ARTICLE IN PRESS

Information Sciences xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A modification of the Lloyd algorithm for *k*-anonymous quantization

David Rebollo-Monedero *, Jordi Forné, Esteve Pallarès, Javier Parra-Arnau

Dept. of Telematics Engineering, Universitat Politécnica de Catalunya, C. Jordi Girona 1-3, E-08034 Barcelona, Spain

ARTICLE INFO

Article history:
Received 9 September 2009
Received in revised form 15 August 2012
Accepted 19 August 2012
Available online xxxx

Keywords:
Microdata anonymization
k-Anonymity
Lloyd algorithm
k-Means method
k-Anonymous quantization

ABSTRACT

We address the problem of designing quantizers that cluster data while satisfying a *k*-anonymity requirement. A general data compression perspective is adopted, which considers both discrete and continuous probability distributions, and corresponding constraints on both cell sizes and quantizer index probabilities. Potential applications of this problem extend well beyond the important case of microdata anonymization, to include also optimized task allocation under workload constraints. Our contribution is twofold. First and most importantly, we present a theoretical analysis showing the optimality conditions which probability-constrained quantizers must satisfy, thereby theoretically characterizing optimal *k*-anonymous aggregation as a special case. As a second contribution, inspired by our theoretical analysis, we propose an alternating optimization algorithm for the design of this type of quantizers. Our algorithm is conceptually motivated by the popular Lloyd–Max algorithm for quantization design, originally intended for data compression, also known as the *k*-means method. Experimental results for synthetic and real data, with mean squared error as a distortion measure, confirm that our method outperforms MDAV, a popular fixed-size microaggregation algorithm for statistical disclosure control. This performance improvement is in terms of data utility, for the exact same *k*-anonymity constraint, but does come at the expense of higher computational sophistication.

1. Introduction

A microdata set is a database table whose records carry information concerning individual respondents, either people or companies. This database commonly contains a set of attributes that may be classified into *identifiers*, *key attributes* and *confidential attributes*. Firstly, identifiers allow to unequivocally identify individuals. This is the case of social security numbers or full names, which would be removed before the publication of the microdata set. Secondly, key attributes, also called *quasi-identifiers*, are those attributes that, in combination, may be linked with external information to reidentify the respondents to whom the records in the microdata set refer. Examples include job, address, age, gender, height and weight. A notorious fact is that 87% of the population in the United States may be reidentified solely on the basis of their ZIP code, gender, and date of birth, according to 1990 census data [43]. Finally, the dataset contains *confidential attributes* with sensitive information on the respondent, such as salary, religion, political affiliation or health condition. The classification of attributes as key or confidential may ultimately rely on the specific application and the privacy requirements the microdata set is intended for.

Intuitively, perturbation of the key attributes enables us to preserve privacy to a certain extent, at the cost of losing some of the data utility with respect to the unperturbed version. *k*-Anonymity is the requirement that each tuple of key-attribute

* Corresponding author. Tel.: +34 93 401 7027.
E-mail addresses: david.rebollo@entel.upc.edu (D. Rebollo-Monedero), jforne@entel.upc.edu (J. Forné), esteve@entel.upc.edu (E. Pallarès), javier.parra@entel.upc.edu (J. Parra-Arnau).

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.08.022

values be shared by at least k records in the dataset. This may be achieved through the microaggregation approach illustrated by the example depicted in Fig. 1, where age and nationality are regarded as key attributes, and health condition as a confidential attribute. Rather than making the original table available, we publish a k-anonymous *quantized* version containing aggregated records, in the sense that all key-attribute values within each group are replaced by a common representative tuple. As a result, a record cannot be unambiguously linked to the corresponding record in the original table or, more generally, to any public database containing identifier attributes. Despite the fact that k-anonymity as a measure of privacy is not without shortcomings, its simplicity makes it a widely popular criterion in the *statistical disclosure control* (SDC) literature.

1.1. Contribution

The object of this paper is to tackle the problem of designing k-anonymous clusters from a data compression perspective. More specifically, we present a general formulation that considers, on the one hand, both discrete and continuous probability distributions rather than simply tables. On the other hand, this formulation contemplates quantization with cell-probability constraints as a generalization of the concept of k-anonymity. Our contribution is twofold:

- First and most importantly, we present a theoretical analysis that proves the optimality conditions probability-constrained quantizers must satisfy. Part of the importance of this analysis lies in the fact that it provides a novel, theoretical characterization of optimal k-anonymous aggregation, although the theory is applicable to a wider range of problems, including resource allocation.
- As a second contribution, inspired by our theoretical analysis, we propose an alternating optimization algorithm for the design of this type of quantizers. Our algorithm is conceptually motivated by the popular Lloyd–Max algorithm for quantization design, originally intended for data compression, also known as the k-means method.

Because the scope of this work is necessarily limited, and specifically focused on the above two contributions of theoretical nature, the supporting experimentation provided here cannot be extremely exhaustive. Nevertheless, our experimental results do contemplate various statistical distributions, dataset lengths and anonymity constraints, in order to compare distortion, anonymity and running time. These results show consistent improvement over the state of the art in terms of the anonymity-utility trade-off, at the expense of higher computational complexity. Precisely, experimental results for Gaussian statistics and real data, using mean squared error as a distortion measure, confirm that our method is capable of significantly outperforming MDAV, a widely popular, fixed-size microaggregation algorithm for data anonymization. This improvement is in terms of data utility, for the same exact anonymity constraints. We shall also see that although higher running times are required, they scale with the dataset size roughly quadratically for a fixed cell-size constraint, just as for MDAV. In spite of the fact that our proposal is a fixed-size algorithm and any variable-size improvements are left for future research, we verify that it also outperforms VMDAV for the standardized dataset used.

1.2. Contents

This paper is organized as follows: Section 2 reviews the state of the art in k-anonymity-based privacy models. Mathematical conventions and a brief review of quantization are provided in Section 3. Section 4 gives a formulation of the problem of probability-constrained quantization, a data compression generalization of the problem of k-anonymous microaggregation. Section 5 contains a theoretical analysis of the solution to this problem, which inspires a practical algorithm, presented in Section 6. Empirical results are reported in Section 7. Conclusions are drawn in Section 8.

Identifier | Key Attributes | Confidential Attribute

Name	Age	Nationality	Health Condition
William	45	US	AIDS
Emmanuel	42	French	AIDS
Syme	47	Indian	AIDS
Naoto	31	Japanese	Diabetes
Katharine	30	US	Heart Disease
Julia	36	British	Heart Disease

(a) Original table

Aggregated Key Attributes | Confidential Attribute

	Age	Nationality	Health Condition
k-Aggregated Records	40 – 50	*	AIDS
	40 – 50	*	AIDS
	40 – 50	*	AIDS
	< 40	*	Diabetes
	< 40	*	Heart Disease
	< 40	*	Heart Disease

(b) Perturbed table

Fig. 1. k-Anonymous quantization, that is, aggregation of key-attribute values to attain k-anonymity.

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022

2. State of the art on microdata anonymization

The concept of k-anonymity, proposed by the SDC community [39,38], is a widely popular privacy criterion, partly due to its mathematical tractability. In Refs. [9,12,15,13], the original formulation of this privacy criterion, based on generalization and recording of key attributes, was modified into the microaggregation-based approach already commented on, and illustrated in Fig. 1.

Multivariate microaggregation has been shown to be NP-hard [35]. A number of heuristic methods have been proposed, which can be categorized into fixed-size and variable-size methods, according to whether all aggregated groups but one have exactly k elements. The maximum distance (MD) algorithm [12] and its less computationally demanding variation, the maximum distance to average vector (MDAV) algorithm [15,11,20,44], are fixed-size algorithms that perform particularly well in terms of the distortion they introduce, for many data distributions. Popular variable-size algorithms include the μ-approx [13], the minimum spanning tree (MST) [23], the variable MDAV (VMDAV) [41] and the two fixed reference points (TFRPs) [7] algorithms. Efforts to circumvent the complexity of multivariate microaggregation exploit projections onto one dimension but are reported to yield a much higher disclosure risk [34].

Research on microaggregation algorithms has continued recently. In particular, an approach recommends creating clusters of k records according to their densities [26]. Still in the case of perturbative algorithms, Matatov et al. [31] contemplate the partition of the original dataset into several projections such that each projection satisfies the k-anonymity requirement, with the help of genetic algorithms. A well-known alternative to perturbative algorithms is the generation of synthetic data that preserves some pre-established statistics of the original dataset. The combination of perturbed and synthetic data is exactly the approach followed by Domingo-Ferrer and González-Nicolás [10], who propose a method for the generation of hybrid data through microaggregation.

Further recent investigation has also been conducted on the scenario of online data collection, where a data miner queries a set of users, each of whom responds with a piece of data. For example, Zhong et al. [47] propose a cryptographic method that allows users to submit their data anonymously. Namely, the authors present a protocol that eliminates the restriction of using unidentified communication channels and allows users to include identifying information in their responses. Still in this context, Domingo-Ferrer et al. [14] present a set of protocols and methods aimed to protect the privacy of users that query Web search engines. Lastly, in the scenario of streaming data, Cao et al. [6] propose a cluster-based approach that k-anonymizes data streams and, in addition, guarantees the freshness of the anonymized data by imposing a restriction on the delay.

Unfortunately, while k-anonymity prevents identity disclosure, it may fail to protect against attribute disclosure. Precisely, the definition of this privacy criterion establishes that complete reidentification is unfeasible within a group of records sharing the same tuple of perturbed key attributes. However, if the records in the group also share a common value of a confidential attribute, the association between an individual linkable to the group of perturbed key attributes and the corresponding confidential attribute remains disclosed. Concretely, consider the example depicted in Fig. 1 and suppose that a privacy attacker knows Emmanuel's key attributes. If the attacker learns that Emmanuel is included in the released table, then the attacker may conclude that this patient suffers from AIDS even though the attacker is unable to ascertain which record belongs to this individual. This is known as *homogeneity attack*. Now suppose that the adversary strives to infer the confidential attribute value of Naoto, who belongs to a group in which the distribution of this confidential attribute value is not completely homogeneous. Even in this case, the adversary could exploit the fact that the Japanese have a low incidence of heart disease and hence they could deduce that this individual is more likely to have diabetes. Such attack is known as *background knowledge attack*. From a statistical perspective, the main issue with k-anonymity as a privacy criterion is its vulnerability against the exploitation of the difference between the prior distribution of confidential data in the entire population, and the posterior conditional distribution of a group given the observed, perturbed key attributes. For example, imagine that the proportion of respondents with heart disease is much higher than that in the overall dataset. This is normally referred to as a *skewness attack*.

These vulnerabilities motivated the proposal of enhanced privacy criteria, some of which we proceed to sketch briefly, along with algorithm modifications. A restriction of k-anonymity called *p-sensitive* k-anonymity was presented in [45,42]. In addition to the k-anonymity requirement, it is required that there be at least p different values for each confidential attribute within the group of records sharing the same tuple of perturbed key attributes. Clearly, large values of p may lead to huge data utility loss. A slight generalization called *l-diversity* [29,21] was defined with the same purpose of enhancing k-anonymity. The difference with respect to p-sensitivity is that group of records must contain at least l "well-represented" values for each confidential attribute. Depending on the definition of well-represented, l-diversity can reduce to p-sensitive k-anonymity or be more restrictive. We would like to stress that neither of these enhancements succeeds in completely removing the vulnerability of k-anonymity against skewness attacks. Furthermore, both are still susceptible to *similarity attacks*, in the sense that while confidential attribute values within a cluster of aggregated records might be p-sensitive or l-diverse, they might also very well be semantically similar for the practical purposes of the attacker. Consider for example confidential attributes indicating prostate cancer or bladder cancer.

A privacy criterion aimed at overcoming similarity and skewness attacks is *t-closeness* [25]. An aggregated microdata set satisfies t-closeness if for each group, a predefined measure of discrepancy between the posterior distribution of the confidential attributes within the group and the prior distribution of the overall population does not exceed a threshold t. As ar-

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022

gued in [16], to the extent to which the within-group distribution of confidential attributes resembles the distribution of those attributes for the entire dataset, skewness attacks will be thwarted. In addition, since the within-group distribution of confidential attributes mimics the distribution of those attributes over the entire dataset, no semantic similarity can occur within a group that does not occur in the entire dataset.

The main limitation of the original t-closeness work [25] is that no computational procedure to reach t-closeness was specified. An information-theoretic privacy criterion, inspired by t-closeness, was proposed in [36,37]. In the latter work, privacy risk is defined as the conditional Kullback–Leibler divergence between the posterior and the prior distributions. This criterion is also tightly related to the concept of *equivocation* introduced by Shannon in 1949 [40], namely the conditional entropy of a private message given an observed cryptogram.

In conclusion, we would like to emphasize that despite the shortcomings of k-anonymity and its enhancements as a measure of privacy, it is still a widely popular criterion for SDC, mainly because of its simplicity. More generally, we acknowledge that the formulation of any privacy-utility problem relies on the appropriateness of the criteria optimized. These criteria depend, in turn, on the specific application, on the statistics of the data, on the degree of data utility we are willing to compromise, and last but not least, on the adversarial model and the mechanisms against privacy contemplated. No privacy criterion, including k-anonymity in its numerous varieties, is the be-all and end-all of database anonymization [4].

3. Mathematical preliminaries and background on conventional quantization

3.1. Notation and mathematical preliminaries

Throughout the paper, the measurable space in which a random variable (r.v.) takes on values will be called an *alphabet*. The cardinality of a set $\mathscr{X}$ is denoted by $|\mathscr{X}|$. We shall follow the convention of using uppercase letters for r.v.'s, and lowercase letters for particular values they take on. Probability density functions (PDFs) and probability mass functions (PMFs) are denoted by p and subindexed by the corresponding r.v. The expectation operator is denoted by E. Expectation can model the special case of averages over a finite set of data points $\{x_1, \ldots, x_n\}$, simply by defining an r.v. X uniformly distributed over this set, so that, for instance, $EX = \frac{1}{n}\sum_{i=1}^{n} x_i$.

A *halfspace* is either of the two parts into which a hyperplane divides the Euclidean space. A (possibly unbounded) *convex polytope* may be defined as the finite intersection of halfspaces. It is a convex set in the sense that the segment connecting any two points in the set is entirely contained in it, and may be regarded as a generalization of the two-dimensional concept of (convex) polygon.

3.2. Conventional quantization

A *quantizer* is a function that partitions a range of values x of an r.v. X, commonly continuous, approximating each resulting cell by a value $\hat{x}$ of a discrete r.v. $\widehat{X}$. The quantizer map $\hat{x}(x)$ may be broken down into two steps. Firstly, an assignment of *source data* X to a *quantization index* Q in a finite alphabet $\mathscr{Q} = \{1, 2, \ldots, |\mathscr{Q}|\}$, by means of a clustering function $q(x)$, and secondly, a *reconstruction function* $\hat{x}(q)$ mapping the index Q into a value $\widehat{X}$ that approximates the original data, so that $\hat{x}(x) = \hat{x}(q(x))$. Both $\hat{x}(x)$ and $q(x)$ are often referred to as quantizer. This is represented in Fig. 2, along with an example where the r.v. X takes on values in $\mathbb{R}^2$.

In the context of source coding, quantizers are required due to the need to represent the data in a countable alphabet, such as the set of finite bit strings, suitable for storage and transmission in computer systems. We reflect this requirement mathematically by assuming that Q is a finite-alphabet r.v. The size of this alphabet, that is, the number of quantization cells, is usually given as an application requirement. Clearly, quantization comes at the price of introducing a certain amount of distortion between the original data X and its reconstructed version $\widehat{X}$. In mathematical terms, we define a nonnegative function $d(x, \hat{x})$ called *distortion measure*, and consider the expected *distortion* $\mathcal{D} = Ed(X, \widehat{X})$. A common measure of distortion is the *mean squared error* (MSE), that is, $\mathcal{D} = E\|X - \widehat{X}\|^2$, popular due to its mathematical tractability. When the r.v. X models a

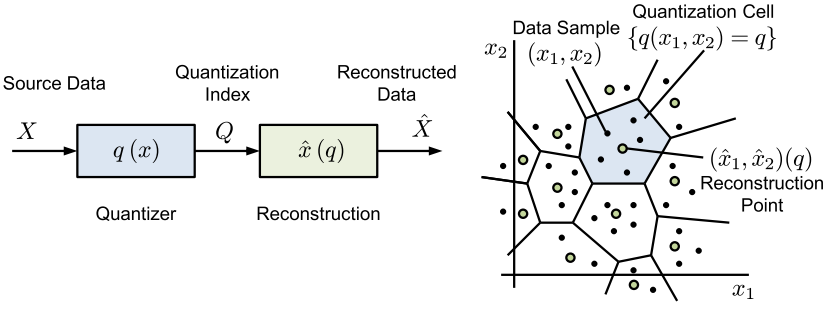


Fig. 2. Example of a two-dimensional quantizer.

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022

Table 1
Some key terms in data compression have their equivalence in the fields of statistics and SDC.

Data compression	Statistics	Statistical disclosure control (SDC)
Lloyd/Lloyd–Max algorithm	k-Means method	–
Source data	Sample/dataset	Microdata
Sample/point	Point/value	Record
Data component	Random variable (r.v.)	Attribute
Cell	Cluster	k-Anonymous group
Quantization	Clustering	Microaggregation/k-anonymization
Reconstruction point/centroid	Mean	Perturbed (value of the) key attribute
Quadratic distortion (per sample)	Mean squared error (MSE)	Sum of squared errors (SSEs)

set of points $\{x_1, \ldots, x_n\}$, the expected distortion or MSE is just $\mathcal{D} = \frac{1}{n}\sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2$. In the microaggregation literature, often an unnormalized measure is taken, namely the sum of squared errors (SSEs), i.e., $n\mathcal{D}$, occasionally divided by the total variance of the sample set before microaggregation, called sum of squares total (SST).

Optimal quantizers are those of minimum distortion for a given number of possible indices. It is well known [18,19] that optimal quantizers must satisfy the following conditions:

- *Nearest-neighbor condition.* Given a reconstruction function $\hat{x}(q)$, the optimal quantizer $q^*(x)$ is given by

$$q^*(x) = \arg\min_{q \in \{1,2,\ldots,|\mathcal{Q}|\}} d(x, \hat{x}(q)), \tag{1}$$

that is, each value x of the data is assigned to the index corresponding to the nearest reconstruction value.
- *Centroid condition.* In the special case when MSE is used as a distortion measure, given a quantizer $q(x)$, the optimal reconstruction function $\hat{x}^*(q)$ is given by

$$\hat{x}^*(q) = E[X|q], \tag{2}$$

that is, each reconstruction value is the *centroid* of a quantization cell.

Each necessary condition may be interpreted as the solution to a Bayesian decision problem. These optimality conditions are exploited in the *Lloyd–Max algorithm* [27,32], a quantizer design algorithm based on the alternating optimization of $q(x)$ given $\hat{x}(q)$ and vice versa, according to (1) and (2). The Lloyd algorithm for nondistributed quantization has been rediscovered several times in the statistical clustering literature [19], under names such as the k-means method. While the algorithm does not guarantee convergence to an optimal solution in general, the sequence of distortions obtained at each step is nonincreasing, and for appropriate initializations it commonly provides a useful upper bound on the optimal distortion [19]. Lastly, Table 1 shows the equivalence relation between the key terms in data compression and the terminology used in the fields of statistics and SDC.

4. Formulation of the problem of probability-constrained quantization

4.1. Formal problem statement

We consider the design of minimum-distortion quantizers satisfying cell-probability constraints, with the same block structure depicted in Fig. 2. (Tuples of) key-attribute values are modeled by an r.v. X in an arbitrary alphabet $\mathcal{X}$, possibly discrete or continuous. The quantizer $q(x)$ assigns X to a quantization index Q in a finite alphabet $\mathcal{Q} = \{1, \ldots, |\mathcal{Q}|\}$ of a predetermined size. The reconstruction function $\hat{x}(q)$ maps Q into the aggregated key-attribute value $\widehat{X}$, which may be regarded as an approximation to the original data, defined in an arbitrary alphabet $\hat{\mathcal{X}}$, commonly but not necessarily equal to the original data alphabet $\mathcal{X}$.

For any nonnegative (measurable) function $d(x, \hat{x})$, called distortion measure, define the associated expected distortion $\mathcal{D} = Ed(X, \widehat{X})$, that is, a measure of the discrepancy between the key-attribute values and their aggregation values, which reflects the loss in data utility. An important example of distortion measure is $d(x, \hat{x}) = \|x - \hat{x}\|^2$, for which $\mathcal{D}$ becomes the MSE. $p_Q(q)$ denotes the PMF corresponding to the cell probabilities. The privacy requirement in the aggregation problem is established by means of the cell-probability constraints $p_Q(q) = p_0(q)$, for any predetermined PMF $p_0(q)$.

Given a distortion measure $d(x, \hat{x})$ and probability constraints expressed by means of $p_0(q)$ (along with the number of quantization cells $|\mathcal{Q}|$), we wish to design an optimal quantizer $q^*(x)$ and an optimal reconstruction function $\hat{x}^*(q)$, in the sense that they *jointly* minimize the distortion $\mathcal{D}$ while satisfying the probability constraints.

4.2. Cases and applications

The motivating application of this paper is the problem of microdata k-anonymization. In this important albeit special case, the r.v. X would be given by the empirical distribution of the key attributes in the original microdata set, that is,

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022

$p_X(x)$ would be the PMF corresponding to the relative frequency of occurrences of the different tuples of key attributes. Let n be the number of records in the microdata set. The k-anonymity constraint could be translated into probability constraints by setting $|\mathcal{Q}| = \lfloor n/k \rfloor$ and $p_0(q) = 1/|\mathcal{Q}|$, which ensures that $np_0(q) \geqslant k$.

In addition, probability-constrained quantization may find applications in a variety of resource allocation problems. We may of course be only interested in the *clustering* portion $q(x)$ of the quantization process, in the sense that the reconstruction values may be only used as a means to compute a measure of similarity, and not as a replacement for the data clustered. The following are applications of similarity-based allocation of resources or workload according to predetermined volume constraints:

- Assignment of clients or tasks to customer agents or centers based on demographic similarity. Agents can handle roughly similar volumes of work, and maintaining similarity would contribute to efficiency through specialization.
- Automatic (pre) assignment of documents to technical reviewers based on context similarity. Data here could be the number of occurrences of certain keywords.
- Determination of areas of coverage of hospitals for a given volume of patients, taking into account population density. In this application centroids are hospital locations, thereby given fixed values. Similarity is estimated traveling distance or merely geographic distance.

Finally, we would like to remark that our formulation uses a single PMF $p_X(x)$ of X in the computation of the distortion $\mathcal{D}$ and the cell probabilities $p_Q(q)$ involved in the probability constraints. However, it is mathematically immediate to generalize our theoretical study to a problem with two PMFs, one for the computation of the distortion, and another for the cell probabilities. The former would still represent a measure of frequency of key attributes. The latter could be regarded as a measure of diversity with respect to confidential attribute values.

5. Necessary optimality conditions and quantizer geometry

We now establish necessary conditions for optimal, probability-constrained quantizers and reconstruction functions, analogous to the nearest neighbor (1) and centroid condition (2) found in conventional quantization. In the next section, we shall modify the conventional Lloyd algorithm by applying the underlying alternating optimization principle to these optimality conditions.

5.1. Optimality of the centroid condition

Finding the optimal reconstruction function $\hat{x}^*(q)$ for a given quantizer $q(x)$ is a problem identical to that in conventional quantization, simply a centroid (2) in the MSE case. Accordingly, we refer the proof of the following proposition to the standard literature on quantization theory [18,19].

Proposition 1 (*Centroid condition*). *Given a quantizer $q(x)$, the optimal reconstruction function $x^*(q)$ is given by the generalized centroids of each cell, i.e.,*

$$\hat{x}^*(q) = \arg\min_{\hat{x}\in\widehat{\mathcal{X}}} E[d(X,\hat{x})|q]. \tag{3}$$

In the special case when MSE is used as distortion measure, $\hat{x}^*(q) = E[X|q]$.

On the other hand, we may not apply the nearest-neighbor condition (1) in conventional quantization directly, if we wish to guarantee the probability constraints $p_Q(q) = p_0(q)$. We introduce a cell *cost function* $c : \mathcal{Q} \to \mathbb{R}$, a real-valued function of the quantization indices, which assigns an additive cost $c(q)$ to a cell indexed by q. The intuitive purpose of this function is to shift the cell boundaries appropriately to satisfy the probability constraints. Specifically, given a reconstruction function $\hat{x}(q)$ and a cost function $c(q)$, we define the associated quantizer

$$q^*(x) = \arg\min_{q\in\mathcal{Q}} \{d(x,\hat{x}(q)) + c(q)\}. \tag{4}$$

Two important results are provided in the following theorems. The first theorem states that the cells of the quantizer $q^*(x)$ thus defined are convex polytopes, just as those of optimal conventional quantizers. The second theorem asserts that, if there exists a cost function $c(q)$ such that $q^*(x)$ satisfies the probability constraints, then $q^*(x)$ is optimal with respect to the given reconstruction function $\hat{x}(q)$.

5.2. Quantizer cells as convex polytopes

Before proving the result concerning the shape of quantization cells, we must provide a quick, preliminary lemma showing that the region of points closer to one of two centroids is a halfspace, by expressing this region in terms of a vector projection. Next, a quantization cell is shown to be the intersection of such halfspaces. The symbols $\langle\cdot,\cdot\rangle$ denote the usual inner product. Define

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022

$$\mathscr{H}_{qr} = \{x | \|x - \hat{x}(q)\|^2 + c(q) \leqslant \|x - \hat{x}(r)\|^2 + c(r)\}.$$

Lemma 2. *$\mathscr{H}_{qr}$ is a halfspace. More specifically,*

$$\mathscr{H}_{qr} = \left\{ x \middle| \left\langle x - \frac{\hat{x}(q) + \hat{x}(r)}{2}, \frac{\hat{x}(r) - \hat{x}(q)}{\|\hat{x}(r) - \hat{x}(q)\|} \right\rangle \leqslant \frac{c(r) - c(q)}{2\|\hat{x}(r) - \hat{x}(q)\|} \right\}.$$

Proof.

$$\|x\|^2 - 2\langle x, \hat{x}(q)\rangle + \|\hat{x}(q)\|^2 + c(q) \leqslant \|x\|^2 - 2\langle x, \hat{x}(r)\rangle + \|\hat{x}(r)\|^2 + c(r),$$
$$2\langle x, \hat{x}(r) - \hat{x}(q)\rangle \leqslant \|\hat{x}(r)\|^2 - \|\hat{x}(q)\|^2 + c(r) - c(q) = \langle \hat{x}(q) + \hat{x}(r), \hat{x}(r) - \hat{x}(q)\rangle + c(r) - c(q),$$
$$\left\langle x - \frac{\hat{x}(q) + \hat{x}(r)}{2}, \hat{x}(r) - \hat{x}(q) \right\rangle \leqslant \frac{c(r) - c(q)}{2}. \quad \square$$

Lemma 2 confirms the intuition that in the special case of a two-region quantizer, the boundary between the two quantization regions is a line perpendicular to the segment connecting the two centroids $\hat{x}(q)$ and $\hat{x}(r)$, shifted with respect to the midpoint $\frac{\hat{x}(q)+\hat{x}(r)}{2}$, according to the cost difference $c(r) - c(q)$. We now turn to Theorem 3, which considers the case of an arbitrary number of centroids.

Theorem 3 (*Quantization cell convexity*). *Provided that MSE is used as a distortion measure, for any reconstruction function $\hat{x}(q)$ and any cost function $c(q)$, the quantization cells of the associated quantizer $q(x)$ are convex polytopes.*

Proof. The statement of the theorem is an immediate consequence from the fact that the quantization cell $\mathscr{C}_q$ is an intersection of the halfspaces $\mathscr{H}_{qr}$ of Lemma 2:

$$\mathscr{C}_q = \{x | q(x) = q\} = \{x | \|x - \hat{x}(q)\|^2 + c(q) \leqslant \|x - \hat{x}(r)\|^2 + c(r) \text{ for all } r\} = \bigcap_r \mathscr{H}_{qr} \quad \square$$

5.3. Optimality of the modified nearest-neighbor condition

We prove the necessity of the modified nearest-neighbor condition, stated in the following theorem, Theorem 4. Recall that the theorem assumes the existence of a cost function such that the associated quantizer satisfies the probability constraint. The proof compares the distortion of this optimal quantizer with that introduced by any other quantizer. The intuition behind our proof, valid for general alphabets and any number of centroids, capitalizes on the following observation in the simple case of finite alphabets and two quantization indices. Consider a point inside the optimal quantization region 1 assigned to region 2 by the other quantizer. Because the number of points in the two regions involved must be preserved, there must be a point assigned to region 2 by the optimal quantizer but assigned to 1 by the other quantizer. It is not hard to see that exchanging the assignment of these points will in general reduce the distortion introduced by the other quantizer.

Theorem 4 (***Modified nearest-neighbor condition***). *Given a reconstruction function $\hat{x}(q)$, suppose there exists a cell cost function $c(q)$ such that the associated quantizer $q^*(x)$ defined by (4) satisfies the cell-probability constraints $p_{Q^*}(q) = p_0(q)$. Then, $q^*(x)$ minimizes the distortion among all quantizers from $\mathscr{X}$ to $\mathscr{Q}$ satisfying the same probability constraints.*

Proof. Let $\mathcal{D}^*$ be the distortion introduced by the optimal quantizer $q^*(x)$ in the theorem in question. Consider any other quantizer $q(x)$ from $\mathscr{X}$ to $\mathscr{Q}$ satisfying the same probability constraints, with associated distortion $\mathcal{D}$. We need to prove that $\mathcal{D} \geqslant \mathcal{D}^*$.

We denote by $\mathbf{1}_{\mathscr{S}}$ the indicator r.v. corresponding to the event $X \in \mathscr{S}$, that is, the binary r.v. taking the value 1 if X belongs to the (measurable) set $\mathscr{S}$, and 0 otherwise. Recall that if $\{\mathscr{S}_i\}_i$ is any (measurable) partition of $\mathscr{X}$ and $f(x)$ any (measurable) function of X, then $Ef(X) = \sum_i E\mathbf{1}_{\mathscr{S}_i} f(X)$. In addition, the expectation of an indicator of a set is simply the probability of that set, i.e., $E\mathbf{1}_{\mathscr{S}} = P(\mathscr{S})$.

For all pairs of quantization indices $q, r = 1, \ldots, |\mathscr{Q}|$, define $\mathscr{S}_{qr} = \{x | q^*(x) = q, q(x) = r\}$, i.e., the set of values of X mapped to the quantization index q by the optimal quantizer $q^*(x)$, but mapped to r by $q(x)$. Decompose the distortion introduced by $q^*(x)$ according to the partition $\{\mathscr{S}_{qr}\}_{q,r}$ of $\mathscr{X}$ as

$$\mathcal{D}^* = Ed(X, \hat{x}(q^*(X))) = \sum_{q,r} E\mathbf{1}_{\mathscr{S}_{qr}} d(X, \hat{x}(q)),$$

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022

and similarly for $q(x)$,

$$\mathcal{D} = Ed(X, \hat{x}(q(X))) = \sum_{q,r} E\mathbf{1}_{\mathscr{S}_{qr}}\ d(X, \hat{x}(r)).$$

By construction $q^*(x)$ satisfies

$$d(x, \hat{x}(q^*(x))) + c(q^*(x)) \leqslant d(x, \hat{x}(r)) + c(r)$$

for any value x and any quantization index r. Consequently,

$$\mathcal{D}^* \leqslant \sum_{q,r} E\mathbf{1}_{\mathscr{S}_{qr}}(d(X, \hat{x}(r)) + c(r) - c(q)) = \mathcal{D} + \sum_{q,r} P(\mathscr{S}_{qr})(c(r) - c(q)) = \mathcal{D} + \sum_s c(s)\left(\sum_t P(\mathscr{S}_{ts}) - \sum_t P(\mathscr{S}_{st})\right).$$

But both $q^*(x)$ and $q(x)$ satisfy the same cell-probability constraints. Hence, for all s,

$$\sum_t P(\mathscr{S}_{ts}) = P\{q(X) = s\} = p_Q(s) = p_{Q^*}(s) = P\{q^*(X) = s\} = \sum_t P(\mathscr{S}_{st}),$$

and, finally, $\mathcal{D}^* \leqslant \mathcal{D}$. □

Theorem 4 naturally leads to the question of how to find a cost function $c(q)$ such that the probability constraints $p_Q(q) = p_0(q)$ are satisfied, given a reconstruction function $\hat{x}(q)$. We remark that for discrete probability distributions of X it is easy to see that it may not be possible to find such $c(q)$. In the continuous case, we propose the following method, which proved to be very successful in all of our experiments, including those reported in Section 7. Specifically, we propose an application of the Levenberg–Marquardt algorithm [24,30,33], an algorithm to solve systems of nonlinear equations numerically, or similarly but slightly more simply, a Tychonov regularization of the Gauss–Newton algorithm [2]. Appendix A outlines the application of these numerical methods to the computation of $c(q)$ in our formulation. It is important to stress the differentiability assumptions inherent in the numerical methods for cost adjustment mentioned, which estimate the Jacobian of the function that models the dependence between costs and cell probabilities.

6. Modified Lloyd algorithm for probability-constrained quantization

In the previous section, we proved two optimality conditions that a probability-constrained quantizer must necessarily satisfy. We stress that, exactly as in conventional quantization, each of these optimality conditions obtained merely characterizes an optimal quantizer for a given reconstruction, and viceversa. Additionally, even if both conditions hold simultaneously, they are still necessary conditions, not sufficient. Ideally, we wish to find a pair of quantizers and reconstruction functions that *jointly* minimize the distortion. We mentioned in Section 3.2 that the Lloyd algorithm is essentially an alternating optimization algorithm that iterates between the nearest-neighbor and the centroid optimality conditions, hoping to approximate a jointly optimal pair $q^*(x)$, $\hat{x}^*(q)$, but only guaranteeing that the sequence of distortions is nonincreasing. Experimentally, the Lloyd algorithm very often shows excellent performance [19, Sections II.E and III].

6.1. Our algorithm as the alternation between optimality conditions

The very same alternating optimization principle serving as the basis of the Lloyd algorithm is applied in this work to the design of probability-constrained quantizers. Roughly speaking, we fix one functional block, quantizer or reconstruction, and optimize the other, repeatedly, obtaining a sequence of distortion improvements over the initial configuration. The main difference with respect to the Lloyd algorithm for conventional quantization is the more sophisticated nearest-neighbor condition (4), which we proved to be necessary for optimality in Theorem 4.

Below, we define, in greater detail, our modification of the Lloyd algorithm for probability-constrained quantization, which we shall call *probability-constrained Lloyd (PCL) algorithm*, henceforth:

1. Choose an initial reconstruction function $\hat{x}(q)$ and initial cost function $c(q)$.
2. Numerically adjust $c(q)$ in order to satisfy the probability constraints $p_Q(q) = p_0(q)$, given the current $\hat{x}(q)$. To this end, apply the regularized Gauss–Newton or Levenberg–Marquardt method described in Appendix A, and set the initial cost function as the cost function at the beginning of this step.
3. Update the current quantizer to the optimal one $q^*(x)$, corresponding to the current $\hat{x}(q)$ and the current $c(q)$, according to (4).
4. Find the optimal $\hat{x}^*(q)$ corresponding to the current $q(x)$, according to (3).
5. Go back to 2, until an appropriate convergence condition is satisfied, for example, a slowdown in the improvement of the sequence of distortion values obtained, or a given number of iterations reflecting a limit on the computation time.

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022

The initial reconstruction values may simply be chosen as $|\mathcal{Q}|$ random points drawn according to the probability distribution $p_X(x)$ of X, but the experiments in Section 7 will show that those computed by MDAV yield excellent results. A simple yet effective cost function initialization is $c(q) = 0$, because it ensures that the corresponding quantizer cells cannot have zero volume (provided no two reconstruction points coincide).

In practice, when the data is a finite set of points, the cost adjustment in Step 2 requires that the number of points per cell be large enough for this dependence to be smooth. As we shall see in the experiments in Section 7, in this work we used at least 500 points per cell, although the robust estimation technique of the Jacobian exploiting its seminegative definiteness, described in the appendix, enables to lower this number to almost 100 points for many statistics. In any case, the current, derivative-based cost adjustment method seems to represent a limitation on the applicability of the algorithm in k-anonymous clustering to relatively large values of k only. The fact remains that this is not a limitation of PCL, but of a step of the algorithm. Precisely, we could in principle explore discrete optimization algorithms to adjust costs, thereby making PCL suitable for any sort of microaggregation. An alternative approach to be explored in future research consists in adding noisy points around the original data points, with decreasing variance that will gradually vanish as the algorithm iterates. Note that the numerical computation of $c(q)$ in Step 2 should benefit from better and better initializations, that is, $c(q)$ in the previous iteration, as the reconstruction points become stable.

Just as in the conventional Lloyd algorithm, either the quantization or the reconstruction is improved at a time, leading to an improvement on the distortion. Strictly speaking, assuming that the numerical update of the cost function were exact, the sequence of distortions would be nonincreasing. Even though a nonnegative, nonincreasing sequence has a limit, rigorously speaking, the convergence of the sequence of distortions does not guarantee that the sequence of quantizers generated by PCL will tend to a stable configuration, less so to an optimal one. This is an issue inherited from the conventional form of the Lloyd algorithm on which we based our extension [19]. It is the excellent experimental performance of the Lloyd algorithm and not the lack of theoretical guarantees which motivated our second contribution, namely the proposal of PCL, based on the optimality conditions developed as our main contribution. In practice, as we shall see in Section 7, we do not need to solve the numerical computation update exactly to see monotonic distortion improvements. And although convergence to a *jointly* optimal solution is not guaranteed, our experimental results show that our algorithm outperforms MDAV, one of the best algorithms for microdata k-anonymization.

It is important to stress that any clustering algorithm, including sophisticated variable-size aggregation algorithms with excellent anonymity-distortion performance, may be used to initialize PCL. Indeed, not only the initial centroids $\hat{x}(q)$, but also the cell probabilities (relative sizes) $p_0(q)$ may be replicated from the partition produced by a starting algorithm such as MDAV, VMDAV, or μ-approx, mentioned in Section 2, possibly chosen according to its suitability to the dataset at hand. The optimal nearest-neighbor condition of Theorem 4 guarantees that, if cell costs $c(q)$ can be adjusted for those constraints, the very first PCL iteration along Steps 2 and 3 can only improve the starting distortion, while respecting the minimum anonymity constraint. Together with the iteration of Step 4, based on Theorem 1, this means that consequent distortions can only improve further over the distortion attained by whichever clustering algorithm the initialization and probability constraints of PCL were based on.

6.2. Further details and modifications in our algorithm

Next, we provide additional details and modifications of PCL that, while technical, are key to its robust implementation. An issue that may possibly arise after numerically adjusting the costs $c(q)$ in Step 2 to then obtain the optimized quantizer $q^*(x)$ in Step 3, is the creation of empty cells, or cells that are too small to provide a reliable estimation of a Jacobian in the method described in Appendix A, which loosely speaking assumes a certain smoothness or differentiability in the dependence between costs and cell sizes or probabilities. An effective mechanism to counter this consists in resetting the reconstruction points of cells affected by this issue to a value similar to the reconstruction point of a large cell, and resetting the cost to the same value. In other words, we split the largest cells, while keeping the allowed number of centroids, rather than throwing away centroids of empty or near empty cells, because the latter can only lead to a distortion increase. We would like to remark that we can only justify the specifics of this purely heuristic mechanism by its excellent empirical behavior in the experiments of Section 7.

A further modification to the conventional Lloyd algorithm, simple in appearance but quite beneficial in practice, consists in decelerating the reconstruction update of Step 4 by a speed factor $s \in (0, 1]$, which roughly speaking weighs optimal updates against previous reconstructions, and would yield the conventional aggressive update (3) for $s = 1$, namely $s\,\hat{x}^*(q) + (1 - s)\,\hat{x}(q)$. It follows almost directly from the definition of convexity that, for distortion measures $d(x, \hat{x})$ convex in the second argument $\hat{x}$, such as the squared norm used in MSE, this strategy will in general lead to an improvement in expected distortion $\mathcal{D}$, albeit suboptimal, thus maintaining the nonincreasing monotonicity of the sequence of distortions obtained with PCL. Intuitively, as cells experience only slight updates after sufficient iterations, the centroid update should tend to the optimal solution. While supported theoretically, ultimately, this is a heuristic rule, justified empirically by a better convergence to smaller distortion values. The reason is that the numerical method to perform the cost adjustment that ensures the probability constraints uses the costs in the previous iteration as initialization. Intuitively, centroid updates make these initializations more meaningful. Experimentally, this provides higher robustness in the numerical computation of the costs $c(q)$ in Step 2, and helps counter the occasional issue of empty cells. In all the experiments of Section 7 we used $s = 1/2$; large enough to approach optimal reconstructions quickly, but small enough to facilitate cost adjustment.

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022

Additional considerations dealing with repetition of points and inexact fulfillment of the cell size constraints are commented on in the experiments of Section 7.

7. Experimental results

This section will illustrate our PCL algorithm, inspired by the theoretical analysis of Section 5 and described in Section 6, with experimental results for both synthetic and real data, namely, a multivariate Gaussian distribution, and a well-known dataset containing certain census data.

We would like to emphasize that the main focus of this work is the proposal and the theoretical analysis of an algorithm for probability-constrained quantization, in principle applicable to a number of scenarios, including those of Section 4.2. Concordantly, we shall content ourselves with the empirical intuition provided by such synthetic data and single case of real data. On the one hand, it is far from difficult to find real-world data roughly fitting a jointly Gaussian model. This is certainly the case of the height and (the logarithm of) the weight of adult men, with correlation coefficient 0.48 according to [5]. On the other hand, the strong decay in probability density away from the mean of this synthetic distribution should pose an interesting challenge to PCL or any cell size constrained clustering algorithm, as cells of drastically different volume will be required to engulf the same number of points.

Finally, since the scope of this work is necessarily limited and mainly theoretical, our experimentation on real data is far from exhaustive, but it is insightful enough to serve our purposes. That is, to recognize that PCL outperforms the state of the art in microdata anonymization for at least certain statistics, and to confirm the desirable optimality properties established theoretically in previous sections. In more practical terms, we provide evidence to suggest PCL as a strong candidate to consider, along with any other state-of-the-art microdata anonymization algorithm, when seeking the best utility–anonymity performance, for mild running-time constraints.

We choose MDAV [15,11,20,44], mentioned in Section 2, as the contender against PCL, because of its acknowledged, state-of-the-art performance. In addition to quadratic distortion, our experiments compare the asymptotic complexity of MDAV and PCL, by means of a simple, doubly logarithmic regression analysis of running time versus number of data points. Although PCL is a fixed cell size algorithm, we shall further challenge it against VMDAV [41], a modification of MDAV for variable cell size, known to occasionally yield small improvements in distortion, while respecting the minimum cell size allowed, by heuristically exploiting high skewness in the point density of some datasets.

7.1. Quadratic-Gaussian case

We shall start with a simple, intuitive probability distribution. Precisely, we assume that the data X to be quantized may be modeled by a multivariate zero-mean Gaussian distribution, with a covariance matrix Σ of the form

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & & \ddots & & \vdots \\ \rho & \dots & \rho & 1 & \rho \\ \rho & \dots & \rho & \rho & 1 \end{pmatrix},$$

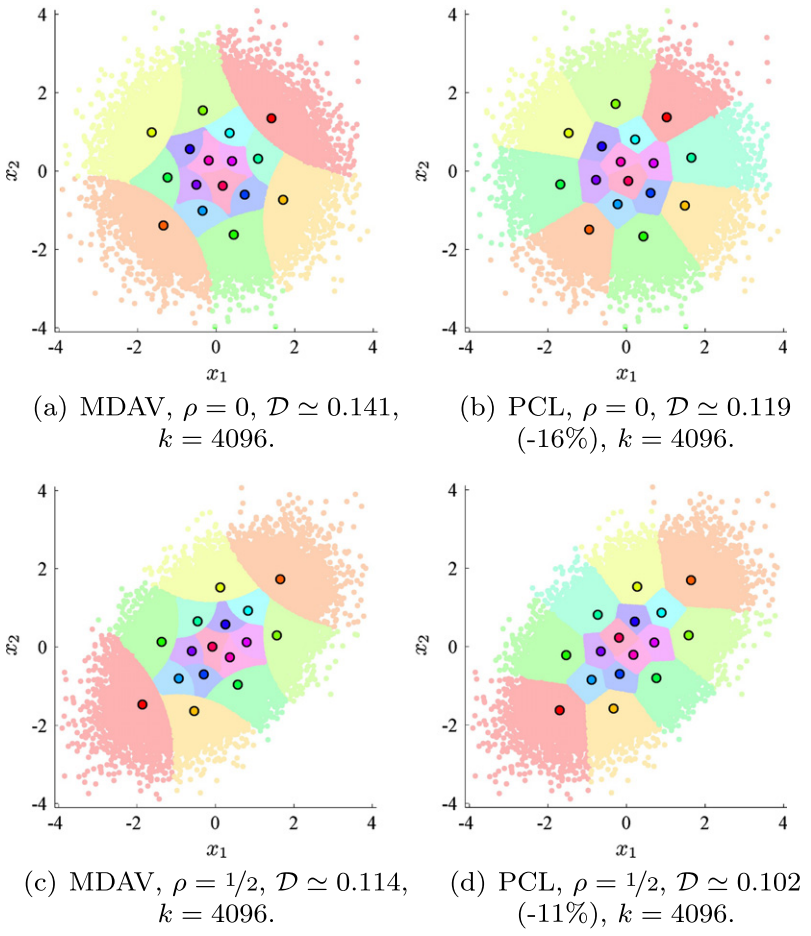
where the correlation coefficient ρ may take on the values 0, corresponding to independent entries, and 1/2.

A total of $n = 2^{16} = 65{,}536$ points are drawn according to these statistics, and the algorithm executed with the probability constraints $p_Q(q) = p_0$, with $p_0 = k/n = 1/|\mathcal{Q}|$, where as in Section 4.2, k denotes the k-anonymity requirement, and $|\mathcal{Q}|$ the number of quantization cells. Experiments were carried out for 1, 2, 3, and 4-dimensional points of X, and for $|\mathcal{Q}| = 2^0, \dots, 2^6$ cells, which corresponds to $k = 2^{16}, \dots, 2^{10}$, respectively. The choice of a large number k of points per cell facilitates the operation of the cost-optimization algorithm described in Appendix A, namely the Levenberg–Marquardt method, which relies on the smoothness of the objective function. If small values of k had been used instead, a discrete-optimization method would have undoubtedly been more suited. MSE normalized per dimension d is used as distortion measure, that is, $\mathcal{D} = \frac{1}{d} E\|X - \widehat{X}\|^2$, where the expectation is of course taken according to the empirical distribution of the n points: $\mathcal{D} = \frac{1}{d\,n} \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2$. Centroid initialization was based on that computed by MDAV, and costs were initialized to zero. The centroid speed factor was set to $s = 1/2$.

The same combinations of parameters and the same distortion criterion were chosen to compare the performance of our PCL algorithm with the popular MDAV for k-anonymous microaggregation. The results reported exclude the scalar case, which unsurprisingly yields identical performance for PCL and MDAV, simply corresponding to trivial quantizers with contiguous intervals of size k. As for two dimensions, the resulting clusters are shown in Fig. 3, which compares MDAV and PCL for $|\mathcal{Q}| = 16$ cells and $\rho = 0, 1/2$. For the 2D experiments depicted, PCL carried out a total of 40 iterations adjusting cells $q(x)$ and centroids $\hat{x}(q)$. Each of these outer iterations implemented up to 300 inner iterations of the numerical method in Appendix A to adjust the costs $c(q)$ to the probability constraints, while leaving the centroids fixed.

Interestingly, MDAV built cells resembling circular sectors as it ate away pairs of groups of k points from the remaining dataset, including diametrically opposed points. On the other hand, PCL constructed convex polytopes, confirming Theorem

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022



(a) MDAV, $\rho = 0$, $\mathcal{D} \simeq 0.141$, $k = 4096$.

(b) PCL, $\rho = 0$, $\mathcal{D} \simeq 0.119$ (-16%), $k = 4096$.

(c) MDAV, $\rho = 1/2$, $\mathcal{D} \simeq 0.114$, $k = 4096$.

(d) PCL, $\rho = 1/2$, $\mathcal{D} \simeq 0.102$ (-11%), $k = 4096$.

Fig. 3. Clustering of $n = 65{,}536$ Gaussian points into $|\mathcal{Q}| = 16$ equiprobable cells using MDAV and PCL.

3. It is important to notice that the quantization cells produced by PCL for the highest values of $|\mathcal{Q}|$ roughly resembled a hexagonal lattice, a lattice known to minimize the distortion among all two-dimensional lattices [8], up to local congruence transformations. In keeping with the centroid condition (3) in Proposition 1, reconstruction points $\hat{x}(q)$ are at the center of gravity of the cells. On account of the modified nearest-neighbor condition (4) used in Theorem 4, cell boundaries are segments orthogonal to the lines connecting reconstruction points, shifted according to the cost $c(q)$ difference between cells. Distortion was improved by PCL by a 16% and a 11% with respect to MDAV, for $\rho = 0$ and 1/2 respectively, for the same exact anonymity constraint $k = 4096$.

All of our experiments showed that PCL substantially and consistently outperformed MDAV, occasionally at the cost of a negligible probability constraint error. Finally, due to the differentiability assumption inherent in the cost adjustment methods used, when PCL is applied to a finite set of data points the cell size constraints may be attained only within a small margin of error. Occasionally, cells constraints were met by plus or minus 1 or 2 points, out of thousands. Posterior reassignment of these points, taking into account simple considerations of centroid proximity, enabled us to satisfy the constraints perfectly, with numerically negligible impact on the distortion. Of course, the distortions reported here take into account this small correction.

Fig. 4 plots both $\mathcal{D}$ and $p_{\min}$ for several dimensions, relative to MDAV. Note that negative increments are in fact improvements, that is, a lower MSE distortion for the same exact anonymity constraint. As these plot report, distortion improvements seem to increase with the number of cells $|\mathcal{Q}|$, and slightly decrease with higher correlation and dimension. PCL exhibits a potential distortion reduction over MDAV of up to 19%.

Concerning convergence behavior, we observed in all of our experiments that each optimization step decreased the distortion while approximately respecting the probability constraints. This experimental finding supports the fact that both the centroid condition (3) and our modified nearest-neighbor condition (4) are optimal. A representative case is shown in Fig. 5, where the distortion $\mathcal{D}$ of PCL is plotted for each of 50 outer iterations of the algorithm, for $|\mathcal{Q}| = 32$, and compared to the distortion introduced by MDAV. This plot is representative of the fact that a few outer iterations commonly suffice to reach a better distortion. The convergence behavior observed in all of our experiments seems very promising, and similar to that

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022

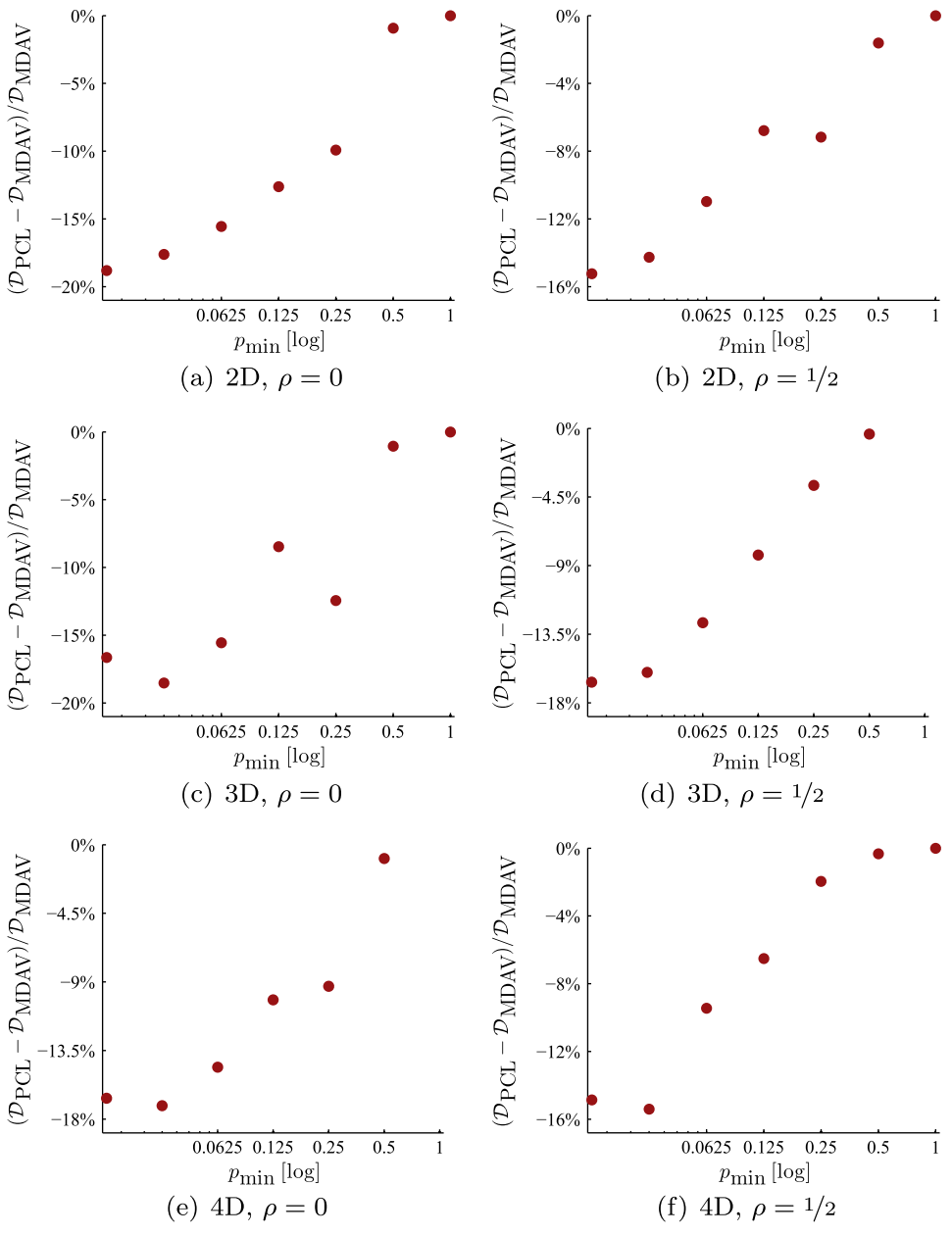


Fig. 4. Relative MSE increments $(\mathcal{D}_{\mathrm{PCL}} - \mathcal{D}_{\mathrm{MDAV}})/\mathcal{D}_{\mathrm{MDAV}}$ of PCL with respect to MDAV, versus probability constraints $p_{\min} = 2^0, 2^{-1}, \ldots, 2^{-6}$, for $n = 2^{16}$ Gaussian points in 2–4 dimensions, and correlations $\rho = 0, 1/2$. Negative increments represent improvements.

often exhibited by the conventional Lloyd algorithm. Namely, often the same low-distortion solution is found regardless of the initialization, in a small number of iterations.

Regarding running time, we must mention that our proof-of-concept implementations of PCL and MDAV were written in a commercial fourth-generation programming environment, with an integrated C routine to speed up nearest-neighbor search, and run on a modern computer equipped with an Intel Xeon CPU @ 2.67 GHz and Windows 7 64-bit. Roughly 44 s were required for each of the experiments of Fig. 3. Recall that the complexity of MDAV is theoretically proved to be $\Theta(n^2)$, that is, its running time asymptotically grows with the square of the number of data points n, for fixed anonymity k and dimension of the data. Although a thorough theoretical analysis of the asymptotic complexity of PCL is out of the scope of this paper, the following simple experimental analysis suggests that its running time is also nonlinear; the results point towards an approximate asymptotic dependence with the square of the number of data points. Concretely, Fig. 6 compares the running times of MDAV and PCL for a synthetic dataset generated according to the Gaussian statistics described, uncorrelated ($\rho = 0$) and in two dimensions, and n from 30,000 to 100,000 points, in steps of 5000. The k-anonymity requirement

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022

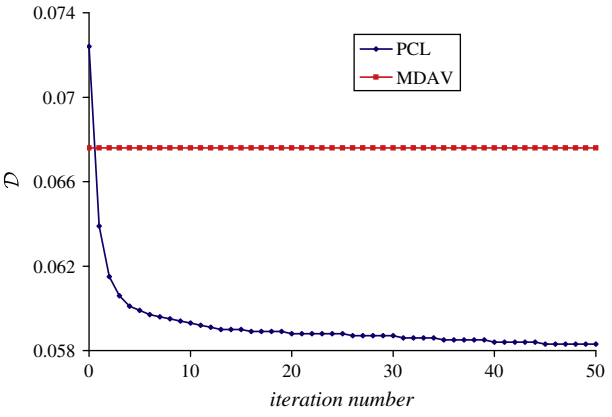


Fig. 5. Distortion optimization for $|\mathcal{Q}| = 32$ and $\rho = 1/2$.

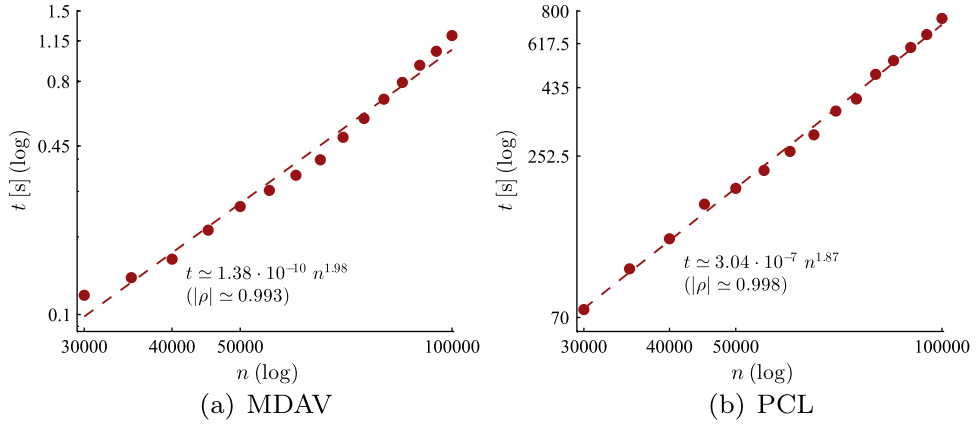


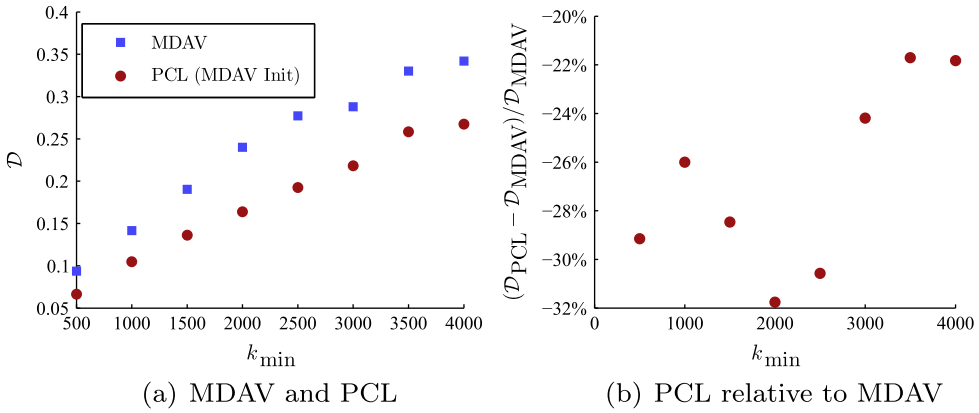
Fig. 6. Running time t in seconds vs. dataset size n for MDAV and PCL. A total of $n = 30{,}000, 35{,}000, 40{,}000, 45{,}000, \ldots, 100{,}000$ 2D points of uncorrelated, zero-mean, unit-variance Gaussian coordinates were generated. The anonymity constraint k was fixed at 1000 points, and the number of PCL iterations, set to 100. The linear regression in the doubly logarithmic axes points towards an approximately quadratic dependence, with very high absolute regression coefficient $|\rho|$ (not to be confused with the correlation of the data points).

was fixed at 1000 points. The number of outer iterations of PCL was set to 100, more than sufficient to attain distortion convergence in all cases, speeding up as cell changes and cost adjustments became gradually less pronounced. A regression analysis in the doubly logarithmic representation[1] confirms the quadratic dependence of MDAV with an estimated exponent of 1.98, and suggests a roughly similar polynomial dependence for PCL, with an estimated exponent of 1.87. Even though the same experiments were repeated for higher dimensions, this quadratic dependence remained approximately unchanged; e.g., 2.00 for MDAV and 1.91 for PCL, in 10D.

Although the asymptotic behavior seems to be the same, from a more practical standpoint, and even though PCL consistently outperformed MDAV in terms of distortion for the same exact anonymity requirement, MDAV required seconds where PCL took minutes. For those SDC applications where the data collection procedure itself may require weeks, months or years, one may be more than willing to pay the price. Note also that the iterative nature of PCL lends itself to dynamically updated datasets. In more time-demanding scenarios, a principle explored by [22], to speed up the conventional Lloyd algorithm when the amount of data is significantly large, consists in prepartitioning the data, and then applying the algorithm to each partition individually rather than to the entire dataset, hopefully at the cost of a small distortion loss. The general principle of prepartitioning, or hierarchical clustering, is directly applicable to PCL, in order to keep in check the nonlinear cost of the algorithm with the number of data points, but this remains the object of future research.

[1] Suppose $t \simeq \alpha n^{\beta}$. Then, $\log t \simeq \log\alpha + \beta \log n$, i.e., $\log t$ is approximately an affine function of $\log n$. Note however that complexities of the form $t \simeq \alpha n^{\beta} \log^{\gamma} n$ would also be approximated in the same fashion, for large n.

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022



(a) MDAV and PCL (b) PCL relative to MDAV

Fig. 7. MSEs versus anonymity constraints $k_{\min} = 500, 1000, 1500, \ldots, 4000$, for the UCI Adult dataset. $\mathcal{D}_{\mathrm{MDAV}}$ and $\mathcal{D}_{\mathrm{PCL}}$ denote the MSE of MDAV and PCL, respectively. Relative MSE increments $(\mathcal{D}_{\mathrm{PCL}} - \mathcal{D}_{\mathrm{MDAV}})/\mathcal{D}_{\mathrm{MDAV}}$ are also reported. Negative increments represent improvements.

7.2. UCI adult dataset

Although, as we have stressed, the main focus of this work is theoretical, we would like to complete our experiments with a quick example based on real data, specifically on the UCI Adult dataset [46]. This is a dataset commonly used to infer whether incomes exceed 50 K/yr based on census data, but it is perfectly suited for microaggregation.[2] All 48,842 records, corresponding to individual respondents in the USA in 1994, were included (training and testing), but out of the 14 attributes available, many categorical, only three numerical attributes were considered: age, education number and hours per week. Ages range from 17 to 90 years old. The education quantity is related to the number of years of education, 1 being preschool, 2 representing 1st–4th grades, and so on, all the way to 16 for a doctorate. Hours per week range from 1 to 99. The respective means are 38.6, 10.1 and 40.4, with standard deviations 13.7, 2.57 and 12.4. The corresponding matrix of correlation coefficients

$$\begin{pmatrix} 1 & 0.0309 & 0.0716 \\ 0.0309 & 1 & 0.144 \\ 0.0716 & 0.144 & 1 \end{pmatrix}$$

shows a weak correlation between these three attributes. As customary in SDC, each column underwent a zero-mean, unit-variance normalization, prior to carrying out the microaggregation process. Because there are only 9953 distinct combinations among all 48,842 points, all points were slightly perturbed by independent, zero-mean, uniformly distributed noise of negligible variance, in order to facilitate the setting of quantization cell boundaries in accordance to precise cell size constraints. Needless to say, the assignments from perturbed points to quantization cells made by PCL were later applied to the original, unperturbed points, and the distortion recomputed to verify that an utterly negligible variation was obtained ($\leqslant$0.1%).

Just as for the Gaussian experiments reported in Section 7.1, MSE per dimension was used as distortion measure $\mathcal{D}$, precisely, SSE normalized by dimension (3) and number of points (48,842). Because of the unit-variance normalization, the SST is equal to the dimension, thus our normalized distortion measure coincides with the usual in the literature: $\mathcal{D} = \mathrm{SSE}/\mathrm{SST}$.

The centroids of PCL were initialized from MDAV, and the costs initially set to zero. The speed rate in the centroid update was also $s = 1/2$. The anonymity constraints were $k = 500, 1000, 1500, \ldots, 4000$. The number of iterations of PCL ranged from 35, for the largest k, to 70, for the smallest, and execution times from 26 s, to 5 min and 25 s.

As Fig. 7 reports, PCL outperformed MDAV by at least a 22% reduction in distortion for the largest values of k, and up to a 32% reduction for the intermediate value $k = 2000$. The anonymity constraints were met in all cases but $k = 500$, for which the correction procedure used in the Gaussian experiments had to move 1 sample from a cell with excess samples to a cell with only 499, procedure that left the distortion virtually unchanged (0.2% variation).

We repeated our experiments on this standardized dataset comparing PCL with VMDAV in lieu of MDAV, a slightly more challenging test due to the mentioned fact that VMDAV may occasionally adjust cell sizes to local skewness in the density of the data points, to improve the distortion while respecting the minimum anonymity k. We closely followed the description of the VMDAV algorithm in [41]. The distortion shown for VMDAV corresponds to the lowest among those obtained for γ from 0 to 2 in steps of 0.1, a parameter of VMDAV roughly construed as the aggressiveness with which groups initially formed absorb nearby points in a later step, behaving almost identically to MDAV when $\gamma = 0$.[3] Other than replacing MDAV

[2] We thank an anonymous reviewer for recommending carrying out experiments for this particular dataset, and comparing PCL not only with MDAV but also VMDAV.

[3] We are grateful to the authors of the paper for verifying that our implementation of VMDAV was thoroughly faithful to their description, that the values of γ were appropriate, and that the distortions obtained matched their own implementation.

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022

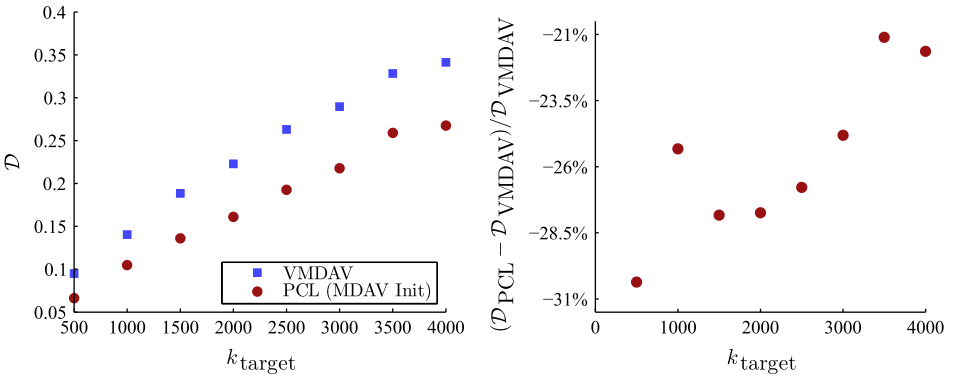


Fig. 8. MSEs versus anonymity constraints $k_{\min} = 500, 1000, 1500, \ldots, 4000$, for the UCI Adult dataset. $\mathcal{D}_{\text{VMDAV}}$ and $\mathcal{D}_{\text{PCL}}$ denote the MSE of VMDAV and PCL, respectively. Relative MSE increments $(\mathcal{D}_{\text{PCL}} - \mathcal{D}_{\text{VMDAV}})/\mathcal{D}_{\text{VMDAV}}$ are also reported. Negative increments represent improvements.

by VMDAV, no changes were made in the experimental procedure described above for the UCI Adult dataset, not even the initialization method for PCL, as we preferred to leave any possible variable-size improvements to our algorithm for future research. We would like to stress that instead of basing the initialization of PCL on MDAV, we could have replicated the initial centroids and the probability constraints of PCL from VMDAV, as explained towards the end of Section 6.1. Despite the fixed-size initialization and operation of PCL, as shown in Fig. 8, VMDAV hardly led to any significant improvements over MDAV and was clearly outperformed by our algorithm, by a distortion reduction ranging from 21% to 31%.

8. Concluding remarks

We study the problem of designing quantizers of minimum distortion satisfying arbitrary cell-probability constraints, for both discrete and continuous probability distributions, from a data compression perspective. This includes the problem of clustering satisfying the k-anonymity requirement, but also addresses applications of similarity-based, workload-constrained resource allocation.

Our contribution is twofold. First and most importantly, we present a theoretical analysis that proves the optimality conditions probability-constrained quantizers must satisfy. Part of the importance of this analysis lies in the fact that it provides a novel, theoretical characterization of optimal k-anonymous aggregation. As a second contribution, inspired by our theoretical analysis, we propose an alternating optimization algorithm for the design of this type of quantizers, which we call PCL. Our algorithm is conceptually motivated by the popular Lloyd–Max algorithm for quantization design, originally intended for data compression, also known as the k-means method. The centroid condition remains the same, but the nearest-neighbor condition is expressed in terms of an additive cost function that shifts cell boundaries to satisfy the probability constraint. The resulting cells are shown to be convex polytopes. Costs are updated by solving a system of nonlinear equations with a Tychonov regularized version of the Gauss–Newton method, or with the Levenberg–Marquardt algorithm. Although this is the computationally most expensive part of our quantizer design algorithm, fortunately, the Jacobian of the function modeling the dependence between costs and cell probabilities possesses a sparse structure, and the estimation of each of its entries does not require the requantization of all data points. A few sophistications are mentioned for robustness, such as a slowdown in the centroid update, and a Frobenius approximation of the Jacobian to a nonpositive definite matrix.

Because PCL is based on the alternation of two optimality conditions, there exists a theoretical guarantee that the sequence of distortions produced will be nonincreasing. A convenient consequence of this fact is that, in practice and in general, it will lead to a performance improvement over any other algorithm used for initialization, MDAV in our experiments.

It is important to stress the differentiability assumptions inherent in the numerical methods for cost adjustment mentioned, which estimate the Jacobian of the function that models the dependence between costs and cell probabilities. In practice, when PCL is applied to a finite set of points, this cost adjustment method requires that the number of points per cell be large enough for this dependence to be smooth. The slow centroid update, the mechanism against empty cells, and the robust estimation technique of the Jacobian exploiting its seminegative definiteness, enable to lower this number to almost 100 points for many statistics. We could in principle explore discrete optimization algorithms to adjust costs, thereby making PCL suitable for any sort of microaggregation. An alternative approach to be explored in future research consists in adding noisy points around the original data points, with decreasing variance that will gradually vanish as the algorithm iterates.

Experimental results regarding k-anonymous clustering (not microaggregation) for Gaussian statistics and the UCI Adult dataset, with MSE distortion, confirm that our method significantly outperforms MDAV, one of the best algorithms for microdata k-anonymization, in terms of data utility, for the same exact anonymity constraint. In addition to quadratic distortion, our experiments compare the asymptotic complexity of MDAV and PCL by means of a doubly logarithmic regression analysis

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022

of running time versus number of data points, which points towards a quadratic dependence of the running time with the number of data points, suggesting that PCL scales similarly to MDAV with dataset size. For the same dataset size, however, the distortion reduction gained by PCL does come at the expense of higher running time and mathematical sophistication. The variable-size modification of MDAV known as VMDAV is capable of occasionally exploiting skewness of the data point density of certain datasets, in order to slightly lower the distortion. Yet our experiments show that VMDAV is also significantly outperformed by PCL, for the standardized dataset tested, in spite of the fact that possible variable-size improvements of PCL itself are left for future research.

In addition to its greater generality and lower distortion, our framework enables us to represent a quantizer unambiguously and compactly, simply as a list of reconstruction values and costs, one per cell, rather than an arbitrary clustering of a large cloud of points. This is particularly useful when a model of the data is given by means of a PDF, for which a probability-constrained quantizer is to be designed only once, but later on applied repeatedly to dynamic sets of points distributed according to the original model.

Because the scope of this work is necessarily limited and mainly theoretical, our experimentation on real data is far from exhaustive, but it is insightful enough to serve our purposes. That is, to recognize that PCL outperforms the state of the art in microdata anonymization for at least certain statistics, and to confirm the desirable optimality properties established theoretically in previous sections. In more practical words, admittedly, the experimental evidence provided does not guarantee the best utility–anonymity trade-off for any and all datasets or statistics corresponding to specific applications of SDC. However, it does suffice to suggest PCL as a strong candidate to consider, along with any other state-of-the-art microdata anonymization algorithms, when seeking optimal performance in terms of distortion and k-anonymity. It is also fair to add that PCL comes at the cost of higher, albeit tractable, computational complexity and implementation sophistication. The general principle of prepartitioning the data, and then clustering each prepartition individually, could very well be applied to PCL, in order to keep in check the roughly quadratic scaling of the algorithm with the number of data points, similar to MDAV.

Acknowledgments

We would like to thank the anonymous reviewers for numerous, invaluable comments, which greatly improved the manuscript; particularly, the experiments with the standardized dataset UCI Adult and the comparison with VMDAV owe to these suggestions. We are also grateful to Dr. J. Nin for thorough comments that helped us improve the clarity of our presentation. Last but not least, our appreciation extends to Prof. A. Solanas and Prof. T. Martínez-Ballesté, the authors of VMDAV, for their help in verifying our own implementation and the results obtained.

This work was partly supported by the Spanish Government through projects Consolider Ingenio 2010 CSD2007-00004 "ARES", TEC2010-20572-C02-02 "Consequence", and by the Government of Catalonia under Grant 2009 SGR 1362. D. Rebollo-Monedero is the recipient of a Juan de la Cierva postdoctoral fellowship, JCI-2009-05259, from the Spanish Ministry of Science and Innovation.

Appendix A. Numerical computation of the cost function

Given a reconstruction function $\hat{x}(q)$, we address the problem of the numerical computation of the cost function $c(q)$ such that the associated quantizer $q^*(x)$ (4) satisfies the probability constraints $p_Q(q) = p_0(q)$. We assume that the distribution of X is continuous, or a fine discretization of a continuous one.

First, we note that a choice of cost function $c(q)$ completely determines the cell PMF $p_Q(q)$ (recall $\hat{x}(q)$ is given). Accordingly we concisely denote $c(q)$ and $p_Q(q) - p_0(q)$ by the vectors c and p in $\mathbb{R}^{|\mathcal{Q}|}$, respectively, and define the corresponding function $p(c)$. In this notation, we wish to solve $p(c) = 0$, which we regard as a differentiable system of nonlinear equations. We denote by $J(c)$ the Jacobian of $p(c)$ at c.

Observe that solving $p(c) = 0$ is equivalent to minimizing $\|p(c)\|^2$. The minimization problem may be solved via an iterative descent method, which involves finding a descent direction and carrying a line search along it, at each step. For the line search portion, we used the popular backtracking search method with Armijo's rule [28]. To find an appropriate descent direction, we tried, with similar success, a Tychonov-regularized version of the Gauss–Newton method [2], and a variation called the Levenberg–Marquardt method [24,30,33], which we proceed to describe.

Let the current estimate of the solution for the cost be c_0. The Gauss–Newton method considers the minimization of the linearization $\|p(c_0) + J(c_0)(c - c_0)\|^2$ of the objective $\|p(c)\|^2$. The Jacobian $J(c_0)$ is estimated by introducing small perturbations to the cost vector around c_0. The choice of the appropriate size of the small increment for each entry of the cost vector may be done reliably using a one-dimensional search algorithm, under the constraint that the number of points in the cell decrease by a given percentage, say 5%. To this end, we used a standard implementation of Brent's root-finding method [3], based on [17]. Brent's method is a sophisticated combination of bisection, secant, and inverse quadratic interpolation methods.

A potential issue occurs when the Jacobian $J(c_0)$ is singular or ill-conditioned. To counter it, we adopt the common strategy of Tychonov's regularization, in other words, we seek to minimize

$$\|p(c_0) + J(c_0)(c - c_0)\|^2 + \mu\|c - c_0\|^2$$

as a function of c, for a small damping parameter $\mu > 0$. Recall that in the limit of small μ, this gives the least-norm solution $c - c_0$ to the possibly underdetermined system $p(c_0) + J(c_0)(c - c_0) = 0$. Conceptually, we seek the smallest perturbation of c_0 solving the linearized minimization problem. μ may be chosen simply by trial and error, or using any of the common heuristics for the Levenberg–Marquardt algorithm. The solution to the regularized least-squares problem satisfies the normal equation

$$(J(c_0)^{\mathrm{T}} J(c_0) + \mu I)(c - c_0) = -J(c_0)^{\mathrm{T}} p(c_0), \tag{5}$$

where it turns out that $J(c_0)^{\mathrm{T}} p(c_0)$ is the gradient of $\frac{1}{2}\|p(c)\|^2$ at $c = c_0$, and $J(c_0)^{\mathrm{T}} J(c_0)$ its Hessian matrix. Intuitively, μ slightly rotates the Newton search direction towards the steepest descent direction, increasing the robustness of the method at the cost of convergence speed near the solution. The solution $c - c_0$ is taken as the search direction for the current step. It is a descent direction because it may be written as the product of a positive definite matrix by the negative gradient. The backtracking line search determines the new value for c_0. A heuristic improvement, proposed by Marquardt, consists in replacing the identity matrix in (5) by a matrix containing only the diagonal elements of $J(c_0)^{\mathrm{T}} J(c_0)$. In this fashion, for large values of the damping parameter μ, the algorithm does not simply follow the direction of the gradient, but it does take into account the local curvature along the canonical directions, which may accelerate its convergence.

Finally, we would like to remark that the estimation of the Jacobian $J(c_0)$ is carried out by slightly increasing each of the coordinates of c_0 at a time, exploiting the fact that only the coordinates of p corresponding to neighboring cells may be changed. More precisely, note that if $c(q)$ is slightly increased, then only points initially assigned to q by the quantizer may now be assigned to other regions. Therefore, there is no need to compute the new quantizer completely to obtain the perturbed $p_Q(q)$. On the other hand, directly from the definition of eigenvectors, it can be shown that the Jacobian is symmetric and negative semidefinite, in the case of continuous input data. In the practical case of a discrete cloud of data points, the Jacobian estimate can be refined by the closest negative semidefinite matrix, in the Frobenius norm, simply by computing the symmetric portion $\frac{1}{2}(J(c_0) + J(c_0)^{\mathrm{T}})$ of our Jacobian estimate, and then resetting positive eigenvalues to zero in its spectral decomposition. In practice, this provides a more reliable Jacobian estimate and cost adjustment. Further, the sparsity and definiteness of $J(c_0)$ may not only exploited to speed up its estimation, but also to solve the normal Eq. (5) more efficiently, for instance applying the conjugate gradient method [1].

The following summary outlines the algorithm to compute the cost function $c(q)$ such that $p_Q(q) = p_0(q)$, which is denoted more concisely here as c satisfying $p(c) = 0$. We consider $\mu > 0$ and the initial c_0 to be the input of the algorithm, and the final c_0 its output. We already commented on the initialization of μ. As for the initialization of c_0, the matter is addressed when the cost algorithm is incorporated into our PCL algorithm in Section 6.

1. Estimate the sparse matrix $J(c_0)$ by slightly increasing each of the coordinates of c_0 at a time and analyzing the changes with respect to $p(c_0)$. Since only neighboring cells will be affected, recomputing $q(x)$ entirely is unnecessary. For better robustness, optionally compute the negative semidefinite approximation in Frobenius norm.
2. Solve the normal Eq. (5) for c. Use the conjugate gradient method if $|\mathcal{Q}|$ is very large.
3. Perform backtracking line search along $c - c_0$ using Armijo's rule. Set c_0 as the new solution.
4. Go back to 1, unless the number of iterations exceeds a certain limit, or until an appropriate convergence condition is satisfied.

References

[1] R. Barrett, M. Berry, T.F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, H.V. der Vorst, Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, SIAM, Philadelphia, PA, 1994.
[2] A. Björck, Numerical Methods for Least Squares Problems, SIAM, Philadelphia, PA, 1996.
[3] R.P. Brent, Algorithms for Minimization without Derivatives, Prentice-Hall, Englewood Cliffs, NJ, 1973.
[4] J. Brickell, V. Shmatikov, The cost of privacy: destruction of data-mining utility in anonymized data publishing, in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery, Data Mining (KDD), Las Vegas, NV, August 2008.
[5] D.E. Burmaster, D.M. Murray, A trivariate distribution for the height, weight, and fat of adult men, Risk Analysis 18 (4) (1998) 385–389.
[6] J. Cao, B. Carminati, E. Ferrari, K. Tan, CASTLE: continuously anonymizing data streams, IEEE Transaction on Dependable and Secure Computation 99 (2009).
[7] C. Chin-chen, L. Yu-chiang, H. Wen-huang, TFRP: an efficient microaggregation algorithm for statistical disclosure control, Journal of Systems and Software 80 (11) (2007) 1866–1878.
[8] J.H. Conway, N.J.A. Sloane, Sphere Packings, Lattices and Groups, Springer, Berlin, Germany, 1993.
[9] D. Defays, P. Nanopoulos, Panels of enterprises and confidentiality: the small aggregates method, in: Proceedings of the Symposium on Design, Anal. Longitudinal Surveys at Statistics Canada, Ottawa, Canada, 1993, pp. 195–204.
[10] J. Domingo-Ferrer, U. González-Nicolás, Hybrid microdata using microaggregation, Information Sciences 180 (15) (2010) 2834–2844.
[11] J. Domingo-Ferrer, A. Martínez-Ballesté, J.M. Mateo-Sanz, F. Sebé, Efficient multivariate data-oriented microaggregation, The VLDB Journal 15 (4) (2006) 355–369.
[12] J. Domingo-Ferrer, J.M. Mateo-Sanz, Practical data-oriented microaggregation for statistical disclosure control, IEEE Transactions on Knowledge and Data Engineering 14 (1) (2002) 189–201.
[13] J. Domingo-Ferrer, F. Sebé, A. Solanas, A polynomial-time approximation to optimal multivariate microaggregation, Computers and Mathematics with Applications 55 (4) (2008) 714–732.
[14] J. Domingo-Ferrer, A. Solanas, J. Castellà-Roca, $h(k)$-private information retrieval from privacy-uncooperative queryable databases, Online Informations Review 33(4) (2009) 720–744.

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022

[15] J. Domingo-Ferrer, V. Torra, Ordinal, continuous and heterogenerous k-anonymity through microaggregation, Data Mining and Knowledge Discovery 11 (2) (2005) 195–212.
[16] J. Domingo-Ferrer, V. Torra, A critique of k-anonymity and some of its enhancements, in: Proceedings of Workshop Privacy and Security, Artificial Intelligence (PSAI), Barcelona, Spain, 2008, pp. 990–993.
[17] G.E. Forsythe, M.A. Malcolm, C.B. Moler, Computer Methods for Mathematical Computations, Prentice-Hall, Englewood Cliffs, NJ, 1976.
[18] A. Gersho, R.M. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publishers, Boston, MA, 1992.
[19] R.M. Gray, D.L. Neuhoff, Quantization, IEEE Transactions on Information Theory 44 (1998) 2325–2383.
[20] A. Hundepool, R. Ramaswamy, P.-P. DeWolf, L. Franconi, R. Brand, J. Domingo-Ferrer, μ-ARGUS version 4.1 software and user's manual, Voorburg, Netherlands, 2007 <http://neon.vb.cbs.nl/casc>.
[21] H. Jian-min, C. Ting-ting, Y. Hui-qun, An improved V-MDAV algorithm for l-diversity, in: Proceedings of IEEE International Symposium on Information Processes (ISIP), Moscow, Russia, May 2008, pp. 733–739.
[22] M. Käärik, K. Pärna, On the quality of k-means clustering based on grouped data, Journal of Statistical Planning and Inference 139 (11) (2009) 3836–3841.
[23] M. Laszlo, S. Mukherjee, Minimum spanning tree partitioning algorithm for microaggregation, IEEE Transactions on Knowledge and Data Engineering 17 (7) (2005) 902–911.
[24] K. Levenberg, A method for the solution of certain problems in least-squares, Quarterly of Applied Mathematics 2 (1944) 164–168.
[25] N. Li, T. Li, S. Venkatasubramanian, t-Closeness: privacy beyond k-anonymity and l-diversity, in: Proceedings of IEEE International Conference on Data Engineering (ICDE), Istanbul, Turkey, April 2007, pp. 106–115.
[26] J.L. Lin, T.H. Wen, J.C. Hsieh, P.C. Chang, Density-based microaggregation for statistical disclosure control, Expert Systems with Applications 37 (4) (2010) 3256–3263.
[27] S.P. Lloyd, Least squares quantization in PCM, IEEE Transactions on Information Theory IT-28 (1982) 129–137.
[28] D.G. Luenberger, Y. Ye, Linear and Nonlinear Programming, third ed., Springer, New York, 2008.
[29] A. Machanavajjhala, J. Gehrke, D. Kiefer, M. Venkitasubramanian, l-Diversity: privacy beyond k-anonymity, in: Proceedings of IEEE International Conference on Data Engineering (ICDE), Atlanta, GA, April 2006, p. 24.
[30] D. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, SIAM Journal of Applied Mathematics (SIAP) 11 (1963) 431–441.
[31] N. Matatov, L. Rokach, O. Maimon, Privacy-preserving data mining: a feature set partitioning approach, Information Sciences 180 (14) (2010) 2696–2720.
[32] J. Max, Quantizing for minimum distortion, IEEE Transactions on Information Theory 6 (1) (1960) 7–12.
[33] J.J. Moré, The Levenberg–Marquardt algorithm: implementation and theory, in: G.A. Watson (Ed.), Numerical Analysis, ser. Lecture Notes Math, vol. 630, Springer-Verlag, 1977, pp. 105–116.
[34] J. Nin, J. Herranz, V. Torra, On the disclosure risk of multivariate microaggregation, Data and Knowledge Engineering 67 (3) (2008) 399–412.
[35] A. Oganian, J. Domingo-Ferrer, On the complexity of optimal microaggregation for statistical disclosure control, UNECE Statistical Journal 18 (4) (2001) 345–354.
[36] D. Rebollo-Monedero, J. Forné, J. Domingo-Ferrer, From t-closeness to PRAM and noise addition via information theory, in: Privacy and Statistical Databases (PSDs), ser. Lecture Notes on Computational Sciences (LNCS), Springer-Verlag, Istambul, Turkey, 2008, pp. 100–112.
[37] D. Rebollo-Monedero, J. Forné, J. Domingo-Ferrer, From t-closeness-like privacy to postrandomization via information theory, IEEE Transactions on Knowledge Data Engineering 22(11) (2010) 1623–1636 <http://doi.ieeecomputersociety.org/10.1109/TKDE.2009.190>.
[38] P. Samarati, Protecting respondents' identities in microdata release, IEEE Transactions on Knowledge and Data Engineering 13 (6) (2001) 1010–1027.
[39] P. Samarati, L. Sweeney, Protecting Privacy When Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression, SRI Int., Tech. Rep., 1998.
[40] C.E. Shannon, Communication theory of secrecy systems, Bell System Technical Journal (1949).
[41] A. Solanas, A. Martínez-Ballesté, J. Domingo-Ferrer, VMDAV: a multivariate microaggregation with variable group size, in: Proceedings of Computational Statistics (COMPSTAT), Springer-Verlag, Rome, Italy, 2006.
[42] X. Sun, H. Wang, J. Li, T.M. Truta, Enhanced p-sensitive k-anonymity models for privacy preserving data publishing, Transactions of Data Privacy 1 (2) (2008) 53–66.
[43] L. Sweeney, Uniqueness of Simple Demographics in the US Population, Carnegie Mellon Univ., Sch. Comput. Sci., Data Priv. Lab., Pittsburgh, PA, Tech. Rep. LIDAP-WP4, 2000.
[44] M. Templ, Statistical disclosure control for microdata using the R-package sdcMicro, Transactions of Data Privacy 1(2) (2008) 67–85 <http://cran.r-project.org/web/packages/sdcMicro>.
[45] T.M. Truta, B. Vinay, Privacy protection: p-sensitive k-anonymity property, in: Proceedings of the International Workshop of Privacy Data Management (PDM), Atlanta, GA, 2006, p. 94.
[46] UCI adult dataset, 1996 <http://archive.ics.uci.edu/ml/datasets/Adult>.
[47] S. Zhong, Z. Yang, T. Chen, k-Anonymous data collection, Information Sciences 179 (172) (2009) 2948–2963.

Please cite this article in press as: D. Rebollo-Monedero et al., A modification of the Lloyd algorithm for k-anonymous quantization, Inform. Sci. (2012), http://dx.doi.org/10.1016/j.ins.2012.08.022