Bernoulli **24**(1), 2018, 271–296
DOI: 10.3150/16-BEJ876

Sparse oracle inequalities for variable selection via regularized quantization

CLÉMENT LEVRARD

Université Paris Diderot, 8 place Aurélie Nemours, 75013 Paris.
E-mail: levrard@math.univ-paris-diderot.fr

We give oracle inequalities on procedures which combines quantization and variable selection via a weighted Lasso k-means type algorithm. The results are derived for a general family of weights, which can be tuned to size the influence of the variables in different ways. Moreover, these theoretical guarantees are proved to adapt the corresponding sparsity of the optimal codebooks, suggesting that these procedures might be of particular interest in high dimensional settings. Even if there is no sparsity assumption on the optimal codebooks, our procedure is proved to be close to a sparse approximation of the optimal codebooks, as has been done for the Generalized Linear Models in regression. If the optimal codebooks have a sparse support, we also show that this support can be asymptotically recovered, providing an asymptotic consistency rate. These results are illustrated with Gaussian mixture models in arbitrary dimension with sparsity assumptions on the means, which are standard distributions in model-based clustering.

Keywords: clustering; high dimension; k-means; Lasso; oracle inequalities; sparsity; variable selection

1. Introduction

Let P be a distribution over $\mathbb{R}^d$. Quantization is the problem of replacing P with a finite set of points, without loosing too much information. To be more precise, if k denotes an integer, a k points quantizer Q is defined as a map from $\mathbb{R}^d$ into a finite subset of $\mathbb{R}^d$ with cardinality k. In other words, a k-quantizer divide $\mathbb{R}^d$ into k groups, and assigns each group a representative, providing both a compression and a classification scheme for the distribution P.

The quantization theory was originally developed as a way to answer signal compression issues in the late 1940s (see, e.g., [8]). However, unsupervised classification is also in the scope of its application. Isolating meaningful groups from a cloud of data is a topic of interest in many fields, from social science to biology.

Assume that P has a finite second moment, and let Q be a k points quantizer. The performance of Q in representing P is measured by the distortion

$$R(Q) = P\|x - Q(x)\|^2,$$

where Pf means integration of f with respect to P. It is worth pointing out that many other distortion functions can be defined, using $\|x - Q(x)\|^r$ or more general distance functions (see, e.g., [7] or [9]). However, the choice of the Euclidean squared norm is convenient, since it allows to fully take advantage of the Euclidean structure of $\mathbb{R}^d$, as described in [12]. Moreover, from a practical point of view, the k-means algorithm (see [14]) is designed to minimize this squared-norm distortion and can be easily implemented.

Since the distortion is based on the Euclidean distance between a point and its image, it is well known that only nearest-neighbor quantizers are to be considered (see, e.g., [9] or [19]). These quantizers are quantizers of the type $x \mapsto \operatorname{argmin}_{j=1,\dots,k} \|x - c_j\|$, where the c_i's are elements of $\mathbb{R}^d$ and are called code points. A vector of code points $(c_1, \dots, c_k)$ is called a codebook, so that the distortion takes the form

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

It has been proved in [18] that, whenever $P\|x\|^2 < \infty$, there exists optimal codebooks, denoted by $\mathbf{c}^*$.

Let $X_1, \dots, X_n$ denote an independent and identically distributed sample drawn from P, and denote by P_n the associated empirical distribution, namely, for every measurable subset A, $P_n(A) = 1/n|\{i|X_i \in A\}|$. The aim is to design a codebook from this n-sample, whose distortion is as close as possible to the optimum $R(\mathbf{c}^*)$. The k-means algorithm provides the empirical codebook $\hat{\mathbf{c}}_n$, defined by

$$\hat{\mathbf{c}}_n = \operatorname{argmin} \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,k} \|X_i - c_j\|^2 = \operatorname{argmin} P_n \min_{j=1,\dots,k} \|x - c_j\|^2.$$

Unfortunately, if $P^{(p)} \neq 0$, where $P^{(p)}$ denotes the marginal distribution of P on the pth coordinate, then $\hat{\mathbf{c}}_n^{(p)} = (\hat{c}_1^{(p)}, \dots, \hat{c}_k^{(p)})$ may not be zero, even if the pth coordinate has no influence on the classification provided by the k-means. For instance, if $\mathbf{c}^{*,(p)} = 0$, and $P^{(p)}$ has a density, then $\hat{\mathbf{c}}_n^{(p)} \neq 0$ almost surely. This suggests that the k-means algorithm does not provide sparse codebooks, even in the case where some variables plays no role in the classification, which can be detrimental to the computational tractability and to the interpretation of the corresponding clustering scheme in high-dimensional settings.

Consequently, when d is large, a variable selection procedure is usually performed preliminary to the k-means algorithm. The variable selection can be achieved using penalized BCCS strategies, as exposed in [5] or [29]. Though these procedures offer good performance in classifying the sample $X_1, \dots, X_n$, under the assumption that the marginal distributions $P^{(p)}$ are independent, no theoretical result on the prediction performance has been given. An other way to perform variable selection can be to select coordinates whose empirical variances are larger than a determined ratio of the global variance, following the idea of [21]. This algorithm has shown good results on practical examples, such as curve clustering (see, e.g., [1]). However, there is no theoretical result on the prediction performance of the selected coordinates.

Algorithms combining variable selection through PCA and clustering via k-means, like RKM (Reduced k-means, introduced in [6]) and FKM (Factorial k-means, introduced in [28]), are also very popular in practice. Some results on the performance in classifying the sample $X_1, \dots, X_n$ have been derived in [25] under strong conditions on P. In addition, some asymptotic prediction results on these procedures have been established in [23] and [24], showing that both the resulting codebook and its distortion converge almost surely to respectively a minimizer of the distortion constrained on a lower-dimensional subspace of $\mathbb{R}^d$ and the distortion of the latter, following the approach of [18]. However, these methods could be unsuitable for interpreting which variables are relevant for the clustering. In addition, no bounds on the excess distortion are available to our knowledge, and the choice of the dimension of the reduction space remains a hard issue, tackled in our procedure by a L_1-type penalization.

In fact, excess risk bounds for procedures combining dimensionality reduction and clustering are mostly to be found in the model-based clustering literature (see, e.g., [16] for a L_0-type penalization method, and [17] for a L_1-type penalization method). This approach, consisting in modeling P via a Gaussian mixture with sparse means through density estimation via constrained Maximum Likelihood Estimators, is clearly connected to ours. In fact, most of the derivation for the oracle inequalities stated in this paper use the same tools, drawn from empirical process theory. Nevertheless, no results on the convergence of the estimated means (i.e., model consistency) have been derived in this framework, and this model-based approach theoretically fails when P is not continuous, unlike k-means one (see, e.g., [12]).

This paper exposes a theoretical study of a weighted Lasso type procedure adapted to k-means, as suggested in [22]. Results are given for a general family of weights, encompassing the weights proposed in [22] as well as those proposed in [27] in a Generalized Linear Models for regression setting. To be more precise, we provide non-asymptotic excess distortion bounds along with model consistency results, under weaker conditions than ones required in [22] (for instance, the coordinates are not assumed to be independant), and adapting the sparsity of the optimal codebooks. From these non-asymptotic bounds, some asymptotic rates of convergence are derived when both the dimension and the sample size are large, showing that these Lasso type procedures may be suitable for high dimensional quantization. Interestingly, the excess distortion bounds are valid in the case where it may exist several optimal codebooks, contrary to results in [22] and [27]. These results are illustrated with Gaussian mixture distributions, often encountered in model-based clustering literature, showing at the same time that optimal codebooks can be proved to be unique for this type of distributions, under some conditions on the variances of the components of the mixture.

The paper is organized as follows. Some notation are introduced in Section 2, along with the Lasso k-means procedure and the different assumptions. The consistency and prediction results are gathered in Section 3, and the proof of these results are exposed in Section 4. At last, the proofs of some auxiliary results are deferred to the supplementary material [13].

2. Notation

Let x be in $\mathbb{R}^d$, then the pth coordinate of x will be denoted by $x^{(p)}$. Throughout this paper, it is assumed that, for every $p = 1, \dots, d$, there exists a sequence M_p, such that $|x^{(p)}| \leq M_p$ P-almost surely. In other words P is assumed to have bounded marginal distributions $P^{(p)}$. To shorten notation, the Euclidean coordinate-wise product $\prod_{p=1}^{d} [-M_p, M_p]$ will be denoted by C. To frame quantization as a contrast minimization issue, let us introduce the following contrast function

$$\gamma : \begin{cases} \left(\mathbb{R}^d\right)^k \times \mathbb{R}^d \longrightarrow \mathbb{R}, \\ (\mathbf{c}, x) \longmapsto \min\limits_{j=1,\dots,k} \|x - c_j\|^2, \end{cases}$$

where $\mathbf{c} = (c_1, \dots, c_k)$ denotes a codebook, that is a kd-dimensional vector. The risk $R(\mathbf{c})$ then takes the form $R(\mathbf{c}) = R(Q) = P\gamma(\mathbf{c}, \cdot)$, where we recall that Pf denotes the integration of the function f with respect to P. Similarly, the empirical risk $\hat{R}_n(\mathbf{c})$ can be defined as $\hat{R}_n(\mathbf{c}) = P_n\gamma(\mathbf{c}, \cdot)$, where P_n is the empirical distribution associated with $X_1, \dots, X_n$, in other words $P_n(A) = 1/n|\{i|X_i \in A\}|$, for every measurable subset $A \subset \mathbb{R}^d$. The usual k-means codebook $\hat{\mathbf{c}}_n$ is then defined as a minimizer of $\hat{R}_n(\mathbf{c})$.

It is worth pointing out that, since the support of P is bounded, then there exist such minimizers $\hat{\mathbf{c}}_n$ and $\mathbf{c}^*$ (see, e.g., Corollary 3.1 in [7]). In the sequel, the set of minimizers of the risk $R(\cdot)$ will be denoted by $\mathcal{M}$. Then, for any codebook $\mathbf{c}$, the loss $\ell(\mathbf{c}, \mathbf{c}^*)$ may be defined as the excess distortion, namely $\ell(\mathbf{c}, \mathbf{c}^*) = R(\mathbf{c}) - R(\mathbf{c}^*)$, for $\mathbf{c}^*$ in $\mathcal{M}$.

From now on, we assume that $k \geq 2$. Let $c_1, \dots, c_k$ be a sequence of code points. A central role is played by the set of points which are closer to c_i than to any other c_j's. To be more precise, the Voronoi cell, or quantization cell associated with c_i is the closed set defined by

$$V_i(\mathbf{c}) = \left\{x \in \mathbb{R}^d | \forall j \neq i \, \|x - c_i\| \leq \|x - c_j\|\right\}.$$

It may be noted that $(V_1(\mathbf{c}), \dots, V_k(\mathbf{c}))$ does not form a partition of $\mathbb{R}^d$, since $V_i(\mathbf{c}) \cap V_j(\mathbf{c})$ may be non empty. To address this issue, the Voronoi partition associated with $\mathbf{c}$ is defined as the sequence of subsets $W_i(\mathbf{c}) = V_i(\mathbf{c}) \setminus (\bigcup_{i>j} V_j(\mathbf{c}))$, for $i = 1, \dots, k$. It is immediate that the $W_i(\mathbf{c})$'s form a partition of $\mathbb{R}^d$, and that for every $i = 1, \dots, k$,

$$\bar{W}_i(\mathbf{c}) = V_i(\mathbf{c}),$$

where $\bar{W}_i(\mathbf{c})$ denotes the closure of the subset $W_i(\mathbf{c})$. The open Voronoi cell is defined the same way by

$$\overset{\circ}{V}_i(\mathbf{c}) = \left\{x \in \mathbb{R}^d | \forall j \neq i \, \|x - c_i\| < \|x - c_j\|\right\},$$

and the following inclusion holds, for i in $\{1, \dots, k\}$,

$$\overset{\circ}{V}_i(\mathbf{c}) \subset W_i(\mathbf{c}) \subset V_i(\mathbf{c}).$$

The risk $R(\mathbf{c})$ then takes the form

$$R(\mathbf{c}) = \sum_{i=1}^{k} P\left(\|x - c_i\|^2 \mathbb{1}_{W_i(\mathbf{c})}(x)\right),$$

where $\mathbb{1}_A$ denotes the indicator function associated with A. In the case where $P(W_i(\mathbf{c})) \neq 0$, for every $i = 1, \dots, k$, it is clear that

$$P\left(\|x - c_i\|^2 \mathbb{1}_{W_i(\mathbf{c})}(x)\right) \geq P\left(\|x - \eta_i\|^2 \mathbb{1}_{W_i(\mathbf{c})}(x)\right),$$

with equality only if $c_i = \eta_i$, where η_i denotes the conditional expectation of P over the subset $W_i(\mathbf{c})$, that is

$$\eta_i = \frac{P(x \mathbb{1}_{W_i(\mathbf{c})}(x))}{P(W_i(\mathbf{c}))}.$$

Moreover, it is proved in Proposition 1 of [10] that, for every Voronoi partition $W(\mathbf{c}^*)$ associated with an optimal codebook $\mathbf{c}^*$, and every $i = 1, \dots, k$, $P(W_i(\mathbf{c}^*)) \neq 0$. Consequently, any optimal codebook satisfies the so-called centroid condition (see, e.g., Section 6.2 of [8]), that is

$$\mathbf{c}_i^* = \frac{P(x \mathbb{1}_{W_i(\mathbf{c}^*)}(x))}{P(W_i(\mathbf{c}^*))}.$$

As a remark, the centroid condition ensures that $\mathcal{M} \subset C^k$, and, for every $\mathbf{c}^*$ in $\mathcal{M}$, $i \neq j$,

$$\begin{aligned} P\big(V_i(\mathbf{c}^*) \cap V_j(\mathbf{c}^*)\big) &= P\big(\big\{x \in \mathbb{R}^d | \forall i' \|x - c_i^*\| = \|x - c_j^*\| \leq \|x - c_{i'}^*\|\big\}\big) \\ &= 0. \end{aligned}$$

A proof of this statement can be found in Proposition 1 of [10]. According to [12], for every $\mathbf{c}^*$ in $\mathcal{M}$, the following set is of special interest:

$$N_{\mathbf{c}^*} = \bigcup_{i \neq j} V_i(\mathbf{c}^*) \cap V_j(\mathbf{c}^*).$$

To be more precise, the key quantity is the margin function, which is defined as

$$p(t) = \sup_{\mathbf{c}^* \in \mathcal{M}} P\big(N_{\mathbf{c}^*}(t)\big),$$

where $N_{\mathbf{c}^*}(t)$ denotes the t-neighborhood of $N_{\mathbf{c}^*}$. As shown in [12], bounds on this margin function (see Assumption 2 below) can provide interesting results on the convergence rate of the k-means codebook, along with basic properties of optimal codebooks.

In order to perform both variable selection and quantization, we introduce the Lasso k-means codebook $\hat{\mathbf{c}}_{n,\lambda}$ as follows

$$\hat{\mathbf{c}}_{n,\lambda} \in \operatorname*{argmin}_{\mathbf{c} \in C^k} P_n \gamma(\mathbf{c}, \cdot) + \lambda I_{\hat{w}}(\mathbf{c}), \tag{1}$$

where $\hat{w}$ is a possibly random sequence of weights of size d, and $I_{\hat{w}}()$ denotes the penalty function

$$I_{\hat{w}}(\mathbf{c}) = \sum_{p=1}^{d} \hat{w}_p \|\mathbf{c}^{(p)}\|. \tag{2}$$

Let us recall here that $\mathbf{c}^{(p)}$ denote the vector $(c_1^{(p)}, \dots, c_k^{(p)})$ made of the pth coordinates of the different codepoints. The results exposed in the following section are illustrated with three sequences of weights, corresponding to different codebooks: the *plain Lasso* codebook, defined by the deterministic sequence $\hat{w}_p = 1$, the *normalized Lasso* codebook, defined by $\hat{w}_p = \hat{\sigma}_p$, and the *threshold Lasso* codebook, which is a slight modification of the original Lasso-type procedure mentioned in [22] and is defined by $\hat{w}_p = 1/(\delta \vee \|\hat{\mathbf{c}}_n^{(p)}\|)$, where $\hat{\mathbf{c}}_n$ denotes the k-means codebook and δ a parameter to be tuned. It is likely that other families of weights may be of special interest, for instance combining normalization and threshold. Consequently, the results are derived for an arbitrary family of weights satisfying some convergence conditions.

These L_1-type penalties have been designed to drive the irrelevant (p)th coordinates $c_1^{(p)}, \dots, c_k^{(p)}$ together to zero (see, e.g., [2]), according to different criterions. Note that this kind of penalties is well-adapted to centered distributions. In practice, centering the data provides codebooks of the form $(\hat{c}_{n,\lambda,1} + \bar{X}, \dots, \hat{c}_{n,\lambda,k} + \bar{X})$ for the non centered distribution, where $\bar{X}$ denotes the empirical mean and $\hat{\mathbf{c}}_{n,\lambda}$ is hopefully sparse. From a theoretical point of view, deriving how close the codebook $\hat{\mathbf{c}}_{n,\lambda}$ computed on the centered data is to a codebook $\mathbf{c}^* - m$ would require a bound on $\|\bar{X} - m\|$, where m is the mean of P. In our framework, such a bound is typically of order $\sqrt{d/n}$ (see, e.g., Figure 1), hence might be unsuited for high dimensional settings. However, in some particular cases (for instance when the mean is sparse), other estimators of the means that are adapted to the high dimensional framework could be combined with our procedure.

To describe the influence of the different coordinates, the following notation are adopted. Let $S \subset \{1, \dots, d\}$ denote a subset of coordinates, then for any vector x in $(\mathbb{R}^d)^\ell$ and set $A \subset (\mathbb{R}^d)^\ell$, ℓ being a positive integer, x_S will denote the vector in $(\mathbb{R}^{|S|})^\ell$ corresponding to the coefficients of x on variables in S, and A_S will denote the set of such x_S, for x in A. Moreover, let P^S denote the marginal distribution of P over the set $\mathbb{R}^{|S|}$. We may then define the restricted distortions and variances as follows:

$$\begin{cases} \sigma_S^2 = P^S\|x\|^2, \\ \hat{\sigma}_S^2 = P_n^S\|x\|^2, \\ R_S^* = \min\limits_{\mathbf{c} \in C_S} P^S\gamma(\mathbf{c}, \cdot), \\ \hat{R}_S = \min\limits_{\mathbf{c} \in C_S} P_n^S\gamma(\mathbf{c}, \cdot), \end{cases}$$

where the vector x is element of $\mathbb{R}^{|S|}$. Elementary properties of the distortion show that, if $S = S_1 \cup S_2$, with empty intersection, then

$$\begin{cases} \sigma_S^2 = \sigma_{S_1}^2 + \sigma_{S_2}^2, \\ \hat{\sigma}_S^2 = \hat{\sigma}_{S_1}^2 + \hat{\sigma}_{S_2}^2, \\ R_S^* \geq R_{S_1}^* + R_{S_2}^*, \\ \hat{R}_S \geq \hat{R}_{S_1} + \hat{R}_{S_2}. \end{cases} \tag{3}$$

These elementary properties will be of importance when choosing which coordinate to select. A special attention will be paid to the subsets of variables formed by the support of codebooks. To be more precise, for every codebook $\mathbf{c}$ in C^k, we define the support $S(\mathbf{c})$ of $\mathbf{c}$ by $S(\mathbf{c}) = \{j \in \{1, \dots, d\} | \mathbf{c}^{(j)} \neq 0\}$. The following proposition gives a first glance at which variables are in $S(\hat{\mathbf{c}}_{n,\lambda})$.

Proposition 2.1. *Let p be in $\{1, \dots, d\}$. If*

$$\sqrt{\hat{\sigma}_p^2 - \hat{R}_p} < \frac{\hat{w}_p\lambda}{2},$$

then

$$\hat{\mathbf{c}}_{n,\lambda}^{(p)} = \left(\hat{c}_{n,\lambda,1}^{(p)}, \dots, \hat{c}_{n,\lambda,k}^{(p)}\right) = (0, \dots, 0).$$

According to Proposition 2.1, the Lasso k-means procedures may be thought of as a multimodularity test on every coordinate, in the spirit of [11]. This result ensures that, if the distortion of the codebook $(0, \dots, 0)$ is close to the optimal empirical distortion, on the pth coordinate, then the Lasso k-means will drive the pth variable to 0. For the plain Lasso, the differences $\sqrt{\hat{\sigma}_p^2 - \hat{R}_p}$ are uniformly thresholded, whereas for the normalized Lasso, the threshold in λ is applied on the ratios $\hat{R}_p/\hat{\sigma}_p^2$. This point suggests that the normalized Lasso may succeed in recovering informative variables with small ranges.

We introduce now the assumptions which will be required to derive theoretical results on the performance of the Lasso codebooks. To deal with the case of possibly several optimal codebooks, we introduce the following structural assumption on P.

Assumption 1. *For every* $\mathbf{c}^*$ *in* $\mathcal{M}$ *and* $\mathbf{c}$ *in* C^k, *if* $S(\mathbf{c}) \subsetneq S(\mathbf{c}^*)$, *then* $R(\mathbf{c}) > R(\mathbf{c}^*)$.

Assumption 1 roughly requires that no optimal codebook has a support strictly contained in the support of another optimal codebook. This is obviously the case if P has a unique optimal codebook, up to relabeling.

Assumption 2 (Margin condition). *There exists* $r_0 > 0$ *such that*

$$\forall t \le r_0 \qquad p(t) \le c_0(P)t, \tag{4}$$

where $c_0(P)$ *is a fixed constant, defined in* [12].

As exposed in [12], Assumption 2 may be thought of as a margin condition for squared distance based quantization. Some examples of distributions satisfying (4) are given in [12], including Gaussian mixtures under some conditions. Roughly, if P is well concentrated around k poles, then (4) will hold. It is also worth mentioning that the condition required in [22] seems stronger than the condition required in Assumption 2, since it requires P to have a unique optimal codebook, to be a mixture of components centered on the different optimal code points, and that the Hessian matrix of the risk function located at the optimal codebook is positive definite.

Moreover, Assumption 2 is a sufficient condition to ensure that some elementary properties that are often assumed are satisfied, as described in the following proposition.

Proposition 2.2. *If* P *satisfies Assumption* 2, *then*

(i) $\mathcal{M}$ *is finite*,
(ii) *there exists* $\kappa_0' > 0$ *such that, for every* $\mathbf{c}$ *in* C^k, $\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2 \le \kappa_0' \ell(\mathbf{c}, \mathbf{c}^*)$,

where $\mathbf{c}^*(\mathbf{c}) \in \operatorname{argmin}_{\mathbf{c}^*} \|\mathbf{c} - \mathbf{c}^*\|$.

Moreover, if P satisfies Assumption 1, then there exists a constant κ_0'' such that, for every $\mathbf{c}^*$ in $\mathcal{M}$ and $S(\mathbf{c}) \subsetneq S(\mathbf{c}^*)$, we have

$$\left\|\mathbf{c}-\mathbf{c}^*\right\|^2 \leq \kappa_0'' \ell\left(\mathbf{c}, \mathbf{c}^*\right).$$

The two first statements of Proposition 2.2 are to be found in Proposition 2.2 of [12], the proof of the third statement is given in Section 4.2. Proposition 2.2 may be thought of as a generalization of the positive Hessian matrix condition of [19] to the non-continuous case. It also allows to deal with the case where P has several optimal codebooks. In the following, we denote by κ_0 the quantity $\kappa_0' \vee \kappa_0''$, whenever Assumptions 2 and 1 are satisfied.

In addition to Assumption 2, we assume that the weights $\hat{w}_p$ satisfy a uniform concentration inequality around some deterministic weights, as stated below.

Assumption 3 (Weights concentration). *There exist deterministic weights $w_p > 0$, $p = 1, \ldots, d$, and a constant $0 \leq \kappa_1 < 1$ such that*

$$\mathbb{P}\left(\sup_{p=1,\ldots,d}\left|\frac{\hat{w}_p}{w_p} - 1\right| > \kappa_1\right) := r_1(n) \underset{n\to\infty}{\longrightarrow} 0. \tag{5}$$

Assumption 3 is obviously satisfied for the plain Lasso ($\hat{w}_p = 1$). The following proposition ensures that this statement remains true for the two other examples of weights. For any sequence w_p, we denote by $T(w)$ the quantity $\sup_{p=1,\ldots,d} M_p/w_p$. With a slight abuse of notation, $T(\sigma)$ and $T(\delta)$ will refer to the sequences σ_p and $1/(\|\mathbf{c}^{*,(p)}\| \vee \delta)$, where the latter is well defined when P has a unique optimal codebook.

Proposition 2.3. *For $\hat{w}_p = \hat{\sigma}_p$, if $1 > \kappa_1 > \frac{T^2(\sigma)\sqrt{\log(d)}}{\sqrt{2n}}$, then Assumption 3 holds with $w_p = \sigma_p$ and $r_1(n) = e^{-(\frac{\sqrt{2n}\kappa_1}{T^2(\sigma)} - \sqrt{\log(d)})^2}$.*

For $\hat{w}_p = 1/(\|\hat{\mathbf{c}}_n^{(p)}\| \vee \delta)$, let M be defined as $M = \sqrt{M_1^2 + \cdots + M_d^2}$. If $1 > \kappa_1 > C_0\frac{M\sqrt{k}}{\sqrt{n\delta}}$, for a fixed constant C_0, Assumption 2 is satisfied, and $\mathbf{c}^*$ is unique (up to relabeling), then Assumption 3 holds with $w_p = 1/(\|\mathbf{c}^{*,(p)}\| \vee \delta)$ and $r_1(n) = e^{-(\frac{n\delta^2\kappa_1^2}{C_0^2M^2} - k)}$.

The proof of Proposition 2.3 follows from standard concentration inequalities, and can be found in the supplemental article [13]. At first sight, the assumption that $\mathbf{c}^*$ is unique seems quite restrictive. However, as exposed in Section 3.4, it can be shown that Gaussian mixtures satisfy this property, provided that the variances of the components are small enough. In fact, if P has several optimal codebooks, there is no intuition about toward which one $\hat{\mathbf{c}}_n$ will converge, hence the difficulty of defining deterministic limit weights for $\hat{w}_p$.

At last, we define the following quantities λ_0 and λ_1 which will play the role of minimal values for the regularization parameter λ, as exposed in [27].

$$\begin{cases} \lambda_0 = 8\sqrt{2\pi}\sqrt{\dfrac{k\log(kd)}{n}}T(w), \\ \lambda_1(x) = e\lambda_0\left(1+\sqrt{\dfrac{u+x}{k\log(kd)}}\right), \end{cases} \tag{6}$$

where $x > 0$ and $u = \log(\frac{\|w\|_2^2\sqrt{n}}{\sqrt{\log(kd)}})$. These two quantities come from empirical process theory, their roles are explained in Section 4. Roughly, λ_0 is the minimal value of the regularization parameter which ensures that the empirical risk is close to the true risk uniformly on C^k, and $\lambda_1(x)$ is the minimal value which ensures that the deviation between empirical and true risk may be compared to the norm I_w uniformly on C^k.

3. Results

We recall here that $k \geq 2$. The case $k = 1$ may be treated as a special case of the standard Lasso estimator for linear regression (see, e.g., Chapter 2 of [4]).

3.1. Sparsity adaptive slow rate of convergence for the distortion

Following the approach of [15], Lasso type procedures may be thought of as model selection procedures over L_1 balls. Theorem 3.1 below is the adaptation of this idea for the Lasso k-means procedures.

Theorem 3.1. *Suppose that Assumption 3 is satisfied, for some constant* $\kappa_1 < 1$, *and choose*

$$\lambda \geq \frac{\lambda_1(x)}{1-\kappa_1},$$

for some $x > 0$, *where* λ_1 *is defined in* (6). *Then, with probability larger than* $1 - r_1(n) - e^{-x}$, *for every* $\mathbf{c}^*$ *in* $\mathcal{M}$, *we have*

$$\ell\big(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*\big) \leq \inf_{r>0} \inf_{I_w(\mathbf{c})\leq r} \big(\ell\big(\mathbf{c}, \mathbf{c}^*\big) + (3-\kappa_1)\lambda(r \vee \lambda_0)\big).$$

A direct implication of Theorem 3.1 is that $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq 4\lambda(I_w(\mathbf{c}^*) \vee \lambda_0)$. Hence, choosing $\lambda \sim \lambda_1(x)$ gives a convergence rate for $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*)$ of order $T(w)/\sqrt{n}$, up to a $\log(n)$ factor. If $T(w)$ is fixed, that is, does not depend on n, this rate is roughly the same as the rate of convergence of the k-means codebook without margin assumption, as shown in [3].

Besides, some asymptotic results for $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*)$ when both d and n are large may also be deduced from Theorem 3.1, as stated by the following corollary.

Corollary 3.1. *Let* $\mathbf{c}^*$ *be in* $\mathcal{M}$ *and denote by* d^* *the quantity* $|S(\mathbf{c}^*)|$. *Assume that* $\max_{p=1,\ldots,d} M_p = O(1)$, $n^{-1}\log(d) \to 0$, *and* $n^{-1}\lambda^{-2}\log(d\sqrt{n}) \to 0$.

For $\hat{w}_p = 1$, $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) = O_P(\lambda d^*)$.

For $\hat{w}_p = \hat{\sigma}_p$, *if we further assume* $\max_{p=1,\ldots,d} \sigma_p = O(1)$ *and* $1 = O(\min_{p=1,\ldots,d} \sigma_p)$, *then* $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) = O_P(\lambda d^*)$.

This result may be compared for instance with Theorem 4.1 of [20], in the framework of high dimensional regression. In this case an asymptotic convergence rate of $d^*\lambda$ may be similarly derived under the same assumptions (up to a $\log(n)$ factor) that $\log(d)n^{-1} \to 0$ and $\lambda^{-2}n^{-1} \to 0$. This shows that the optimal distortion may be asymptotically attained for dimension d of order e^{n^κ}, with $\kappa < 1$, choosing λ of order $n^{\frac{\kappa'-1}{2}}$, with $\kappa < \kappa' < 1$.

Moreover, Corollary 3.1 can provide a convergence rate of order $O(d^* \log(d) n^{-1/2})$ for the excess distortion of these Lasso-type procedures, up to a $\log(n)$ factor, hence adapting the sparsity of the optimal codebooks. In comparison to the $O(dn^{-1/2})$ rate that can be derived for the excess distortion of the k-means codebook (see, e.g., [3]), this suggests that regularized k-means might outperform standard k-means whenever $d^* \ll d$ and d is large. Some numerical illustration of this point is given below.

Numerical illustration: We consider the Gaussian mixture distributions with 4 components, each of them having covariance matrix I_d (identity matrix on $\mathbb{R}^d$), and with the following means:

$$\mu_1 = \big(\overbrace{0.8,\ldots,0.8}^{5}, \overbrace{-0.8,\ldots,-0.8}^{5}, \overbrace{0,\ldots,0}^{d-10}\big), \qquad \mu_3 = -\mu_1,$$

$$\mu_2 = \big(\overbrace{0.8,\ldots,0.8}^{10}, \overbrace{0,\ldots,0}^{d-10}\big), \qquad \mu_4 = -\mu_2.$$

The weights of the mixture are chosen as $(0.3, 0.2, 0.2, 0.3)$. For d growing from 10 to 500, we compute the plain Lasso k-means codebooks with regularization parameter $\lambda(d) = 1.5 \times \log(d)/\sqrt{n}$, in the cases $n = 50$ and $n = 200$. Note that, since Gaussian mixture distributions have not a bounded support, this example does not fall in the scope of Theorem 3.1. This issue might be bypassed considering truncated Gaussian mixture distributions, as exposed in Section 3.4.

Following the approach of *Algorithm 1* of [22], the codebooks are computed using a Lloyd's-type algorithm: for any initial codebook, we update the assignments of data points to the closest code point and then update the code points to minimize the penalized squared distances to the previously assigned data points, using the Karush–Kuhn–Tucker condition that is necessary and sufficient when assignments are fixed. This procedure is repeated until convergence. Since every iteration decreases the penalized empirical distortion, the outcome of such an algorithm is clearly a local minimum of the penalized empirical distortion. This suggests that an effective global minimization of the penalized empirical distortion could be achieved by the comparison of the outcomes of several executions of the latter procedure with different initializations, as for the classical k-means implementation.

We choose as initial code points the ones given by the standard k-means algorithm, as suggested in [22], and the means of the mixture, leading to few iterations before convergence, based on our limited experience. Then the best of these two codebooks in terms of penalized empirical

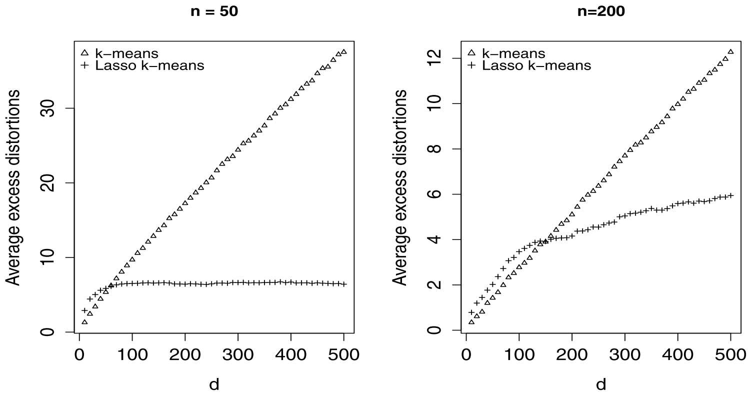

Figure 1. Average excess distortions of the plain Lasso k-means and k-means codebooks over 100 replications.

distortion is chosen. In full generality, the choice of a good initialization for such an algorithm is likely to be a crucial issue, and is beyond the scope of this paper. It may also be noted that the choice of the constant 1.5 is based on experimental observations. The calibration of such a constant might be more generally performed using cross-validation, as done in [22].

Figure 1 depicts the average distortions of both plain Lasso k-means and k-means codebook, over 100 replications. The both panels show that the excess distortion of the k-means codebook grows linearly with respect to the dimension, whereas the excess distortion of the plain Lasso k-means codebook exhibits a dependence on the dimension that looks like sub logarithmic. In fact, in the case $n = 50$, the Lasso k-means codebook turns out to be the zero codebook when d is larger than about 100, hence its constant excess distortion.

Under the sole assumptions of Corollary 3.1, no results on the threshold Lasso may be stated, since Assumption 3 cannot be checked.

3.2. Convergence towards a sparse codebook and fast rate of convergence for the distortion

If P satisfies Assumptions 2 and 1, further results may be derived, following the approach of [27]. To this aim, we defined, for a fixed codebook $\mathbf{c}^*$ and a weight family w, the set of w-sparse approximations of $\mathbf{c}^*$ at order λ by

$$\mathcal{M}_\lambda(\mathbf{c}^*) = \left\{ \underset{S(\mathbf{c})\subset S(\mathbf{c}^*)}{\operatorname{argmin}} \; 3R(\mathbf{c}) + 8\kappa_0\lambda^2\|w_{S(\mathbf{c})}\|^2 \right\},$$

where $\kappa_0 = \kappa_0' \vee \kappa_0''$, as defined below Proposition 2.2. Then, the closest w-sparse approximation of $\mathbf{c}^*$ may be defined as $\mathbf{c}^*_\lambda(\mathbf{c}^*) \in \operatorname{argmin}_{\mathbf{c}\in\mathcal{M}_\lambda(\mathbf{c}^*)} \|\mathbf{c} - \mathbf{c}^*\|$. With a slight abuse of notation, the w-sparse approximation of a codebook $\mathbf{c}$ is defined by $\mathbf{c}^*_\lambda(\mathbf{c}) = \mathbf{c}^*_\lambda(\mathbf{c}^*(\mathbf{c}))$. It is immediate that, for the plain Lasso, $\|w_{S(\mathbf{c})}\| = |S(\mathbf{c})| = \|\mathbf{c}\|_0$, whereas for the normalized Lasso, $\|w_{S(\mathbf{c})}\| = \sigma_{S(\mathbf{c})}$.

For the threshold Lasso, $\|w_{S(\mathbf{c})}\|$ has the slightly more intricate expression

$$\|w_{S(\mathbf{c})}\|^2 = \frac{1}{\delta^2}\left|S(\mathbf{c}) \cap \left\{j \,|\, \|\mathbf{c}^{*,(j)}\| \leq \delta\right\}\right| + \sum_{S(\mathbf{c}) \cap \{j \,|\, \|\mathbf{c}^{*,(j)}\| > \delta\}} \frac{1}{\|\mathbf{c}^{*,(j)}\|^2}.$$

However, it is easy to see that $\|w_{S(\mathbf{c})}\| \leq |S(\mathbf{c})|/\delta$. If we assume that the optimal codebooks $\mathbf{c}^*$ are sparse, some guarantees on the support of the $\mathbf{c}^*_\lambda(\mathbf{c}^*)$'s may be given. To be more precise, the following subset of variables is introduced

$$\left(S^+\right)^c = \bigcap_{\mathbf{c}^* \in \mathcal{M}} S\left(\mathbf{c}^*\right)^c.$$

S^+ may be thought of as the generalized support over optimal codebooks, extending the definition of these sets from the unique optimal codebook case. If P has a unique optimal codebook, it is immediate that $S^+ = S(\mathbf{c}^*)$. However, even if all the codebook are sparse, S^+ may not be sparse. For instance, if $d = k = 2$ and P is a pointwise distribution with support $(-1,-1), (-1,1), (1,-1), (1,1)$ equally weighted, then every optimal codebook has at least one zero coordinate, whereas $S^+ = \{1, 2\}$.

From its definition, it is straightforward that $S(\mathbf{c}^*_\lambda(\mathbf{c}^*)) \subset S^+$, for $\mathbf{c}^*$ in $\mathcal{M}$. Nevertheless, in the case where $S^+ = \{1, \ldots, d\}$, the $\mathbf{c}^*_\lambda(\mathbf{c}^*)$'s may still have zero coordinates. In fact, $\mathbf{c}^*_\lambda(\mathbf{c}^*)$ may be thought of as tradeoff between distortion and size of the support, the latter being measured by $\|w\|_S^2$. As in the empirical case of Proposition 2.1, this tradeoff property may be illustrated in the following way.

Proposition 3.1. *Let p be in $\{1, \ldots, d\}$. If*

$$\sigma_p^2 - R_p^* < \frac{8\lambda^2 \kappa_0 w_p^2}{3},$$

then, for every $\mathbf{c}^*$ in $\mathcal{M}$,

$$\mathbf{c}_\lambda^{*,(p)}\left(\mathbf{c}^*\right) = (0, \ldots, 0).$$

The proof of Proposition 3.1 is given in Section 4.4. Proposition 3.1, as well as Proposition 2.1, may be thought of as a comparison between the risk of optimal codebooks and the risk of the codebook $\mathbf{0}$, on the pth variable. It is worth noticing that, for the plain Lasso and threshold Lasso, the comparison is based on the difference $\sigma_p^2 - R_p^*$, whereas for the normalized Lasso only the ratio R_p^*/σ_p^2 is to be considered. Once more, this point suggests that the sparse w-approximation may not be sensitive to coordinate-wise dilations in this case. We are now in position to state sharper oracle results, on both the excess distortion and the convergence of the Lasso k-means codebooks.

Theorem 3.2. *Suppose that Assumptions 1, 2 and 3 are satisfied. If*

$$\lambda \geq \frac{2\lambda_1(x)}{1 - \kappa_1},$$

where λ_1 is defined in (6), *then, with probability larger than* $1 - r_1(n) - e^{-x}$, *we have*

$$\ell\big(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*\big) + \lambda(1-\kappa_1) I_w\big(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda(\hat{\mathbf{c}}_{n,\lambda})\big) \leq 3\ell\big(\mathbf{c}^*_\lambda, \mathbf{c}^*\big) + 8\kappa_0 \lambda^2 \|w_{S(\mathbf{c}^*_\lambda(\hat{\mathbf{c}}_{n,\lambda}))}\|^2 \vee 3\lambda\lambda_0. \quad (7)$$

A consequence of Theorem 3.2 is $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq 8\kappa_0\lambda^2 \|w_{S(\mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda}))}\|^2 \vee 3\lambda\lambda_0$, which provides an oracle inequality adapting the sparsity of the $\mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda})$'s. For instance, considering the plain Lasso, provided that $\mathbf{c}^* \neq \mathbf{0}$ for some $\mathbf{c}^*$, (7) leads to $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq 8(\kappa_0 \vee 1)|S(\mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda}))|\lambda^2 \leq 8(\kappa_0 \vee 1)|S^+|\lambda^2$. However, Theorem 3.2 also deals with the case where the $\mathbf{c}^*$'s are not sparse, comparing the Lasso k-means codebook $\hat{\mathbf{c}}_{n,\lambda}$ to the closest sparse w-approximations, for which Proposition 3.1 yields a reduced support whenever λ is large enough.

Theorem 3.2 may be considered as an application of Theorem 2.1 in [26] to the k-means case, with a slight improvement in the analysis of the complexity term (the details are to be found in the supplementary material [13]). The numerical constants in Theorem 3.2 have been arbitrarily fixed for clarity sakeness, note however that a more general version of Theorem 3.2 can be derived the same way as Theorem 2.1 in [26].

At last, it is worth pointing out that the inequality $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq 8\kappa_0\lambda^2 \|w_{S(\mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda}))}\|^2 \vee 3\lambda\lambda_0$ provides a convergence rate in $1/n$, up to a $\log(n)$ factor, when $\lambda \sim \lambda_1$ and w does not depend on n. Interestingly, this rate is the convergence rate of the k-means codebook $\hat{\mathbf{c}}_n$, when P satisfies a margin condition, as described in [12].

Similarly to Theorem 3.1, Theorem 3.2 may provide asymptotic convergence rates for both distortion and distance to optimal codebooks.

Corollary 3.2. *For the plain and normalized Lasso, assume that P satisfies Assumptions* 1, 2 *so that $\kappa_0 = O(1)$, and the requirements of Corollary* 3.1. *If $n^{-1}\log(d) \to 0$ and $n^{-1}\lambda^{-2}\log(d\sqrt{n}) \to 0$, then*

$$\ell\big(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*\big) + \lambda(1-\kappa_1) I_w\big(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda(\hat{\mathbf{c}}_{n,\lambda})\big) = O_P\big(|S^+|\lambda^2\big).$$

For the threshold Lasso, assume that $\mathbf{c}^*$ is unique, P satisfies Assumption 2 *so that $\kappa_0 = O(1)$, and $\max_{p=1,\ldots,d} M_p = O(1)$. If δ and λ are chosen so that $\delta \to 0$, $dn^{-1}\delta^{-2} \to 0$, and $\lambda^{-1}\log(n)^{1/2}n^{-1/2} \to 0$, then*

$$\ell\big(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*\big) + \lambda(1-\kappa_1) I_w\big(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda(\mathbf{c}^*)\big) = O_P\big(d^*\lambda^2\big).$$

As a consequence, in every case, $\|\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda})\| = O_P(\sqrt{|S^+|}\lambda)$.

According to Theorem 3.1 of [12], the excess distortion of the k-means codebook is of order $O(dn^{-1})$, under the assumptions of Corollary 3.2. In comparison, Corollary 3.2 yields convergence rates of order $O(|S^+|\log(d)n^{-1})$, up to a $\log(n)$ factor, for the plain and normalized Lasso. Again, this suggests that these procedures might outperform standard k-means procedures in terms of distortion in high dimensional settings, when $|S^+| \ll d$.

The requirement $\kappa_0 = O(1)$ may be thought of as an assumption on the local strong convexity of the excess distortion. This condition is similar to a uniform lower bound on the Hessian matrix of the excess distortion, as required for the asymptotic results in [22]. This asymptotic framework also allows for further comparison between our results and Theorem 1 of [22], which states that, when choosing $\hat{w}_p = \|\hat{\mathbf{c}}_n^{(p)}\|^{-1}$, provided that $\mathbf{c}^*$ is unique, $n^{1/2}\lambda d \to 0$, and $n^{-2}\lambda^{-2}d \to 0$, $\|\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*\| = O_P(n^{1/2}\lambda d^{-1})$.

If we choose λ close to the given lower bounds, for instance, $\lambda = n^{-1/2}\log(d\sqrt{n})^{-1/2}u_n$ for our plain Lasso, and $\lambda = \sqrt{d}n^{-1}u_n$ in the setting of [22], with $u_n \to \infty$, then Theorem 1 of [22] yields $\|\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*\| = O_P(u_n n^{-1/2}d^{-1/2})$, whereas Corollary 3.2 gives a slightly worse bound, $\|\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*\| = O_P(u_n\sqrt{|S^+|}\log(d\sqrt{n})n^{-1/2})$. However, Theorem 1 of [22] is valid for $d = o(\sqrt{n})$, when Corollary 3.2 only requires $\log(d) = o(n)$ for the plain Lasso.

For the threshold Lasso, when $\mathbf{c}^*$ is unique, Corollary 3.2 requires $d = o(n)$ and $\lambda = u_n\log(n)^{1/2}n^{-1/2}$ to get bounds on $\|\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*\|$ of order $\sqrt{d^*}\log(n)^{1/2}n^{-1/2}u_n$, hence still worse than $u_n n^{-1/2}d^{-1/2}$. It is interesting to note that these two very similar procedures give almost the same rate for $\|\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*\|$, but with possibly very different choices of λ (of order n^{-1} in [22], and of order $n^{-1/2}$ here). Some explanation can be given, noting that our results are intended to provide bounds on the distance between $\hat{\mathbf{c}}_{n,\lambda}$ and $\mathbf{c}^*_\lambda(\mathbf{c}^*)$, where $\mathbf{c}^*_\lambda(\mathbf{c}^*)$ is possibly different from $\mathbf{c}^*$. Thus, the empirical processes involved in our derivations are not the same than those of [22]. Besides, Corollary 3.2 states bounds in terms of the I_w distance, rather than the Euclidean one. This I_w distance is of particular interest, especially for coordinates which are not in $S(\mathbf{c}^*)$. For instance, if j is not in $S(\mathbf{c}^*)$, and if we choose $\delta = d^{1/2}n^{-1/2}u_n$ along with $\lambda = u_n\log(n)^{1/2}n^{-1/2}$, then Corollary 3.2 ensures that $\|\hat{\mathbf{c}}_{n,\lambda}^{(j)}\| = O_P(d^*\sqrt{d}\log(n)^{1/2}n^{-1}u_n^2)$. This convergence rate turns out to be faster in terms of n than the one which can be derived from $\|\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*\|$. In this particular case of a unique optimal codebook, further consistency results may be given, as described in the following subsection.

3.3. Consistency of the threshold lasso k-means

Throughout this subsection, we assume that there exists a unique optimal codebook $\mathbf{c}^*$, up to relabeling. Let j be in $S(\mathbf{c}^*)^c$. Then Theorem 2 in [22] established that $\hat{\mathbf{c}}_{n,\lambda}^{(j)} \to 0$ in probability under strong assumptions on P. To be more precise, it is assumed in [22] that $P_{|V_j^*} = c_j^* + \varepsilon_j$, where $P_{|V_j^*}$ denotes the conditional law of P on the optimal Voronoi cell centered at the jth optimal code point c_j^*, ε_j has independent coordinates, and the ε_j's are independent. Theorem 3.3 below gives a generalization of this result, along with a convergence rate for $\mathbb{P}(\hat{\mathbf{c}}_{n,\lambda}^{(j)} \neq 0)$.

Theorem 3.3. *Suppose that Assumption 2 is satisfied, and that d is fixed. For $\hat{w}_p = 1/(\delta \vee \|\hat{\mathbf{c}}_n^{(p)}\|)$, if $n^{-1}\log(n)\delta^{-2} \to 0$ and $\delta \to 0$, choose $\lambda \sim \delta$. Then, for every j in $S(\mathbf{c}^*)^c$, we have*

$$\mathbb{P}\big(\hat{\mathbf{c}}_{n,\lambda}^{(j)} \neq 0\big) \underset{n\to\infty}{=} O\big(e^{-n\delta^2}\big). \tag{8}$$

Moreover, for every j in $S(\mathbf{c}^*)$, we have

$$\mathbb{P}\big(\hat{\mathbf{c}}_{n,\lambda}^{(j)} = 0\big) \underset{n\to\infty}{=} O\big(e^{-n\delta^2}\big). \tag{9}$$

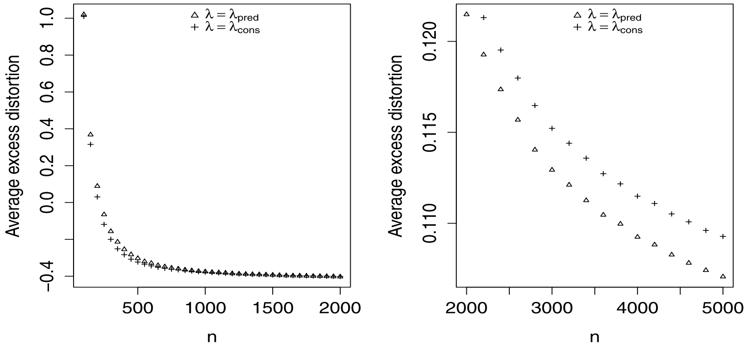

Figure 2. Average excess distortion for $\lambda \sim \lambda_1(\log(n))$ and $\lambda \sim \delta(n)$, over 100 replications.

Note that the two consistency rates given by (8) and (9) are in fact of order $o(n^{-1})$. The choice $\lambda \sim \delta$ has been made to optimize the consistency rate. However, this choice may lead to suboptimal convergence rate for $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*)$ in Corollary 3.2. For instance, if we choose $\delta = n^{-\alpha}$, $0 < \alpha < 1/2$, then this choice of λ leads to $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) = O_P(n^{-2\alpha})$. In fact, we only need $\lambda^{-1}\log(n)^{1/2}n^{-1/2} \to 0$, as in Corollary 3.2, to ensure that this model consistency result holds. Thus, the choice $\lambda \sim n^{-1/2}\log(n)^{(1/2+\varepsilon)}$, for a positive ε, provides both model consistency and almost optimal convergence of $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*)$. Some numerical illustration of this point is given below.

The way Theorem 3.3 is derived makes use of the Vapnik–Chervonenkis dimension of the Voronoi cells associated with the codebooks. Provided that a sharp bound on this dimension can be given, some asymptotic results when both d and n tend to infinity could also be stated.

Numerical illustration: To illustrate this consistency result, we consider the same Gaussian mixture distribution as in Section 3.1, but with fixed dimension $d = 100$ and sample size n growing from 100 to 5000. The threshold δ is chosen as $\delta(n) = n^{-1/6}$, and the threshold Lasso k-means codebooks are computed for two sequences of regularization parameters, namely $\lambda_{\text{pred}}(n) = 0.12 \times \log(n)/\sqrt{n}$ and $\lambda_{\text{cons}}(n) = 0.12 \times \delta(n)$. The choice of the constant 0.12 is based on experimental observations and is clearly suboptimal for the small values of n, however it renders the comparison between the two regularization parameters easier.

The right-hand side panel of Figure 2 shows that the threshold Lasso with parameter λ_{pred} asymptotically outperforms the threshold Lasso with parameter λ_{cons} in terms of distortion, as expected. However, for values of n below 1000, the λ_{cons} regularization gives the best excess distortion, as shown by the left-hand side panel. The intuition behind Figure 2 is that for small values of n, λ_{pred} under-penalizes irrelevant coordinates $p \geq 11$ compared to λ_{cons}, hence provides a too large support. On the other hand, for large values of n, the optimal support $\{1, \dots, 10\}$ is recovered by both regularization strategies, and in this case the milder penalization λ_{pred} leads to better excess distortion. This intuition is confirmed by Figure 3.

The left-hand side panel of Figure 3 illustrates the fact that the optimal support is more efficiently recovered through the λ_{cons} regularization strategy. In the case $n \geq 2000$ corresponding to the right-hand side panel, the support of the threshold Lass k-means codebook with regular-

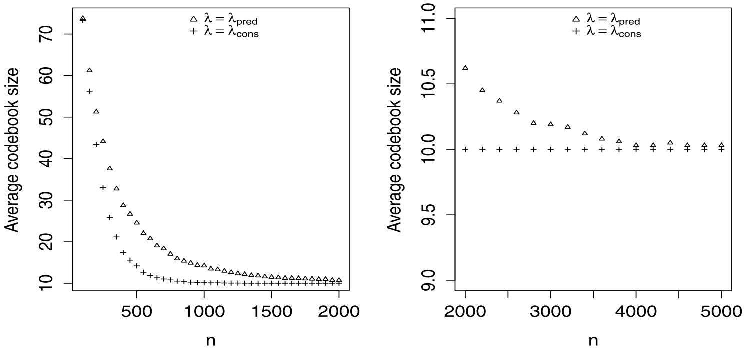

Figure 3. Average support size for $\lambda \sim \lambda_1(\log(n))$ and $\lambda \sim \delta(n)$, over 100 replications.

ization parameter λ_{cons} is exactly the optimal support in each of the replications, whereas the support corresponding to λ_{pred} may sometimes contain some irrelevant coordinates. This illustrates that the probability of exact support recovery given by Theorem 3.3 is in fact larger for the λ_{cons} strategy than for the λ_{pred} strategy.

To apply Theorem 3.3, checking that $\mathbf{c}^*$ is unique remains a hard issue when $d > 1$. The following section gives an example where this problem can be tackled using straightforward consequences of Assumption 2.

3.4. Quasi-Gaussian mixture example

The aim of this section is to provide some theoretical background for the application of Theorem 3.3 to the Gaussian mixture distributions. In general, a Gaussian mixture distribution $\tilde{P}$ may be defined by its density

$$\tilde{f}(x) = \sum_{i=1}^{\tilde{k}} \frac{\theta_i}{(2\pi)^{d/2}\sqrt{|\Sigma_i|}} e^{-\frac{1}{2}(x-m_i)^t \Sigma_i^{-1}(x-m_i)},$$

where $\tilde{k}$ denotes the number of components of the mixture, and the θ_i's denote the weights of the mixture, which satisfy $\sum_{i=1}^{k} \theta_i = 1$. Moreover, the m_i's denote the means of the mixture, so that $m_i \in \mathbb{R}^d$, and the Σ_i's are the $d \times d$ variance matrices of the components. In this case, we define the active set $\tilde{S}$ as $S(\mathbf{m})$, where $\mathbf{m}$ denotes the codebook with code points the m_i's. For convenience it is assumed that $\tilde{S} = \{1, \ldots, d'\}$, with $d' < d$. We restrict ourselves to the case where the number of components $\tilde{k}$ is known, and match the size k of the codebooks.

Since the support of a Gaussian random variable is not bounded, we define the "quasi-Gaussian" mixture model as follows, truncating each Gaussian component. Let the density f

of the distribution P be defined by

$$f(x) = \sum_{i=1}^{k} \frac{\theta_i}{(2\pi)^{d/2} N_i \sqrt{|\Sigma_i|}} e^{-\frac{1}{2}(x-m_i)^t \Sigma_i^{-1}(x-m_i)} \mathbb{1}_{\mathcal{B}(0,M)},$$

where N_i denotes a normalization constant for each Gaussian variable. To ensure that this model is close to the Gaussian mixture model, M has to be chosen large enough, say $M \geq 2\sup_{j=1,\ldots,k} \|m_j\|$ for instance. Let σ^2 and σ_-^2 denote the largest and smallest eigenvalues of the Σ_i's. Then the following proposition states that, provided that σ is small enough, a model consistency result can be derived for the threshold Lasso.

Proposition 3.2. *Assume that $\sigma_- \geq c^- \sigma$, for some constant c^-. Then there exists a constant σ^+ such that, if $\sigma \leq \sigma^+$, then Assumption 2 holds, and $\mathbf{c}^*$ is unique.*

Moreover, if $(\Sigma_i)_{pq} = 0$, for every (i, p, k) in $\{1, \ldots, k\} \times \{1, \ldots, d'\} \times \{d'+1, \ldots, d\}$, then, for every $j \geq d'+1$, the threshold Lasso with $n^{-1}\log(n)\delta^{-2} \to 0$ satisfies

$$\mathbb{P}\big(\hat{\mathbf{c}}_{n,\lambda}^{(j)} \neq 0\big) \underset{n\to\infty}{=} O\big(e^{-n\delta^2}\big).$$

The assumption on the covariance matrices requires that the variable in $\tilde{S}$ are independent from the variables in $\tilde{S}^c$. As shown in Section 4.8, this strong requirement ensures in fact that the support of the optimal codebook is contained in $\tilde{S}$. Proposition 3.2 then directly follows from Theorem 3.3.

The first part of Proposition 3.2 extends the results of Proposition 3.2 in [12] to arbitrary dimension d. It also enhances the results of this Proposition, showing that there exists a unique optimal codebook instead of finitely many ones.

It is worth noting that Proposition 3.2 is valid when $k = \tilde{k}$. When k differs from $\tilde{k}$, Assumption 2 may not be satisfied. For instance, if P is rotationally symmetric and $k, d \geq 2$, then $\mathcal{M}$ cannot be finite, which contradicts Proposition 2.2.

4. Proofs

4.1. Proof of Proposition 2.1

Let $W_1, \ldots, W_k$ be the Voronoi partition associated with $\hat{\mathbf{c}}_{n,\lambda}$, and let $L(\hat{\mathbf{c}}_{n,\lambda})$ be the matrix of assignments, defined by

$$L(\hat{\mathbf{c}}_{n,\lambda})_{i,j} = \mathbb{1}_{X_i \in W_j}.$$

Suppose that $\hat{\mathbf{c}}_{n,\lambda}^{(p)} \neq 0$, where $\hat{\mathbf{c}}_{n,\lambda}^{(p)}$ denotes the vector $(\hat{c}_{n,\lambda,1}^{(p)}, \ldots, \hat{c}_{n,\lambda,k}^{(p)})^t$, and denote by $X^{(p)}$ the vector $(X_1^{(p)}, \ldots, X_n^{(p)})^t$. Then the Karush–Kuhn–Tucker condition, for this penalized empirical risk minimization strategy, ensures that (see, e.g., the proof of Theorem 2 in [22])

$$\frac{-2}{\sqrt{n}} L(\hat{\mathbf{c}}_{n,\lambda})^t \big(X^{(p)} - L(\hat{\mathbf{c}}_{n,\lambda}) \hat{\mathbf{c}}_{n,\lambda}^{(p)} \big) + \sqrt{n} \lambda \frac{\hat{w}_p \hat{\mathbf{c}}_{n,\lambda}^{(p)}}{\| \hat{\mathbf{c}}_{n,\lambda}^{(p)} \|} = 0. \tag{10}$$

Since $L(\hat{\mathbf{c}}_{n,\lambda})^t (X^{(p)} - L(\hat{\mathbf{c}}_{n,\lambda}) \hat{\mathbf{c}}_{n,\lambda}^{(p)})$ is the following vector of size k

$$\bigg(\sum_{X_i \in W_1} \big(X_i^{(p)} - \hat{c}_{n,\lambda,1}^{(p)} \big), \dots, \sum_{X_i \in W_k} \big(X_i^{(p)} - \hat{c}_{n,\lambda,k}^{(p)} \big) \bigg),$$

it may be noted that

$$\big\| L(\hat{\mathbf{c}}_{n,\lambda})^t \big(X^{(p)} - L(\hat{\mathbf{c}}_{n,\lambda}) \hat{\mathbf{c}}_{n,\lambda}^{(p)} \big) \big\|^2 = \sum_{j=1}^k n_j^2 \big(\bar{c}_j^{(p)} - \hat{c}_{n,\lambda,j}^{(p)} \big)^2,$$

where n_j denotes the number of sample vectors X_i's in W_j, and $\bar{c}_j$ denotes the empirical mean of the sample over the set W_j, that is $\bar{c}_j = \frac{1}{n_j} \sum_{X_i \in W_j} X_i$. Denote by $\hat{p}_j$ the empirical weight of W_j, that is, $\hat{p}_j = n_j / n$, then

$$\frac{1}{n^2} \big\| L(\hat{\mathbf{c}}_{n,\lambda})^t \big(X^{(p)} - L(\hat{\mathbf{c}}_{n,\lambda}) \hat{\mathbf{c}}_{n,\lambda}^{(p)} \big) \big\|^2 \leq \sum_{j=1}^k \hat{p}_j \big(\bar{c}_j^{(p)} - \hat{c}_{n,\lambda,j}^{(p)} \big)^2,$$

where $\hat{p}_j \leq 1$ has been used. Let Q_1 be the quantizer which maps W_j to $\bar{c}_j$, then it is easy to see that

$$\sum_{j=1}^k \hat{p}_j \big(\bar{c}_j^{(p)} - \hat{c}_{n,\lambda,j}^{(p)} \big)^2 = \hat{R}_p(\hat{\mathbf{c}}_{n,\lambda}) - \hat{R}_p(Q_1),$$

where we recall that $\hat{R}_p(Q) = P_n^{(p)} \|x - Q(x)\|^2$, for any quantizer Q.

Denote by $\hat{\mathbf{c}}_{n,\lambda}^{(-p)}$ the codebook that has pth coordinate 0 and the same coordinates as $\hat{\mathbf{c}}_{n,\lambda}$ otherwise, and let Q_2 be the quantizer which maps W_j to $\hat{c}_{n,\lambda,j}^{(-p)}$. Then, by definition, $\hat{R}(\hat{\mathbf{c}}_{n,\lambda}) + I_{\hat{w}}(\hat{\mathbf{c}}_{n,\lambda}) \leq \hat{R}(\hat{\mathbf{c}}_{n,\lambda}^{(-p)}) + \lambda I_{\hat{w}}(\hat{\mathbf{c}}_{n,\lambda}^{(-p)})$, and $\hat{R}(\hat{\mathbf{c}}_{n,\lambda}^{(-p)}) \leq \hat{R}(Q_2)$. Thus, direct calculation leads to $\hat{\sigma}_p^2 \geq \hat{R}_p(\hat{\mathbf{c}}_{n,\lambda})$, according to (3). Since $\hat{R}_p(\hat{\mathbf{c}}_{n,\lambda}) - \hat{R}_p(Q_1) \leq \hat{\sigma}_p^2 - \hat{R}_p$, (10) ensures that

$$\frac{\lambda \hat{w}_p}{2} \leq \sqrt{\hat{\sigma}_p^2 - \hat{R}_p}.$$

4.2. Proof of Proposition 2.2

As mentioned below Proposition 2.2, a proof of the two first statements can be found in the proof of Proposition 2.2 in [12]. The last statement follows from the compactness of $\{\mathbf{c} \in C^k | S(\mathbf{c}) \subsetneq$

$S(\mathbf{c}^*)\}$ and the fact that $\mathcal{M}$ is finite, knowing that $R()$ is continuous (see, e.g., Lemma 4.1 in [12]). This ensures that $\inf_{\mathbf{c}^* \in \mathcal{M}} \inf_{S(\mathbf{c}) \subsetneq S(\mathbf{c}^*)} \ell(\mathbf{c}, \mathbf{c}^*) \geq c > 0$, for some constant c, whenever Assumption 1 is satisfied. Since C^k is bounded, $\sup_{\mathbf{c}^* \in \mathcal{M}, S(\mathbf{c}) \subsetneq S(\mathbf{c}^*)} \|\mathbf{c} - \mathbf{c}^*\|^2 / \ell(\mathbf{c}, \mathbf{c}^*)$ is finite.

4.3. Proof of Theorem 3.1

We recall here that $T(w)$ denotes the quantity $T(w) = \max_{p=1,\dots,d} M_p / w_p$. Let also $\bar{M}(w)$ be defined as $\sqrt{k}\|w\|^2 T(w)$. It is immediate that, for every $\mathbf{c}$ in C^k, $I_w(\mathbf{c}) \leq \bar{M}(w)$. Moreover, we define $\bar{\gamma}$ by

$$\bar{\gamma}(\mathbf{c}, x) = \min_{j=1,\dots,k} -2\langle x, c_j \rangle + \|c_j\|^2,$$

for every $\mathbf{c}$ in C^k and x in $\mathbb{R}^d$. The prediction results of this paper are based on the following concentration inequality, which connects the deviation of the empirical processes $(P - P_n)\bar{\gamma}(\mathbf{c}, \cdot)$ to the I_w norm. Note that, since $\bar{\gamma}$ is continuous and $(\mathbb{R}^d)^{k+1}$ is a separable metric space, the empirical processes introduced throughout the derivation are measurable.

Proposition 4.1. *Suppose that w is a deterministic sequence of weights. Denote by u the quantity $\log(\frac{\|w\|^2 \sqrt{n}}{\sqrt{\log(kd)}})$. If we denote by*

$$\lambda_0 = 8\sqrt{2\pi}\sqrt{\frac{k \log(kd)}{n}} T(w),$$

then, for every $x > 0$, denoting by

$$\lambda_1 = e\lambda_0 \left(1 + \sqrt{\frac{u + x}{k \log kd}}\right),$$

we have, for any fixed $\mathbf{c}'$ in C^k, with probability larger than $1 - e^{-x}$,

$$\sup_{I_w(\mathbf{c} - \mathbf{c}') \leq 2\bar{M}(w)} \frac{|(P - P_n)(\bar{\gamma}(\mathbf{c}, \cdot) - \bar{\gamma}(\mathbf{c}', \cdot))|}{I_w(\mathbf{c} - \mathbf{c}') \vee \lambda_0} \leq \lambda_1. \tag{11}$$

Proposition 4.1 may be thought of as a slight generalization of inequality (7) in [26]. Its proof, given in the supplemental article [13], differs from the original one by the use of Gaussian complexities rather than Talagrand's generic chaining principle to derive a multivariate contraction principle. We are now in position to prove Theorem 3.1.

We recall here that $\lambda \geq \frac{\lambda_1}{1 - \kappa_1}$. From Assumption 3, it easily follows that $(1 - \kappa_1) I_w(\mathbf{c}) \leq I_{\hat{w}}(\mathbf{c}) \leq (1 + \kappa_1) I_w(\mathbf{c})$, with probability larger than $1 - r_1(n)$. On this event, we have, for every $\mathbf{c}$ in C^k,

$$P_n \bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda}, \cdot) + \lambda(1 - \kappa_1) I_w(\hat{\mathbf{c}}_{n,\lambda}) \leq P_n \bar{\gamma}(\mathbf{c}, \cdot) + \lambda(1 + \kappa_1) I_w(\mathbf{c}).$$

Using (11), with $\mathbf{c}' = \mathbf{0}$, it follows that

$$\begin{aligned} P\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda}, \cdot) \leq P\bar{\gamma}(\mathbf{c}, \cdot) + \lambda(1+\kappa_1) I_w(\mathbf{c}) + \lambda_1\big(I_w(\mathbf{c}) \vee \lambda_0\big) \\ + \lambda_1\big(I_w(\hat{\mathbf{c}}_{n,\lambda}) \vee \lambda_0\big) - \lambda(1-\kappa_1) I_w(\hat{\mathbf{c}}_{n,\lambda}). \end{aligned} \tag{12}$$

If $I_w(\hat{\mathbf{c}}_{n,\lambda}) > \lambda_0$, adding $-P\bar{\gamma}(\mathbf{c}^*, \cdot)$ on the both sides leads to

$$\ell\big(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*\big) \leq \ell\big(\mathbf{c}, \mathbf{c}^*\big) + \lambda(1+\kappa_1) I_w(\mathbf{c}) + \lambda_1\big(I_w(\mathbf{c}) \vee \lambda_0\big).$$

Hence,

$$\ell\big(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*\big) \leq \inf_{r>0} \inf_{I_w(\mathbf{c}) \leq r} \ell\big(\mathbf{c}, \mathbf{c}^*\big) + 2\lambda(r \vee \lambda_0).$$

If $I_w(\hat{\mathbf{c}}_{n,\lambda}) \leq \lambda_0$, (12) may be written

$$\ell\big(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*\big) \leq \ell\big(\mathbf{c}, \mathbf{c}^*\big) + 2\lambda\big(I_w(\mathbf{c}) \vee \lambda_0\big) + \lambda(1-\kappa_1)\lambda_0,$$

hence

$$\ell\big(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*\big) \leq \inf_{r>0} \inf_{I_w(\mathbf{c}) \leq r} \ell\big(\mathbf{c}, \mathbf{c}^*\big) + (3-\kappa_1)\lambda(r \vee \lambda_0).$$

4.4. Proof of Proposition 3.1

Let S be a subset of $\{1, \dots, d\}$, and let p be in S such that

$$\sigma_p^2 - R_p^* < \frac{8\kappa_0\lambda^2 w_p^2}{3}.$$

Denote by $\mathbf{c}_S^*$ an optimal codebook with support S, that is

$$\mathbf{c}_S^* = \operatorname*{argmin}_{S(\mathbf{c})=S} R(\mathbf{c}).$$

Then, according to (3), we may write

$$\begin{aligned} R\big(\mathbf{c}^*_{S\setminus\{p\}}\big) - R\big(\mathbf{c}^*_S\big) &\leq R^*_{S\setminus\{p\}} + \sigma^2_{(S\setminus\{p\})^c} - \big(R^*_{S\setminus\{p\}} + R^*_p\big) - \sigma^2_{S^c} \\ &\leq \sigma_p^2 - R_p^*. \end{aligned}$$

Therefore

$$3R\big(\mathbf{c}^*_{S\setminus\{p\}}\big) + 8\lambda^2\kappa_0 \|w_{S\setminus\{p\}}\|^2 < 3R\big(\mathbf{c}^*_S\big) + 8\lambda^2\kappa_0\|w_S\|^2.$$

4.5. Proof of Theorem 3.2

Let $\mathbf{c}$ be a fixed codebook in C^k, and $\mathbf{c}'$ be another codebook in C^k. Following the notation of [26], with a slight abuse of notation, we denote by $I_{w,1}(\mathbf{c}'-\mathbf{c})$ and $I_{w,2}(\mathbf{c}'-\mathbf{c})$ the quantities

$$\begin{cases} I_{w,1}(\mathbf{c}'-\mathbf{c}) = I_w\big((\mathbf{c}'-\mathbf{c})_{S(\mathbf{c})}\big), \\ I_{w,2}(\mathbf{c}'-\mathbf{c}) = I_w\big((\mathbf{c}'-\mathbf{c})_{S^c(\mathbf{c})}\big). \end{cases}$$

The following result is derived from Lemma A.4 in [27].

Lemma 4.1. *Let* $\mathbf{c}$ *and* $\mathbf{c}'$ *be in* C^k, *and denote by* $\mathbf{c}^* = \mathbf{c}^*(\mathbf{c}')$. *If* $S(\mathbf{c}) \subsetneq S(\mathbf{c}^*)$ *or* $\mathbf{c}^*(\mathbf{c}) = \mathbf{c}^*$, *for any* $\delta > 0$, *we have*

$$2\lambda I_{w,1}(\mathbf{c}'-\mathbf{c}) \leq \frac{1}{\delta}\ell(\mathbf{c},\mathbf{c}^*) + \frac{1}{\delta}\ell(\mathbf{c}',\mathbf{c}^*) + 2\delta\kappa_0\lambda^2\|w_{S(\mathbf{c})}\|^2. \tag{13}$$

The proof of Lemma 4.1 can be found in [27]. For the sake of completeness, it is briefly recalled here.

Proof of Lemma 4.1. Using Cauchy–Schwarz inequality, it is easy to see that

$$\begin{aligned} 2\lambda I_{w,1}(\mathbf{c}'-\mathbf{c}) &\leq 2\lambda\sqrt{\sum_{p\in S(\mathbf{c})} w_p^2}\,\|\mathbf{c}'-\mathbf{c}\| \\ &\leq 2\lambda\sqrt{\sum_{p\in S(\mathbf{c})} w_p^2}\big(\|\mathbf{c}'-\mathbf{c}^*\| + \|\mathbf{c}-\mathbf{c}^*\|\big). \end{aligned}$$

Using the inequality $2ab \leq \kappa_0\delta a^2 + \frac{1}{\delta\kappa_0}b^2$ and Proposition 2.2 leads to

$$2\lambda I_{w,1}(\mathbf{c}'-\mathbf{c}) \leq \frac{1}{\delta}\big(\ell(\mathbf{c},\mathbf{c}^*) + \ell(\mathbf{c}',\mathbf{c}^*)\big) + 2\delta\kappa_0\lambda^2\|w_{S(\mathbf{c})}\|^2. \qquad \square$$

Equipped with this lemma, we are in position to prove Theorem 3.2. By definition, $\hat{\mathbf{c}}_{n,\lambda}$ satisfies

$$P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},\cdot) + \lambda I_{\hat{w}}(\hat{\mathbf{c}}_{n,\lambda}) \leq P_n\bar{\gamma}(\mathbf{c},\cdot) + \lambda I_{\hat{w}}(\mathbf{c}).$$

Using Proposition 4.1, we get

$$P\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},\cdot) + \lambda I_{\hat{w},2}(\hat{\mathbf{c}}_{n,\lambda}-\mathbf{c}) \leq P\bar{\gamma}(\mathbf{c},\cdot) + \lambda I_{\hat{w},1}(\hat{\mathbf{c}}_{n,\lambda}-\mathbf{c}) + \lambda_1\big(\lambda_0 \vee I_w(\hat{\mathbf{c}}_{n,\lambda}-\mathbf{c})\big). \tag{14}$$

If $I_w(\hat{\mathbf{c}}_{n,\lambda}-\mathbf{c}) > \lambda_0$, then adding $-P\bar{\gamma}(\mathbf{c}^*,\cdot)$ on the both sides an using Assumption 3 leads to

$$\ell(\hat{\mathbf{c}}_{n,\lambda},\mathbf{c}^*) + (1-\kappa_1)\lambda I_{w,2}(\hat{\mathbf{c}}_{n,\lambda}-\mathbf{c}) \leq \ell(\mathbf{c},\mathbf{c}^*) + (1+\kappa_1)\lambda I_{w,1}(\hat{\mathbf{c}}_{n,\lambda}-\mathbf{c}) + \lambda_1 I_w(\hat{\mathbf{c}}_{n,\lambda}-\mathbf{c}).$$

Hence, if $\mathbf{c}^*(\mathbf{c}) = \mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda}) = \mathbf{c}^*$ or $S(\mathbf{c}) \subsetneq S(\mathbf{c}^*)$,

$$\begin{aligned}\ell\big(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*\big) + \big[(1-\kappa_1)\lambda - \lambda_1\big] I_w(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) &\leq \ell\big(\mathbf{c}, \mathbf{c}^*\big) + 2\lambda I_{w,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) \\ &\leq \frac{3}{2}\ell\big(\mathbf{c}, \mathbf{c}^*\big) + \frac{1}{2}\ell\big(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*\big) + 4\kappa_0\lambda^2 \|w_{S(\mathbf{c})}\|^2,\end{aligned}$$

according to Lemma 4.1, with $\delta = 2$. Noting that $\lambda_1 \leq (1-\kappa_1)\lambda/2$ yields

$$\ell\big(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*\big) + (1-\kappa_1)\lambda I_w(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) \leq 3\ell\big(\mathbf{c}, \mathbf{c}^*\big) + 8\kappa_0\lambda^2 \|w_{S(\mathbf{c})}\|^2.$$

If $I_w(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) \leq \lambda_0$, then combining Assumption 3 with (14) entails

$$\begin{aligned}\ell\big(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*\big) + (1-\kappa_1)\lambda I_w(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) &\leq \ell\big(\mathbf{c}, \mathbf{c}^*\big) + 2\lambda I_{w,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) + \lambda_1\lambda_0 \\ &\leq \ell\big(\mathbf{c}, \mathbf{c}^*\big) + 3\lambda\lambda_0.\end{aligned}$$

Since $\mathbf{c}^*_\lambda(\hat{\mathbf{c}}_{n,\lambda})$ satisfies $S(\mathbf{c}^*_\lambda(\hat{\mathbf{c}}_{n,\lambda})) \subsetneq S(\mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda}))$ or $\mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda}) = \mathbf{c}^*(\mathbf{c}^*_\lambda(\hat{\mathbf{c}}_{n,\lambda}))$, choosing $\mathbf{c} = \mathbf{c}^*_\lambda$ gives the result.

4.6. Proofs of Corollaries 3.1 and 3.2

For the plain Lasso, $\log(d)n^{-1} \to 0$ and $\sup_{p=1,\dots,d} M_p = O(1)$ ensures that $\lambda_0 \to 0$. $n^{-1}\lambda^{-2}\log(d\sqrt{n})$ yields $\lambda_1(\log(d\sqrt{n})) = O(\lambda)$. Applying Theorems 3.1 or 3.2 gives the results for $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*)$ and $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + \lambda(1-\kappa_1)I_w(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda(\hat{\mathbf{c}}_{n,\lambda}))$, since $\kappa_0 = O(1)$. The result on $\|\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda})\|$ follows from Proposition 2.2.

Similarly, for the normalized Lasso, the additional requirement $1 = O(\min_{p=1,\dots,d}\sigma_p)$ entails $\lambda_0 \to 0$, and Assumption 3 is satisfied, according to Proposition 2.3. In turn, $\max_{p=1,\dots,d}\sigma_p = O(1)$ leads to $\lambda_1(\log(d\sqrt{n})) = O(\lambda)$, hence the results.

At last, for the threshold Lasso, $\sup_{p=1,\dots,d} M_p = O(1)$, $\delta \to 0$ and $dn^{-1}\delta^{-2}$ imply that Assumption 3 is satisfied, $\lambda_0 \to 0$, and $d = O(n)$. Since $T(\delta) = O(1)$, combining $d = o(n)$ with $\lambda^{-1}\log(n)^{1/2}n^{-1/2} = o(1)$ leads to $\lambda_1(\log(n)) = o(\lambda)$. Noting that $\kappa_0 = O(1)$ and applying Theorem 3.2 gives the result.

4.7. Proof of Theorem 3.3

Let j be in $S(\mathbf{c}^*)^c$, and denote by $f(n)$ the quantity $n\delta^2 \sim n\lambda^2$. Since $n^{-1}\delta^{-2}\log(n) \to 0$, then $\log(n) = o(f(n))$. Hence Assumption 3 holds, for some $0 < \kappa_1 < 1$, with $r_1(n) = O(e^{-n\delta^2})$. Theorem 3.3 follows from the Karush–Kuhn–Tucker condition, as in [22], combined with some standard deviation bounds, which are listed below. Throughout this derivation, K will denote a generic positive constant not depending on n.

Proposition 4.2. *For every x in $\mathbb{R}^d$, let $G^{(j)}(x, \mathbf{c}^*)$ denote the k dimensional vector*

$$\big(x^{(j)}\mathbb{1}_{W_1(\mathbf{c}^*)}(x), \dots, x^{(j)}\mathbb{1}_{W_k(\mathbf{c}^*)}(x)\big).$$

Then, *we have*

$$\mathbb{P}\big[\big\|(P-P_n)G^{(j)}(\cdot,\mathbf{c}^*)\big\| \geq Kn^{-1/2}f(n)^{1/2}\big] = O\big(e^{-f(n)}\big),$$
$$\mathbb{P}\Bigg[\sup_{\mathbf{c}\in C^k}\Bigg|(P_n-P)\sum_{i\neq p}\mathbb{1}_{W_i(\mathbf{c})\cap W_p(\mathbf{c}^*)}\Bigg| \geq Kn^{-1/2}f(n)^{1/2}\Bigg] = O\big(e^{-f(n)}\big).$$

The proof of Proposition 4.2 is deferred to the supplementary material [13]. Assume that $\hat{\mathbf{c}}_{n,\lambda}^{(j)} \neq 0$, then the KKT condition yields

$$\frac{2}{\sqrt{n}}\big\|\hat{L}\hat{X}^{(j)}\big\| = \Bigg\|\sqrt{n}\lambda\hat{w}_j\frac{\hat{\mathbf{c}}_{n,\lambda}^{(j)}}{\|\hat{\mathbf{c}}_{n,\lambda}^{(j)}\|} + \frac{2}{\sqrt{n}}\hat{L}^t\hat{L}\hat{\mathbf{c}}_{n,\lambda}^{(j)}\Bigg\| \geq \sqrt{n}\lambda\hat{w}_j,$$

since $\hat{L}^t\hat{L}$ is positive. According to Proposition 2.3, it follows that $\sqrt{n}\lambda\hat{w}_j \geq (1-\kappa_1)n^{1/2}$, with probability larger than $1-O(e^{-f(n)})$, when n is large enough. On the other hand, we have

$$\begin{aligned}\big\|L(\hat{\mathbf{c}}_{n,\lambda})X^{(j)}\big\| &\leq \big\|L\big(\mathbf{c}^*\big)X^{(j)}\big\| + \big\|\big(L(\hat{\mathbf{c}}_{n,\lambda}) - L\big(\mathbf{c}^*\big)\big)X^{(j)}\big\| \\ &\leq n\big\|P_nG^{(j)}\big(\cdot,\mathbf{c}^*\big)\big\| + M^{(j)}nP_n\sum_{i\neq p}\mathbb{1}_{W_i(\hat{\mathbf{c}}_{n,\lambda})\cap W_p(\mathbf{c}^*)}.\end{aligned}$$

According to the centroid condition, $PG^{(j)}(\cdot,\mathbf{c}^*)=0$. Thus, it follows from Proposition 4.2 that $\|P_nG^{(j)}(\cdot,\mathbf{c}^*)\| \leq Kn^{-1/2}f(n)^{1/2}$ with probability $1-O(e^{-f(n)})$. Lemma 4.2 in [12] ensures that

$$P\sum_{i\neq p}\mathbb{1}_{W_i(\hat{\mathbf{c}}_{n,\lambda})\cap W_p(\mathbf{c}^*)} \leq p\big(K\big\|\hat{\mathbf{c}}_{n,\lambda}-\mathbf{c}^*\big\|\big) \leq K\big\|\hat{\mathbf{c}}_{n,\lambda}-\mathbf{c}^*\big\|,$$

according to Assumption 2. Besides, since $\log(n)=o(f(n))$, we may write $\lambda \sim \lambda_1(f(n))$. Taking into account that $\|w_{S(\mathbf{c}^*)}\|$ tends to $\sum_{i\in S(\mathbf{c}^*)}1/\|\mathbf{c}^{*,(i)}\|$, Theorem 3.2 yields $\ell(\hat{\mathbf{c}}_{n,\lambda},\mathbf{c}^*) \leq K\lambda^2$, with probability larger than $1-O(e^{-f(n)})$, for n large enough. On the same event, Proposition 2.2 gives $\|\hat{\mathbf{c}}_{n,\lambda}-\mathbf{c}^*\| \leq K\lambda$, which leads to

$$P_n\sum_{i\neq p}\mathbb{1}_{W_i(\hat{\mathbf{c}}_{n,\lambda})\cap W_p(\mathbf{c}^*)} \leq K\lambda,$$

with probability larger than $1-O(e^{-f(n)})$, according to Proposition 4.2. Then the KKT condition entails

$$(1-\kappa_1)n^{1/2} \leq Kn^{1/2}\lambda,$$

with probability larger than $1-O(e^{-f(n)})$, which is impossible when n is large enough.

Conversely, if j is in $S(\mathbf{c}^*)$ and $\hat{\mathbf{c}}_{n,\lambda}^{(j)}=0$, then the KKT condition yields

$$\frac{2}{n}\big\|L(\hat{\mathbf{c}}_{n,\lambda})\hat{X}^{(j)}\big\| \leq \lambda\hat{w}_j.$$

Since w_j tends to $1/\|\mathbf{c}^{*,(j)}\|$ and $\|PG^{(j)}(\cdot,\mathbf{c}^*)\| > 0$, according to the centroid condition, using the same concentration bounds as above leads to a contradiction.

4.8. Proof of Proposition 3.2

The first part of Proposition 3.2 is derived from the following lemma, which extends Proposition 3.2 of [12]. We denote by $\tilde{B}$ the quantity $\inf_{i\neq j}\|m_i - m_j\|$.

Lemma 4.2. *Denote by $\eta = \sup_{j=1,\dots,k} 1 - N_i$. Then the risk $R(\mathbf{m})$ may be bounded as follows.*

$$R(\mathbf{m}) \leq \frac{\sigma^2 k \theta_{\max} d}{(1-\eta)}, \tag{15}$$

where $\theta_{\max} = \max_{j=1,\dots,k}\theta_j$. For any $0 < \tau < 1/2$, let $\mathbf{c}$ be a codebook with a code point c_i such that $\|c_i - m_j\| > \tau\tilde{B}$, for every j in $\{1,\dots,k\}$. Then we have

$$R(\mathbf{c}) > \frac{\tau^2\tilde{B}^2\theta_{\min}}{4}\left(1 - \frac{2\sigma\sqrt{d}}{\sqrt{2\pi}\tau\tilde{B}}e^{-\frac{\tau^2\tilde{B}^2}{4d\sigma^2}}\right)^d, \tag{16}$$

where $\theta_{\min} = \min_{j=1,\dots,k}\theta_j$. At last, if $\sigma^- \geq c_-\sigma$, for any τ' such that $2\tau + \tau' < 1/2$, we have

$$\forall t \leq \tau' \qquad \tilde{B}p(t) \leq t\frac{2k^2\theta_{\max}M^{d-1}S_{d-1}}{(2\pi)^{d/2}(1-\eta)c_-^d\sigma^d}e^{-\frac{[\frac{1}{2}-(2\tau+\tau')]^2\tilde{B}^2}{2\sigma^2}}, \tag{17}$$

where S_{d-1} denotes the Lebesgue measure of the unit ball in $\mathbb{R}^{d-1}$.

The proof of Lemma 4.2 follows from direct calculation, as in the proof of Proposition 3.2 of [12]. For the sake of completeness, it is given in the supplemental article [13].

Let τ and τ' be positive quantities satisfying $2\tau + \tau' < 1/2$, and $\tau' > 8\sqrt{2}M\tau/(1-2\tau)\tilde{B}$. According to (15) and (16), if σ is small enough, then every optimal codebook $\mathbf{c}^*$ satisfies $\sup_{j=1,\dots,k}\|m_j - c_j^*\| \leq \tau\tilde{B}$, up to relabeling code points.

On the other hand, (17) ensures that, for σ small enough, P satisfies Assumption 2 with radius $r_0 \geq \tau'\tilde{B}$. Let $\mathbf{c}^*$ be an optimal codebook. According to (i) of Proposition 2.2 in [12], no other optimal codebook can be found in a ball of radius $(1-2\tau)\tilde{B}\tau'/4\sqrt{2}M$ centered at $\mathbf{c}^*$. Since $(1-2\tau)\tilde{B}\tau'/4\sqrt{2}M > 2\tau$, this proves that $\mathbf{c}^*$ must be unique.

To apply Theorem 3.3, we need to show that $S(\mathbf{c}^*) \subset \tilde{S}$. Suppose that there exists $j \geq d'+1$ such that $\mathbf{c}^{*,(j)} \neq 0$. Let s denote the orthogonal transformation defined by $s(x_1,\dots,x_{d'},x_{d'+1},\dots,x_d) = (x_1,\dots,x_{d'},-x_{d'+1},\dots,-x_d)$. Since $(\Sigma_i)_{p,q} = 0$, for every (i,p,k) in $\{1,\dots,k\}\times\{1,\dots,d'\}\times\{d'+1,\dots,d\}$, P is invariant through composition by s. Hence $s(\mathbf{c}^*)$ is an optimal codebook, and $\mathbf{c}^* \neq s(\mathbf{c}^*)$, which contradicts the fact that $\mathbf{c}^*$ is unique.

Acknowledgement

The author is supported by the ANR project TopData ANR-13-BS01-0008.

Supplementary Material

Appendix: Remaining proofs (DOI: 10.3150/16-BEJ876SUPP; .pdf). Due to space constraints, we relegate technical details of the remaining proofs to the supplement [13].

References

[1] Antoniadis, A., Brossat, X., Cugliari, J. and Poggi, J.-M. (2013). Clustering functional data using wavelets. *Int. J. Wavelets Multiresolut. Inf. Process.* **11** 1350003, 30. MR3038615

[2] Bach, F.R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9** 1179–1225. MR2417268

[3] Biau, G., Devroye, L. and Lugosi, G. (2008). On the performance of clustering in Hilbert spaces. *IEEE Trans. Inform. Theory* **54** 781–790. MR2444554

[4] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data. Springer Series in Statistics*. Heidelberg: Springer. MR2807761

[5] Chang, X., Wang, Y., Li, R. and Xu, Z. (2014). Sparse k-means with ℓ_∞/ℓ_0 penalty for high-dimensional data clustering. Available at arXiv:1403.7890.

[6] De Soete, G. and Carroll, J.D. (1994). k-means clustering in a low-dimensional Euclidean space. In *New Approaches in Classification and Data Analysis* (E. Diday, Y. Lechevallier, M. Schader, P. Bertrand and B. Burtschy, eds.). *Studies in Classification, Data Analysis, and Knowledge Organization*. 212–219. Heidelberg: Springer.

[7] Fischer, A. (2010). Quantization and clustering with Bregman divergences. *J. Multivariate Anal.* **101** 2207–2221. MR2671211

[8] Gersho, A. and Gray, R.M. (1991). *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer Academic.

[9] Graf, S. and Luschgy, H. (2000). *Foundations of Quantization for Probability Distributions. Lecture Notes in Math.* **1730**. Berlin: Springer. MR1764176

[10] Graf, S., Luschgy, H. and Pagès, G. (2007). Optimal quantizers for Radon random vectors in a Banach space. *J. Approx. Theory* **144** 27–53. MR2287375

[11] Jin, J. and Wang, W. (2014). Important feature PCA for high dimensional clustering. Available at arXiv:1407.5241.

[12] Levrard, C. (2015). Nonasymptotic bounds for vector quantization in Hilbert spaces. *Ann. Statist.* **43** 592–619. MR3316191

[13] Levrard, C. (2016). Supplement to "Sparse oracle inequalities for variable selection via regularized quantization." DOI:10.3150/16-BEJ876SUPP.

[14] Lloyd, S.P. (1982). Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28** 129–137. MR0651807

[15] Massart, P. and Meynet, C. (2011). The Lasso as an ℓ_1-ball model selection procedure. *Electron. J. Stat.* **5** 669–687. MR2820635

[16] Maugis-Rabusseau, C. and Michel, B. (2013). Adaptive density estimation for clustering with Gaussian mixtures. *ESAIM Probab. Stat.* **17** 698–724. MR3126158

[17] Meynet, C. (2013). An ℓ_1-oracle inequality for the Lasso in finite mixture Gaussian regression models. *ESAIM Probab. Stat.* **17** 650–671. MR3126156

[18] Pollard, D. (1981). Strong consistency of k-means clustering. *Ann. Statist.* **9** 135–140. MR0600539

[19] Pollard, D. (1982). A central limit theorem for k-means clustering. *Ann. Probab.* **10** 919–926. MR0672292

[20] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771. MR2816337

[21] Steinley, D. and Brusco, M.J. (2008). Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika* **73** 125–144. MR2395296

[22] Sun, W., Wang, J. and Fang, Y. (2012). Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electron. J. Stat.* **6** 148–167. MR2879675

[23] Terada, Y. (2014). Strong consistency of reduced k-means clustering. *Scand. J. Stat.* **41** 913–931. MR3277030

[24] Terada, Y. (2015). Strong consistency of factorial k-means clustering. *Ann. Inst. Statist. Math.* **67** 335–357. MR3315263

[25] Timmerman, M.E., Ceulemans, E., Kiers, H.A.L. and Vichi, M. (2010). Factorial and reduced k-means reconsidered. *Comput. Statist. Data Anal.* **54** 1858–1871. MR2608979

[26] van de Geer, S. (2013). Generic chaining and the ℓ_1-penalty. *J. Statist. Plann. Inference* **143** 1001–1012. MR3029225

[27] van de Geer, S.A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. MR2396809

[28] Vichi, M. and Kiers, H.A.L. (2001). Factorial k-means analysis for two-way data. *Comput. Statist. Data Anal.* **37** 49–64. MR1862479

[29] Witten, D.M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *J. Amer. Statist. Assoc.* **105** 713–726. MR2724855

Received April 2015 and revised May 2016