Published in final edited form as:

Priv Stat Databases. 2016 ; 9867: 41–53. doi:10.1007/978-3-319-45381-1_4.

A Second Order Cone Formulation of Continuous CTA Model

Goran Lesaja[1], **Jordi Castro**[2], **Anna Oganian**[1,3]

Goran Lesaja: goran@georgiasouthern.edu; Jordi Castro: jordi.castro@upc.edu; Anna Oganian: aoganyan@cdc.gov

[1] Department of Mathematical Sciences, Georgia Southern University, P.O. Box 8093, Statesboro, GA 30460-8093, U.S.A.

[2] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Jordi Girona 1–3, 08034 Barcelona, Catalonia.

[3] National Center for Health Statistics 3311 Toledo Rd, Hyatsville, MD, 20782, U.S.A.

Abstract

In this paper we consider a minimum distance Controlled Tabular Adjustment (CTA) model for statistical disclosure limitation (control) of tabular data. The goal of the CTA model is to find the closest safe table to some original tabular data set that contains sensitive information. The measure of closeness is usually measured using ℓ_1 or ℓ_2 norm; with each measure having its advantages and disadvantages. Recently, in [4] a regularization of the ℓ_1-CTA using Pseudo-Huber function was introduced in an attempt to combine positive characteristics of both ℓ_1-CTA and ℓ_2-CTA. All three models can be solved using appropriate versions of Interior-Point Methods (IPM). It is known that IPM in general works better on well structured problems such as conic optimization problems, thus, reformulation of these CTA models as conic optimization problem may be advantageous. We present reformulation of Pseudo-Huber-CTA, and ℓ_1-CTA as Second-Order Cone (SOC) optimization problems and test the validity of the approach on the small example of two-dimensional tabular data set.

Keywords

Statistical disclosure limitation (control); controlled tabular adjustment models; pseudo-Huber function; convex optimization; second-order cone optimization; interior-point methods

1 Introduction

The *statistical disclosure limitation (control)* is the term that describes the theory and methods of protecting sensitive information when releasing statistical microdata or tabular data. An up-to-date overview of theory and methods of this field can be found in the monograph [19] and, for tabular data only, in the survey [8]. An excellent reference is also [27].

Minimum-distance controlled tabular adjustment (CTA) methodology was first introduced in [7,15]. As indicated in [4] CTA can be formulated as the following problem: Given a table with sensitive cells, compute the closest safe table in which sensitive cells are modified to avoid recomputation, and the remaining cells are minimally adjusted to satisfy the table equations. The closeness of the original and modified table is measured by the weighted

distance between the tables with respect to a certain norm. Most commonly used norms are ℓ_1 and ℓ_2 norms. Thus, the problem can be formulated as a minimization problem with the objective function being a particular weighted distance function and constraints being table equations and lower and upper bounds on the cell values.

In general, CTA is Mixed Integer Optimization Problem (MIOP) which is a difficult problem to solve especially for the large dimension problems. A priori fixing the values of binary variables reduces the problem to the continuous optimization problem which is easier to solve, however, the quality of the solution may be reduced. In addition, the values of the binary variables have to be assigned carefully otherwise the problem may become infeasible [12, 13].

The objective function in continuous CTA is based on either the ℓ_1-norm or ℓ_2-norm. The formulation of ℓ_2 -CTA leads to the Quadratic Programing (QP) problem, while ℓ_1-CTA can be formulated as the Linear Programming (LP) problem. However, the resulting LP has the number of variables that is twice the number of cells of the table as opposed to ℓ_2-CTA where the resulting QP problem has a number of variables equal to the number of cells. In general, the QP of ℓ_2-CTA is usually more efficiently solved than the LP of ℓ_1-CTA [4,7].

In [4] the Pseudo-Huber regularization of the ℓ_1-CTA is proposed. The Pseudo-Huber approximation of the ℓ_1-norm objective function leads to the convex optimization problem. However, the advantage is that the number of variables in Pseudo-Huber formulation of the ℓ_1-CTA remains the same as the number of cells. In [4] it is shown that Pseudo-Huber-CTA can be more efficiently solved than LP ℓ_1-CTA for certain types of tables and using an appropriate method that takes into account the structure of the problem.

All these models are solved using appropriate versions of the Interior-Point Method (IPM). These methods have been developed in recent years to efficiently solve different types of, often large, nonlinear (convex) optimization problems. It has been shown both theoretically and numerically that IPMs perform better on problems that have a certain structure, such as Conic Optimization (CO) problems, which are LP problems where variables are elements of cones. Most common cones are non-negative orthant, second order (quadratic) cone and semidefinite cone [2, 3, 25].

Hence, motivated by the above comment, in this paper we develop a new Second Order (Quadratic) Cone (SOC) formulation of the ℓ_1 and Pseudo-Huber-CTA. It is shown on the small example of a two-dimensional table that SOC CTA models are more efficiently solved than the original models. It is expected that the same will be the case for larger and more complex tables. Extensive numerical testing on various types of tables is beyond the scope of this paper; however, it is needed and it is forthcoming as a part of future research.

The paper is organized as follows. In Section 2 the general MIOP and then continuous CTA are formulated. Then the ℓ_1 and ℓ_2 continuous CTA are derived. The Pseudo-Huber-CTA formulation is considered in Section 3. The new SOC formulations of both Pseudo-Huber and ℓ_1 CTA are developed in Section 4. In Section 5 the SOC CTA models are applied to the small example of two-dimensional table and these instances are solved using MOSEK SOC solver. The concluding remarks are given in Section 6.

2 Formulation of the General CTA Model

The following CTA formulation is given in [4]: Given the following set of parameters:

- **i.** A set of cells a_i, $i \in \mathcal{N} = \{1, \ldots, n\}$. The vector $a = (a_1, \ldots, a_n)^T$ satisfies certain linear system $Aa = b$ where $A \in \mathbb{R}^{m \times n}$ is an $m \times n$ matrix and and $b \in \mathbb{R}^m$ is m-vector.
- **ii.** A lower, and upper bound for each cell, $l_{a_i} \leq a_i \leq u_{a_i}$ for $i \in \mathcal{N}$, which are considered known by any attacker.
- **iii.** A set of indices of sensitive cells, $\mathcal{S} = \{i_1, i_2, \ldots, i_s\} \subseteq \mathcal{N}$.
- **iv.** A lower and upper protection level for each sensitive cell $i \in \mathcal{S}$ respectively, lpl_i and upl_i, such that the released values must be outside of the interval $(a_i - lpl_i, a_i + upl_i)$.
- **v.** A set of weights, w_i, $i \in \mathcal{N}$ used in measuring the deviation of the released data values from the original data values.

A CTA problem is a problem of finding values z_i, $i \in \mathcal{N}$, to be released, such that z_i, $i \in \mathcal{S}$ are safe values and the weighted distance between released values z_i and original values a_i, denoted as $\|z - a\|_{l(w)}$, is minimized, which leads to solving the following optimization problem

$$\begin{aligned} &\min_{z} \|z - a\|_{l(w)} \\ &s.t.\ Az = b, \\ &\quad l_{a_i} \leq z_i \leq u_{a_i},\ i \in \mathcal{N}, \\ &\quad z_i,\ i \in \mathcal{S} \text{ are safe values.} \end{aligned} \tag{1}$$

As indicated in the assumption (iv) above, safe values are the values that satisfy

$$z_i \leq a_i - lpl_i \text{ or } z_i \geq a_i + upl_i,\ i \in \mathcal{S}\,. \tag{2}$$

By introducing a vector of binary variables $y \in \{0,1\}^s$ the constraint (2) can be written as

$$\begin{aligned} z_i &\geq -M(1 - y_i) + (a_i + upl_i)y_i,\ i \in \mathcal{S}, \\ z_i &\leq My_i + (a_i - lpl_i)(1 - y_i), \quad i \in \mathcal{S}, \end{aligned} \tag{3}$$

where $M \gg 0$ is a large positive number. Constraints (3) enforce the upper safe value if $y_i = 1$ or the lower safe value if $y_i = 0$.

Replacing the last constraint in the CTA model (1) with (3) leads to a mixed integer convex optimization problem (MIOP) which is in general a difficult problem to solve; however, it provides solutions with high data utility [11]. The alternative approach is to fix binary variables up front which leads to a CTA that is acontinuous convex optimization problem.

The continuous CTA may be easier to solve; however, the obtained solution may have a lower data utility. Furthermore, a wrong assignment of binary variables may result in the problem being infeasible. Strategies on how to avoid this difficulty are discussed in [12, 13].

In this paper we consider a continuous CTA where binary variables are fixed and vector z is replaced by the vector of *cell deviations*

$$x = z - a\,. \tag{4}$$

The CTA (1) with constraints (3) reduces to the following convex optimization problem:

$$\begin{aligned} &\min_{x} \|x\|_{l(w)} \\ &s\,.t\,.\ Ax = 0, \\ &\quad l \le x \le u, \end{aligned} \tag{5}$$

where upper and lover bounds for x_i, $i \in \mathcal{N}$ are defined as follows:

$$l_i = \begin{cases} upl_i & \text{if } i \in \mathcal{S} \text{ and } y_i = 1 \\ l_{a_i} - a_i & \text{if } (i \in \mathcal{N}\backslash\mathcal{S}) \text{ or } \left(i \in \mathcal{S} \text{ and } y_i = 0\right) \end{cases} \tag{6}$$

$$u_i = \begin{cases} -lpl_i & \text{if } i \in \mathcal{S} \text{ and } y_i = 0 \\ u_{a_i} - a_i & \text{if } (i \in \mathcal{N}\backslash\mathcal{S}) \text{ or } \left(i \in \mathcal{S} \text{ and } y_i = 1\right). \end{cases} \tag{7}$$

The two most commonly used norms in problem (5) are the ℓ_1 and ℓ_2 norms. For the ℓ_2-norm the problem, (5) reduces to the following ℓ_2-CTA model which is a QP problem:

$$\begin{aligned} &\min_{x} \sum_{i=1}^{n} w_i x_i^2 \\ &s\,.t\,.\ Ax = 0, \\ &\quad l \le x \le u\,. \end{aligned} \tag{8}$$

For the ℓ_1-norm the problem, (5) reduces to the following ℓ_1-CTA model:

$$\begin{aligned} &\min_{x} \sum_{i=1}^{n} w_i |x_i| \\ &s.t.\ Ax = 0, \\ &l \le x \le u. \end{aligned} \tag{9}$$

The above ℓ_1-CTA model (9) is a convex optimization problem; however, the objective function is not differentiable at $x = 0$. Since most of the algorithms, including IPMs, require differentiability of the objective function, problem (9) needs to be reformulated. The reformulations that have been considered in [4] are reviewed in the next section.

3 LP and Pseudo-Huber Formulation of ℓ_1-CTA

The ℓ_2-CTA model (8) is a standard QP problem that can be efficiently solved using IPM or other methods. However, as noted at the end of the previous section, the ℓ_1-CTA model (9) needs reformulation in order to be efficiently solved by IPM or some other method. The standard reformulation is the transformation of model (9) to the following LP model:

$$\begin{aligned} &\min_{x^-, x^+} \sum_{i=1}^{n} w_i \left(x_i^+ + x_i^-\right) \\ &\text{s.t. } A\left(x_i^+ - x_i^-\right) = 0, \\ &l^+ \le x^+ \le u^+, \\ &l^- \le x^- \le u^-, \end{aligned} \tag{10}$$

where

$$x^+ = \begin{cases} x & \text{if } x \ge 0 \\ 0 & \text{if } x < 0, \end{cases} \qquad x^- = \begin{cases} 0 & \text{if } x > 0 \\ -x & \text{if } x \le 0, \end{cases} \tag{11}$$

and lower and upper bounds for x_i^- and x_i^+, $i \in \mathcal{N}$ are as follows:

$$
\begin{aligned}
l_i^+ &= \begin{cases} upl_i & \text{if } i \in \mathcal{S} \text{ and } y_i = 1 \\ 0 & \text{if } (i \in \mathcal{N}\backslash\mathcal{S}) \text{ or } \left(i \in \mathcal{S} \text{ and } y_i = 0\right) \end{cases} \\
u_i^+ &= \begin{cases} 0 & \text{if } i \in \mathcal{S} \text{ and } y_i = 0 \\ u_{a_i - a_i} & \text{if } (i \in \mathcal{N}\backslash\mathcal{S}) \text{ or } \left(i \in \mathcal{S} \text{ and } y_i = 1\right) \end{cases} \\
l_i^- &= \begin{cases} lpl_i & \text{if } i \in \mathcal{S} \text{ and } y_i = 0 \\ 0 & \text{if } (i \in \mathcal{N}\backslash\mathcal{S}) \text{ or } \left(i \in \mathcal{S} \text{ and } y_i = 1\right) \end{cases} \\
u_i^- &= \begin{cases} 0 & \text{if } i \in \mathcal{S} \text{ and } y_i = 1 \\ a_i - l_{a_i} & \text{if } (i \in \mathcal{N}\backslash\mathcal{S}) \text{ or } \left(i \in \mathcal{S} \text{ and } y_i = 0\right). \end{cases}
\end{aligned} \tag{12}
$$

Problem ℓ_1-CTA (10) is an LP problem; however, it has twice the number of variables as the QP problem (8) and twice the number of box constraints. As indicated in [4], the splitting of the variables $x = x^+ - x^-$ and the increased dimension of the model may cause problems. In order to overcome these difficulties in [4] it was suggested to use a regularization of problem (9) by approximating absolute value with the Pseudo-Huber function that has the same number of variables as in the QP formulation (8).

The original Huber function $\varphi_\delta : \mathbb{R} \to \mathbb{R}_+$ is defined as

$$
\varphi_\delta(x_i) = \begin{cases} \dfrac{x_i^2}{2\delta} & |x_i| \le \delta \\ |x_i| - \dfrac{\delta}{2} & |x_i| \ge \delta. \end{cases} \tag{13}
$$

It approximates $|x_i|$ for small values of $\delta > 0$; the smaller the δ, the better the approximation.The Huber function is continuously differentiable; however, the second derivative is not continuous at $|x_i| = \delta$ which may cause problems when this function is used in second order optimization algorithms, such as IPMs. Hence, it is better to consider the Pseudo-Huber function $\phi_\delta : \mathbb{R} \to \mathbb{R}_+$

$$
\phi_\delta(x_i) = \sqrt{\delta^2 + x_i^2} - \delta \tag{14}
$$

whose first and second derivatives are bounded and Lipschitz continuous [17]. Again, the smaller the δ the better the approximation.

Now, the ℓ_1-CTA problem (9) can be approximated by the following convex optimization problem

$$
\begin{aligned}
&\min_x \sum_{i=1}^n w_i \phi_\delta(x_i) \\
&s.t.\ Ax = 0, \\
&l \le x \le u.
\end{aligned} \tag{15}
$$

The advantage of the Pseudo-Huber-CTA model (15) is that it has the same number of variables as ℓ_2 - CTA and the same feasible region, the only difference is that the quadratic objective function is replaced by a strictly convex function.

Optimization problems (8), (10) and (15) can be solved with appropriate versions of the Interior-Point Methods (IPM). Since IPMs are the methods of choice to solve different CTA models, in the rest of the section we describe the main ideas of IPMs, only on a conceptual level, and then we discuss their application on given CTA models.

IPMs have in many ways revolutionized the optimization theory and practice in the past three decades since the appearance of the Karmarkar's breakthrough paper [20]. Since then, the field of IPMs has been a very active area of research with literary thousands of papers published as well as numerous excellent monographs and textbooks. The general theory of IPMs for convex optimization problems can be found in the seminal monograph of Nesterov and Nemirovskii [26]. In addition to this monograph, and without any attempt to be complete, we mention a few other relevant references [29, 28, 22]. The reason for such an interest is that IPMs have proven to be very efficient in solving large linear and non-linear (convex) optimization problems which were previously hard to solve. Now-days almost every relevant optimization software, whether commercial or open source, contains an IPM solver which is capable of solving at least LP problems and in many cases QP problems, and, less frequently, conic optimization problems. In the case of LP there are plenty of numerical studies showing that IPMs are at least as efficient, if not more, as the classical Simplex Method (SM) on large scale LP problems.

The basic idea of path-following IPMs, that are most commonly used and studied, is centered around approximately following the parametric trajectory that is called *central path* which leads to the solution of the problem when a parameter is approaching zero. The points on the central path are called *μ-centers* and are obtained as solutions of the Karush-Kuhn-Tucker (KKT) optimality conditions of the problem where a (the) complementarity equation(s) is (are) perturbed by a positive parameter $\mu > 0$. In particular, the perturbed KKT system for Pseudo-Huber-CTA is explicitly listed in [4].

The solution of the problem, which is obtained when $\mu = 0$, is found by tracing the central path while gradually reducing μ to zero. However, tracing the central path exactly would be prohibitively inefficient. The main achievement of IPMs have been to show that it is sufficient to trace the central path approximately; as long as the iterates are in the certain neighborhood of the central path, it is still possible to prove global convergence and, moreover, show that the -approximate solution of the problem, according to the appropriate proximity measure, can be obtained in polynomial number of iterations with the best theoretical upper bound being $O\left(\sqrt{n}\log\frac{n}{\epsilon}\right)$, where n represents the number of variables of the problem at hand.

However, practical behavior of IPM heavily depends on many factors, such as the structure of the problem, the starting point, the accuracy needed, etc. As reported in [4], Pseudo-Huber-CTA (15) can be difficult to solve with a general convex optimization solver even for small instances if the solver is not 'appropriately tuned'. However, for problems that exhibit

a special structure such as 3-D tables whose constraints have a block-angular structure, the specialized block-angular IPM of J. Castro [5, 9, 10] solves Pseudo-Huber-CTA more efficiently than ℓ_1-CTA while ℓ_2-CTA has by far the best CPU time. Hence, Pseudo-Huber-CTA is a viable option for solving ℓ_1-CTA; however, the IPM have to be implemented with care and, in addition, the specialized IPM may not work efficiently for other types of tables. As indicated in [4], modifications and tuning of the Block-angular IPM so it can handle large and complex tables of different types is a direction for future research.

Another direction in searching how to efficiently solve Pseudo-Huber-CTA and ℓ_1-CTA is to investigate whether these models can be transformed into the conic optimization (CO) problems. The motivation for such investigation comes from the fact that it has been established both theoretically and numerically that IPMs perform better on the well structured problems such as CO problems than on general convex optimization problems [2, 3, 25]. CO problems are LP problems over cones, that is, variables belong to certain types of cones. Most common cones are either non-negative orthant, second-order (quadratic) cone or semidefinite cone definitions; of which are listed in the next section. Thus, formulating Pseudo-Huber and ℓ_1-CTA as CO problems would be advantageous. In the next section we develop SOC formulation of both Pseudo-Huber and ℓ_1 CTA.

4 SOC Formulation of Pseudo-Huber and ℓ_1 CTA

In this section we investigate how Pseudo-Huber and ℓ_1 CTA can be formulated as SOC models. The CO problems can be formulated as

$$\begin{aligned} &\min_x c^T x \\ &s.t.\ Ax = b, \\ &x \in \mathscr{K}, \end{aligned} \tag{16}$$

where $\mathscr{K}$ is a cone of the following three types:

1. The linear cone or non-negative orthant:

$$\mathscr{K} = \mathbb{R}^n_+ := \left\{x \in \mathbb{R}^n : x_i \geq 0,\ i = 1, \ldots, n\right\}.$$

2. The positive semidefinite cone:

$$\mathscr{K} = \mathbf{S}^n_+ := \left\{X \in \mathbf{S}^n : X \succeq 0\right\},$$

where $\succeq$ means that X is positive semidefinite matrix and $\mathbf{S}^n$ is a set of symmetric n-dimensional matrices.

3. The quadratic or second-order cone:

$$\mathscr{K} = \mathscr{L}^n = \{x \in \mathbb{R}^n : x_i \geq \sqrt{x_1^2 + \cdots + x_{i-1}^2 + x_{i+1}^2 + \cdots + x_n^2}\}.$$

More generally, $\mathcal{K}$ can be a Cartesian product of the above mentioned cones. It is also worth mentioning that the cones defined above are examples of symmetric cones, thus problem (16) can be considered in a more general framework of Symmetric Optimization (SO) problems, see [16,18, 24] and references therein.

In what follows, we present a reformulation of Pseudo-Huber-CTA problem (15) as a SOC problem. Consider Pseudo-Huber Function (14)

$$\phi_\delta(x_i) = \sqrt{\delta^2 + x_i^2} - \delta .$$

Let's define

$$t_i := \sqrt{\delta^2 + x_i^2} \quad \text{and} \quad y_i := \delta, \quad i = 1, \ldots, n . \tag{17}$$

Hence, we have

$$t_i = \sqrt{x_i^2 + y_i^2}$$

which is the boundary of the second-order (quadratic) cone

$$\mathcal{K}_i = \left\{ (x_i, y_i, t_i) \in \mathbb{R}^3 \ : \ t_i \geq \sqrt{x_i^2 + y_i^2} \right\} .$$

Now, the reformulation of the Pseudo-Huber-CTA (15) as a SOC problem follows

$$\begin{aligned} \max_x \ & \sum_{i=1}^n w_i (t_i - y_i) \\ s.t. \ & Ax = 0, \\ & y_i = \delta; \quad i = 1, \ldots, n, \\ & (x_i, y_i, t_i) \in \mathcal{K}_i; \quad i = 1, \ldots, n, \\ & l \leq x \leq u . \end{aligned} \tag{18}$$

This model is valid even for $\delta = 0$. In that case we obtain a SOC formulation of the $l1$-CTA (9)

$$\begin{aligned} \min_x \ & \sum_{i=1}^n w_i t_i \\ s.t. \ & Ax = 0, \\ & (x_i, t_i) \in \mathcal{K}_i; \quad i = 1, \ldots, n, \\ & l \leq x \leq u . \end{aligned} \tag{19}$$

This model could have been obtained directly from ℓ_1-CTA (9) because the absolute value has an obvious second-order cone representation since the epigraph of the absolute value function is exactly second-order cone, that is,

$$t_i = |x_i| \quad \rightarrow \quad \mathcal{K}_i = \left\{(x_i, t_i) \in \mathbb{R}^2 : t_i \geq \sqrt{x_i^2}\right\}.$$

It is well known that the solutions of SOC problems (18) and (19) achieve solutions at the boundary of the cones, hence, equations (17) will hold at the solution [2,3]. Thus, it is not necessary to enforce these equations in SOC models; in fact, their inclusion would lead to noncovex problems that would be difficult to solve.

An IPM for SOC can now be used to find an -approximate solutions to SOC Pseudo-Huber and ℓ_1 CTA models. We have used MOSEK SOC solver [1] that is considered one of the best, if not the best, SOC solver available on the market today.

5 Numerical Results for the Small Example

In this section an example of the small two-dimensional table stated in Figure 3 in [4] is considered. The table is listed in Figure 1 below as the table (a).

The continuous CTA model based on the table (a) is formulated in the following way:

- The linear constraints are obtained from the requirement that the sum of the elements in each row (or column) remains constant and is equal to the corresponding component in the last column (or row) of table (a).
- The sensitive cells are cells a_1 and a_{12}. For both of them the upper safe values are enforced, which are listed in the parentheses in the lower right corners of the cells, $upl_1 = 3$ and $upl_{12} = 5$ respectively. Hence, in the transformed tables the upper safe value of the cell a_1 should be 13 or above and for a_{12} the upper safe value should be 18 or above.
- For the nonsensitive cells the lower and upper bounds are set to be zero and positive infinity respectively, that is, $l_{ai} = 0$ and $u_{ai} = \text{inf}$ for $i = 2, \ldots, 11$.
- The weights in the objective function are set to have the value one, that is, $w_i = 1$ for $i = 1, \ldots, 12$.

From this basic CTA model different CTA models discussed in the paper were formulated and then these models were solved using appropriate IPM solvers. The results are listed in Figure 1. Below is the summary of the IPM solvers used.

1. The ℓ_2-CTA (8) instance was solved in [4] using IPM based MOSEK QP solver. Table (d).
2. The LP-ℓ_1-CTA (10) instance was solved in [4] using MOSEK LP solver. The IPM solver with crossover option was used. Table (b).
3. The Pseudo-Huber-CTA (15) instance was solved in [4] using Block-angular IPM of J. Castro [5, 9, 10]. Table (c).

4. The SOC ℓ_1-CTA (19) instance was solved for the first time in this paper using IPM based MOSEK SOC solver. Table (e).

5. The SOC Pseudo-Huber-CTA (18) instance was solved for the first time in this paper using IPM based MOSEK SOC solver. Table (f).

In [4] it was observed that ℓ_2-CTA had the fastest execution. Hence, we replicated the solution of the ℓ_2-CTA instance of the example and compared its performance with SOC models instances. The calculations were carried out on a Lenovo ThinkPad W530 computer with Intel(R) CORE i7–3740QM 2.70GHz processor. The results are given in Table 1.

From Table 1 we can observe that SOC versions are comparable to the ℓ_2 version both in number of iterations and CPU time; SOC ℓ_1 was slightly faster than ℓ_2 while SOC Pseudo-Huber was slightly slower, which is the expected result. Hence, the SOC models are more effective than the LP ℓ_1 and Pseudo-Huber-CTA models for this example.

Furthermore, for LP ℓ_1, Pseudo-Huber $\phi_{0.001}$, SOC ℓ_1, and SOC Pseudo-Huber $\phi_{0.001}$ CTA instances the optimal values of their respective objective functions are the same, namely, the value is 20, while for ℓ_2-CTA instance it is 20.69. Thus, the objective values for SOC Pseudo Huber and ℓ_1-CTA instances are the same as for the original non-SOC instances, namely 20, which was expected.

These results are in line with plenty of other evidence that it is advantageous to solve the SOC formulation of the problem by IPM, rather than using IPM to the original formulation of the problem (see for example [2, 3, 25, 23]). We are confident that the advantages of the SOC models will be even more visible when applied to larger tabular data sets. Moreover, the SOC IPM is robust and flexible enough to handle different types of tables.

6 Concluding Remarks

The main goal of the paper is mainly theoretical, that is, to present a Second Order Cone (SOC) formulation of the Pseudo-Huber and the ℓ_1 CTA models, (18) and (19) respectively as an alternative to the original Pseudo-Huber and LP ℓ_1 CTA models, (15) and (10) respectively. The application of the SOC models to the small example in Section 5 shows promise to be an effective alternative to the application of the original models to the small example. More numerical testing is needed and is forthcoming as a future research topic where SOC models would be implemented and tested on the different types of tables of large dimensions mentioned in the Conclusion of [4].

From Figure 1, it can be observed that the resulting tables for all the models except LP ℓ_1 change most of the cells of the original table (a) that are not fixed, even the ones that are not sensitive cells. The reason lays in the nature of IPMs. In these methods, the iterates approximately follow the central path that converges to the analytic center of the optimal set which implies that most of the cells will be changed, while the IPM with crossover or alternatively the Simplex Method, for LP ℓ_1 finds the basic solution which implies fewer cells will be changed. Hence, if there is a requirement to minimize the number of non-sensitive cells that are changed, then the LP ℓ_1 models solved with SM or IPM with crossover is the right approach. However, if the number of nonsensitive cells changed is not

an issue such as for certain types of magnitude tables, then the suggested approach is to use either the SOC ℓ_1 model or the ℓ_2 model because they are faster. Unless prior regularization of the ℓ_1 model is necessary, which then leads to the Pseudo-Huber model and related SOC Pseudo-Huber model, it is more efficient to use the SOC ℓ_1 model directly.

As noted in [4], it has been empirically shown that CTA in general exhibits a low disclosure risk [6] and, at the same time, high data utility [14, 13] (see also [21]). However, the study of the disclosure risk and data utility of tables protected by the Pseudo-Huber-CTA model and the SOC CTA models is lacking and is certainly an interesting future research topic.

Acknowledgments

The first author would like to thank Erling Andersen and Florian Jarre for the constructive discussion regarding the SOC model and Iryna Petrenko for her help in performing the calculations described in Section 5.

The authors would like to express their appreciation to Donald Malec for his careful reading of the paper and many useful suggestions.

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors only and do not necessarily reflect the views of the Centers for Disease Control and Prevention.

References

1. Andersen ED, MOSEK solver, https://mosek.com/resources/doc, 2016
2. Alizadeh F, and Goldfarb D, Second-order cone programming. Math. Programming, 95(1):3–51, 2003.
3. Andersen ED, Roos C and Terlaky T, On implementing a primal-dual interior-point method for conic quadratic optimization. Math. Programming, 95(2):249–277, 2003.
4. Castro J, A CTA Model Based on the Huber Function Privacy in Statistical Databases 2014, LNCS, 8744:79–88, 2014.
5. Castro J, An interior-point approach for primal block-angular problems. Computational Optimization and Applications, 36:195–219, 2007.
6. Castro J, On assessing the disclosure risk of controlled adjustment methods for statistical tabular data. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 20:921–941, 2012.
7. Castro J, Minimum-distance controlled perturbation methods for large-scale tabular data protection. European Journal of Operational Research, 171:39–52, 2006.
8. Castro J, Recent advances in optimization techniques for statistical tabular data protection. European Journal of Operational Research, 216:257–269, 2012.
9. Castro J and Cuesta J, Quadratic regularization in an interior-point method for primal block-angular problems. Mathematical programming, 130:415–445, 2011.
10. Castro J and Cuesta J, Solving ℓ_1-CTA in 3D tables by an interior-point method for primal block-angular problems. TOP, 21:25–47, 2013.
11. Castro J and Gonzalez JA, Assessing the information loss of controlled adjustment methods in two-way tables Privacy in Statistical Databases 2014, LNCS, 8744:79–88, 2014.
12. Castro J and Gonzalez JA, A fast CTA method without complicating binary decisions. Documents of the Joint UNECE / Eurostat Work Session on Statistical Data Confidentiality, Statistics Canada, Ottawa, 1–7, 2013.
13. Castro J and Gonzalez JA A multiobjective LP approach for controlled tabular adjustment in statistical disclosure control. Working paper, Department of Statistics and Operations Research, Universitat Politecnica de Catalunya, 2014.

14. Castro J and Giessing S, Testing variants of minimum distance controlled tabular adjustment In Monographs of Official Statistics, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 333–343, 2006.
15. Dandekar RA and Cox LH, Synthetic tabular Data: an alternative to complementary cell suppression Manuscript, Energy Information Administration, U.S. 2002.
16. Faraut J and Koranyi A, Analysis on Symmetric Cones. Oxford University Press, New York, USA, 1994.
17. Fountoulakis K and Gondzio J A second-order method for strongly convex L1-regularization problems. Technical Report ERGO-14–005, School of Mathematics, The University of Edinburgh, 2014.
18. Gu G, Interior-Point Methods for Symmetric Optimization. Ph.D. Thesis, TU Delft, 2009.
19. Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Schulte Nordholt E,Spicer K and DeWolf P-P Statistical Disclosure Control. Wiley, Chichester, United Kingdom, 2012.
20. Karmarkar N, A polynomial-time algorithm for linear programming. Combinatorica, 4:373–395, 1984.
21. Karr AF, Kohnen CN, Oganian A, Reiter JP, and Sanil AP, A frameworkfor evaluating the utility of data altered to protect confidentiality. The American Statistician, 60(3):224–232, 2006.
22. Lesaja G, Introducing Interior-Point Methods for Introductory Operations Research Courses and/or Linear Programming Courses. The Open Operational Research Journal, 3:1–12, 2009.
23. Lesaja G and Slaughter V, Interior-point algorithms for a class of convex optimization problems. Yugoslav Journal of Operations Research, 19(3):239–248, 2009.
24. Lesaja G and Roos C, Kernel-based interior-point methods for monotone linear complementarity problems over symmetric cones, J. Optim. Theory Appl 150(3):444–474, 2011.
25. Ben-Tal A and Nemirovski A, Lectures in Modern Convex Optimization: Analysis, Algorithms and Engineering Applications MPS/SIAM Series in Optimization, SIAM, Philadelphia, 2001.
26. Nesterov Y and Nemirovski A, Interior-Point Polynomial Algorithms in Convex Programming SIAM Studies in Applied Mathematics, Volume 13, SIAM, Philadelphia, 1994.
27. Oganian A, Security and Information Loss in Statistical Database Protection. PhD thesis, Universitat Politecnica de Catalunya, 2003.
28. Roos C, Terlaky T, and Ph J Vial, Theory and Algorithms for Linear Optimization An Interior-Point Approach. Springer Science, 2005.
29. Wright SJ, Primal-Dual Interior-Point Methods. SIAM, Philadelphia, 1996.

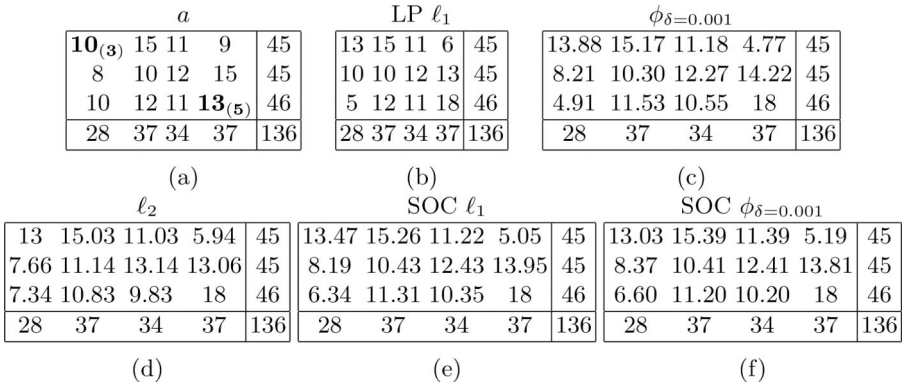

Fig. 1.
Results of the small example (rounded to two decimal places).

Table 1.

Results for ℓ_2 and SOC CTA

CTA Model	Obj. Funct.	It. No.	CPU
ℓ_2	20.69	6	0.08
SOC-ℓ_1	20	7	0.07
SOC Pseudo-Huber	20	9	0.09