A CTA Model Based on the Huber Function*

Jordi Castro

Department of Statistics and Operations Research,
Universitat Politècnica de Catalunya,
Jordi Girona 1–3, 08034 Barcelona, Catalonia
jordi.castro@upc.edu

Abstract. Minimum distance controlled tabular adjustment (CTA) is an emerging perturbative method of statistical disclosure control for tabular data. The goal of CTA is to find the closest safe table to some original tabular data with sensitive information. Closeness is usually measured by ℓ_1 or ℓ_2 distances. Distance ℓ_1 provides solutions with a smaller ℓ_0 norm than ℓ_2 (i.e., with a lesser number of changes with respect to the original table). However the optimization problem formulated with ℓ_2 requires half the number of variables than that for ℓ_1, and it is more efficiently solved. In this work a pseudo-Huber function (which is a continuous nonlinear approximation of the Huber function) is considered to measure the distance between the original and protected tables. This pseudo-Huber function approximates ℓ_1 but can be formulated with the same number of variables than ℓ_2. It results in a nonlinear convex optimization problem which, theoretically, can be solved in polynomial time. Some preliminary results using the Huber-CTA model are reported.

Keywords: Statistical disclosure control, controlled tabular adjustment, nonlinear optimization, convex optimization, interior-point methods, Huber loss function, pseudo-Huber loss function.

1 Introduction

The statistical disclosure control field aims at protecting sensitive information when releasing statistical microdata or tabular data. A description of the state-of-the-art in this field can be found in the monograph [17] and—only for tabular data—in the survey [5].

Minimum-distance controlled tabular adjustment (CTA), introduced in [3,14], is one of the available post-tabular perturbation approaches for tabular data. The purpose of CTA is, given a table with sensitive cells, to compute the closest safe table (i.e., sensitive cells are modified to avoid re-computation, the remaining cells are minimally adjusted to satisfy the table equations) through the solution of an optimization problem using some particular distance in its objective

* Supported by grants MTM2012-31440 of the Spanish Ministry of Economy and Competitiveness, SGR-2014-542 of the Government of Catalonia, and DwB INFRA-2010-262608 of the FP7 European Union Program.

J. Domingo-Ferrer (Ed.): PSD 2014, LNCS 8744, pp. 79–88, 2014.

function. CTA is considered an emerging technology for tabular data [17]. It has been empirically shown that CTA in general exhibits a low disclosure risk [6] and, at the same time, a high data utility [10,11].

CTA was originally formulated as a mixed integer linear programming (MILP) problem [14], while the minimum distance formulation of [3] was continuous (either a linear programming (LP) or a quadratic programming (QP) problem). Continuous formulations, which can be obtained by a priori fixing the value of the binary variables, provide faster optimizations, at the expense of reducing the quality of the solution. A wrong assignment of binary variables may result in an infeasible problem. The approach of [12,13] solves this situation by allowing small changes in three different type of CTA constraints. Together with the original objective, this results in a four-objective problem, which can be solved by multiobjective optimization methods [12,13].

In this work we focus on the continuous formulation of CTA. Using ℓ_1 as the distance in the objective function we obtain a LP whose number of variables is twice the number of cells of the table. For ℓ_2 we obtain a QP with a number of variables equal to the number of cells, which is in general more efficiently solved than the LP of ℓ_1-CTA [3]. (This does not hold if binary variables are considered: the MIQP ℓ_2-CTA is significantly harder than the MILP ℓ_1-CTA, as noted in [9].) On the other hand, ℓ_1-CTA solutions have a lesser ℓ_0 norm (where $\|x\|_{\ell_0}$ is the number of nonzero elements of x), i.e., the number of changes in cell values with respect to the original table is smaller. The purpose of this work is to present a new CTA model using a different objective function, whose optimization problem is of the same dimension than the one formulated by ℓ_2-CTA, but with a solution similar to that obtained with ℓ_1-CTA. We will see that the pseudo-Huber function guarantees both properties.

The paper is organized as follows. Section 2 reviews the CTA formulation without binary variables for ℓ_1 and ℓ_2. Section 3 presents a CTA variant based on a pseudo-Huber function, and provides some of its properties. Section 4 discusses the solution of the convex optimization problem formulated by the Huber-CTA model by an interior-point polynomial time algorithm. Finally, Section 5 reports very preliminary computational results with some midsize three-dimensional tables.

2 The CTA Formulation

Any CTA instance can be formulated from the following parameters: (i) 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); (ii) 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; (iii) nonnegative cell weights $w_i, i \in \mathcal{N}$, used for the distance between the original and the perturbed released cell values; (iv) a set $\mathcal{S} = \{i_1, i_2, \ldots, i_s\} \subseteq \mathcal{N}$ of indices of sensitive cells; (v) and 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 out of the interval$(a_i - lpl_i, a_i + upl_i)$.

CTA attempts to find the closest values $z_i, i \in \mathcal{N}$ —according to some distance $\ell(w)$, weighted by w— that make the released table safe. This involves the solution of the following optimization problem:

$$\begin{aligned} \min_{z} \quad & ||z-a||_{\ell(w)} && \text{(1a)}\\ \text{s. to} \quad & Az = b && \text{(1b)}\\ & l_{a_i} \le z_i \le u_{a_i} \quad i \in \mathcal{N} && \text{(1c)}\\ & z_i \; i \in \mathcal{S} \text{ are safe values.} && \text{(1d)} \end{aligned}$$

The formulation of (1d) depends on the particular controlled adjustment variant considered. For instance, in the standard CTA approach, this constraint is

$$(z_i \le a_i - lpl_i) \text{ or } (z_i \ge a_i + upl_i) \quad i \in \mathcal{S}, \tag{2}$$

which, by introducing a vector of binary variables $y \in \mathbb{R}^s$ can be written as

$$\begin{array}{ll} z_i \ge -M(1-y_i) + (a_i + upl_i)y_i & i \in \mathcal{S},\\ z_i \le My_i + (a_i - lpl_i)(1-y_i) & i \in \mathcal{S},\\ y_i \in \{0,1\} & i \in \mathcal{S}, \end{array} \tag{3}$$

$0 \ll M \in \mathbb{R}$ being a large positive value. Constraints (3) impose either "upper protection sense" $z_i \ge a_i + upl_i$, when $y_i = 1$, or "lower protection sense" $z_i \le a_i - lpl_i$ when $y_i = 0$. The CTA problem (1a)–(1c), (3) is a (in general difficult) MILP, but it provides solutions with a high data utility [11].

Formulating problem (1) in terms of cell deviations $x = z - a$, $x \in \mathbb{R}^n$, and fixing the binary variables, the resulting continuous CTA problem can be formulated as the general convex optimization problem

$$\begin{aligned} \min_{x} \quad & ||x||_{\ell(w)}\\ \text{s. to} \quad & Ax = 0\\ & l \le x \le u, \end{aligned} \tag{4}$$

where

$$\begin{aligned} l_i &= \begin{cases} upl_i & \text{if } i \in \mathcal{S} \text{ and } y_i = 1\\ l_{a_i} - a_i & \text{if } (i \in \mathcal{N} \setminus \mathcal{S}) \text{ or } (i \in \mathcal{S} \text{ and } y_i = 0) \end{cases}\\ u_i &= \begin{cases} -lpl_i & \text{if } i \in \mathcal{S} \text{ and } y_i = 0\\ u_{a_i} - a_i & \text{if } (i \in \mathcal{N} \setminus \mathcal{S}) \text{ or } (i \in \mathcal{S} \text{ and } y_i = 1), \end{cases} \end{aligned} \tag{5}$$

for $i \in \mathcal{N}$.

Problem (4) can be specialized for several norms, ℓ_1 and ℓ_2 being the two most relevant. For ℓ_1, defining $x = x^+ - x^-$, we obtain the following LP:

$$\begin{aligned} \min_{x^+, x^-} \quad & \sum_{i=1}^{n} w_i(x_i^+ + x_i^-)\\ \text{s. to} \quad & A(x^+ - x^-) = 0\\ & l^+ \le x^+ \le u^+\\ & l^- \le x^- \le u^-, \end{aligned} \tag{6}$$

$x^+ \in \mathbb{R}^n$ and $x^- \in \mathbb{R}^n$ being the vectors of positive and negative deviations in absolute value, and $l^+, l^-, u^+, u^- \in \mathbb{R}^n$ lower and upper bounds for the positive and negative deviations defined as

$$\begin{aligned}
l_i^+ &= \begin{cases} upl_i & \text{if } i \in \mathcal{S} \text{ and } y_i = 1 \\ 0 & \text{if } (i \in \mathcal{N} \setminus \mathcal{S}) \text{ or } (i \in \mathcal{S} \text{ and } y_i = 0) \end{cases} \\
u_i^+ &= \begin{cases} 0 & \text{if } i \in \mathcal{S} \text{ and } y_i = 0 \\ u_{a_i} - a_i & \text{if } (i \in \mathcal{N} \setminus \mathcal{S}) \text{ or } (i \in \mathcal{S} \text{ and } y_i = 1) \end{cases} \\
l_i^- &= \begin{cases} lpl_i & \text{if } i \in \mathcal{S} \text{ and } y_i = 0 \\ 0 & \text{if } (i \in \mathcal{N} \setminus \mathcal{S}) \text{ or } (i \in \mathcal{S} \text{ and } y_i = 1) \end{cases} \\
u_i^- &= \begin{cases} 0 & \text{if } i \in \mathcal{S} \text{ and } y_i = 1 \\ a_i - l_{a_i} & \text{if } (i \in \mathcal{N} \setminus \mathcal{S}) \text{ or } (i \in \mathcal{S} \text{ and } y_i = 0), \end{cases}
\end{aligned} \tag{7}$$

for $i \in \mathcal{N}$. For ℓ_2, problem (4) can be directly recast as the following QP without introducing additional variables:

$$\begin{aligned}
\min_x \ & \sum_{i=1}^{n} w_i x_i^2 \\
\text{s. to } & Ax = 0 \\
& l \le x \le u.
\end{aligned} \tag{8}$$

Infeasibilities in continuous models (6), (8) due to pre-fixing the binary variables can be dealt as in [12,13]. Problem (8) requires half the number of variables than (6). In addition, the splitting of variables $x = x^+ - x^-$ may create difficulties to some optimization methods. On the other hand the ℓ_1 solutions are known to change fewer cells than ℓ_2 solutions. The next Section introduces a new nonlinear CTA model with the same number of variables that (8) and similar solutions to those of (6).

3 Using a Pseudo-Huber Function as Objective Function

The Huber function [16] $\varphi_\delta : \mathbb{R} \to \mathbb{R}$, defined as

$$\varphi_\delta(x_i) = \begin{cases} \frac{x_i^2}{2\delta} & |x_i| \le \delta \\ |x_i| - \frac{\delta}{2} & |x_i| \ge \delta \end{cases} \tag{9}$$

approximates $|x_i|$ for small values of $\delta > 0$ (the smaller δ the better the approximation). φ_δ is a continuous and first-order differentiable function; but second derivatives are not continuous at points $|x_i| = \delta$.

To avoid this discontinuity in second derivatives, we may consider the pseudo-Huber function $\phi_\delta : \mathbb{R} \to \mathbb{R}$:

$$\phi_\delta(x_i) = \sqrt{\delta^2 + x_i^2} - \delta. \tag{10}$$

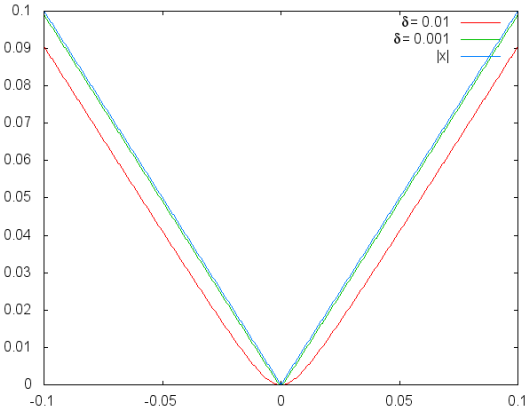

Fig. 1. Pseudo-Huber function for some δ, and $|x|$

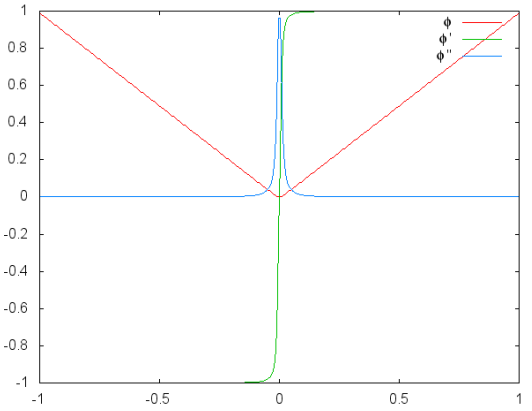

Fig. 2. Graph of ϕ_δ, ϕ'_δ and ϕ''_δ for $\delta = 0.01$

This function has been recently successfully used in other ℓ_1-regularization problems [15]. $\phi_\delta \in \mathcal{C}^2$, with first and second derivatives

$$\phi'_\delta(x_i) = \frac{x_i}{\sqrt{\delta^2 + x_i^2}} \qquad \phi''_\delta(x_i) = \frac{\delta^2}{(\delta^2 + x_i^2)^{3/2}}. \tag{11}$$

As shown in Figure 1, ϕ_δ is a better approximation of $|x_i|$ as δ approaches 0. Figure 2 plots the graph of ϕ_δ, ϕ'_δ and ϕ''_δ for $\delta = 0.01$. As shown in [15], the first and second derivatives are bounded and Lipschitz continuous.

a

10$_{(3)}$	15	11	9	45
8	10	12	15	45
10	12	11	**13**$_{(5)}$	46
28	37	34	37	136

(a)

ℓ_1

13	15	11	6	45
10	10	12	13	45
5	12	11	18	46
28	37	34	37	136

(b)

ϕ

13.88	15.17	11.18	4.77	45
8.21	10.30	12.27	14.22	45
4.91	11.53	10.55	18	46
28	37	34	37	136

(c)

ℓ_2

13	15.03	11.03	5.94	45
7.66	11.14	13.14	13.06	45
7.34	10.83	9.83	18	46
28	37	34	37	136

(d)

Fig. 3. Results with ℓ_1, $\phi_{0.001}$, and ℓ_2 (tables (b), (c) and (d)) for the small two-dimensional small table (a) (rounded to two decimal positions). The optimal value of $||x||_1$ for ℓ_1 and $\phi_{0.001}$ is 20, while for ℓ_2 is 20.69.

Therefore we can replace $||x||_{\ell_1}$ by $f(x) = \sum_{i=1}^{n} \phi_\delta(x_i)$, and the ℓ_1-CTA problem (6) can be approximately solved by the convex optimization problem

$$\begin{array}{ll} \min\limits_{x} & f(x) = \sum_{i=1}^{n} \phi_\delta(x_i) \\ \text{s. to} & Ax = 0 \\ & l \leq x \leq u. \end{array} \tag{12}$$

This optimization problem has the same space of variables and feasible region than (8), but with a strictly convex nonlinear function instead of a quadratic one.

Figure 3 shows the solutions obtained with ℓ_1, ℓ_2 and $\phi_{\delta=0.001}$ with a small two-dimensional table. In this small table both ϕ and ℓ_2 changed most of the cells, whereas ℓ_1 only changed a few of them. However, the optimal objective functions with ℓ_1 and $\phi_{\delta=0.001}$ were exactly the same ($||x||_1 = 20$), whereas $||x||_1 = 20.69$ for ℓ_2. The ϕ_δ function thus provided the same objective function that ℓ_1, but cell deviations were distributed among more cells. This is explained by the different optimization algorithms used for the solution of ℓ_1 and ϕ_δ (which needs a nonlinear optimization method, as discussed in next Section). A more extensive study with larger and more complex tables is out of the scope of this work.

4 Solution of the Huber-CTA Model

The Huber-CTA model (12) is a nonlinear convex optimization problem. In theory, this kind of problems are polynomially solved with interior-point methods [18,19], with a best bound of $O(\sqrt{n} \log 1/\epsilon)$, n being the number of variables and ϵ the optimality tolerance (discussed below). The complexity of CTA with ℓ_1 or ϕ_δ is thus the same if solved by an interior-point algorithm.

Broadly speaking, interior-point methods attempt to solve a perturbation of the first-order optimality conditions (named Karush-Kuhn-Tucker or KKT conditions) of (12):

$$\begin{aligned} Ax &= b \\ A^\top \lambda + \lambda_l - \lambda_u - \nabla f(x) &= 0 \\ (X - L)\Lambda_l e &= \mu e \\ (U - X)\Lambda_u e &= \mu e \\ u \geq x \geq l, \quad (\lambda_l, \lambda_u) &\geq 0, \end{aligned} \tag{13}$$

where $\lambda \in \mathbb{R}^m$, $\lambda_l, \lambda_u \in \mathbb{R}^n$ are the Lagrange multipliers of respectively the equality constraints and lower and upper bounds, $e \in \mathbb{R}^n$ is a vector of 1's, and matrices $X, \Lambda_l, \Lambda_u, L, U \in \mathbb{R}^{n \times n}$ are diagonal matrices made from vectors $x, \lambda_l, \lambda_u, l, u$. The set of unique solutions of (13) for each μ value is known as the central path, and when $\mu \to 0$ these solutions converge to those of (12). The nonlinear system (13) is usually solved by a damped version of Newton's method, reducing the μ parameter at each iteration, until $\mu \leq \epsilon$, ϵ being the required optimality tolerance. This procedure is known as the path-following interior-point algorithm. An excellent discussion about the theoretical and practical properties of this interior-point algorithm can be found in [20].

Although theoretically the same interior-point path-following algorithm should be as efficient for ℓ_1 than for ϕ_δ, in practice the Huber function requires a more robust solver. Some early tests with general tables using the convex interior-point algorithm of [2] show that even small instances can be difficult with ϕ_δ if the solver is not appropriately tuned. In this sense, reformulations of the model as a second order conic optimization problem could be preferable [1].

However, for some interior-point methods specialized to particular structures, such as block-angular problems, ϕ_δ may be more efficiently solved than ℓ_1: the technical explanation is that, since the Hessian of ϕ is nonzero, unlike for the LP formulated by ℓ_1, the internal linear systems of equations may require less iterations of the preconditioned conjugate gradient [7]. For instance, this may happen for three-dimensional tables, whose constraints exhibit a block-angular structure [8]. Next Section shows a few preliminary computational results with some three-dimensional tables using such a specialized interior-point solver.

5 Computational Results

Preliminary results have been obtained for a set of eight three-dimensional tables of r rows, c columns and l levels (where rows, columns and levels refer to each of the table dimensions). Table 1 shows the problem dimensions for each instance; n and m denote the number of variables and constraints of problems (12) and (8) for ϕ and ℓ_2 (which are of the same size), and (6) for ℓ_1. Tables were obtained with the same generator used in [8].

These eight instances have been solved with an efficient implementation of the specialized interior-point method described in [4] including the quadratic

Table 1. Dimensions of some 3D CTA optimization problems for pseudo-Huber, ℓ_1 and ℓ_2

			ϕ, ℓ_2		ℓ_1	
r	c	l	n	m	n	m
25	25	25	16250	1875	31875	1875
25	25	50	31875	3125	63125	3125
25	50	25	32500	3125	63750	3125
25	50	50	63750	5000	126250	5000
50	25	25	32500	3125	63750	3125
50	25	50	63750	5000	126250	5000
50	50	25	65000	5000	127500	5000
50	50	50	127500	7500	252500	7500

Table 2. Results for 3D CTA using pseudo-Huber, ℓ_1 and ℓ_2

			ϕ		ℓ_1		ℓ_2	
r	c	l	obj.	CPU	obj.	CPU	obj.	CPU
25	25	25	101096	1.68	101572	4.37	4161290	0.09
25	25	50	104706	4.59	105409	10.94	3915100	0.19
25	50	25	104030	4.54	104720	13.87	3969550	0.27
25	50	50	110537	8.71	111679	9.72	3915150	0.55
50	25	25	107138	4.87	107832	23.9	4107990	0.26
50	25	50	109068	7.67	110199	5.54	3832800	0.54
50	50	25	106173	8.17	107309	4.15	3666090	0.9
50	50	50	113858	15.68	116279	67.91	3678810	1.84

regularization strategy of [7]. Table 2 reports for each of the three CTA variants—using ϕ, ℓ_1 and ℓ_2—, the optimal objective function achieved and the CPU time. All runs were carried out on a Fujitsu Primergy RX300 server with 3.33 GHz Intel Xeon X5680 CPUs, under a GNU/Linux operating system (Suse 11.4), without exploitation of parallelism capabilities. It is clearly seen that ℓ_2 provides the fastest executions; this is consistent with the results of [3] obtained with a generic solver. However, the objective function with ℓ_2 naturally differs from that obtained with ℓ_1. On the other hand both ϕ and ℓ_1 provide very similar objective function values, ϕ being more efficiently solved in six of the eight instances. In particular the largest instance required 67.91 seconds with ℓ_1 and only 15.68 with ϕ.

6 Conclusions

We have presented a CTA model which replaces the usual ℓ_1 distance in the objective by the pseudo-Huber function. Although the resulting problem is convex

and nonlinear, it requires half the number of variables than the ℓ_1-CTA LP. It has been observed that for certain classes of tables (i.e., some three-dimensional tables) the Huber-CTA model can be more efficiently solved than ℓ_1-CTA using and appropriate interior-point solver.

The preliminary results reported in this work are non-conclusive, but just a first step in the solution of the Huber-CTA model. Among the future tasks to be done in this direction we mention: (i) the application of the Huber-CTA model to other classes of structured tables (real-world linked or hierarchical tables); (ii) a more detailed analysis of the disclosure risk and data utility of tables protected by the Huber function, comparing them with tables protected with ℓ_1- and ℓ_2-CTA; (iii) an efficient implementation for general tables, not just three-dimensional ones; (iv) and the tuning or implementation of second-order interior-point solvers for the highly efficient solution of the Huber-CTA problem.

References

1. Andersen, E.D.: Personal communication (2014)
2. Andersen, E.D., Andersen, K.D.: The MOSEK interior point optimizer for linear programming: an implementation of the homogeneous algorithm. In: Frenk, H., Roos, K., Terlaky, T., Zhang, S. (eds.) High Performance Optimization, pp. 197–232. Kluwer (2000)
3. Castro, J.: Minimum-distance controlled perturbation methods for large-scale tabular data protection. European Journal of Operational Research 171, 39–52 (2006)
4. Castro, J.: An interior-point approach for primal block-angular problems. Computational Optimization and Applications 36, 195–219 (2007)
5. Castro, J.: Recent advances in optimization techniques for statistical tabular data protection. European Journal of Operational Research 216, 257–269 (2012)
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., Cuesta, J.: Quadratic regularizations in an interior-point method for primal block-angular problems. Mathematical Programming 130, 415–445 (2011)
8. Castro, J., Cuesta, J.: Solving L_1-CTA in 3D tables by an interior-point method for primal block-angular problems. TOP 21, 25–47 (2013)
9. Castro, J., Frangioni, A., Gentile, C.: Perspective reformulations of the CTA problem with L_2 distances. Operations Research (in press, 2014)
10. Castro, J., Giessing, S.: Testing variants of minimum distance controlled tabular adjustment. Monographs of Official Statistics, pp. 333–343. Eurostat-Office for Official Publications of the European Communities, Luxembourg (2006)
11. Castro, J., González, J.A.: Assessing the information loss of controlled adjustment methods in two-way tables. Lecture Notes in Computer Science: Privacy in Statistical Databases (2014) (Submitted)
12. Castro, J., González, J.A.: A fast CTA method without the complicating binary decisions. Documents of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Statistics Canada, Ottawa, pp. 1–7 (2013)
13. Castro, J., González, J.A.: A multiobjective LP approach for controlled tabular adjustment in statistical disclosure control. Working paper, Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya (2014)

14. Dandekar, R.A., Cox, L.H.: Synthetic tabular Data: an alternative to complementary cell suppression, manuscript, Energy Information Administration, U.S. (2002)
15. Fountoulakis, K., 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)
16. Huber, P.J.: Robust estimation of a location parameter. Annals of Statistics 53, 73–101 (1964)
17. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte-Nordholt, E., Spicer, K., de Wolf, P.P.: Statistical Disclosure Control. Wiley, Chichester (2012)
18. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer, Boston (2004)
19. Nesterov, Y., Nemirovskii, A.: Interior-Point Polynomial Algorithms in Convex Programming. SIAM, Philadelphia (1994)
20. Wright, S.J.: Primal-Dual Interior-Point Methods. SIAM, Philadelphia (1996)