Consolider Computing

Optimization Techniques for Statistical Data Protection

Interior-point methods for large-scale optimization. Application to statistical data protection.

GNOM

Group of Numerical Optimization and Modelling

http://www-eio.upc.es/research/gnom

Jordi Castro

http://www-eio.upc.es/~jcastro

Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya
Barcelona

Main research areas

- Operations Research/Optimization
  - Interior-point methods for large-scale linear and quadratic programming problems
  - Solution of large-scale non-linear optimization problems
  - Solution of large-scale structured problems, in particular stochastic optimization problems, network flows problems...
  - Efficient implementation of algorithms
- Applications:
  - Statistical tabular data protection: real problem of great interest for National Statistical Institutes (NSIs)

Tools (and problems)

- Direct methods for the solution of large and sparse linear positive and semidefinite positive systems and indefinite systems of IP methods.

Tools (and problems)

- Direct methods for the solution of large and sparse linear positive and semidefinite positive systems and indefinite systems of IP methods.
  - Which is the best package?

Tools (and problems)

- Direct methods for the solution of large and sparse linear positive and semidefinite positive systems and indefinite systems of IP methods.
  - Which is the best package?
- For very large problems, iterative methods (e.g., preconditioned conjugate gradients) for the solution of definite and quasidefinite systems of IP methods.

Tools (and problems)

- Direct methods for the solution of large and sparse linear positive and semidefinite positive systems and indefinite systems of IP methods.
  - Which is the best package?
- For very large problems, iterative methods (e.g., preconditioned conjugate gradients) for the solution of definite and quasidefinite systems of IP methods.
  - Development of efficient preconditioners

Tools (and problems)

- Direct methods for the solution of large and sparse linear positive and semidefinite positive systems and indefinite systems of IP methods.
  - Which is the best package?
- For very large problems, iterative methods (e.g., preconditioned conjugate gradients) for the solution of definite and quasidefinite systems of IP methods.
  - Development of efficient preconditioners
- Combinatorial optimization tools for the optimal solution of some statistical tabular data protection techniques.

Tools (and problems)

- Direct methods for the solution of large and sparse linear positive and semidefinite positive systems and indefinite systems of IP methods.
  - Which is the best package?
- For very large problems, iterative methods (e.g., preconditioned conjugate gradients) for the solution of definite and quasidefinite systems of IP methods.
  - Development of efficient preconditioners
- Combinatorial optimization tools for the optimal solution of some statistical tabular data protection techniques.
  - Optimal CTA (described later) is an open problem

Tools (and problems)

- Direct methods for the solution of large and sparse linear positive and semidefinite positive systems and indefinite systems of IP methods.
  - Which is the best package?
- For very large problems, iterative methods (e.g., preconditioned conjugate gradients) for the solution of definite and quasidefinite systems of IP methods.
  - Development of efficient preconditioners
- Combinatorial optimization tools for the optimal solution of some statistical tabular data protection techniques.
  - Optimal CTA (described later) is an open problem
- Solution of the LP relaxations of CTA

Tools (and problems)

- Direct methods for the solution of large and sparse linear positive and semidefinite positive systems and indefinite systems of IP methods.
  - Which is the best package?
- For very large problems, iterative methods (e.g., preconditioned conjugate gradients) for the solution of definite and quasidefinite systems of IP methods.
  - Development of efficient preconditioners
- Combinatorial optimization tools for the optimal solution of some statistical tabular data protection techniques.
  - Optimal CTA (described later) is an open problem
- Solution of the LP relaxations of CTA
  - IP methods are far more efficient than simplex. Developing efficient IP methods.

Disclosure in tabular data: External attacker

Table of average salary by ZIP code and Age

	Z_1	Z_2	Z_3	TOTAL
E_1	20	24	28	72
E_2	38	38	40	116
E_3	40	39	42	121
TOTAL	98	101	110	309

Table of individuals by ZIP code and Age

	Z_1	Z_2	Z_3	TOTAL
E_1	20	15	30	65
E_2	15	20	1	36
E_3	8	9	8	25
TOTAL	43	44	39	126

Disclosure in tabular data: Internal attacker

Table of average salary by ZIP code and Age

	Z_1	Z_2	Z_3	TOTAL
E_1	20	24	28	72
E_2	38	38	40	116
E_3	40	39	42	121
TOTAL	98	101	110	309

Table of individuals by ZIP code and Age

	Z_1	Z_2	Z_3	TOTAL
E_1	20	15	30	65
E_2	15	20	2	37
E_3	8	9	8	25
TOTAL	43	44	40	127

Modelling tables

- Set of cells $a_i, i = 1, \ldots, n$, that satisfy $Aa = b$.
- Usually positive tables: $a \geq 0$.
- Real tables:
  - any structure (A)
  - n is large. E.g.: $n = 800$ millions cells for bussiness data of Germany

Current methods used by NSIs

- Cell Suppression Problem
- Minimum-distance Controlled Tabular Adjustment

Cell Suppression Problem

ORIGINAL TABLE

	Z_1	Z_2	Z_3	TOTAL
E_1	20	24	28	72
E_2	38	38	40	116
E_3	40	39	42	121
TOTAL	98	101	110	309

Cell Suppression Problem

ORIGINAL TABLE

	Z_1	Z_2	Z_3	TOTAL
E_1	20	24	28	72
E_2	38	38	40	116
E_3	40	39	42	121
TOTAL	98	101	110	309

PROTECTED TABLE

	Z_1	Z_2	Z_3	TOTAL
E_1	20	24	28	72
E_2	38	38		116
E_3	40	39	42	121
TOTAL	98	101	110	309

Cell Suppression Problem

ORIGINAL TABLE

	Z_1	Z_2	Z_3	TOTAL
E_1	20	24	28	72
E_2	38	38	40	116
E_3	40	39	42	121
TOTAL	98	101	110	309

PROTECTED TABLE

	Z_1	Z_2	Z_3	TOTAL
E_1		24		72
E_2		38		116
E_3	40	39	42	121
TOTAL	98	101	110	309

Algorithms for Cell Suppression Problem

- The MILP formulation is very large:

Algorithms for Cell Suppression Problem

- The MILP formulation is very large:
  - Table of 8000 cells, 800 sensitive cells, and 4000 linear relations: MILP of 8000 binary variables, 12,800,000 continuous variables, and 32,000,000 constraints.

Algorithms for Cell Suppression Problem

- The MILP formulation is very large:
  - Table of 8000 cells, 800 sensitive cells, and 4000 linear relations: MILP of 8000 binary variables, 12,800,000 continuous variables, and 32,000,000 constraints.
- Exact algorithms for general tables
- Heuristic algorithms for some structured tables

Minimum Distance Controlled Tabular Adjustment

ORIGINAL TABLE

	Z_1	Z_2	Z_3	TOTAL
E_1	20	24	28	72
E_2	38	38	40	116
E_3	40	39	42	121
TOTAL	98	101	110	309

Minimum Distance Controlled Tabular Adjustment

ORIGINAL TABLE

	Z_1	Z_2	Z_3	TOTAL
E_1	20	24	28	72
E_2	38	38	40	116
E_3	40	39	42	121
TOTAL	98	101	110	309

We want: $40 \geq 45$ or $40 \leq 35$, for instance

Minimum Distance Controlled Tabular Adjustment

ORIGINAL TABLE

	Z_1	Z_2	Z_3	TOTAL
E_1	20	24	28	72
E_2	38	38	40	116
E_3	40	39	42	121
TOTAL	98	101	110	309

We want: $40 \geq 45$ or $40 \leq 35$, for instance

PROTECTED TABLE

	Z_1	Z_2	Z_3	TOTAL
E_1	25	24	23	72
E_2	33	38	45	116
E_3	40	39	42	121
TOTAL	98	101	110	309

Minimum Distance Controlled Tabular Adjustment

ORIGINAL TABLE

	Z_1	Z_2	Z_3	TOTAL
E_1	20	24	28	72
E_2	38	38	40	116
E_3	40	39	42	121
TOTAL	98	101	110	309

We want: $40 \geq 45$ or $40 \leq 35$, for instance

PROTECTED TABLE

	Z_1	Z_2	Z_3	TOTAL
E_1	15	24	33	72
E_2	43	38	35	116
E_3	40	39	42	121
TOTAL	98	101	110	309

Algorithms for Minimum Distance CTA

- CTA is a recent method

Algorithms for Minimum Distance CTA

- CTA is a recent method
- CTA is being considered by European NSIs

Algorithms for Minimum Distance CTA

- CTA is a recent method
- CTA is being considered by European NSIs
- MILP problem
- Approximate solutions using LP subproblems

Algorithms for Minimum Distance CTA

- CTA is a recent method
- CTA is being considered by European NSIs
- MILP problem
- Approximate solutions using LP subproblems
- Specialized IP algorithms useful:
  - Example of $100 \times 100 \times 50$ table (500K cells)
    - ★ CPLEX: 900 seconds
    - ★ Specialized IP algorithm: 7 seconds

Software used by European NSIs

- TAU-ARGUS: http://neon.vb.cbs.nl/casc/tau.html
- Developed within CASC European Union Project
- D.EIO-UPC has contributed to TAU-ARGUS with a heuristic for Cell Suppression
- CTA being developed within a national project
- CTA has still to be added to TAU-ARGUS

Research opportunities - Actions

- Exact algorithms for MILP CTA formulation
- Heuristic algorithms for MILP CTA formulation
- Solution of large scale LP relaxations
- Models for correlated tables
- And eventually, software for NSI's reals problems

THANKS

Jordi Castro

http://www-eio.upc.es/~jcastro