Provided that $\mathcal{F}[\eta]$ exists and is strictly positive, a deconvolution kernel is defined as:

$$\begin{aligned} \mathcal{K}_\eta \quad : \quad & \mathbb{R}^d \to \mathbb{R} \\ & t \mapsto \mathcal{K}_\eta(t) = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\mathcal{K}](\cdot)}{\mathcal{F}[\eta](\cdot/\lambda)}\right](t). \end{aligned} \tag{4}$$

Given this deconvolution kernel, we introduce a deconvolution kernel estimator of the form:

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_i - x}{\lambda}\right), \tag{5}$$

where $\lambda \in \mathbb{R}_+^d$ is a bandwidth parameter. In the sequel, with a slight abuse of notations, we write $1/\lambda = (1/\lambda_1, \ldots, 1/\lambda_d)$.

Originally presented in Loustau and Marteau (2012), the idea of Noisy k-means is to plug the deconvolution kernel estimator $\hat{f}_n(x)$ into the distortion (1). It gives rise to the following deconvolution empirical distortion:

$$\tilde{W}_n(\mathbf{c}) = \int_K \min_{j=1,\dots,k} \|x - c_j\|^2 \hat{f}_n(x)dx, \tag{6}$$

where $\hat{f}_n(x)$ is the kernel deconvolution estimator (5) and $K \subset \mathbb{R}^{dk}$ is a compact domain. Finally, we denote by $\tilde{\mathbf{c}}_n^\star$ the solution of the following stochastic minimization:

$$\tilde{\mathbf{c}}_n^\star = \arg \min_{\mathbf{c} \in \mathbb{R}^{dk}} \tilde{W}_n(\mathbf{c}).$$

The statistical performances of $\tilde{\mathbf{c}}_n^\star$ in terms of distortion (1) has been studied recently in Loustau (2013a). It is based on uniform law of large numbers applied to the noisy empirical measure $\tilde{P}_n$. In particular, the consistency and the precise rates of convergence of the excess distortion can be stated as follows:

Theorem 1 (Loustau (2013a)) *Given an integer $s \in \mathbb{N}^*$, suppose f has partial derivatives up to order $s-1$, such that all the partial derivatives of order $s-1$ are lipschitz. Suppose Pollard's regularity assumption are satisfied (see Pollard (1982)). Then, $\tilde{\mathbf{c}}_n^\star$ with $\bar{\lambda} = n^{-1/(2s+2\sum_{v=1}^d \beta_v)}$ is consistent and satisfies:*

$$W_P(\tilde{\mathbf{c}}_n^\star) - W_P(\mathbf{c}^\star) \leq Cn^{-s/s+\sum_{v=1}^d \beta_v},$$

where $\beta = (\beta_1, \dots, \beta_v)$ is related with the asymptotic behaviour of the characteristic function of η as follows:

$$|\mathcal{F}[\eta](t)| \approx \Pi_{v=1}^d \left(\frac{t_v^2 + 1}{2} \right)^{-\beta_v/2}.$$

Remark 2 (Consistency) *Theorem 1 ensures the consistency of the stochastic minimization (6) based on a noisy dataset Z_i, $i = 1, \dots, n$. The rates of convergence depend on the regularity of the density f and η. The assumption over the smoothness of f is a particular case of the more classical Hölder regularity. It is standard in the deconvolution literature (see for instance Meister (2009)) and more generally in the nonparametric statistical inference (see Tsybakov (2004)). The assumption over the characteristic function of η is also extensively used in the inverse problem literature (see for instance Cavalier (2008)).*

Remark 3 (Bias-variance decomposition) *The proof of this result is based on a decomposition of the quantity $W_P(\tilde{\mathbf{c}}_n^\star) - W_P(\mathbf{c}^\star)$ into two terms: a bias term and a variance term. The variance term is controlled by using the theory of empirical processes, adapted to the noisy set-up (see Loustau (2013b)). The regularity assumption over the density f allows us to control the bias term and to get the proposed rates of convergence.*

Remark 4 (Choice of the bandwidth) *The result of Theorem 1 holds for a particular choice of $\hat{f}_n$ in (6), namely with a particular bandwidth $\bar{\lambda}$ in (5) such that:*

$$\bar{\lambda} = n^{-1/(2s+2\sum_{v=1}^d \beta_v)}. \tag{7}$$

This choice trades off the bias term and the variance term in the proof. It depends explicitely on the regularity of the density f, throught its Hölder exponent s. From practical viewpoint, a data-driven choice of the bandwidth λ is an open problem (see Chichignoud and Loustau (2013) for a theoretical point of view).

3. Noisy k-means algorithm

When we consider direct data $X_1, \ldots, X_n$, we want to minimize the empirical distortion associated with the empirical mesure P_n defined in (2), over $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^{dk}$ the set of k possible centers. This leads to the well-known k-means or Lloyd algorithm presented in Section 1. Similarly, in the noisy case, when considering indirect data $Z_1, \ldots, Z_n$, a deconvolution empirical distortion is defined as:

$$\tilde{W}_n(\mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \int \min_{j=1,\ldots,k} \|x - c_j\|^2 \hat{f}_n(x) dx.$$

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.

3.1 First order conditions

Let us consider an observed corrupted data sample $Z_1, \ldots, Z_n \in \mathbb{R}^d$ which is generated from the additive measurement error model of Section 2 as follows:

$$Z_i = X_i + \epsilon_i, \ \forall i \in \{1, \ldots, n\}. \tag{8}$$

The following theorem gives the first order conditions to minimize the empirical distortion (6). In the sequel, $\nabla f(x)$ denotes the gradient of f at point $x \in \mathbb{R}^{dk}$.

Theorem 5 *Suppose assumptions of Theorem 1 are satisfied. Then, for any $\lambda > 0$:*

$$\mathbf{c}_{\ell,j} = \frac{\sum_{i=1}^{n} \int_{V_j} x_\ell \mathcal{K}_\eta \left(\frac{Z_i - x}{\lambda} \right) dx}{\sum_{i=1}^{n} \int_{V_j} \mathcal{K}_\eta \left(\frac{Z_i - x}{\lambda} \right) dx}, \ \forall \ell \in \{1, \ldots, d\}, \forall j \in \{1, \ldots, k\} \Rightarrow \nabla \tilde{W}_n(\mathbf{c}) = 0_{\mathbb{R}^{dk}}, \tag{9}$$

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 $\mathbf{c}$:

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

Remark 6 (Comparison with the k-means) *It is easy to see that a similar result is available in the direct case of the k-means. Indeed, a necessary condition to minimize the standard empirical distortion $W_n(\cdot)$ is as follows:*

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

where δ_{X_i} is the Dirac function at point X_i. Theorem 5 proposes a same kind of condition in the errors-in-variable case replacing the Dirac function by a deconvolution kernel.

Remark 7 (A simpler representation) *We can remark that by switching the integral with the sum in equation (9), the first order conditions on* $\mathbf{c}$ *can be rewritten as follows :*

$$\mathbf{c}_{\ell,j} = \frac{\int_{V_j} x_\ell \hat{f}_n(x)dx}{\int_{V_j} \hat{f}_n(x)dx}, \forall \ell \in \{1,\dots,d\}, \forall j \in \{1,\dots,k\}, \tag{10}$$

where $\hat{f}_n(x) = 1/n \sum_{i=1}^n \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_i - x}{\lambda}\right)$ *is the kernel deconvolution estimator of the density* f*. This property is at the core of the algorithm presented in Section 3.2.*

Proof The proof is based on the first order conditions for the deconvolution empirical distortion defined in (6) 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 $(\mathbf{c} + h\mathbf{c}')_j$:

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

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 $h \leq h(x)$, $\mathbb{1}_{V(h)^C}(x) = 0$, 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 $\mathbf{c}' \in \{e_{ij} = (0, \ldots, 0, 1, \ldots, 0) | i = 1 \ldots d, j = 1 \ldots k\}$ 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:

$$c_{\ell,j} = \frac{1/n \sum_{i=1}^n \int_{V_j} x_\ell \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{Z_i - x}{\lambda} \right) dx}{1/n \sum_{i=1}^n \int_{V_j} \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{Z_i - x}{\lambda} \right) dx}. \tag{11}$$

■

3.2 The noisy k-means algorithm

In the same spirit of the k-means algorithm of Figure 1, 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 : (1) a deconvolution estimation step in order to estimate the density f of direct data from the corrupted data and (2) an iterative Newton's procedure according to (10). This second step can be repeated several times until a stable solution is available.

3.2.1 Estimation step

In this step, the purpose is to estimate the density f from the model (8) in which the $X_1, \ldots, X_n$ are unobserved. Let us denote by f_Z the density of corrupted data Z. Then, according to (8), 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:

$$\mathcal{F}[\hat{f}] = \widehat{\mathcal{F}}[f_Z]/\mathcal{F}[\eta], \tag{12}$$

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 (5) presented in Section 1. 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 (Wand, 1994) 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 (Wand, 1994). Then, the discrete Fourier transform of f is obtained from equation (12) and an estimation of f is found by an inverse Fourier transform.

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 (10).
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 4: The algorithm of Noisy k-means.

3.2.2 Newton's iterations step

The center of the jth group on the ℓth component can therefore be computed from (10) as follows :

$$c_{ij} = \frac{\int_{V_j} x_\ell \hat{f}_n(x)dx}{\int_{V_j} \hat{f}_n(x)dx},$$

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

Consequently, the estimation procedure needs two different steps : an estimation step to compute the kernel density estimator, and an iterative procedure to converge to the first order conditions. The noisykmeans algorithm is summed up in Figure 4.

4. Experimental validation

Evaluation of clustering algorithms is not an easy task (see von Luxburg et al. (2009)). In this section, we choose to highlight some important phenomena related with the inverse problem we have at hand. These phenomena are of different nature and show the usefulness of the deconvolution step when we deal with noisy data:

- In the first experiment, we show that the noisy k-means algorithm is consistent to discriminate two spherical well-separated gaussian in two dimension, when we observe corrupted sample with increasing variance. It illustrates the particular case of Section

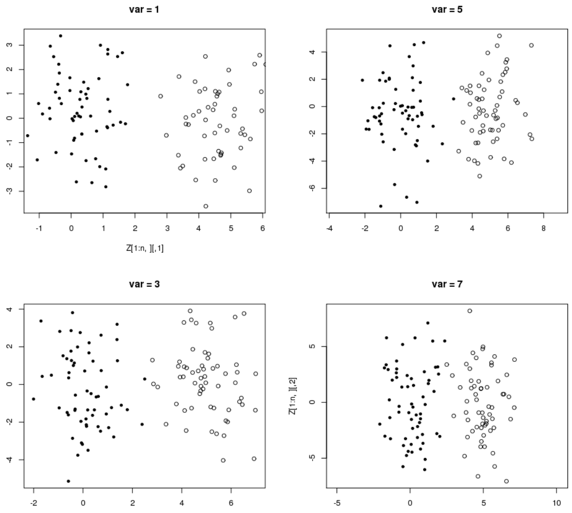

Figure 5: First experiment's setting for $u \in \{1, 3, 5, 7\}$

1 (Figure 3) where Noisy k-means appears as a good alternative to the standard k-means.

- The second experiment is related with Figure 2 of Section 1, where the initialization affects the k-means. In this case, by decreasing the distance between the 3 spherical gaussians, the Noisy k-means algorithm highlights a good resistency.

4.1 First experiment

4.1.1 Simulation setup

In this simulation, we consider, for $u \in \{1, \ldots, 10\}$, the following model, called $\mathbf{Mod1}(u)$:

$$Z_i = X_i + \epsilon_i(u),\ i = 1, \ldots, n, \quad \mathbf{Mod1}(u)$$

where:

- $(X_i)_{i=1}^n$ are i.i.d. with density $f = 1/2 f_{\mathcal{N}(0_2, I_2)} + 1/2 f_{\mathcal{N}((5,0)^T, I_2)}$
- and $(\epsilon_i(u))_{i=1}^n$ are i.i.d. with law $\mathcal{N}(0_2, \Sigma(u))$, where $\Sigma(u)$ is a diagonal matrix with diagonal vector $(0, u)^T$, for $u \in \{1, \ldots, 10\}$.

In this case, the error is concentrated into the vertical axe, and increases with parameter $u \in \{1, \ldots, 10\}$.

We study the behaviour of the Lloyd algorithm of Figure 1 and the noisy k-means algorithm of Figure 4 in $\mathbf{Mod1}(u)$, for $u \in \{1, \ldots, 10\}$. For this purpose, for each u, we run 100 realizations of training set $\{Z_1, \ldots, Z_n\}$ from $\mathbf{Mod1}(u)$ with $n = 100$. At each realization, we run Lloyd algorithm and Noisy k-means with the same random initialization.

The value of $\lambda > 0$ in Noisy k-means is tuned with a grid Λ of 20×20 parameters as follows:

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \sum_{t=1}^{1000} \min_{j=1,2} \|X_t^{\text{tun}} - (\tilde{\mathbf{c}}_n)_j\|^2,$$

where $\tilde{\mathbf{c}}_n = ((\tilde{\mathbf{c}}_n)_1, (\tilde{\mathbf{c}}_n)_2)$ is the solution of Noisy k-means with parameter $\lambda = (\lambda_1, \lambda_2)$ and $(X_t^{\text{tun}})_{t=1}^{1000}$ is an additional i.i.d. sample with density f.
In Figure 6, we are mainly interested in the clustering risk at each realization, defined as:

$$r_n(\hat{\mathbf{c}}) = \frac{1}{100} \sum_{i=1}^{100} \mathbb{1}_{Y_i \neq \hat{\mathbf{c}}(X_i)}, \tag{13}$$

where $\hat{\mathbf{c}}$ denotes either the standard Lloyd algorithm performed on the dataset $\{Z_1, \ldots, Z_n\}$ or the Noisy k-means of Figure 4 with $\lambda = \hat{\lambda}$.

4.1.2 Results of the first experiment

Figure 6 illustrates the result of the first experiment. At first, it shows rather 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. Here is a detailed explanation of Figure 6.

A These boxplots show the evolution of the clustering risk (13) of the two algorithms when the variance increases. When parameter $u \in \{1, \ldots, 5\}$, the results are comparable and standard k-means seems to slightly outperform Noisy k-means. However, when the level of noise in the vertical axe becomes higher (i.e. $u \geq 6$), Lloyd algorithm shows a very bad behaviour. On the contrary, noisy k-means seems to be more robust in these situations.

B Here, we are interested on situations where the studied algrithms fail, i.e. when the clustering errors $r_n(\hat{\mathbf{c}}) > 0.2$. Figure 5.B shows rather well the robustness of the noisy k-means in comparison with the standard k-means. The situation becomes problematic for k-means when $u \geq 6$ where the numbers of failures is bigger than 20 (from 22 to 68 over a total of 100 runs). On the contrary, the maximum of failures of the Noisy k-means does not exceed 19.

C Figure 6.C is a precise illustration of the behaviour of the two algorithms in the particular model **Mod1(7)**. We have plot the clustering errors $r_n(\hat{\mathbf{c}})$ for each run in this model. Of course, here again, Noisy k-means outperforms standard k-means at many runs. However, at some runs, the standard k-means does a good job whereas Noisy k-means completely fails. This can be explained by the dependence on the solution to the random initialization (and then on the non-convexity of the problem).

D Finally, last plot deals with the mean clustering risk in each model **Mod1(u)**, $u \in \{1, \ldots, 10\}$ and the corresponding confidence intervals, calculated thanks to the following formula:

$$\left[\mu(r_n(\hat{\mathbf{c}})) - 1.96 \times \frac{\sigma(r_n(\hat{\mathbf{c}}))}{\sqrt{n}}, \mu(r_n(\hat{\mathbf{c}})) - 1.96 \times \frac{\sigma(r_n(\hat{\mathbf{c}})}{\sqrt{n}}\right],$$

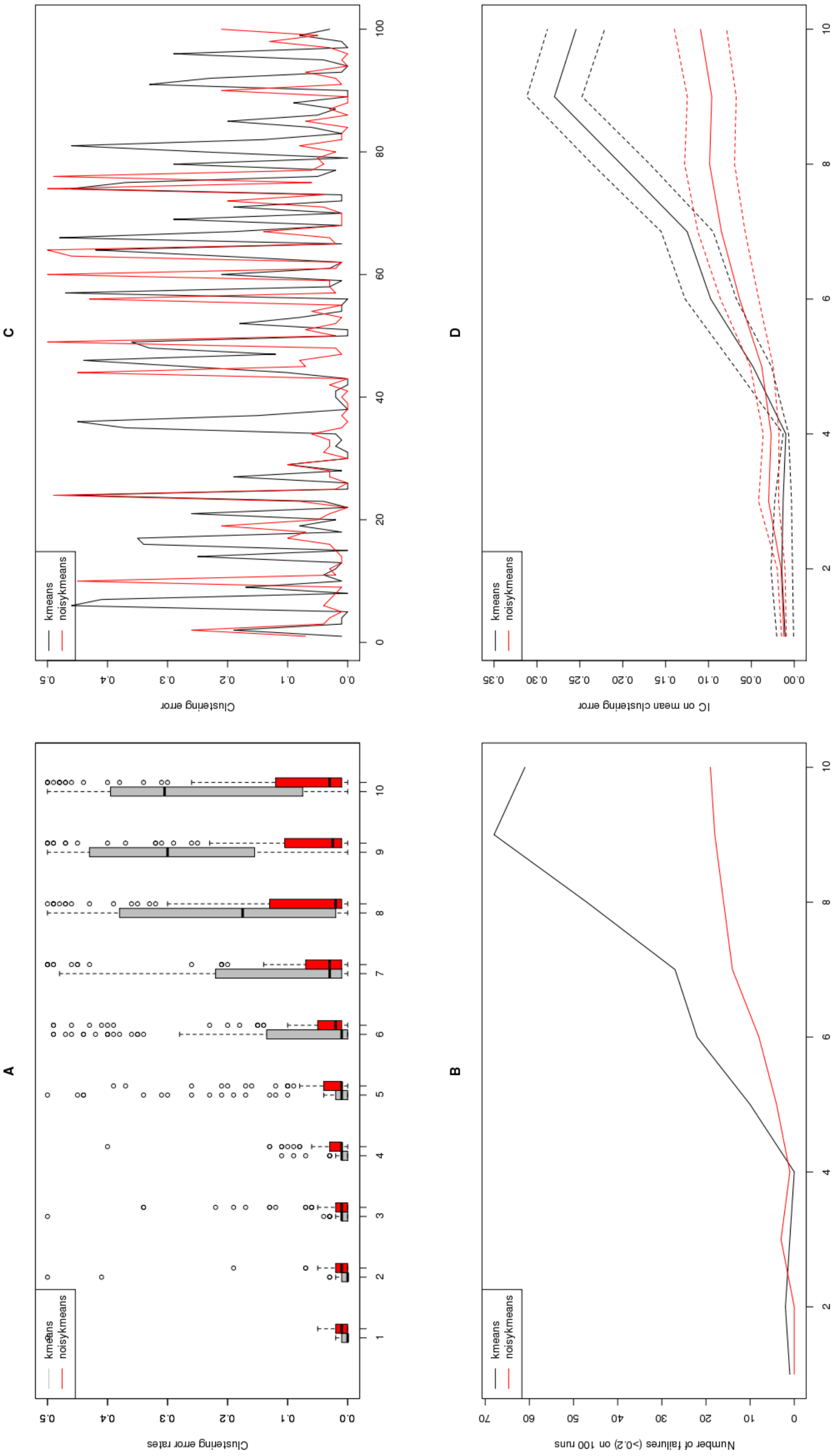

Figure 6: Results of the first experiment

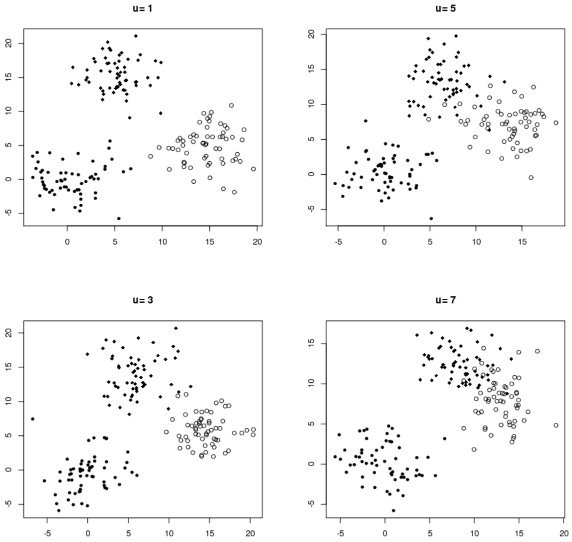

Figure 7: Second experiment's setting for $u \in \{1, 3, 5, 7\}$

where $\mu(r_n(\hat{\mathbf{c}}))$ (and respectively $\sigma(r_n(\hat{\mathbf{c}}))$) is the mean clustering risk over the 100 runs (and respectively the standard deviation).

This study highlights the robustness of the Noisy k-means in comparison with the k-means. Indeed, the associated IC are well-separated when $u \geq 7$.

4.2 Second experiment

4.2.1 Simulation setup

In this simulation, we consider, for $u \in \{1, \ldots, 10\}$, the model $\mathbf{Mod2}(u)$ as follows:

$$Z_i = X_i(u) + \epsilon_i, \; i = 1, \ldots, n, \quad \mathbf{Mod2}(u)$$

where:

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

$$f = 1/3 f_{\mathcal{N}(0_2, I_2)} + 1/3 f_{\mathcal{N}((a,b)^T, I_2)} + 1/3 f_{\mathcal{N}((b,a)^T, I_2)},$$

  where $(a, b) = (15 - (u-1)/2, 5 + (u-1)/2)$, for $u \in \{1, \ldots, 10\}$,

- and $(\epsilon_i)_{i=1}^n$ are i.i.d. with law $\mathcal{N}(0_2, \Sigma)$, where Σ is a diagonal matrix with diagonal vector $(5, 5)^T$.

In this case, the errors in variables is stable but the distance between the clusters is decreasing (see Figure 7).

As in the first experiment, we study the behaviour of both algorithms in $\mathbf{Mod2}(u)$, for $u \in \{1, \ldots, 10\}$. For this purpose, for each u, we run 100 realizations of training set

$\{Z_1, \ldots, Z_n\}$ from $\mathbf{Mod2}(u)$ with $n = 180$. At each realization, we run Lloyd algorithm and Noisy k-means with the same random initialization and the same tuned choice of $\hat{\lambda} > 0$ described in Section 4.1.

4.2.2 Results of the second experiment

Figure 8 illustrates the result of the second experiment. At first, it shows rather well the difficulty of the problem when the distance between the clusters decreases. When parameter u increases, the performances are deteriorated. However, the noisy k-means shows a better robustness for this problem. Here is a detailed explanation of Figure 8.

A These boxplots show the evolution of the clustering risk (13) of the two algorithms when parameter u increases. When $u \in \{1, \ldots, 7\}$, the results are comparable and standard k-means seems to outperform slightly Noisy k-means. However, when $u \geq 8$, Lloyd algorithm shows a very bad behaviour. On the contrary, Noisy k-means seems to be more robust in these difficult situations.

B Here, we are interested on situations where the studied algrithms fail, i.e. when the clustering errors $r_n(\hat{\mathbf{c}}) > 0.2$. Figure 5.B shows rather well the better performances of the Noisy k-means in comparison with the standard k-means when $u \leq 8$. However, the number of failures becomes problematic for k-means and Noisy k-means when $u \geq 9$ where the numbers of failures is bigger than 30 (over a total of 100 runs). It corresponds to very difficult clustering problems.

C Figure 8.C is a precise illustration of the behaviour of the two algorithms in the particular model $\mathbf{Mod2}(4)$. We have plot the clustering errors $r_n(\hat{\mathbf{c}})$ for each run in this model. Here, Noisy k-means outperforms standard k-means at many runs. However, at some runs, the standard k-means does a good job whereas Noisy k-means seems to fail. This can be explained by the dependence on the solution to the random iteration (and then on the non-convexity of the problem).

D Finally, last plot deals with the mean clustering risk in each model $\mathbf{Mod2}(u)$, $u \in \{1, \ldots, 10\}$ with the corresponding confidence intervals (see the previous subsection for the precise formula). With such statistics, the ability of noisy k-means seems to be clearly better than k-means, for any value of $u \in \{1, \ldots, 10\}$.

4.3 Conclusion of the experimental study

This experimental study can be seen as a first step into the study of clustering from a corrupted data. The results of this section are quiet promising and show rather well the importance of the deconvolution step in this inverse statistical learning problem. The first experiments show that standard k-means is not abble to separate two spherical gaussians in the presence of an additive vertical noise. In this case, noisy k-means appears as an interesting solution, mainly when this noise is increasing. Moreover, we also show that in the presence of three spherical gaussians with different separations, noisy k-means outperforms k-means.

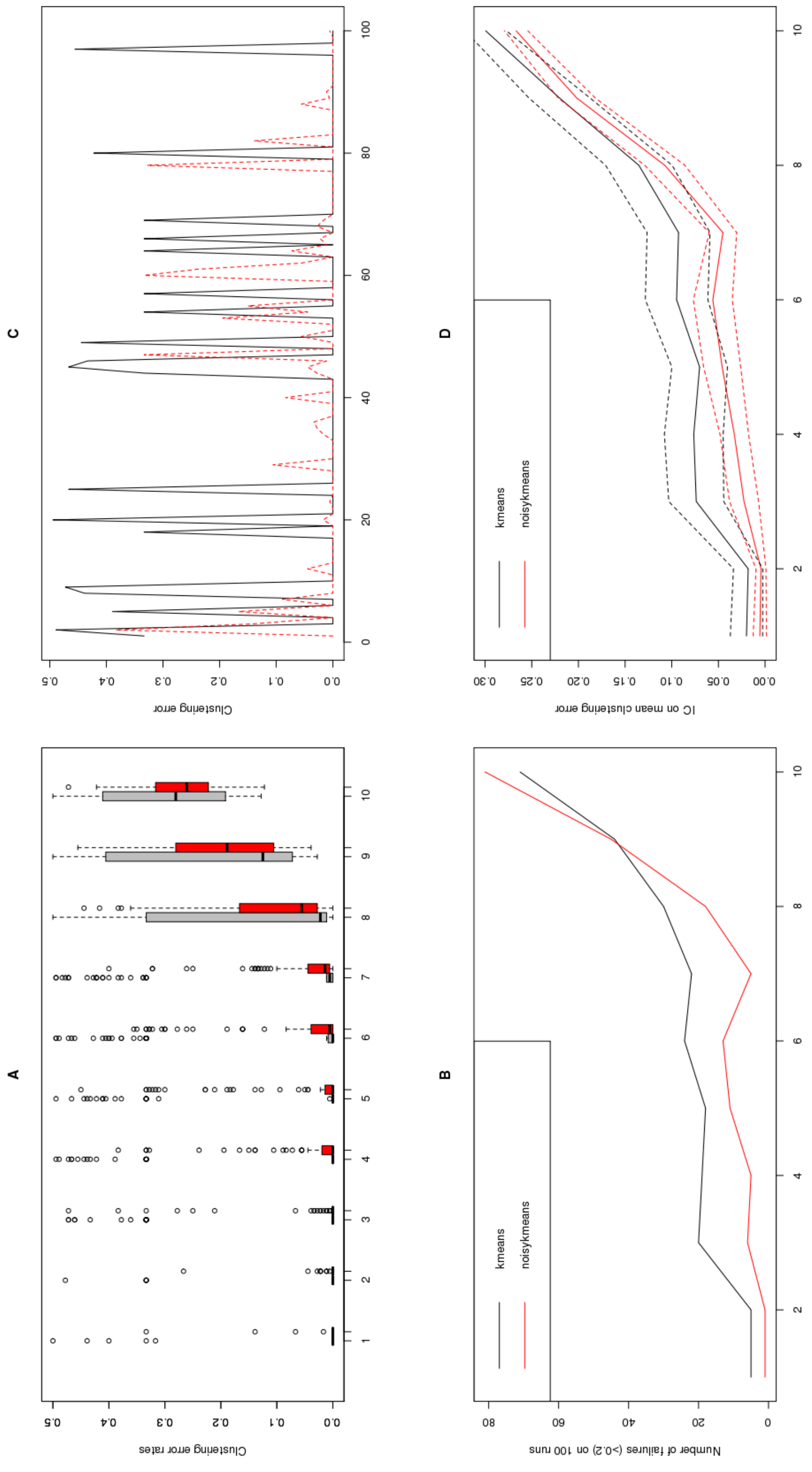

Figure 8: Results of the second experiment

5. Conclusion

This technical report presents a new algorithm to deal with clustering with errors in variables. The procedure mixes deconvolution tools with standard k-means iterations. The design of the algorithm is based on the calculus of the first order conditions of a deconvolution empirical distortion, based on a deconvolution kernel. As a result, the algorithm can be seen as the indirect counterpart of the Lloyd algorithm of the direct case, which appears to do exactly Newton's iterations (see Bubeck (2002)). Due to the deconvolution step, the algorithm extensively uses the two-dimensional FFT (Fast Fourier Transform) algorithm.

We show numerical results in particular simulated examples to illustrate various phenomena. At first, we show that the standard k-means is not abble to separate two spherical gaussian in the presence of a vertical noise. On the contrary, noisy k-means can deal with measurement errors thanks to the deconvolution step. Moreover, we show that noisy k-means is also more suitable to separate three spherical gaussians with different separations.

This algorithm could be considered as a staple into the study of noisy clustering, or more generally classification with errors in variables. As the popular k-means, it suffers from non-convexity and as the popular k-means, the initialization affects the performances. Moreover, due to the inverse problem we have at hand, the dependence on the bandwidth $\lambda > 0$ has to be considered seriously. In this paper, we perform the algorithm with a tuning choice of the bandwidth, where the criterion to choose the bandwidth depends on the density f itself. Of course in practice this choice is not available and the problem of choosing the bandwidth is a challenging future direction. However, this problem is not out of reach. Indeed, recently, Chichignoud and Loustau (2013) proposes an adaptive choice of the bandwidth based on the Lepski's procedure.

Another problem of the algorithm of this paper is the dependence on the law of the noise ϵ. The construction of the deconvolution kernel needs the entire knowledge of the density η, which is used in the algorithm of noisy k-means. This problem is very popular in the statistical inverse problem literature, where various solutions are proposed. The most classical one is to use repeated measurements to estimate the law of ϵ. In this direction, we have compiled an adaptive (to the noise) algorithm to deal with repeated measurements. This algorithm estimates the Fourier transform of η thanks to the repeated measurements. We omit this part for concisions.

Finally, applications to real-datasets is still in progress. To the best of our knowledge, the problem of clustering with noisy data is rather new and benchmark datasets are not easily available. However, there is nice hope that existing and popular datasets could be considered in future works. In this paper, we highlight good robustness for noisy k-means when the different clusters are not well separated. An interesting direction for future works is to consider difficult datasets which can be fitted to the model of clustering with errors in variables. Indeed ,to perform the algorithm, the question is the following : is it a model of clustering with errors in variables ? Of course, the answer is not possible without a detailed knowledge of the experimental set-up and eventually repeated measurements. We argue that this knowledge could be a way of improving classification rates for many problems.

References

S. Bubeck. How the initialization affects the k means. *IEEE Tansactions on Information Theory*, 48:2789–2793, 2002.

L. Cavalier. Nonparametric statistical inverse problems. *Inverse Problems*, 24:1–19, 2008.

M. Chichignoud and S. Loustau. Adaptive noisy clustering. Submitted, 2013.

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

J.A. Hartigan. *Clustering algorithms*. Wiley, 1975.

S.P. Lloyd. Least square quantization in pcm. *IEEE Transactions on Information Theory*, 28 (2):129–136, 1982.

S. Loustau. Anisotropic oracle inequalities in noisy quantization. Submitted, 2013a.

S. Loustau. Inverse statistical learning. *Electronic Journal of Statistics*, 7:2065–2097, 2013b.

S. Loustau and C. Marteau. Minimax fast rates for discriminant analysis with errors in variables. In revision to Bernoulli, 2012.

A. Meister. *Deconvolution problems in nonparametric statistics*. Springer-Verlag, 2009.

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

A.B. Tsybakov. *Introduction à l'estimation non-paramétrique*. Springer-Verlag, 2004.

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

M.P. Wand. Fast computation of multivariate kernel estimators. *Journal of Computational and Graphical Statistics*, 3 (4):433–445, 1994.