Convex Optimization Methods for Dimension Reduction and Coefficient Estimation in Multivariate Linear Regression

Renato D.C. Monteiro
Georgia Tech

FODAVA REVIEW MEETING

Atlanta, USA

Dec. 3, 2009

Motivation

Compressed sensing: Candes et al. (2006) have shown that a sparse signal can be recovered by solving a non-smooth optimization problem of the form

$$\min\{\|\mathbf{x}\|_0 : \mathbf{A}\mathbf{x} = \mathbf{b}\} \qquad (*)$$

where $\mathbf{A}$ is $\mathbf{l} \times \mathbf{p}$-matrix, $\mathbf{x} \in \Re^{\mathbf{p}}$ and $\mathbf{b} \in \Re^{\mathbf{l}}$. Here $\|\mathbf{x}\|_0$ denotes the number of nonzero components of $\mathbf{x}$.

Under some conditions on $\mathbf{A}$, they have shown that $(*)$ is also equivalent to the convex program

$$\min\{\|\mathbf{x}\|_1 : \mathbf{A}\mathbf{x} = \mathbf{b}\} \qquad (**)$$

where $\|\mathbf{x}\|_1 := \sum_{\mathbf{i}=1}^{\mathbf{p}} |\mathbf{x}_\mathbf{i}|$ is the $\mathbf{1}$-norm of $\mathbf{x}$.

There has been a lot of research in the optimization community to develop methods that can solve $(**)$ efficiently.

We are interested in the extensions of these problems, where the variable is now a $\mathbf{p} \times \mathbf{q}$ matrix.

Consider the problem

$$\min_{\mathbf{X}\in\Re^{\mathbf{p}\times\mathbf{q}}} \{\mathbf{rank}(\mathbf{X}) : \mathcal{A}(\mathbf{X}) = \mathbf{b}\}, \qquad (*)$$

where $\mathcal{A} : \Re^{\mathbf{p}\times\mathbf{q}} \to \Re^{\mathbf{l}}$ **is a linear map and** $\mathbf{b} \in \Re^{\mathbf{l}}$**.**

This problem (and its variations) has many applications (e.g., matrix completion problems, netflix problem, dimension reduction in statistics and etc.)

When $\mathbf{X}$ **is restricted to a diagonal (square) matrix, i.e.,** $\mathbf{X} = \mathbf{Diag}(\mathbf{x})$, $\mathbf{x} \in \Re^{\mathbf{p}}$, **then** $(*)$ **reduces to**

$$\min\{\|\mathbf{x}\|_{\mathbf{0}} : \mathbf{A}\mathbf{x} = \mathbf{b}\}$$

A convex approximation of $(*)$ **is**

$$\min_{\mathbf{X}\in\Re^{\mathbf{p}\times\mathbf{q}}} \{\|\mathbf{X}\|_* : \mathcal{A}(\mathbf{X}) = \mathbf{b}\} \qquad (**)$$

where $\|\mathbf{X}\|_*$ **denotes the nuclear norm of** $\mathbf{X}$**:**

$$\|\mathbf{X}\|_* := \mathrm{Trace}[(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{\mathbf{1/2}}] = \sum_{\mathbf{i}=\mathbf{1}}^{\min\{\mathbf{p},\mathbf{q}\}} \sigma_{\mathbf{i}}(\mathbf{X})$$

Under suitable conditions on $\mathcal{A}$**, Recht et al. (2007) have shown that** $(*)$ **and** $(**)$ **are equivalent.**

A relaxation of $(**)$ is the problem

$$\min_{\mathbf{X}\in\Re^{\mathbf{p}\times\mathbf{q}}} \frac{\mathbf{1}}{\mathbf{2}}\|\mathcal{A}(\mathbf{X})-\mathbf{b}\|_{\mathbf{F}}^{\mathbf{2}}+\lambda\|\mathbf{X}\|_{*}$$

where $\|\mathbf{X}\|_{\mathbf{F}}^{\mathbf{2}} := \sum_{\mathbf{i},\mathbf{j}} \mathbf{X}_{\mathbf{ij}}^{\mathbf{2}}$.

Our problem of interest is the following special case of the above problem:

$$\min_{\mathbf{X}\in\Re^{\mathbf{p}\times\mathbf{q}}} \frac{\mathbf{1}}{\mathbf{2}}\|\mathbf{AX}-\mathbf{B}\|_{\mathbf{F}}^{\mathbf{2}}+\lambda\|\mathbf{X}\|_{*}$$

where $\mathbf{A}\in\Re^{\mathbf{n}\times\mathbf{p}}$, $\mathbf{B}\in\Re^{\mathbf{n}\times\mathbf{q}}$ and $\mathbf{A}$ has l.i. columns ($\mathbf{n}>>\mathbf{p}$).

This problem arises in statistics in the context of dimension reduction and coefficient estimate in multivariate linear regression.

Dimension Reduction in Statistics

Assume that $\mathbf{A} \in \Re^{\mathbf{n}\times\mathbf{p}}$ consists of $\mathbf{n}$ observations on $\mathbf{p}$ explanatory variables $\mathbf{a} = (\mathbf{a_1}, \ldots, \mathbf{a_p})'$ and $\mathbf{B} \in \Re^{\mathbf{n}\times\mathbf{q}}$ collects the corresponding $\mathbf{n}$ observations on $\mathbf{q}$ responses $\mathbf{b} = (\mathbf{b_1}, \ldots, \mathbf{b_q})'$. Consider the multivariate linear model

$$\mathbf{B} = \mathbf{A}\mathbf{X} + \mathbf{E},$$

where $\mathbf{X} \in \Re^{\mathbf{p}\times\mathbf{q}}$ is a coefficient matrix, $\mathbf{E} = (\mathbf{e^1}, \ldots, \mathbf{e^n})'$ is the regression noise, and all $\mathbf{e^i}$s are independent samples of $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$.

To estimate $\mathbf{X}$ and accomplish dimension reduction, Yuan et al. proposed to solve

$$\min_{\mathbf{X}} \frac{\mathbf{1}}{\mathbf{2}}\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_{\mathbf{F}}^{\mathbf{2}} + \lambda\|\mathbf{X}\|_* \tag{1}$$

for different $\lambda > \mathbf{0}$ values. The larger the scalar $\lambda > \mathbf{0}$, the more dimension reduction is accomplished.

Reformulations:

- **cone program (includes LP, SDP)**
- **saddle point (min-max convex-concave) problems**

Cone program (CP): **Given a closed convex cone $\mathcal{K} \subseteq \Re^{\mathbf{n}}$, the CP problem is:**

$$\min\{\langle \mathbf{c}, \mathbf{x} \rangle : \mathcal{A}(\mathbf{x}) = \mathbf{b},\ \mathbf{x} \in \mathcal{K}\}$$

where $\mathbf{c} \in \Re^{\mathbf{n}}$, $\mathbf{b} \in \Re^{\mathbf{m}}$ and $\mathcal{A} : \Re^{\mathbf{n}} \to \Re^{\mathbf{m}}$ is a linear map. Its dual is

$$\max\{\langle \mathbf{b}, \mathbf{y} \rangle : \mathbf{c} - \mathcal{A}^*(\mathbf{y}) \in \mathcal{K}^*\}$$

where $\mathcal{K}^* := \{\mathbf{s} \in \Re^{\mathbf{n}} : \langle \mathbf{s}, \mathbf{x} \rangle \geq \mathbf{0},\ \forall \mathbf{x} \in \mathcal{K}\}$.

Remark: **Can be solved by interior-point (second-order) methods or by first-order methods.**

Saddle point or min-max problems

Their general form is

$$\min_{\mathbf{x}\in\mathbf{X}} \left(\mathbf{f}(\mathbf{x}) := \max_{\mathbf{y}\in\mathbf{Y}} \phi(\mathbf{x},\mathbf{y}) \right)$$

where $\mathbf{X}, \mathbf{Y}$ are simple closed convex sets, ϕ is convex in $\mathbf{x}$ and concave in $\mathbf{y}$.

Under the assumption that $\nabla\phi$ is Lipschitz continuous, first-order methods with known iteration-complexity bounds have been developed to solve these problems:

- **Nesterov's smooth or non-smooth methods and their variants;**
- **Korpelevich algorithm or Nemirovski's prox-mirror method**

Cone programming reformulation

Problem (1) can be reformulated as a CP problem as follows. Clearly, (1) is equivalent to

$$\min_{\mathbf{X},\mathbf{t}} \left\{ \frac{1}{2}\|\mathbf{AX}-\mathbf{B}\|^2 + \lambda \mathbf{t} : \|\mathbf{X}\|_* \leq \mathbf{t} \right\}$$

Write $\mathbf{V} \succeq \mathbf{0}$ if $\mathbf{V}$ is symmetric and positive semidefinite. Also, let $\mathcal{S}^{\mathbf{l}}$ denote the space of $\mathbf{l} \times \mathbf{l}$ symm. matrices.

Proposition: Let $\mathbf{X} \in \Re^{\mathbf{p}\times\mathbf{q}}$ and set $\mathbf{k} := \min\{\mathbf{p},\mathbf{q}\}$ and $\mathbf{l} := \mathbf{p}+\mathbf{q}$. For $\mathbf{t} \in \Re$, we have

$$\|\mathbf{X}\|_* \leq \mathbf{t} \;\Leftrightarrow\; \begin{cases} \mathbf{t} - \mathbf{ks} - \mathrm{Trace}(\mathbf{V}) & \geq \mathbf{0}, \\ \mathbf{V} - \mathcal{G}(\mathbf{X}) + \mathbf{sI} & \succeq \mathbf{0}, \\ \mathbf{V} & \succeq \mathbf{0}, \end{cases}$$

for some $\mathbf{V} \in \mathcal{S}^{\mathbf{l}}$ and $\mathbf{s} \in \Re$, where $\mathcal{G} : \Re^{\mathbf{p}\times\mathbf{q}} \to \Re^{\mathbf{l}\times\mathbf{l}}$ is defined as

$$\mathcal{G}(\mathbf{X}) := \begin{pmatrix} \mathbf{0} & \mathbf{X}^{\mathbf{T}} \\ \mathbf{X} & \mathbf{0} \end{pmatrix}$$

SADDLE-POINT REFORMULATIONS

Using the identity

$$\|\mathbf{X}\|_* = \mathbf{k} \max_{\mathbf{W} \in \mathbf{\Omega}} \langle \mathcal{G}(\mathbf{X}), \mathbf{W} \rangle$$

where $\mathbf{k} := \min\{\mathbf{p}, \mathbf{q}\}$ **and**

$$\mathbf{\Omega} := \{\mathbf{W} \in \mathcal{S}^{\mathbf{p}+\mathbf{q}} : \mathbf{0} \preceq \mathbf{W} \preceq \mathbf{I}/\mathbf{k}, \ \mathrm{Trace}(\mathbf{W}) = \mathbf{1}\}$$

problem (1) can be reformulated as

$$\min_{\|\mathbf{X}\|_\mathbf{F} \leq \mathbf{r}} \mathbf{f_p}(\mathbf{X}) := \max_{\mathbf{W} \in \mathbf{\Omega}} \left\{ \frac{\mathbf{1}}{\mathbf{2}} \|\mathbf{AX} - \mathbf{B}\|_\mathbf{F}^\mathbf{2} + \lambda \mathbf{k} \langle \mathcal{G}(\mathbf{X}), \mathbf{W} \rangle \right\},$$

where $\mathbf{r}$ **is an appropriate scalar. (Disadvantage:** $\mathbf{f_p}(\cdot)$ **is non-smooth)**

Instead, we consider the dual of the above problem, namely:

$$\max_{\mathbf{W} \in \mathbf{\Omega}} \mathbf{f_d}(\mathbf{W}) := \min_{\|\mathbf{X}\|_\mathbf{F} \leq \mathbf{r}} \left\{ \frac{\mathbf{1}}{\mathbf{2}} \|\mathbf{AX} - \mathbf{B}\|_\mathbf{F}^\mathbf{2} + \lambda \mathbf{k} \langle \mathcal{G}(\mathbf{X}), \mathbf{W} \rangle \right\},$$

(Advantage: $\mathbf{f_d}(\mathbf{W})$ **has Lipschitz continuous gradient.)**

We then apply a variant of Nesterov's smooth method to solve the latter MAX-MIN reformulation of (1).

Proposition: Given $\epsilon > \mathbf{0}$, Nesterov's smooth method finds an ϵ-optimal solution of the MAX-MIN formulation in a number of iterations which does not exceed

$$\frac{\mathbf{2}\lambda\|(\mathbf{A}^{\mathbf{T}}\mathbf{A})^{-\mathbf{1}/\mathbf{2}}\|}{\sqrt{\epsilon}}\sqrt{\mathbf{k}\log\left(\frac{\mathbf{p}+\mathbf{q}}{\mathbf{k}}\right)},$$

where $\mathbf{k} := \min\{\mathbf{p}, \mathbf{q}\}$.

Note: The complexity of solving the corresponding dual MIN-MAX reformulation of (1) is $\mathcal{O}(\mathbf{1}/\epsilon)$ instead of $\mathcal{O}(\mathbf{1}/\sqrt{\epsilon})$ as above.

Computational Results

The entries of $\mathbf{A} \in \Re^{\mathbf{n} \times \mathbf{p}}$ and $\mathbf{B} \in \Re^{\mathbf{n} \times \mathbf{q}}$, with $\mathbf{p} = \mathbf{2q}$ and $\mathbf{n} = \mathbf{10q}$, were uniformly generated in $[\mathbf{0}, \mathbf{1}]$. The accuracy in the table below is $\epsilon = \mathbf{10^{-1}}$.

Problem	# of Iterations		CPU Time	
(p, q)	MIN-MAX	MAX-MIN	MIN-MAX	MAX-MIN
(200, 100)	610	1	29.60	0.91
(400, 200)	1310	1	432.92	8.36
(600, 300)	2061	1	2155.76	31.23
(800, 400)	2848	1	7831.09	76.75
(1000, 500)	3628	1	21128.70	156.68
(1200, 600)	4436	1	47356.32	276.64
(1400, 700)	5280	1	98573.73	456.61
(1600, 800)	6108	1	176557.49	699.47

Computational Results

The tables below compare the MAX-MIN formulation with the cone programming reformulation. The accuracy is $\epsilon = 10^{-8}$.

Problem (p, q)	# of Iterations MAX-MIN	# of Iterations CONE	CPU Time MAX-MIN	CPU Time CONE
(20,10)	3455	17	3.61	5.86
(40,20)	1696	15	6.90	77.25
(60,30)	1279	15	13.33	506.14
(80,40)	1183	15	25.34	2205.13
(100,50)	1073	19	40.66	8907.12
(120,60)	1017	N/A	62.90	N/A

Problem (p, q)	Memory MAX-MIN	Memory CONE
(20,10)	2.67	279
(40,20)	2.93	483
(60,30)	3.23	1338
(80,40)	3.63	4456
(100,50)	4.23	10445
(120,60)	4.98	> 16109

Summary

We have shown that a (smooth) first-order method applied to a MAX-MIN reformulation of (1) substantially outperforms a first-order method applied to the corresponding dual MIN-MAX reformulation.

We have also shown that it substantially outperforms an interior-point method applied to a CP reformulation of (1).