Problems in Sparse Multivariate Statistics with a Discrete Optimization Lens

Rahul Mazumder

Massachusetts Institute of Technology

(Joint with D. Bertsimas, M. Copenhaver, A. King, P. Radchenko, H. Qin, J. Goetz, K. Khamaru)

August, 2016

Motivation

- Several basic statistical estimation tasks are inherently <u>discrete</u>
- Often dismissed as computationally *infeasible*
- We often "relax" the hard problems:
  – Convex (continuous) optimization plays a **key** role (e.g. Lasso)
  – They work very well in many cases...

Motivation

- Several basic statistical estimation tasks are inherently <u>discrete</u>
- Often dismissed as computationally *infeasible*
- We often "relax" the hard problems:
  – Convex (continuous) optimization plays a **key** role (e.g. Lasso)
  – They work very well in many cases...
- However, often leads to a compromise in statistical performance
- Question: Can we use advances in discrete optimization to *globally solve* nonconvex problems?

Motivation

- We seldom know *a-priori* which method will work for a given application

- "*...A statistician's toolkit should have a whole array of methods, to experiment with...*"

  ...Jerome. H. Friedman

Motivation

- We seldom know *a-priori* which method will work for a given application
- "*...A statistician's toolkit should have a whole array of methods, to experiment with...*"

  ...Jerome. H. Friedman
- Use tools from $\underbrace{\textbf{mathematical optimization}}_{\text{Discrete \& Convex Optimization}}$ to devise estimators:
  - that are flexible
  - have a disciplined computational framework:
    – Obtain almost optimal solutions in seconds/minutes
    – Certify optimality in minutes/hours

Outline

- **Best Subset Selection in Regression** [Mallows '66, Miller '90]
  — Least Squares Variable Selection
  — Discrete Dantzig Selector
  — Grouped Variable Selection and Sparse Additive Models

- **Robust Linear Regression** [Rousseeuw '83]
  — Least Median of Squares Regression

- **Low rank Factor Analysis** [Spearman '04]
  — Least Squares Factor Analysis
  — Maximum Likelihood Factor Analysis

Outline

- **Best Subset Selection in Regression** [Mallows '66, Miller '90]
  — Least Squares Variable Selection
  — Discrete Dantzig Selector
  — Grouped Variable Selection and Sparse Additive Models

- **Robust Linear Regression** [Rousseeuw '83]
  — Least Median of Squares Regression

- **Low rank Factor Analysis** [Spearman '04]
  — Least Squares Factor Analysis
  — Maximum Likelihood Factor Analysis

Outline

- **Best Subset Selection in Regression** [Mallows '66, Miller '90]
  — Least Squares Variable Selection
  — Discrete Dantzig Selector
  — Grouped Variable Selection and Sparse Additive Models

- **Robust Linear Regression** [Rousseeuw '83]
  — Least Median of Squares Regression

- **Low rank Factor Analysis** [Spearman '04]
  — Least Squares Factor Analysis
  — Maximum Likelihood Factor Analysis

Best Subset Regression: Statement

[Bertsimas, King, **M.**, '16, *Annals of Statistics*]

- Usual linear regression model n samples, p regressors
- Want a sparse $\boldsymbol{\beta}$ with good data-fidelity:

$$\min_{\boldsymbol{\beta}} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq k, \qquad (\star)$$

[Miller '90; Foster & George '94; George '00]

- Problem $(\star)$ is NP-hard [Natarajan '95].
- R package `leaps` can handle $n \geq p \leq 31$.
  (branch and leaps [Furnival & Wilson 1974])
- Not surprisingly, advised to stay away from Problem $(\star)$.

Best Subset Regression: Current Approaches & Limitations

- Lasso (ℓ_1) [Tibshirani '96, Chen & Donoho '98] is a very popular and effective proxy:

$$\min_{\boldsymbol{\beta}} \quad \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

- Computation: convex optimization, fast & scalable

- $\ell_1 \implies$ good models, under $\underbrace{\textit{assumptions}}_{\text{difficult to verify}}$

- $\ell_1 \;\not\Longrightarrow\;$ reliable sparse solutions, and $\ell_1 \neq \ell_0$ solutions.
[Buhlmann, Van de Geer '11; Cai, Shen '11; Zhang, Jiang '08....]

Shortcomings of the Lasso: a simple explanation

- In presence of correlated variables, to obtain model with good predictive power, Lasso brings in a large number of nonzero coefficients
- Lasso leads to biased estimates—ℓ_1-norm penalizes large and small coefficients uniformly.
- Upon increasing the degree of regularization, Lasso sets more coefficients to zero—leaves out true predictors from the active set.

Best Subset Regression: ℓ_1 vs ℓ_0

- If $\widehat{\boldsymbol{\beta}}$ denotes the best subset solution, for *any* (fixed) $\mathbf{X}$,

$$\sup_{\|\boldsymbol{\beta}^*\|_0 \leq k} \frac{1}{n}\mathbb{E}(\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2) \lesssim \frac{\sigma^2 k \log p}{n},$$

- If $\widehat{\boldsymbol{\beta}}_{\ell_1}$ denotes a Lasso-based k-sparse estimator, then $\exists\ \mathbf{X}$:

$$\frac{1}{\gamma^2}\frac{\sigma^2 k^{1-\delta} \log p}{n} \lesssim \sup_{\|\boldsymbol{\beta}^*\|_0 \leq k} \frac{1}{n}\mathbb{E}(\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{\ell_1} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2) \lesssim \frac{1}{\gamma^2}\frac{\sigma^2 k \log p}{n},$$

- There is a significant gap between ℓ_0 and ℓ_1-type solutions.

[Bunea et. al. '07; Raskutti et. al. '09; Zhang et. al. '14]

Best Subset Regression: Current Approaches & Limitations

- To circumvent shortcomings, alternatives exist
- Non-convex penalties/ greedy methods
  [Fan, Li '01; Zou '06; Zou, Li '08; Zhang '10 ; Mazumder et. al. '11; Zhang , Zhang '12; Loh, Wainwright '14]
- Problems are non-convex and hard to solve.
- Computational approaches mostly heuristic:
  cannot certify/prove global optimality for arbitrary dataset.
  Exception: [Liu , Yao , Li '16]

Best Subset Regression: Our approach

[Bertsimas, King, **M.**, '16, *Annals of Statistics*]

- Certifiably $\min_{\boldsymbol{\beta}} \;\; \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \;\; \text{s.t.} \;\; \|\boldsymbol{\beta}\|_0 \leq k$

- Main workhorses:

  Tools from different branches of Optimization:

  - Modern Technology of **M**ixed **I**nteger **O**ptimization (MIO)
  - Discrete First Order methods (motivated from convex continuous optimization)

Best Subset Regression: Our approach

- Consider $\min_{\boldsymbol{\beta}} \ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \ \ \text{s.t.} \ \ \|\boldsymbol{\beta}\|_0 \leq k$

- Express as Mixed Integer Optimization problem (MIO)

- Discrete First Order methods for advanced warm-starts

- *Enhancing MIO: Stronger Formulations*

Best Subset Regression: Our approach

- Consider $\min_{\boldsymbol{\beta}} \ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\beta}\|_0 \leq k$
- Express as Mixed Integer Optimization problem (MIO)
- Discrete First Order methods for advanced warm-starts
- *Enhancing MIO: Stronger Formulations*

Brief Background on MIO

Mixed Integer Optimation (MIO)

- MIO: a particular class of discrete optimization problems
- The general form of a Mixed Integer Quadratic Optimization:

$$\begin{aligned} \min \quad & \boldsymbol{\alpha}^T \mathbf{Q} \boldsymbol{\alpha} + \boldsymbol{\alpha}^T \mathbf{a} \\ \text{s.t.} \quad & \mathbf{A}\boldsymbol{\alpha} \leq \mathbf{b} \\ & \alpha_i \in \{0,1\}, \quad \forall i \in \mathcal{I} \\ & \alpha_j \in \Re_+, \quad \forall j \notin \mathcal{I}, \end{aligned}$$

$\mathbf{a} \in \Re^m, \mathbf{A} \in \Re^{k \times m}, \mathbf{b} \in \Re^k$ and $\mathbf{Q} \in \Re^{m \times m}$ (PSD) problem-parameters;

- Special instances: Mixed Integer Linear Optimization, Quadratic/Linear Programming...

Mixed Integer Optimation (MIO)

- MIO optimization methods employ a combination of branch and bound, branch and cut, cutting plane methods, ...
(*not complete enumeration*)

- Foundations deeply rooted in polyhedral theory, combinatorics, discrete geometry/algebra,...

- Worst case: NP hard. Our focus is not worst case analysis.
(Simplex Algorithm, Path Algorithms like LARS, TSP, ...)

Mixed Integer Optimation (MIO)

- MIO optimization methods employ a combination of branch and bound, branch and cut, cutting plane methods, ...
  (*not complete enumeration*)
- Foundations deeply rooted in polyhedral theory, combinatorics, discrete geometry/algebra,...
- Worst case: NP hard. Our focus is not worst case analysis.
  (Simplex Algorithm, Path Algorithms like LARS, TSP, ...)
- Modern MIO is **tractable** (in practice)

Mixed Integer Optimation (MIO)

- MIO optimization methods employ a combination of branch and bound, branch and cut, cutting plane methods, ...
  (*not complete enumeration*)

- Foundations deeply rooted in polyhedral theory, combinatorics, discrete geometry/algebra,...

- Worst case: NP hard. Our focus is not worst case analysis.
  (Simplex Algorithm, Path Algorithms like LARS, TSP, ...)

- Modern MIO is **tractable** (in practice)
  tractability: Ability to solve problems of realistic size in times that are appropriate for the applications we consider.
  (successful applications: production planning, transportation, inventory management, air-traffic control, warehouse location, matching assignments,...)

Progress of MIO

- Algorithms and Software have undergone huge improvements over past 25+ years (1991 - 2016).
- Algorithms speed-up: $\sim$1.4 million times
  (Combined speedup: CPLEX 1.2 to 11 & Gurobi 1.0 to 6.5)
- Hardware speed-up: $\sim$1.6 million times
  (Peak Supercomputer performance)
- Total speed-up: **2.2 trillion times!**
- Commercial packages: Xpress, Gurobi, Cplex,...
  Non-commercial packages: GLPK, lpsolve, CBC, SCIP,...
  Interfaces: Matlab, R, Python, Julia (JuMP)

Back to Formulation

Vanilla MIO formulation

For problem: $\min_{\boldsymbol{\beta}} \ \ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \ \ \text{s.t.} \ \ \|\boldsymbol{\beta}\|_0 \leq k,$

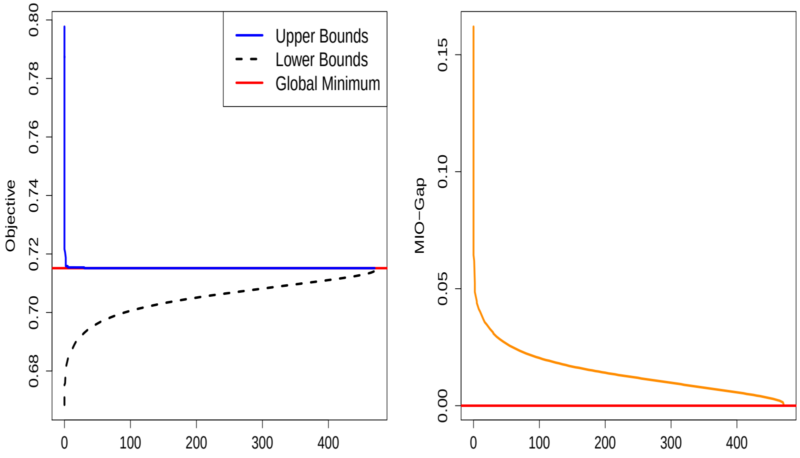
A simple (natural) MIO formulation is given by

$$
\begin{array}{ll}
\min\limits_{\boldsymbol{\beta},\mathbf{z}} & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\
\text{s.t.} & |\beta_i| \leq M \cdot z_i, i = 1, \ldots, p \\
& \sum\limits_{i=1}^{p} z_i \leq k \\
& z_i \in \{0, 1\}, i = 1, \ldots, p,
\end{array}
$$

where, M ("Big-M") is a parameter
— $M \geq \|\boldsymbol{\beta}\|_\infty$
— M controls the strength of the MIO formulation

Diabetes Dataset, $n = 350, p = 64, k = 6$

Typical behavior of Overall Algorithm



Time (secs) Time (secs)

Our approach

- Consider $\min_{\boldsymbol{\beta}} \;\; \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \;\; \text{s.t.} \;\; \|\boldsymbol{\beta}\|_0 \leq k$
- Express best-subset as a Mixed Integer Optimization problem (MIO)
- Discrete First Order methods for advanced warm-starts
- *Enhancing MIO: Stronger Formulations*

Discrete First Order Method

- Stylized *gradient* based method for

$$\min_{\boldsymbol{\beta}} \quad g(\boldsymbol{\beta}) \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq k,$$

- $g(\boldsymbol{\beta})$ convex and $\|\nabla g(\boldsymbol{\beta}) - \nabla g(\boldsymbol{\beta}_0)\| \leq \ell \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|$.
- This implies that for all $L \geq \ell$

$$g(\boldsymbol{\beta}) \leq Q(\boldsymbol{\beta}) = g(\boldsymbol{\beta}_0) + \langle \nabla g(\boldsymbol{\beta}_0), \boldsymbol{\beta} - \boldsymbol{\beta}_0 \rangle + \frac{L}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$$

- For the purpose of finding feasible solutions, we propose

$$\min_{\boldsymbol{\beta}} \quad Q(\boldsymbol{\beta}) \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq k$$

[Related work: Blumensath, Davis '08; Donoho, Johnstone '95]

Solution

- Equivalent to

$$\min_{\boldsymbol{\beta}} \ \frac{L}{2} \left\| \boldsymbol{\beta} - \left(\boldsymbol{\beta}_0 - \frac{1}{L} \nabla g(\boldsymbol{\beta}_0) \right) \right\|_2^2 \text{ s.t. } \|\boldsymbol{\beta}\|_0 \leq k$$

- Reducing to

$$\min_{\boldsymbol{\beta}} \ \|\boldsymbol{\beta} - \mathbf{u}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq k$$

- Optimal solution is $\boldsymbol{\beta}^* \in \mathbf{H}_k(\mathbf{u})$, where $\mathbf{H}_k(\mathbf{u})$ is the hard-thresholding operator (retains the top k entries of $\mathbf{u}$ in absolute value).
  [Donoho & Johnstone '95]

Discrete First Order Algorithm (DFA)

Algorithm to get *feasible* solutions for:

$$\min_{\boldsymbol{\beta}} \quad g(\boldsymbol{\beta}) \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq k.$$

1. Initialize with a solution $\boldsymbol{\beta}_0$; $m = 0$.
2. $m := m + 1$.
3. $\widetilde{\boldsymbol{\beta}}_{m+1} \in \mathbf{H}_k \left(\boldsymbol{\beta}_m - \frac{1}{L}\nabla g(\boldsymbol{\beta}_m)\right)$.
4. Perform a line search to get $\boldsymbol{\beta}_{m+1}$
5. Repeat Steps 2-4 until $\|\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m\| \leq \epsilon$.

Convergence properties

Theorem. (Bertsimas, King, **M.** '16)

Let $\boldsymbol{\beta}_m, m \geq 1$ be generated by DFA:

(a) For any $L \geq \ell$, the sequence $g(\boldsymbol{\beta}_m)$ is decreasing and converges.

(b) If $L > \ell$ and *under some minor regularity properties*

- $\|\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m\|_2^2 \leq \epsilon$ in at most $O(\frac{1}{\epsilon})$ many iterations.
- Supp$(\boldsymbol{\beta}_m)$ stabilizes after finitely many iterations and $\boldsymbol{\beta}_m$ converges to a first order stationary point.

Our approach

- Consider $\min_{\boldsymbol{\beta}} \;\; \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 \;\; \text{s.t.} \;\; \|\boldsymbol{\beta}\|_0 \leq k$
- Express best-subset as a Mixed Integer Optimization problem (MIO)
- Discrete First Order methods for advanced warm-starts
- *Enhancing MIO: Stronger Formulations*

Special Ordered Sets-formulation

$$\min_{\boldsymbol{\beta}} \quad \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq k$$

is equivalent to

$$\begin{aligned} \min_{\boldsymbol{\beta},\mathbf{z}} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\ \text{s.t.} \quad & (\beta_i, 1 - z_i) : \textsf{SOS type-1}, i = 1, \ldots, p \\ & \sum_{i=1}^{p} z_i \leq k \\ & z_i \in \{0, 1\}, \quad i = 1, \ldots, p. \end{aligned}$$

Implied Constraints

$$\min_{\boldsymbol{\beta}} \quad \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq k$$

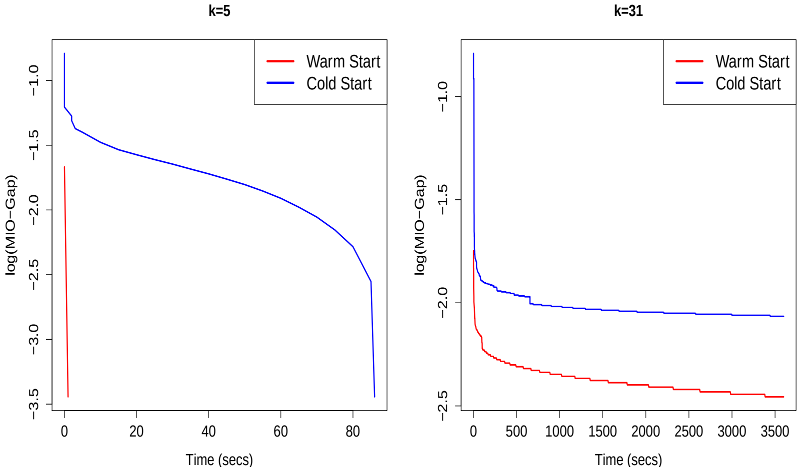
is equivalent to

$$\begin{aligned}
\min_{\boldsymbol{\beta}} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\
\text{s.t.} \quad & \|\boldsymbol{\beta}\|_0 \leq k \\
& \|\boldsymbol{\beta}\|_\infty \leq \delta_{11}, \;\; \|\boldsymbol{\beta}\|_1 \leq \delta_{21} \\
& \|\mathbf{X}\boldsymbol{\beta}\|_\infty \leq \delta_{12}, \;\; \|\mathbf{X}\boldsymbol{\beta}\|_1 \leq \delta_{22}
\end{aligned}$$

for constants $\delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}$ (which can be computed from data).

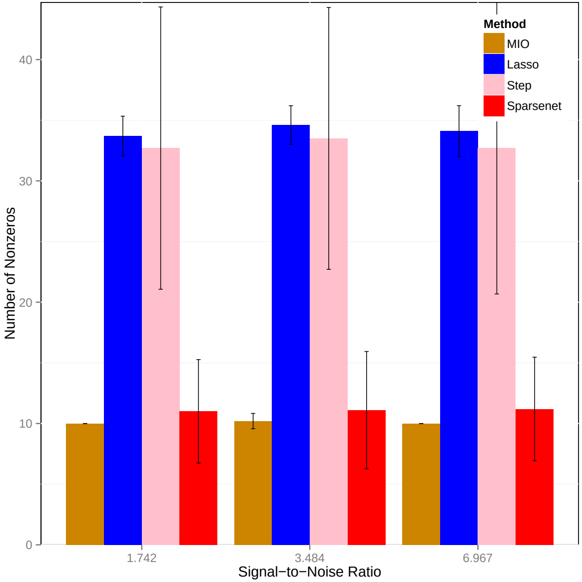
Behavior with user-guided intelligence

Diabetes data: $n = 350, \; p = 64.$

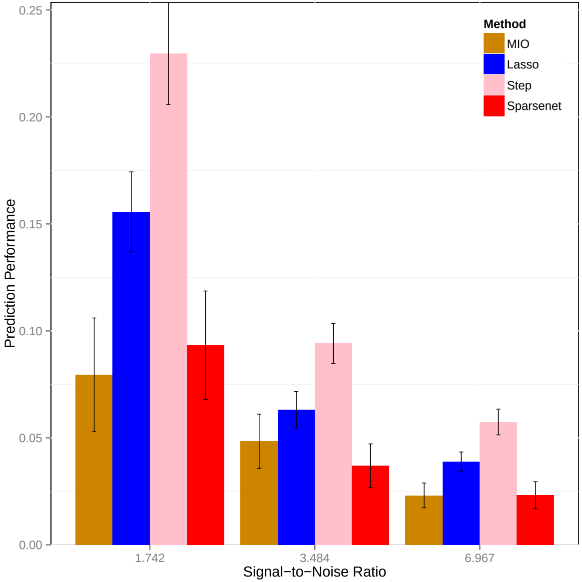


Statistical Behavior

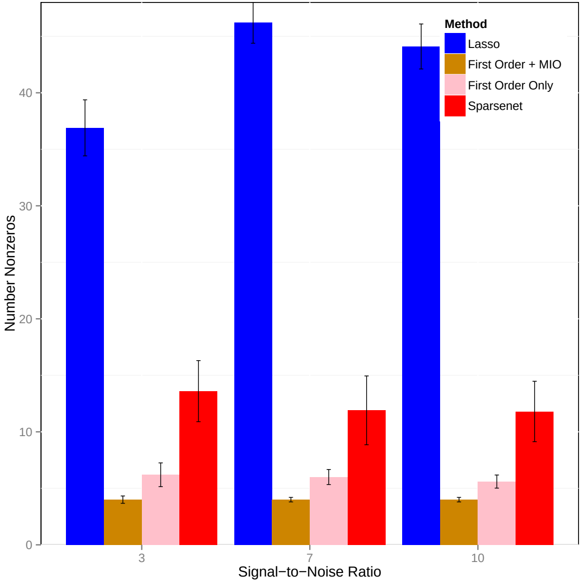
Sparsity Detection for $n = 500$, $p = 100$



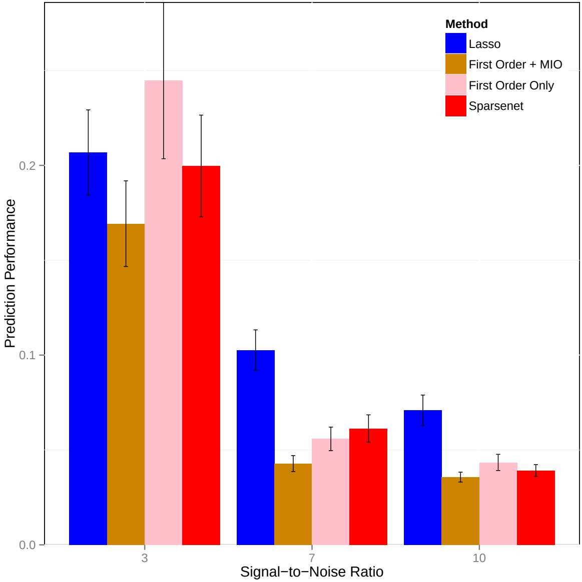
Prediction Error $= \|\mathbf{X}\boldsymbol{\beta}_{alg} - \mathbf{X}\boldsymbol{\beta}_{\text{true}}\|_2^2 / \|\mathbf{X}\boldsymbol{\beta}_{\text{true}}\|_2^2$



Sparsity Detection for $n = 50$, $p = 2000$



Prediction Error for $n = 50$, $p = 2000$



What did we learn?

- For the case $n > p$, MIO+intelligence finds provably optimal solutions for $n = 500s$, $p = 100s$ in minutes.
- For the case $n < p$, MIO+intelligence finds solutions for $n = 50s$, $p = 1000s$ in minutes and proving (approx)-optimality in hours.
- MIO solutions have a significant edge in sparsity and improved prediction accuracy.
- Modern optimization (MIO+user guided intelligence) is capable of tackling large instances.

Outline

- **Best Subset Selection in Regression** [Mallows '66, Miller '90]
  — Least Squares Variable Selection
  — Discrete Dantzig Selector
  — Grouped Variable Selection and Sparse Additive Models

- **Robust Linear Regression** [Rousseeuw '83]
  — Least Median of Squares Regression

- **Low rank Factor Analysis** [Spearman '04]
  — Least Squares Factor Analysis
  — Maximum Likelihood Factor Analysis

The Discrete Dantzig Selector

[**M.** & Radchenko '16+]

- The Dantzig Selector [Candes, Tao '07]:

$$\hat{\boldsymbol{\beta}}^{\mathsf{DS}}_{\ell_1} \in \operatorname{argmin} \quad \|\boldsymbol{\beta}\|_1 \quad \text{s.t.} \quad \|\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_\infty \leq \delta$$

- Instead, consider its ℓ_0 analogue:

$$\hat{\boldsymbol{\beta}}^{\mathsf{DS}}_{\ell_0} \in \operatorname{argmin} \quad \|\boldsymbol{\beta}\|_0 \quad \text{s.t.} \quad \|\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_\infty \leq \delta$$

- Find the sparsest $\boldsymbol{\beta}$ such that maximal (abs) correlation between covariates and residuals is small.

The Discrete Dantzig Selector

[**M.** & Radchenko '16+]

- The Dantzig Selector [Candes, Tao '07]:

$$\hat{\boldsymbol{\beta}}^{\mathsf{DS}}_{\ell_1} \in \operatorname{argmin} \quad \|\boldsymbol{\beta}\|_1 \quad \text{s.t.} \quad \|\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_\infty \leq \delta$$

- Instead, consider its ℓ_0 analogue:

$$\hat{\boldsymbol{\beta}}^{\mathsf{DS}}_{\ell_0} \in \operatorname{argmin} \quad \|\boldsymbol{\beta}\|_0 \quad \text{s.t.} \quad \|\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_\infty \leq \delta$$

- Find the sparsest $\boldsymbol{\beta}$ such that maximal (abs) correlation between covariates and residuals is small.

- Why is this **important**?
  – Formulation is a Mixed Integer *Linear* Optimization.

  – Mixed Integer *Linear* is a more <u>mature</u> technology than Mixed Integer *Quadratic* Optimization

The Discrete Dantzig Selector

Under a sparse linear model with Gaussian errors: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$

- The errors:
  - $\|\hat{\boldsymbol{\beta}}_{\ell_0}^{\mathsf{DS}} - \boldsymbol{\beta}^*\|_2^2$
  - $\|\hat{\boldsymbol{\beta}}_{\ell_0}^{\mathsf{DS}} - \boldsymbol{\beta}^*\|_1^2$
  - $\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\ell_0}^{\mathsf{DS}} - \boldsymbol{\beta}^*)\|_2^2$

  are much smaller than the convex estimator $\hat{\boldsymbol{\beta}}_{\ell_1}^{\mathsf{DS}}$ (when features are correlated)

- $\#$ Non-zeros $\hat{\boldsymbol{\beta}}_{\ell_0}^{\mathsf{DS}} \quad \ll \quad \#$ Non-zeros $\hat{\boldsymbol{\beta}}_{\ell_1}^{\mathsf{DS}}$

- Statistical properties of $\hat{\boldsymbol{\beta}}_{\ell_0}^{\mathsf{DS}}$ comparable with Least Squares Subset Selection

Some Large Problems

(Synthetic Examples)						
n	p	k^*	Upper Bound	Lower Bound	MIO Gap	Time to Prove Opt
4,000	8,000	20	20	20	0	41.9
3,000	8,000	20	20	20	0	18.3
1,000	10,000	10	10	10	0	14.2
5,000	10,000	10	10	10	0	2.5
10,000	10,000	30	30	27	10%	42.5
(Real Data Examples)						
n	p	k^*	Upper Bound	Lower Bound	MIO Gap	Time to Prove Opt
6,000	4,500	20	20	20	0	5.0
6,000	4,500	40	40	37	10%	12.5

Table: Solutions obtained within 5-10 minutes for all problems. Certifying Optimality takes longer.

Some Large Problems

(Synthetic Examples)						
n	p	k^*	Upper Bound	Lower Bound	MIO Gap	**Time to Prove Opt**
4,000	8,000	20	20	20	0	41.9
3,000	8,000	20	20	20	0	18.3
1,000	10,000	10	10	10	0	14.2
5,000	10,000	10	10	10	0	2.5
10,000	10,000	30	30	27	10%	42.5
(Real Data Examples)						
n	p	k^*	Upper Bound	Lower Bound	MIO Gap	**Time to Prove Opt**
6,000	4,500	20	20	20	0	5.0
6,000	4,500	40	40	37	10%	12.5

Table: Solutions obtained within 5-10 minutes for all problems. Certifying Optimality takes longer (several hours).

Outline

- **Best Subset Selection in Regression** [Mallows '66, Miller '90]
  — Least Squares Variable Selection
  — Discrete Dantzig Selector
  — Grouped Variable Selection and Sparse Additive Models

- **Robust Linear Regression** [Rousseeuw '83]
  — Least Median of Squares Regression

- **Low rank Factor Analysis** [Spearman '04]
  — Least Squares Factor Analysis
  — Maximum Likelihood Factor Analysis

Effect of Outliers in Regression

[Bertsimas, **M.**, '14, *Annals of Statistics*]

- Least Squares (LS) estimator

$$\widehat{\boldsymbol{\beta}}^{(\mathrm{LS})} \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} r_i^2, \quad r_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}$$

has a breakdown point of zero (Dohono & Huber '83; Hampel '75).

- The Least Absolute Deviation (LAD) estimator has a breakdown point of zero

$$\widehat{\boldsymbol{\beta}}^{(\mathrm{LAD})} \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} |r_i|,$$

- M-Estimators (Huber '73) slightly improve the breakdown point

$$\sum_{i=1}^{n} \rho(r_i), \quad \rho(r) \text{ symmetric function}$$

Least Median Regression

- Least Median of Squares (LMS) estimator [Rousseew ('84)]

$$\widehat{\boldsymbol{\beta}}^{(\mathsf{LMS})} \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\underset{i=1,\ldots,n}{\operatorname{median}} \;\; |r_i| \right).$$

- LMS highest possible breakdown point of almost 50%.

- More generally, Least Quantile of Squares (LQS) estimator:

$$\widehat{\boldsymbol{\beta}}^{(\mathsf{LQS})} \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \;\; |r_{(q)}|,$$

  where, $r_{(q)}$ is the qth ordered absolute residual:

$$|r_{(1)}| \leq |r_{(2)}| \leq \ldots \leq |r_{(n)}|.$$

Problem we address

- Solve the following problem:

$$\min_{\boldsymbol{\beta}} \; |r_{(q)}|,$$

where, $r_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}$, q is a quantile.

- Our approach extends to

$$\min_{\boldsymbol{\beta}} \; |r_{(q)}|, \;\; \text{s.t.} \;\; \mathbf{A}\boldsymbol{\beta} \leq \mathbf{b} \;\; \text{(and/or} \;\; \|\boldsymbol{\beta}\|_2^2 \leq \delta)$$

LQS and Subset Selection: A surprising link

- LQS and subset-selection in regression seem to be completely unrelated concepts...
- However, a curious link emerges...

LQS and Subset Selection: A surprising link

- LQS and subset-selection in regression seem to be completely unrelated concepts...
- However, a curious link emerges...
- **Claim:** *LQS is performing an implicit subset search*

LQS and Subset Selection: A surprising link

- LQS and subset-selection in regression seem to be completely unrelated concepts...
- However, a curious link emerges...
- **Claim:** *LQS is performing an implicit subset search*

Theorem [Bertsimas & **M.** '14]: The LQS problem is equivalent to the following:

$$\min_{\boldsymbol{\beta}} \; |r_{(q)}| = \min_{\mathcal{I} \in \Omega_q} \; \left(\min_{\boldsymbol{\beta}} \; \|\mathbf{y}_I - \mathbf{X}_I \boldsymbol{\beta}\|_\infty \right),$$

where, $\Omega_q := \{\mathcal{I} : \; \mathcal{I} \subset \{1, \ldots, n\}, |\mathcal{I}| = q\}$ and $(\mathbf{y}_I, \mathbf{X}_I)$ denotes the subsample $(y_i, \mathbf{x}_i), i \in \mathcal{I}$.

Overview of our approach

- Write the LMS problem as a MIO.
  — Main idea: MIO formulation sorts to express $|r_{(q)}|$
  — Formulation *very* different from best subset selection in regression
- Using Discrete First Order methods we find good feasible solutions.
- Warm-starts and improved behavior with user-guided intelligence

MIO Formulation

Notation:

$$|r_{(1)}| \leq |r_{(2)}| \leq \ldots \leq |r_{(n)}|.$$

Step 1: Introduce binary variables $z_i, i = 1, \ldots, n$ such that:

$$z_i = \begin{cases} 1, & \text{if } |r_i| \leq |r_{(q)}|, \\ 0, & \text{otherwise.} \end{cases}$$

Step 2: Use auxiliary continuous variables $\mu_i, \overline{\mu}_i \geq 0$ such that:

$$|r_i| - \mu_i \leq |r_{(q)}| \leq |r_i| + \overline{\mu}_i, i = 1, \ldots, n,$$

with the conditions:

$$\begin{array}{rlll} & \text{If} & |r_i| \geq |r_{(q)}|, & \text{then } \overline{\mu}_i = 0, \mu_i \geq 0, \\ \text{and} & \text{if} & |r_i| \leq |r_{(q)}|, & \text{then } \mu_i = 0, \overline{\mu}_i \geq 0. \end{array}$$

MIO Formulation

Notation:

$$|r_{(1)}| \leq |r_{(2)}| \leq \ldots \leq |r_{(n)}|.$$

Step 1: Introduce binary variables $z_i, i = 1, \ldots, n$ such that:

$$z_i = \begin{cases} 1, & \text{if } |r_i| \leq |r_{(q)}|, \\ 0, & \text{otherwise.} \end{cases}$$

Step 2: Use auxiliary continuous variables $\mu_i, \overline{\mu}_i \geq 0$ such that:

$$|r_i| - \mu_i \leq |r_{(q)}| \leq |r_i| + \overline{\mu}_i, i = 1, \ldots, n,$$

with the conditions:

$$\left.\begin{array}{rlll} & \text{If} & |r_i| \geq |r_{(q)}|, & \text{then } \overline{\mu}_i = 0, \mu_i \geq 0, \\ \text{and} & \text{if} & |r_i| \leq |r_{(q)}|, & \text{then } \mu_i = 0, \overline{\mu}_i \geq 0. \end{array}\right\} \text{MIO representable}$$

MIO Formulation

$$
\begin{aligned}
\min \quad & \gamma & \\
\text{s.t.} \quad & |r_i| + \overline{\mu}_i \geq \gamma, & i = 1 \ldots, n \\
& \gamma \geq |r_i| - \mu_i, & i = 1 \ldots, n \\
& M_u z_i \geq \overline{\mu}_i, & i = 1, \ldots, n \\
& M_\ell (1 - z_i) \geq \mu_i, & i = 1, \ldots, n \\
& \sum_{i=1}^{n} z_i = q & \\
& \mu_i \geq 0, & i = 1, \ldots, n \\
& \overline{\mu}_i \geq 0, & i = 1, \ldots, n \\
& z_i \in \{0, 1\}, & i = 1, \ldots, n,
\end{aligned}
$$

where $\gamma, z_i, \mu_i, \overline{\mu}_i, i = 1, \ldots, n$ are decision variables and M_u, M_ℓ are Big-M constants.

What do we achieve?

- Prior exact algorithms can solve upto: $n = 50$ and $p = 5$
- We obtain:
  - *near optimal* solutions for problems with $n \approx 200$s and $p \approx 20$s in seconds, proving optimality in minutes.
  - *near optimal* solutions for problems with $n \approx 10,000$ and $p \approx 50$ in minutes.

Outline

- **Best Subset Selection in Regression** [Mallows '66, Miller '90]
  — Least Squares Variable Selection
  — Discrete Dantzig Selector
  — Grouped Variable Selection and Sparse Additive Models

- **Robust Linear Regression** [Rousseeuw '83]
  — Least Median of Squares Regression

- **Low rank Factor Analysis** [Spearman '04]
  — Least Squares Factor Analysis
  — Maximum Likelihood Factor Analysis

Background & Formulation

[Bertsimas, Copenhaver, **M.**, '16+]

Low Rank Factor Analysis (FA) [Spearman 1904]:

- widely used in multivariate statistics, econometrics, psychometrics
- represent correlation structure with few common (latent) factors.

Estimation Problem: $$\mathbf{\Sigma} = \underbrace{\mathbf{L}_1\mathbf{L}_1'}_{\mathbf{\Theta}} + \underbrace{\mathbf{L}_2\mathbf{L}_2'}_{\text{Small}} + \mathbf{\Phi}$$

- $\mathbf{\Sigma} \approx \mathbf{\Theta} + \mathbf{\Phi}$
- $\mathbf{\Phi} = \mathrm{diag}(\Phi_1, \ldots, \Phi_p) \succeq \mathbf{0}$
- $\mathrm{rank}(\mathbf{\Theta}) \leq r, \mathbf{\Theta} \succeq \mathbf{0}$
- $\mathbf{\Sigma} - \mathbf{\Theta} \succeq \mathbf{0}; \mathbf{\Sigma} - \mathbf{\Phi} \succeq \mathbf{0}$

Background & Formulation

Low Rank Factor Analysis (FA) [Spearman 1904]:

- widely used in multivariate statistics, econometrics, psychometrics
- represent correlation structure with few common (latent) factors.

Estimation Problem: $$\boldsymbol{\Sigma} = \underbrace{\mathbf{L}_1\mathbf{L}_1'}_{\boldsymbol{\Theta}} + \underbrace{\mathbf{L}_2\mathbf{L}_2'}_{\text{Small}} + \boldsymbol{\Phi}$$

$$\left.\begin{array}{ll} - & \boldsymbol{\Sigma} \approx \boldsymbol{\Theta} + \boldsymbol{\Phi} \\ - & \boldsymbol{\Phi} = \mathrm{diag}(\Phi_1, \ldots, \Phi_p) \succeq \mathbf{0} \\ - & \mathrm{rank}(\boldsymbol{\Theta}) \leq r, \boldsymbol{\Theta} \succeq \mathbf{0} \\ - & \boldsymbol{\Sigma} - \boldsymbol{\Theta} \succeq \mathbf{0}; \boldsymbol{\Sigma} - \boldsymbol{\Phi} \succeq \mathbf{0} \end{array}\right\} \quad \begin{array}{ll} \min & \|\boldsymbol{\Sigma} - (\boldsymbol{\Theta} + \boldsymbol{\Phi})\| \\ \text{s.t.} & \mathrm{rank}(\boldsymbol{\Theta}) \leq r \\ & \boldsymbol{\Sigma} - \boldsymbol{\Phi} \succeq \mathbf{0} \end{array}$$

Our Approach

$$
\begin{array}{lll}
\min & \|\boldsymbol{\Sigma} - (\boldsymbol{\Theta} + \boldsymbol{\Phi})\| & \\
\text{s.t.} & \operatorname{rank}(\boldsymbol{\Theta}) \leq r & (\dagger) \\
 & \boldsymbol{\Sigma} - \boldsymbol{\Theta} \succeq \mathbf{0} &
\end{array}
$$

Our Approach

$$
\begin{array}{lccc}
\min & \|\boldsymbol{\Sigma} - (\boldsymbol{\Theta} + \boldsymbol{\Phi})\| & \leftarrow \text{Sum of Singular Values} & \\
\text{s.t.} & \operatorname{rank}(\boldsymbol{\Theta}) \le r & \leftarrow \text{Rank Constraint} & (\dagger) \\
 & \boldsymbol{\Sigma} - \boldsymbol{\Theta} \succeq \mathbf{0} & \leftarrow \text{Semidefinite Constraint} &
\end{array}
$$

Our Approach

$$\begin{array}{lcll} \min & \|\boldsymbol{\Sigma} - (\boldsymbol{\Theta} + \boldsymbol{\Phi})\| & \leftarrow \text{Sum of Singular Values} & \\ \text{s.t.} & \operatorname{rank}(\boldsymbol{\Theta}) \leq r & \leftarrow \text{Rank Constraint} & (\dagger) \\ & \boldsymbol{\Sigma} - \boldsymbol{\Theta} \succeq \mathbf{0} & \leftarrow \text{Semidefinite Constraint} & \end{array}$$

- SDP with rank constraints
- Key Idea: Reformulate $(\dagger)$ equivalently as a SDP (without rank constraint)
  - Nonlinear Optimization techniques for feasible solutions
  - Specialized Branch & Bound methods to certify optimality

Reformulation and tailored B&B

$$\begin{aligned} \min \quad & \|\boldsymbol{\Sigma} - (\boldsymbol{\Theta} + \boldsymbol{\Phi})\| \\ \text{s.t.} \quad & \operatorname{rank}(\boldsymbol{\Theta}) \leq r \\ & \boldsymbol{\Sigma} - \boldsymbol{\Theta} \succeq \mathbf{0} \end{aligned}$$

$$\Updownarrow \qquad \begin{cases} \text{Variational Representation} \\ \text{of Spectral Functions} \end{cases}$$

$$\begin{aligned} \min \quad & \langle \mathbf{W}, \boldsymbol{\Sigma} - \boldsymbol{\Theta} \rangle - \sum_{i=1}^{p} w_{ii} \Phi_i \\ \text{s.t.} \quad & \mathbf{I} \succeq \mathbf{W} \succeq \mathbf{0} \\ & \mathsf{Tr}(\mathbf{W}) = p - r \\ & \boldsymbol{\Sigma} - \boldsymbol{\Theta} \succeq \mathbf{0} \end{aligned}$$

Reformulation and tailored B&B

$$
\begin{aligned}
\min \quad & \|\boldsymbol{\Sigma} - (\boldsymbol{\Theta} + \boldsymbol{\Phi})\| \\
\text{s.t.} \quad & \operatorname{rank}(\boldsymbol{\Theta}) \leq r \\
& \boldsymbol{\Sigma} - \boldsymbol{\Theta} \succeq \mathbf{0}
\end{aligned}
$$

$$\Updownarrow \qquad \left\{\begin{array}{l}\text{Variational Representation} \\ \text{of Spectral Functions}\end{array}\right.$$

$$
\begin{aligned}
\min \quad & \langle \mathbf{W}, \boldsymbol{\Sigma} - \boldsymbol{\Theta} \rangle - \boxed{\sum_{i=1}^{p} w_{ii} \Phi_i} \quad \longleftarrow \left\{\begin{array}{l}\text{Bilinear Form (Nonconvex)} \\ \text{McCormick Hulls/ B\&B}\end{array}\right. \\
\text{s.t.} \quad & \mathbf{I} \succeq \mathbf{W} \succeq \mathbf{0} \\
& \mathrm{Tr}(\mathbf{W}) = p - r \\
& \boldsymbol{\Sigma} - \boldsymbol{\Theta} \succeq \mathbf{0}
\end{aligned}
$$

What do we learn?

- Several experiments on both real and synthetic datasets, reveal:
  - Upper bounds obtained within few seconds ($p = 100$) to several minutes ($p = 4000$)
  - Certifying optimality takes longer (several hours)

- Global optimality certificates obtained on datasets, where, assumptions required for convex problem to succeed cannot be verified.

ArXiv link: `http://arxiv.org/pdf/1604.06837v1.pdf`

Conclusions

- MIO is an advanced, computationally tractable mathematical programming framework
- Provides a powerful modeling tool for statistical problems
- Leads to a significant *edge* in Sparse Learning problems that are inherently discrete.
- 15.097: PhD class taught at MIT Spring 2016 on related topics.

Thank you!

All papers available at:
`http://www.mit.edu/~rahulmaz/research.html`