Comparing $L_1$ and $L_2$ Distances for CTA*

Jordi Castro

Department of Statistics and Operations Research,
Universitat Politècnica de Catalunya,
Jordi Girona 1–3, 08034 Barcelona, Catalonia
jordi.castro@upc.edu
http://www-eio.upc.edu/~jcastro

Abstract. Minimum distance controlled tabular adjustment (CTA) is a recent perturbative technique of statistical disclosure control for tabular data. Given a table to be protected, CTA looks for the closest safe table, using some particular distance. We focus on the continuous formulation of CTA, without binary variables, which results in a convex optimization problem for distances $L_1$, $L_2$ and $L_\infty$. We also introduce the $L_0$-CTA problem, which results in a combinatorial optimization problem. The two more practical approaches, $L_1$-CTA (linear optimization problem) and $L_2$-CTA (quadratic optimization problem) are empirically compared on a set of public domain instances. The results show that, depending on the criteria considered, each of them is a better option.

Keywords: statistical disclosure control, controlled tabular adjustment, linear optimization, quadratic optimization.

1 Introduction

Controlled tabular adjustment methods (CTA) [1,7] are considered an emerging technology for tabular data [10]. In terms of efficiency and quality of the solution, they usually perform well compared to other techniques [2,3].

CTA was initially [7] only formulated for $L_1$ norms and binary variables for deciding the sense of protection for the sensitive cells, i.e., whether to perturb up or down the original cell value. In [1], $L_2$ and $L_\infty$ were also considered in continuous formulations, i.e., the protection sense was a priori fixed without paying attention to infeasibility issues [4]. Results for the two most practical distances, $L_1$ and $L_2$, were presented in [13], but without a detailed comparison of the reported solutions. In addition, the same cell weights were used for $L_1$ and $L_2$ in the empirical results of [13]; as it will be stated later in this work, the comparison was unfair, since the weights used favored $L_1$. This work tries to fill this void by performing a more exhaustive empirical evaluation of $L_1$-CTA versus $L_2$-CTA. A new variant $L_0$-CTA, closer to $L_1$-CTA than to $L_2$-CTA, will also be formulated.


© Springer-Verlag Berlin Heidelberg 2012
The paper is organized as follows. Section 2 formulates the continuous CTA problem (i.e., a priory fixing the value of binary variables) for \( L_0, L_1, L_2 \) and \( L_\infty \). Section 3 introduces the criteria considered in the comparison of \( L_1 \)-CTA and \( L_2 \)-CTA. Finally, Section 4 reports the computational comparison.

## 2 Formulations of CTA for Several Distances

Any CTA instance can be represented by the following parameters:

- A set of cells \( a_i, i \in \mathcal{N} = \{1, \ldots, n\} \), that satisfy some linear relations \( Aa = b \) (\( a \) being the vector of \( a_i \)'s). The particular structure of the table is defined by equations \( Aa = b \). Each tabular constraint imposes that the inner cells have to be equal to the total or marginal cell. Any type of table can be modeled by these equations.

- A lower and upper bound for each cell \( i \in \mathcal{N} \), respectively \( l_{a_i} \) and \( u_{a_i} \), which are considered to be known by any attacker. If no previous knowledge is assumed for cell \( l_{a_i} = 0 \) (\( l_{a_i} = -\infty \) if \( a \geq 0 \) is not required) and \( u_{a_i} = +\infty \) can be used.

- A set \( S = \{i_1, i_2, \ldots, i_s\} \subseteq \mathcal{N} \) of indices of confidential cells.

- Nonnegative lower and upper protection levels for each confidential cell \( i \in S \), respectively \( lpl_i \) and \( upl_i \), such that the released values satisfy either \( x_i \geq a_i + upl_i \) or \( x_i \leq a_i - lpl_i \).

CTA attempts to find the closest safe values \( x_i, i = 1, \ldots, n \), according to some distance \( L \), that makes the released table safe. This involves the solution of the following optimization problem:

\[
\min_x ||x - a||_L \\
\text{s. to} \quad Ax = b \quad \text{(1)} \\
l_{a_i} \leq x_i \leq u_{a_i} \quad i \in \mathcal{N} \\
(x_i \leq a_i - lpl_i) \text{ or } (x_i \geq a_i + upl_i) \quad i \in S.
\]

Introducing a vector of binary variables \( y \in \mathbb{R}^s \) to model the disjunctive constraints (either “upper protection sense” \( x_i \geq a_i + upl_i \) when \( y_i = 1 \) or “lower protection sense” \( x_i \leq a_i - lpl_i \) when \( y_i = 0 \)), the above problem can be formulated as a mixed integer linear optimization problem (MILP), which can be time consuming for medium-large instances.

A more efficient alternative for the real-time protection in on-line tabular data servers —or in other situations where processing time matters (like when protecting very large sets of linked tables)— would be to a priori fix the binary variables, thus obtaining a CTA formulation with only continuous variables [5]. Possible infeasibilities in the resulting problem could be dealt with the approaches exposed in [6], some of them already used in the context of CTA [4]. Formulating problem (1) in terms of cell deviations \( z = x - a \), and fixing the binary...