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 **discrete**

- 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 discrete

- 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 \textit{a-priori} which method will work for a given application

- "...A statistician's \textit{toolkit} should have a \textit{whole array} of methods, to experiment with..."

  ...Jerome. H. Friedman
Motivation

- We seldom know \textit{a-priori} which method will work for a given application

- “...A statistician's \textit{toolkit} should have a \textit{whole array} of methods, to experiment with...”

  ...Jerome. H. Friedman

- Use tools from \textbf{mathematical optimization} to devise estimators:
  - that are flexible
  - have a disciplined computational framework:
    - Obtain \textit{almost optimal} solutions in \textit{seconds/minutes}
    - \textbf{Certify} optimality in \textit{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


- Usual linear regression model $n$ samples, $p$ regressors

- Want a sparse $\beta$ with good data-fidelity:

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

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

- Problem (⋆) 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 (⋆).
Best Subset Regression: Current Approaches & Limitations

- **Lasso** ($\ell_1$) [Tibshirani '96, Chen & Donoho '98] is a very popular and effective proxy:

\[
\min_{\beta} \quad \frac{1}{2} \| y - X\beta \|^2 + \lambda \| \beta \|_1,
\]

- Computation: convex optimization, fast & scalable

- $\ell_1 \implies$ good models, under **assumptions**
  - difficult to verify

- $\ell_1 \nRightarrow$ reliable sparse solutions, and $\ell_1 \neq \ell_0$ solutions.
  - [Buhlmann, Van de Geer '11; Cai, Shen '11; Zhang, Jiang '08...]

R. Mazumder \hspace{1cm} Sparsity with Discrete Optimization 8
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—$\ell_1$-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: $\ell_1$ vs $\ell_0$

- If $\hat{\beta}$ denotes the best subset solution, for any (fixed) $X$,
  \[
  \sup_{\|\beta^*\|_0 \leq k} \frac{1}{n} \mathbb{E}(\|X\hat{\beta} - X\beta^*\|_2^2) \lesssim \frac{\sigma^2 k \log p}{n},
  \]

- If $\hat{\beta}_{\ell_1}$ denotes a Lasso-based $k$-sparse estimator, then $\exists X$:
  \[
  \frac{1}{\gamma^2} \frac{\sigma^2 k^{1-\delta} \log p}{n} \lesssim \sup_{\|\beta^*\|_0 \leq k} \frac{1}{n} \mathbb{E}(\|X\hat{\beta}_{\ell_1} - X\beta^*\|_2^2) \lesssim \frac{1}{\gamma^2} \frac{\sigma^2 k \log p}{n},
  \]

- There is a significant gap between $\ell_0$ and $\ell_1$-type solutions.

[Bunea et. al. '07; Raskutti et. al. '09; Zhang et. al. '14]
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 \( \min_\beta \frac{1}{2} \| y - X\beta \|^2_2 \) s.t. \( \| \beta \|_0 \leq k \)

- Main workhorses:

  Tools from different branches of Optimization:

  - Modern Technology of **Mixed Integer Optimization** (MIO)
  
  - Discrete First Order methods (motivated from convex continuous optimization)
Best Subset Regression: Our approach

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

- 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
  \[
  \min_{\beta} \quad \frac{1}{2} \| y - X\beta \|^2_2 \quad \text{s.t.} \quad \|\beta\|_0 \leq k
  \]

- 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 Optimization (MIO)

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

\[
\begin{align*}
\text{min} & \quad \alpha^T Q \alpha + \alpha^T a \\
\text{s.t.} & \quad A \alpha \leq b \\
& \quad \alpha_i \in \{0, 1\}, \quad \forall i \in I \\
& \quad \alpha_j \in \mathbb{R}_+, \quad \forall j \notin I,
\end{align*}
\]

\large\text{a} \in \mathbb{R}^m, \ A \in \mathbb{R}^{k \times m}, \ b \in \mathbb{R}^k \text{ and } Q \in \mathbb{R}^{m \times m} \text{ (PSD)} \text{ problem-parameters;}

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

R. Mazumder  Sparsity with Discrete Optimization 17
Mixed Integer Optimization (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 \textit{not} worst case analysis.
  \((\text{Simplex Algorithm, Path Algorithms like LARS, TSP, ...})\)

- Modern MIO is \textit{tractable} (in practice)
Mixed Integer Optimization (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, Ipolve, CBC, SCIP,...
  Interfaces: Matlab, R, Python, Julia (JuMP)
Back to Formulation
Vanilla MIO formulation

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

A simple (natural) MIO formulation is given by

\[
\begin{align*}
\min_{\beta, z} & \quad \frac{1}{2} \| y - X\beta \|_2^2 \\
\text{s.t.} & \quad |\beta_i| \leq M \cdot z_i, i = 1, \ldots, p \\
& \quad \sum_{i=1}^{p} z_i \leq k \\
& \quad z_i \in \{0, 1\}, i = 1, \ldots, p,
\end{align*}
\]

where, \( M \) ("Big-M") is a parameter

- \( M \geq \| \beta \|_\infty \)
- \( M \) controls the strength of the MIO formulation
Diabetes Dataset, \( n = 350, p = 64, k = 6 \)

Typical behavior of Overall Algorithm
Our approach

- Consider \( \min_{\beta} \frac{1}{2} \| y - X\beta \|_2^2 \) s.t. \( \| \beta \|_0 \leq k \)

- 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
  \[
  \min_{\beta} \quad g(\beta) \quad \text{s.t.} \quad \|\beta\|_0 \leq k, \\
  \]

- \( g(\beta) \) convex and \( \|\nabla g(\beta) - \nabla g(\beta_0)\| \leq \ell \cdot \|\beta - \beta_0\| \).

- This implies that for all \( L \geq \ell \)
  \[
  g(\beta) \leq Q(\beta) = g(\beta_0) + \langle \nabla g(\beta_0), \beta - \beta_0 \rangle + \frac{L}{2} \|\beta - \beta_0\|_2^2 \\
  \]

- For the purpose of finding feasible solutions, we propose
  \[
  \min_{\beta} \quad Q(\beta) \quad \text{s.t.} \quad \|\beta\|_0 \leq k \\
  \]

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

R. Mazumder

Sparsity with Discrete Optimization 23
Solution

- Equivalent to

\[
\min_{\beta} \frac{L}{2} \left\| \beta - \left( \beta_0 - \frac{1}{L} \nabla g(\beta_0) \right) \right\|^2_2 \quad \text{s.t. } \|\beta\|_0 \leq k
\]

- Reducing to

\[
\min_{\beta} \|\beta - u\|^2_2 \quad \text{s.t. } \|\beta\|_0 \leq k
\]

- Optimal solution is \( \beta^* \in H_k(u) \), where \( H_k(u) \) is the hard-thresholding operator (retains the top \( k \) entries of \( u \) in absolute value).

[Donoho & Johnstone '95]
Discrete First Order Algorithm (DFA)

Algorithm to get feasible solutions for:

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

1. Initialize with a solution \( \beta_0 \); \( m = 0 \).

2. \( m := m + 1 \).

3. \( \tilde{\beta}_{m+1} \in H_k (\beta_m - \frac{1}{L} \nabla g(\beta_m)) \).

4. Perform a line search to get \( \beta_{m+1} \).

5. Repeat Steps 2-4 until \( \|\beta_{m+1} - \beta_m\| \leq \epsilon \).
Theorem. (Bertsimas, King, M. ’16)

Let $\beta_m, m \geq 1$ be generated by DFA:

(a) For any $L \geq \ell$, the sequence $g(\beta_m)$ is decreasing and converges.

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

- $\|\beta_{m+1} - \beta_m\|_2^2 \leq \epsilon$ in at most $O\left(\frac{1}{\epsilon}\right)$ many iterations.

- $\text{Supp}(\beta_m)$ stabilizes after finitely many iterations and $\beta_m$ converges to a first order stationary point.
Our approach

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

- Express best-subset as a Mixed Integer Optimization problem (MIO)

- Discrete First Order methods for advanced warm-starts

- *Enhancing MIO: Stronger Formulations*
\[
\min_{\beta} \|y - X\beta\|_2^2 \quad \text{s.t.} \quad \|\beta\|_0 \leq k
\]

is equivalent to

\[
\min_{\beta, z} \|y - X\beta\|_2^2 \\
\text{s.t.} \quad (\beta_i, 1 - z_i) : \text{SOS type-1}, i = 1, \ldots, p \\
\sum_{i=1}^{p} z_i \leq k \\
\]

\[
z_i \in \{0, 1\}, \quad i = 1, \ldots, p.
\]
Implied Constraints

\[
\begin{align*}
\min_{\beta} & \quad \| y - X\beta \|^2_2 \\
\text{s.t.} & \quad \|\beta\|_0 \leq k
\end{align*}
\]

is equivalent to

\[
\begin{align*}
\min_{\beta} & \quad \| y - X\beta \|^2_2 \\
\text{s.t.} & \quad \|\beta\|_0 \leq k \\
& \quad \|\beta\|_{\infty} \leq \delta_{11}, \quad \|\beta\|_1 \leq \delta_{21} \\
& \quad \| X\beta \|_{\infty} \leq \delta_{12}, \quad \| X\beta \|_1 \leq \delta_{22}
\end{align*}
\]

for constants \( \delta_{11}, \delta_{12}, \delta_{21}, \delta_{22} \) (which can be computed from data).
Behavior with user-guided intelligence

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

R. Mazumder  Sparsity with Discrete Optimization 30
Statistical Behavior
Sparsity Detection for $n = 500, \ p = 100$

<table>
<thead>
<tr>
<th>Method</th>
<th>Signal-to-Noise Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIO</td>
<td>1.742</td>
</tr>
<tr>
<td>Lasso</td>
<td>3.484</td>
</tr>
<tr>
<td>Step</td>
<td>6.967</td>
</tr>
<tr>
<td>Sparsenet</td>
<td>10</td>
</tr>
</tbody>
</table>
Prediction Error = \|X\beta_{alg} - X\beta_{true}\|^2_2 / \|X\beta_{true}\|^2_2
Sparsity Detection for $n = 50, \ p = 2000$

![Graph showing sparsity detection results for different methods: Lasso, First Order + MIO, First Order Only, and Sparsenet. The x-axis represents the signal-to-noise ratio, and the y-axis represents the number of nonzeros. The graph compares the performance of these methods across various signal-to-noise ratios.](image-url)
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]:
  \[ \hat{\beta}_{\ell_1}^{DS} \in \text{argmin} \| \beta \|_1 \quad \text{s.t.} \quad \| X'(y - X\beta) \|_\infty \leq \delta \]

- Instead, consider its \( \ell_0 \) analogue:
  \[ \hat{\beta}_{\ell_0}^{DS} \in \text{argmin} \| \beta \|_0 \quad \text{s.t.} \quad \| X'(y - X\beta) \|_\infty \leq \delta \]

- Find the sparsest \( \beta \) 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 mature technology than Mixed Integer Quadratic Optimization.
The Discrete Dantzig Selector

[M. & Radchenko '16+]

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

\[ \hat{\beta}^{DS}_{\ell_1} \in \text{argmin} \; \|\beta\|_1 \quad \text{s.t.} \quad \|X'(y - X\beta)\|_\infty \leq \delta \]

▶ Instead, consider its \( \ell_0 \) analogue:

\[ \hat{\beta}^{DS}_{\ell_0} \in \text{argmin} \; \|\beta\|_0 \quad \text{s.t.} \quad \|X'(y - X\beta)\|_\infty \leq \delta \]

▶ Find the sparsest \( \beta \) 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 mature technology than Mixed Integer Quadratic Optimization.
The Discrete Dantzig Selector

Under a sparse linear model with Gaussian errors: \( y = X\beta^* + \epsilon \)

- The errors:
  
  \[- \| \hat{\beta}_{\ell_0}^{DS} - \beta^* \|_2^2 \]
  
  \[- \| \hat{\beta}_{\ell_0}^{DS} - \beta^* \|_1^2 \]
  
  \[- \| X(\hat{\beta}_{\ell_0}^{DS} - \beta^*) \|_2^2 \]

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

- \# Non-zeros \( \hat{\beta}_{\ell_0}^{DS} \ll \# \) Non-zeros \( \hat{\beta}_{\ell_1}^{DS} \)

- Statistical properties of \( \hat{\beta}_{\ell_0}^{DS} \) comparable with Least Squares Subset Selection
Some Large Problems

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>Upper Bound</th>
<th>Lower Bound</th>
<th>MIO Gap</th>
<th>Time to Prove Opt</th>
</tr>
</thead>
<tbody>
<tr>
<td>4,000</td>
<td>8,000</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>41.9</td>
</tr>
<tr>
<td>3,000</td>
<td>8,000</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>18.3</td>
</tr>
<tr>
<td>1,000</td>
<td>10,000</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>14.2</td>
</tr>
<tr>
<td>5,000</td>
<td>10,000</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>2.5</td>
</tr>
<tr>
<td>10,000</td>
<td>10,000</td>
<td>30</td>
<td>30</td>
<td>27</td>
<td>10%</td>
<td>42.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>Upper Bound</th>
<th>Lower Bound</th>
<th>MIO Gap</th>
<th>Time to Prove Opt</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,000</td>
<td>4,500</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>5.0</td>
</tr>
<tr>
<td>6,000</td>
<td>4,500</td>
<td>40</td>
<td>40</td>
<td>37</td>
<td>10%</td>
<td>12.5</td>
</tr>
</tbody>
</table>

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

### (Synthetic Examples)

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p$</th>
<th>$k^*$</th>
<th>Upper Bound</th>
<th>Lower Bound</th>
<th>MIO Gap</th>
<th>Time to Prove Opt</th>
</tr>
</thead>
<tbody>
<tr>
<td>4,000</td>
<td>8,000</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>41.9</td>
</tr>
<tr>
<td>3,000</td>
<td>8,000</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>18.3</td>
</tr>
<tr>
<td>1,000</td>
<td>10,000</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>14.2</td>
</tr>
<tr>
<td>5,000</td>
<td>10,000</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>2.5</td>
</tr>
<tr>
<td>10,000</td>
<td>10,000</td>
<td>30</td>
<td>30</td>
<td>27</td>
<td>10%</td>
<td>42.5</td>
</tr>
</tbody>
</table>

### (Real Data Examples)

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p$</th>
<th>$k^*$</th>
<th>Upper Bound</th>
<th>Lower Bound</th>
<th>MIO Gap</th>
<th>Time to Prove Opt</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,000</td>
<td>4,500</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>5.0</td>
</tr>
<tr>
<td>6,000</td>
<td>4,500</td>
<td>40</td>
<td>40</td>
<td>37</td>
<td>10%</td>
<td>12.5</td>
</tr>
</tbody>
</table>

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

\[ \hat{\beta}^{(LS)} \in \arg\min_{\beta} \sum_{i=1}^{n} r_i^2, \quad r_i = y_i - x_i' \beta \]

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

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

\[ \hat{\beta}^{(LAD)} \in \arg\min_{\beta} \sum_{i=1}^{n} |r_i|, \]

- 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 [Rousseeuw ('84)]

\[
\hat{\beta}^{(\text{LMS})} \in \arg\min_{\beta} \left( \text{median}_{i=1,\ldots,n} |r_i| \right).
\]

- LMS highest possible breakdown point of almost 50%.

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

\[
\hat{\beta}^{(\text{LQS})} \in \arg\min_{\beta} |r(q)|,
\]

where, \( r(q) \) is the \( q \)th ordered absolute residual:

\[
|r(1)| \leq |r(2)| \leq \cdots \leq |r(n)|.
\]
Problem we address

- Solve the following problem:

\[
\min_\beta |r(q)|,
\]

where, \( r_i = y_i - x'_i \beta \), \( q \) is a quantile.

- Our approach extends to

\[
\min_\beta |r(q)|, \quad \text{s.t.} \quad A\beta \leq b \quad (\text{and/or} \quad \|\beta\|_2^2 \leq \delta)
\]
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:

\[
\min_{\beta} |r(q)| = \min_{\mathcal{I} \in \Omega_q} \left( \min_{\beta} \| y_{\mathcal{I}} - X_{\mathcal{I}} \beta \|_\infty \right),
\]

where, \( \Omega_q := \{ \mathcal{I} : \mathcal{I} \subset \{1, \ldots, n\}, |\mathcal{I}| = q \} \) and \((y_{\mathcal{I}}, X_{\mathcal{I}})\) denotes the subsample \((y_i, x_i), i \in \mathcal{I}\).
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:

If \( |r_i| \geq |r_{(q)}| \), then \( \overline{\mu}_i = 0, \mu_i \geq 0 \),

and if \( |r_i| \leq |r_{(q)}| \), then \( \mu_i = 0, \overline{\mu}_i \geq 0 \).
MIO Formulation

Notation:

\[ |r_{(1)}| \leq |r_{(2)}| \leq \cdots \leq |r_{(n)}|. \]

**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 \( \mu_i, \bar{\mu}_i \geq 0 \) such that:

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

with the conditions:

\[
\begin{array}{ll}
\text{If } |r_i| \geq |r_{(q)}|, & \text{then } \bar{\mu}_i = 0, \mu_i \geq 0, \\
\text{and if } |r_i| \leq |r_{(q)}|, & \text{then } \mu_i = 0, \bar{\mu}_i \geq 0.
\end{array}
\]

MIO representable
MIO Formulation

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

where $\gamma, z_i, \mu_i, \overline{\mu}_i, i = 1, \ldots, n$ are decision variables and $M_u, M_\ell$ are Big-M constants.
What do we achieve?

- Prior exact algorithms can solve up to: $n = 50$ and $p = 5$

- We obtain:
  - *near optimal* solutions for problems with $n \approx 200$ and $p \approx 20$ 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:

\[
\Sigma = L_1 L_1' + L_2 L_2' + \Phi \]

- \( \Sigma \approx \Theta + \Phi \)
- \( \Phi = \text{diag}(\Phi_1, \ldots, \Phi_p) \succeq 0 \)
- \( \text{rank}(\Theta) \leq r, \Theta \succeq 0 \)
- \( \Sigma - \Theta \succeq 0; \Sigma - \Phi \succeq 0 \)
Low Rank Factor Analysis (FA) [Spearman 1904]:
- widely used in multivariate statistics, econometrics, psychometrics
- represent correlation structure with few common (latent) factors.

Estimation Problem:
\[ \Sigma = \underbrace{\Theta}_{\text{Small}} + L_1L'_1 + L_2L'_2 + \Phi \]

\[ \begin{align*}
- \Sigma & \approx \Theta + \Phi \\
- \Phi & = \text{diag}(\Phi_1, \ldots, \Phi_p) \succeq 0 \\
- \text{rank}(\Theta) & \leq r, \Theta \succeq 0 \\
- \Sigma - \Theta & \succeq 0; \Sigma - \Phi \succeq 0
\end{align*} \]

\[ \min \| \Sigma - (\Theta + \Phi) \| \quad \text{s.t.} \quad \text{rank}(\Theta) \leq r, \Sigma - \Phi \succeq 0 \]
Our Approach

\[ \min \| \Sigma - (\Theta + \Phi) \| \]

s.t. \[ \text{rank}(\Theta) \leq r \] \hspace{1cm} (†)

\[ \Sigma - \Theta \succeq 0 \]
Our Approach

\[
\begin{align*}
\min & \quad \|\Sigma - (\Theta + \Phi)\| & \text{← Sum of Singular Values} \\
\text{s.t.} & \quad \text{rank}(\Theta) \leq r & \text{← Rank Constraint} \\
& \quad \Sigma - \Theta \succeq 0 & \text{← Semidefinite Constraint}
\end{align*}
\]
Our Approach

\[
\begin{align*}
\min & \quad \| \Sigma - (\Theta + \Phi) \| & \leftarrow \text{Sum of Singular Values} \\
\text{s.t.} & \quad \text{rank}(\Theta) \leq r & \leftarrow \text{Rank Constraint} \\
& \quad \Sigma - \Theta \succeq 0 & \leftarrow \text{Semidefinite Constraint}
\end{align*}
\]

- SDP with rank constraints

- **Key Idea**: Reformulate (†) 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

\[
\begin{align*}
\min & \quad \| \Sigma - (\Theta + \Phi) \| \\
\text{s.t.} & \quad \text{rank}(\Theta) \leq r \\
& \quad \Sigma - \Theta \succeq 0
\end{align*}
\]

\[
\begin{align*}
\uparrow & \quad \downarrow
\end{align*}
\]

\[
\begin{align*}
\left\{ \text{Variational Representation} \right. \\
\left. \text{of Spectral Functions} \right. 
\end{align*}
\]

\[
\begin{align*}
\min & \quad \langle W, \Sigma - \Theta \rangle - \sum_{i=1}^{p} w_{ii} \Phi_i \\
\text{s.t.} & \quad \mathbf{I} \succeq W \succeq 0 \\
& \quad \text{Tr}(W) = p - r \\
& \quad \Sigma - \Theta \succeq 0
\end{align*}
\]
Reformulation and tailored B&B

\[
\begin{align*}
\min & \quad \| \Sigma - (\Theta + \Phi) \| \\
\text{s.t.} & \quad \text{rank}(\Theta) \leq r \\
& \quad \Sigma - \Theta \succeq 0
\end{align*}
\]

\[
\begin{align*}
\Updownarrow & \quad \{ \text{Variational Representation} \\
& \quad \text{of Spectral Functions} \}
\end{align*}
\]

\[
\begin{align*}
\min & \quad \langle W, \Sigma - \Theta \rangle - \sum_{i=1}^{p} w_{ii} \Phi_i \\
\text{s.t.} & \quad I \succeq W \succeq 0 \\
& \quad \text{Tr}(W) = p - r \\
& \quad \Sigma - \Theta \succeq 0
\end{align*}
\]

\[
\begin{align*}
\Updownarrow & \quad \{ \text{Bilinear Form (Nonconvex)} \\
& \quad \text{McCormick Hulls/ B&B} \}
\end{align*}
\]
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.

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