A scalable algorithm for sparse and robust portfolios

Ryan Cory-Wright

November 2018

ORC, Massachusetts Institute of Technology
Joint work with Dimitris Bertsimas
Paper available on Optimization Online

Markowitz's approach to portfolio selection

Seek low-variance high-return, s.t. investment, sparsity constraints:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2\gamma}\boldsymbol{x}^\top \boldsymbol{x} + \frac{\sigma}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}\boldsymbol{x} - \boldsymbol{\mu}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{l} \le \boldsymbol{A}\boldsymbol{x} \le u,\ \boldsymbol{e}^\top \boldsymbol{x} = 1,\ ||\boldsymbol{x}||_0 \le k,\ \boldsymbol{x} \ge 0, \end{aligned}$$

- σ: robustness parameter, controls uncertainty in $\boldsymbol{\mu}$.

Markowitz's approach to portfolio selection

Seek low-variance high-return, s.t. investment, sparsity constraints:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2\gamma}\boldsymbol{x}^\top \boldsymbol{x} + \frac{\sigma}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} - \boldsymbol{\mu}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{l} \leq \boldsymbol{A}\boldsymbol{x} \leq u,\ \boldsymbol{e}^\top \boldsymbol{x} = 1,\ ||\boldsymbol{x}||_0 \leq k,\ \boldsymbol{x} \geq 0, \end{aligned}$$

- σ: robustness parameter, controls uncertainty in $\boldsymbol{\mu}$.
- γ: another robustness parameter, controls uncertainty in $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$.

Markowitz's approach to portfolio selection

Seek low-variance high-return, s.t. investment, sparsity constraints:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2\gamma}\boldsymbol{x}^\top \boldsymbol{x} + \frac{\sigma}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}\boldsymbol{x} - \boldsymbol{\mu}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{l} \leq \boldsymbol{A}\boldsymbol{x} \leq u,\ \boldsymbol{e}^\top \boldsymbol{x} = 1,\ ||\boldsymbol{x}||_0 \leq k,\ \boldsymbol{x} \geq 0, \end{aligned}$$

- σ: robustness parameter, controls uncertainty in $\boldsymbol{\mu}$.
- γ: another robustness parameter, controls uncertainty in $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$.
- k: sparsity parameter, prevents investing in entire market.

Why this problem?

Why include the sparsity constraint?

Three reasons:

- Controlling **transaction costs** is practically **relevant**.
- Managers incur **monitoring costs** for each **non-zero position** held.
- Portfolio optimization without cardinality constraints is viewed by customers as **indexing** while charging **active management** fees.

Why this approach?

The power of existing approaches

Existing exact approaches don't verify optimality for the Russell 1000.

Reference	Solution method	Largest instance (no. securities)
B+Shioda ('04)	Lemke pivot B&B	50
Frangioni and Gentile ('06)	Perspective cut	(1% gap) 200
Vielma et al. ('08)	Lifted B&B	100
Bonami and Lejeune ('09)	Nonlinear B&B	200
Gao and Li ('13)	Lagrangian relaxation B&B	300
Cui et al. ('13)	Lagrangian relaxation B&B	300
Vielma et al. ('17)	Lifted B&B	200

The power of existing approaches

Existing exact approaches don't verify optimality for the Russell 1000.

Reference	Solution method	Largest instance (no. securities)
B+Shioda ('04)	Lemke pivot B&B	50
Frangioni and Gentile ('06)	Perspective cut	(1% gap) 200
Vielma et al. ('08)	Lifted B&B	100
Bonami and Lejeune ('09)	Nonlinear B&B	200
Gao and Li ('13)	Lagrangian relaxation B&B	300
Cui et al. ('13)	Lagrangian relaxation B&B	300
Vielma et al. ('17)	Lifted B&B	200

Existing convex cardinality surrogates don't sparsify over the unit simplex.

How does the new approach work?

A new approach is needed

Our approach:

- Impose sparsity in a non-linear way.
- Apply duality to derive an efficient cutting-plane method.
- Solve the problem to certifiable optimality at scale.

Reformulation I: a regression perspective on portfolio selection

Portfolio selection equivalent to:

$$\begin{aligned}\min_{\boldsymbol{x}} \quad & \frac{1}{2\gamma}\boldsymbol{x}^\top \boldsymbol{x} + \frac{1}{2}||\boldsymbol{X}\boldsymbol{x} - \boldsymbol{Y}||_2^2 + \boldsymbol{d}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{l} \leq \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{u},\ \boldsymbol{e}^\top \boldsymbol{x} = 1,\ ||\boldsymbol{x}||_0 \leq k,\ \boldsymbol{x} \geq 0,\end{aligned}$$

where:

- $\boldsymbol{X} := \sqrt{\boldsymbol{\Sigma}}$,
- $\boldsymbol{Y}$: projection of $\boldsymbol{\mu}$ onto $\boldsymbol{X}$,
- $\boldsymbol{d}$: projection of $\boldsymbol{\mu}$ onto nullspace of $\boldsymbol{X}$.

Reformulation II: Impose l_0 via problem data

Big-M formulation is weak when linear, weaker when quadratic.
Stronger formulation (c.f. Bertsimas+Van Parys, 2017) given by:

$$\begin{aligned}\min_{\boldsymbol{z} | \boldsymbol{e}^\top \boldsymbol{z} \leq k} \min_{\boldsymbol{x}} \quad & \frac{1}{2\gamma}\sum_i z_i x_i^2 + \frac{1}{2}||\boldsymbol{Y} - \sum_i \boldsymbol{X}_i x_i z_i||_2^2 + \sum_i d_i x_i z_i \\ \text{s.t.} \quad & \boldsymbol{l} \leq \sum_i \boldsymbol{A}_i x_i z_i \leq u, \ \sum_i z_i x_i = 1, \ \boldsymbol{x} \geq 0,\end{aligned}$$

where z_i denotes if stock i is selected.

Reformulation III: take dual, reimpose l_0 cleverly

After lots of maths, sparse Markowitz becomes:

$$\min_{\boldsymbol{z} | \boldsymbol{e}^\top \boldsymbol{z} \leq k} \max_{\substack{\lambda, \boldsymbol{\alpha}, \\ \boldsymbol{\beta}_l, \boldsymbol{\beta}_u, \boldsymbol{\pi} \geq \mathbf{0}}} \quad -\frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{\gamma}{2}\sum_i z_i^2 w_i^2 + \boldsymbol{Y}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top \boldsymbol{l} - \boldsymbol{\beta}_u^\top \boldsymbol{u} + \lambda$$

$$\text{s.t.} \quad \boldsymbol{w} = \boldsymbol{X}^\top \alpha + \boldsymbol{\pi} + \lambda \boldsymbol{e} + \boldsymbol{A}^\top (\beta_l - \beta_u) - \boldsymbol{d}.$$

Reformulation III: take dual, reimpose l_0 cleverly

After lots of maths, sparse Markowitz becomes:

$$\min_{\boldsymbol{z} | \boldsymbol{e}^\top \boldsymbol{z} \leq k} \max_{\substack{\lambda, \boldsymbol{\alpha}, \\ \boldsymbol{\beta}_l, \boldsymbol{\beta}_u, \boldsymbol{\pi} \geq \mathbf{0}}} \quad -\frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{\gamma}{2} \sum_i z_i w_i^2 + \boldsymbol{Y}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top \boldsymbol{l} - \boldsymbol{\beta}_u^\top \boldsymbol{u} + \lambda$$

$$\text{s.t.} \quad \boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda \boldsymbol{e} + \boldsymbol{A}^\top (\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d}.$$

We substitute z_i for z_i^2 after taking dual. Without this, won't scale.

Reformulation III: take dual, reimpose l_0 cleverly

After lots of maths, sparse Markowitz becomes:

$$\min_{\boldsymbol{z} | \boldsymbol{e}^\top \boldsymbol{z} \leq k} \max_{\substack{\lambda, \boldsymbol{\alpha}, \\ \boldsymbol{\beta}_l, \boldsymbol{\beta}_u, \boldsymbol{\pi} \geq \boldsymbol{0}}} \quad -\frac{1}{2} \boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{\gamma}{2} \sum_i z_i w_i^2 + \boldsymbol{Y}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top \boldsymbol{l} - \boldsymbol{\beta}_u^\top \boldsymbol{u} + \lambda$$

$$\text{s.t.} \quad \boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda \boldsymbol{e} + \boldsymbol{A}^\top (\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d}.$$

We substitute z_i for z_i^2 after taking dual. Without this, won't scale.

So what?

A cutting-plane method

Our saddle-point representation:

$$\min_{\boldsymbol{z} | \boldsymbol{e}^\top \boldsymbol{z} \leq k} \max_{\substack{\lambda, \boldsymbol{\alpha}, \\ \boldsymbol{\beta}_l, \boldsymbol{\beta}_u, \boldsymbol{\pi} \geq \boldsymbol{0}}} \quad -\frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{\gamma}{2} \sum_i z_i w_i^2 + \boldsymbol{Y}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top \boldsymbol{l} - \boldsymbol{\beta}_u^\top \boldsymbol{u} + \lambda$$

$$\text{s.t.} \quad \boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda \boldsymbol{e} + \boldsymbol{A}^\top (\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d}.$$

- Let $f(\boldsymbol{z})$ be best payoff for support indices $\boldsymbol{z}$.

A cutting-plane method

Our saddle-point representation:

$$\min_{\boldsymbol{z} | \boldsymbol{e}^\top \boldsymbol{z} \leq k} \quad \max_{\substack{\lambda, \boldsymbol{\alpha}, \\ \boldsymbol{\beta}_l, \boldsymbol{\beta}_u, \boldsymbol{\pi} \geq \boldsymbol{0}}} \quad -\frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{\gamma}{2}\sum_i z_i w_i^2 + \boldsymbol{Y}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top \boldsymbol{l} - \boldsymbol{\beta}_u^\top \boldsymbol{u} + \lambda$$

$$\text{s.t.} \quad \boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda \boldsymbol{e} + \boldsymbol{A}^\top (\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d}.$$

- Let $f(\boldsymbol{z})$ be best payoff for support indices $\boldsymbol{z}$.
- Fix $\boldsymbol{z}$, solve a $k \times k$ QP to obtain $f(\boldsymbol{z})$.

A cutting-plane method

Our saddle-point representation:

$$\min_{\boldsymbol{z} \mid \boldsymbol{e}^\top \boldsymbol{z} \le k} \quad \max_{\substack{\lambda, \boldsymbol{\alpha}, \\ \boldsymbol{\beta}_l, \boldsymbol{\beta}_u, \boldsymbol{\pi} \ge \boldsymbol{0}}} \quad -\frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{\gamma}{2}\sum_i z_i w_i^2 + \boldsymbol{Y}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top \boldsymbol{l} - \boldsymbol{\beta}_u^\top \boldsymbol{u} + \lambda$$

$$\text{s.t.} \quad \boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda \boldsymbol{e} + \boldsymbol{A}^\top (\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d}.$$

- Let $f(\boldsymbol{z})$ be best payoff for support indices $\boldsymbol{z}$.
- Fix $\boldsymbol{z}$, solve a $k \times k$ QP to obtain $f(\boldsymbol{z})$.
- $\frac{\partial f(\boldsymbol{z})}{\partial z_i} = \frac{-\gamma}{2} w_i^2$ is valid subgradient, **even if** $z_i = 0$.

A cutting-plane method

Our saddle-point representation:

$$\min_{\boldsymbol{z} \mid \boldsymbol{e}^\top \boldsymbol{z} \leq k} \max_{\substack{\lambda, \boldsymbol{\alpha}, \\ \boldsymbol{\beta}_l, \boldsymbol{\beta}_u, \boldsymbol{\pi} \geq \mathbf{0}}} \quad -\frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{\gamma}{2}\sum_i z_i w_i^2 + \boldsymbol{Y}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top \boldsymbol{l} - \boldsymbol{\beta}_u^\top \boldsymbol{u} + \lambda$$

$$\text{s.t.} \quad \boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda \boldsymbol{e} + \boldsymbol{A}^\top (\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d}.$$

- Let $f(\boldsymbol{z})$ be best payoff for support indices $\boldsymbol{z}$.
- Fix $\boldsymbol{z}$, solve a $k \times k$ QP to obtain $f(\boldsymbol{z})$.
- $\frac{\partial f(\boldsymbol{z})}{\partial z_i} = \frac{-\gamma}{2} w_i^2$ is valid subgradient, **even if** $z_i = 0$.
- $f(z_1) \geq f(z_0) + \nabla f(z_0)^\top (z_1 - z_0)$ is a valid outer-approximation cut, obtained by solving one $k \times k$ subproblem.

A cutting-plane method

Our saddle-point representation:

$$\min_{\boldsymbol{z} | \boldsymbol{e}^\top \boldsymbol{z} \leq k} \max_{\substack{\lambda, \boldsymbol{\alpha}, \\ \boldsymbol{\beta}_l, \boldsymbol{\beta}_u, \boldsymbol{\pi} \geq \boldsymbol{0}}} \quad -\frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{\gamma}{2}\sum_i z_i w_i^2 + \boldsymbol{Y}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top \boldsymbol{l} - \boldsymbol{\beta}_u^\top \boldsymbol{u} + \lambda$$

$$\text{s.t.} \quad \boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda \boldsymbol{e} + \boldsymbol{A}^\top (\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d}.$$

- Let $f(\boldsymbol{z})$ be best payoff for support indices $\boldsymbol{z}$.
- Fix $\boldsymbol{z}$, solve a $k \times k$ QP to obtain $f(\boldsymbol{z})$.
- $\frac{\partial f(\boldsymbol{z})}{\partial z_i} = \frac{-\gamma}{2} w_i^2$ is valid subgradient, **even if** $z_i = 0$.
- $f(z_1) \geq f(z_0) + \nabla f(z_0)^\top (z_1 - z_0)$ is a valid outer-approximation cut, obtained by solving one $k \times k$ subproblem.
- Yields a very efficient cutting-plane method.
    - I.e., branch & cut with lazy constraint generation.
    - Need efficient warm-starts to make fast in practise; see paper.

A cutting-plane method

Our saddle-point representation:

$$\min_{\boldsymbol{z} \mid \boldsymbol{e}^\top \boldsymbol{z} \leq k} \max_{\substack{\lambda, \boldsymbol{\alpha}, \\ \boldsymbol{\beta}_l, \boldsymbol{\beta}_u, \boldsymbol{\pi} \geq \boldsymbol{0}}} \quad -\frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{\gamma}{2}\sum_i z_i w_i^2 + \boldsymbol{Y}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top \boldsymbol{l} - \boldsymbol{\beta}_u^\top \boldsymbol{u} + \lambda$$

$$\text{s.t.} \quad \boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda \boldsymbol{e} + \boldsymbol{A}^\top (\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d}.$$

- Let $f(\boldsymbol{z})$ be best payoff for support indices $\boldsymbol{z}$.
- Fix $\boldsymbol{z}$, solve a $k \times k$ QP to obtain $f(\boldsymbol{z})$.
- $\frac{\partial f(\boldsymbol{z})}{\partial z_i} = \frac{-\gamma}{2} w_i^2$ is valid subgradient, **even if** $z_i = 0$.
- $f(z_1) \geq f(z_0) + \nabla f(z_0)^\top (z_1 - z_0)$ is a valid outer-approximation cut, obtained by solving one $k \times k$ subproblem.
- Yields a very efficient cutting-plane method.
    - I.e., branch & cut with lazy constraint generation.
    - Need efficient warm-starts to make fast in practise; see paper.
- Can also exchange max, min operators to obtain a QCQP which yields a bound gap of $1\% - 5\%$ on real data; see paper.

A cutting-plane method

Our saddle-point representation:

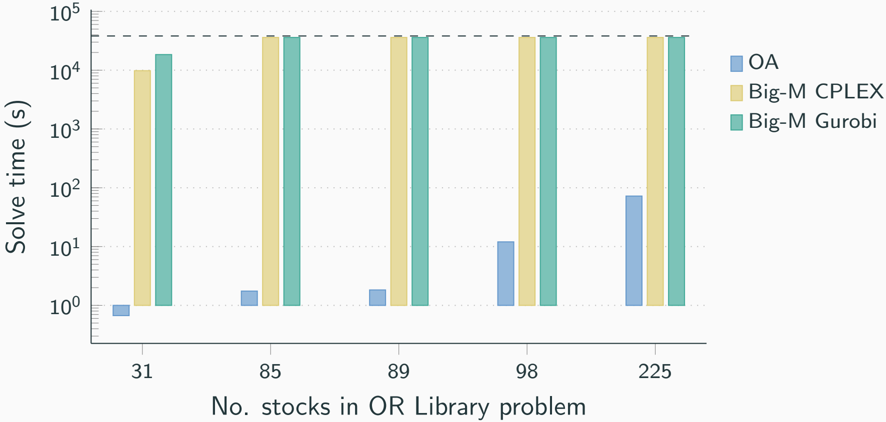
$$\min_{\boldsymbol{z} \mid \boldsymbol{e}^\top \boldsymbol{z} \le k} \max_{\substack{\lambda, \boldsymbol{\alpha}, \\ \boldsymbol{\beta}_l, \boldsymbol{\beta}_u, \boldsymbol{\pi} \ge \boldsymbol{0}}} \quad -\frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{\gamma}{2} \sum_i z_i w_i^2 + \boldsymbol{Y}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top \boldsymbol{l} - \boldsymbol{\beta}_u^\top \boldsymbol{u} + \lambda$$

$$\text{s.t.} \quad \boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda \boldsymbol{e} + \boldsymbol{A}^\top (\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d}.$$

- Let $f(\boldsymbol{z})$ be best payoff for support indices $\boldsymbol{z}$.
- Fix $\boldsymbol{z}$, solve a $k \times k$ QP to obtain $f(\boldsymbol{z})$.
- $\frac{\partial f(\boldsymbol{z})}{\partial z_i} = \frac{-\gamma}{2} w_i^2$ is valid subgradient, **even if** $z_i = 0$.
- $f(z_1) \ge f(z_0) + \nabla f(z_0)^\top (z_1 - z_0)$ is a valid outer-approximation cut, obtained by solving one $k \times k$ subproblem.
- Yields a very efficient cutting-plane method.
  - I.e., branch & cut with lazy constraint generation.
  - Need efficient warm-starts to make fast in practise; see paper.
- Can also exchange $\max, \min$ operators to obtain a QCQP which yields a bound gap of $1\% - 5\%$ on real data; see paper.

How does the approach perform on real data?

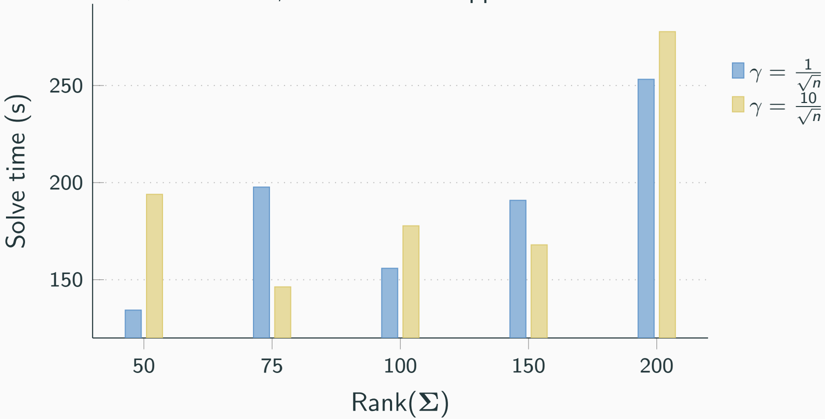
Chang et al. OR library problems; $\gamma = \frac{1}{\sqrt{n}}$, $\sigma = 2$, $k = 20$



- OA is 4 orders of magnitude faster than lifted polyhedral relaxation.
    - Dotted line=times out.

What about the S&P 500? Solve times for $\sigma = 15$, $k = 100$

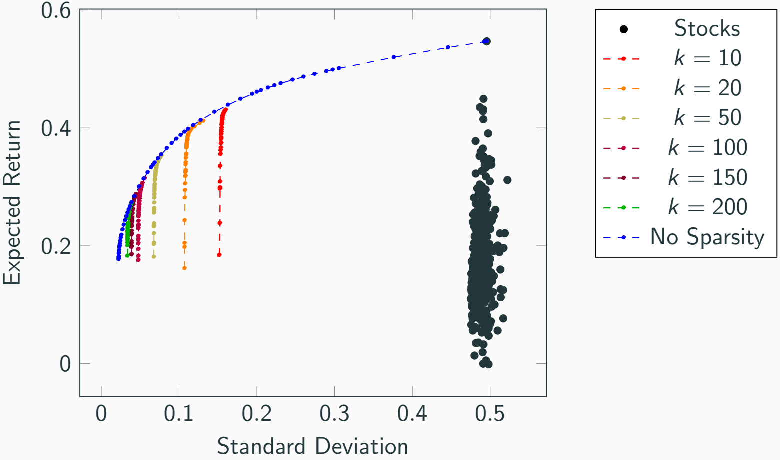
$\boldsymbol{\Sigma}$ unstable+not full-rank, need low-rank approximation.



What about the S&P 500? Efficient frontiers by cardinality

Fix $\gamma = \frac{100}{\sqrt{n}}$, $\text{rank}(\mathbf{\Sigma}) = 200$, vary σ, k.

Provably optimal frontiers+outcomes from investing entirely in one stock:



Comparison of Sparse Markowitz methods as of Nov 2018

Table 1: Largest instance reliably solved, by approach.

Reference	Solution method	Largest instance (no. securities)
B+Shioda ('09)	Lemke pivot B&B	50
Frangioni and Gentile ('06)	Perspective cut	(1% gap) 200
Vielma et al. ('08)	Lifted B&B	100
Bonami and Lejeune ('09)	Nonlinear B&B	200
Gao and Li ('13)	Lagrangian relaxation B&B	300
Cui et al. ('13)	Lagrangian relaxation B&B	300
Vielma et al. ('17)	Lifted B&B	200
B+C ('18)	Dual Branch-and-Cut	3,200

Summary

Contributions:

- A **tractable** nonlinear transformation which decouples the continuous and the discrete, and scales to real-world problems.
- **Scalable** to real-world data sets.
- **Generalizable** to other classes of problems, such as sparse regression with non-negativity constraints (Breiman, 1995).

See Jean's talk for an in-depth look at the method's performance on sparse regression.

Selected references

- Beasley, J.E.: Or-library: distributing test problems by electronic mail. J. Oper. Res. Soc. **41**(11), 1069–1072 (1990)
- Bertsimas, D., Pauphilet, J., Van Parys, B.: Sparse classification and phase transitions: a discrete optimization perspective (2017). J. Mach. Learn. Res.
- Bertsimas, D., Cory-Wright, R.: A scalable algorithm for sparse and robust portfolios. Oper. Res., Under Review (June 2018)
- Bertsimas, D., Van Parys, B.: Sparse high dimensional regression: Exact scalable algorithms and phase transitions (2016). Ann. Statist., Under Review.
- Breiman, L: Better subset regression using the nonnegative garrote. Techno. **37**(4), 373–384 (1995).
- Pilanci, M., Wainwright, M.J., El Ghaoui, L.: Sparse learning via boolean relaxations. Math. Program. **151**(1), 63–87 (2015)
- Vielma, J.P., Ahmed, S., Nemhauser, G.L.: A lifted linear programming branch-and-bound algorithm for mixed-integer conic quadratic programs. INFORMS J. Comput. **20**(3):438–450 (2008)

Thanks for listening!

Questions?

Appendix

Why don't existing approaches scale?

In existing approaches, sparsity is imposed via Big-M constraints.

Why don't existing approaches scale?

In existing approaches, sparsity is imposed via Big-M constraints.

But Big-M yields weak relaxations. Consider a simple example:

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & ||\boldsymbol{x}||_2^2 \\ \text{s.t.} \quad & \boldsymbol{e}^\top \boldsymbol{x} = 1, \\ & ||\boldsymbol{x}||_0 \leq k, \\ & \boldsymbol{x} \geq \boldsymbol{0}. \end{aligned}$$

This has an optimal value of $\frac{1}{k}$.

However, imposing and relaxing big-M ($M = 1$) gives a bound of $\frac{1}{n}$.

Naive B&B doesn't improve on this bound without complete enumeration.

Corollary: big-M MINLP methods won't scale.

Thanks for listening!

Questions?