A New Perspective on Low-Rank Optimization

Ryan Cory-Wright
Goldstine Postdoctoral Fellow @ IBM Research
Incoming Assistant Professor @ Imperial Biz & Imperial-X (July '23) ryancorywright.github.io
r.cory-wright@imperial.ac.uk

Joint work with
Dimitris Bertsimas (MIT)
Jean Pauphilet (LBS)

Motivation: What do these problems have in common?

Problem I: Sparse Linear Regression

- Given data about diabetes patients
- Predict each patient's hemoglobin measure in 1 year's time

ID	response	age	sex	bmi	map	tc
1	-0.0003157	0.00156562	-0.0027648	0.00290403	-0.0032677	-0.000206
2	-0.0025163	-0.0014594	-0.0027648	0.00347708	-0.0023934	-0.0026119
3	-0.0015465	0.00070132	0.00296107	-0.0012347	-0.0001743	0.00228293
4	-0.0030011	0.00329423	-0.0027648	-0.0041	-0.0040746	-4.01E-05

Dependent variable

Independent variables

- To avoid overfitting: restrict complexity, impose regularization

Motivation: What do these problems have in common?

Sparse Linear Regression

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \quad \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{1}{2\gamma}\|\boldsymbol{w}\|_2^2 + \mu\|\boldsymbol{w}\|_0$$

Restrict complexity

Explain data well on average

Regularize

Decision variables/Problem data

β: Sparse coefficient vector

Y: n obs of 1-dimensional outputs

X: n obs of p-dimensional inputs

Motivation: What do these problems have in common?

Problem II: Reduced Rank regression

- Predict weekly log-returns of all securities in S&P 500
- Given factors as inputs, e.g., gas prices, supply chain bottlenecks

- To avoid overfitting: restrict complexity, impose regularization

Motivation: What do these problems have in common?

Reduced Rank Regression

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p \times n}} \quad \frac{1}{2m} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_F^2 + \mu \cdot \mathrm{Rank}(\boldsymbol{\beta}),$$

Restrict complexity

Explain data well on average

Regularize

Decision variables and Problem data

β: Low-rank coefficient matrix

Y: m obs of n-dimensional outputs

X: m obs of p-dimensional inputs

Motivation: What do these problems have in common?

Sparse Linear Regression

$$\min_{\boldsymbol{w} \in \mathbb{R}^p} \quad \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{1}{2\gamma}\|\boldsymbol{w}\|_2^2 + \mu\|\boldsymbol{w}\|_0$$

Complexity is small

Reduced Rank Regression

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p \times n}} \quad \frac{1}{2m}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_F^2 + \mu \cdot \mathrm{Rank}(\boldsymbol{\beta}),$$

Decision variables/Problem data

β: Sparse coefficient vector

Y: n obs of 1-dimensional outputs

X: n obs of p-dimensional inputs

Decision variables and Problem data

β: Low-rank coefficient matrix

Y: m obs of n-dimensional outputs

X: m obs of p-dimensional inputs

The literature: Very little in common. Addressed

- in different application domains- medicine vs. finance
- by different communities- integer optimization vs. statistics
- using different algorithms- branch and cut vs. alternating minimization

Overview: A Tale of Two Constraints

Rank Constraints

Parsimony rank

Modeling constraint **X=YX**

Non-convex set $\mathbf{Y}^2 = \mathbf{Y}$ (Y projection matrix)

To be explicit:

$$\mathrm{Rank}(\mathbf{X}) \leq k \iff \exists \mathbf{Y} \in \mathcal{Y}_n : \mathrm{tr}(\mathbf{Y}) \leq k, \ \mathbf{X} = \mathbf{Y}\mathbf{X}$$

$$\mathcal{Y}_n := \{\mathbf{P} \in S^n : \mathbf{P}^2 = \mathbf{P}\}$$

Sparsity Constraints

Parsimony sparsity

Modeling constraint $x = zx$ ($x = 0$ if $z = 0$)

Non-convex set $z^2 = z$ (z binary)

To be explicit:

$$\|\boldsymbol{x}\|_0 \leq k \iff \exists \boldsymbol{z} \in \mathcal{Z}_n : \boldsymbol{e}^\top \boldsymbol{z} \leq k, \boldsymbol{x} = \boldsymbol{z} \circ \boldsymbol{x},$$

$$\mathcal{Z}_n := \{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{z} \circ \boldsymbol{z} = \boldsymbol{z}\}$$

Overview: A Tale of Two Constraints

Rank Constraints	Sparsity Constraints
Parsimony rank	Parsimony sparsity
Modeling constraint **X=YX**	Modeling constraint $x = zx$ ($x = 0$ if $z = 0$)
Non-convex set $\mathbf{Y}^2 = \mathbf{Y}$ (Y projection matrix)	Non-convex set $z^2 = z$ (z binary)
Applications rank regression, matrix completion, factor analysis, non-negative factorization	Applications sparse PCA, sparse portfolio selection, network design, unit commitment
Convex Relaxation matrix perspective, …?	Convex Relaxation perspective, 2x2 convexifications,…

Main contribution of talk: Build bridge from MIO to rank constraints, leverage MIO marketplace of ideas to design strong low-rank relaxations

Main message from talk: Projection matrices are key ingredient to, for first time, develop strong lower bounds for low-rank problems & even solve them to optimality

Linear Regression and Relaxations Revisited

Sparse Linear Regression: Fit interpretable model using small number of features

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \quad \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2+\frac{1}{2\gamma}\|\boldsymbol{w}\|_2^2+\mu\|\boldsymbol{w}\|_0$$

Perspective Reformulation (Frangioni and Gentile 2006, Günlük and Linderoth 2010)-strong & scalable

$$\min_{\boldsymbol{w},\boldsymbol{\rho}\in\mathbb{R}^p,\boldsymbol{z}\in\{0,1\}^p} \quad \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2+\frac{1}{2\gamma}\boldsymbol{e}^\top\boldsymbol{\rho}+\mu\cdot\boldsymbol{e}^\top\boldsymbol{z} \quad \text{s.t.} \quad z_i\rho_i\geq w_i^2 \quad \forall i\in[p].$$

Allows exact solutions with $p=10^7$ features (Bertsimas and van Parys 2020, Hazimeh and Mazumder 2021)

Further improvements seem possible, e.g., convexifications by Atamturk/Gomez, De Rosa/Khajavirad

Can we play same game in low-rank case?

Literature Review

Exact methods

Branch and bound: Lee and Zou (2014), Kocuk, Dey and Sun (2017), Bertsimas, Copenhaver and Mazumder (2017)
Complementarity: Bi, Pan and Sun (2020)
Sum-of-Squares: d'Aspremont (2004), Naldi (2018)

Convex relaxations

Nuclear norm: Shapiro (1982), Fazel (2002), Candès and Recht (2009), Recht, Fazel and Parrilo (2010)
Log determinant: Fazel (2002)
Nuclear plus Frobenius norm: Mazumder, Hastie and Tibshirani (2010), Cai, Candès and Shen (2010)
Nuclear plus L1 norm: Chandrasekaran, Sanghavi, Parrilo and Willsky (2011), Agarwal, Negahban and Wainwright (2012)
Second-order cone: Kim and Kojima (2003), Lavaei and Low (2012), Ahmadi and Majumdar (2019)

Heuristics

Rounding: Goemans and Williamson (1995), Nesterov (1998), Nemirovski, Roos and Terlaky (1999), So, Ye and Zhang (2007)
Alternating minimization: Burer and Monteiro (2003, 2005), Jain (2013), Boumal, Voroninski and Banderia (2016), Waldspurger and Waters (2020)
Augmented Lagrangian: Yurtsever, Tropp, Fercoq, Udell and Cevher (2021)
Stochastic gradient descent: Recht and Ré (2013)
Frank-Wolfe: Freund, Grigas and Mazumder (2017)
Sketching: Tropp, Yurtsever, Udell and Cevher (2017)
Subgradient: Charisopoulos, Chen, Davis, Diaz, Ding and Drusvyatskiy (2021)
Non-convex penalties: Mazumder, Saldana and Weng (2020), Sagan and Mitchell (2021)

no clear generalization to reduced rank regression in literature

Summary of State of Literature

- With heuristics, obtain high-quality solutions quickly
- But-excluding special cases-no guarantees on quality

  All known algorithms which provide exact solutions [for matrix completion] require time doubly exponential in the dimension n of the matrix in both theory and practice-Candès and Recht (2009)

  - Translation: Completely intractable even for n=10
  - Corollary: Solving low-rank matrix completion problems at all would be very impressive!

- Moreover, “convex relaxations” don’t give valid lower bounds
  - They involve replacing a rank term in the objective with a nuclear norm.
- Can we do better?

Rank Regression and Relaxations

Reduced Rank Regression: Fit **interpretable** model using small number of singular values

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p\times n}} \quad \frac{1}{2m}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_F^2 + \mu\cdot \mathrm{Rank}(\boldsymbol{\beta})$$

Matrix Perspective Relaxation (new): Apply Matrix Perspective Reformulation Technique

Bertsimas, C., and Pauphilet (2021) Equation (6)

The following matrix perspective relaxation is a valid relaxation for reduced rank regression:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p\times n}, \boldsymbol{W}\in\mathcal{S}^n_+, \boldsymbol{\theta}\in\mathcal{S}^p_+} \quad \frac{1}{2m}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \frac{1}{2\gamma}\mathrm{tr}(\boldsymbol{\theta}) + \mu\cdot\mathrm{tr}(\boldsymbol{W}) \quad \text{s.t.} \quad \boldsymbol{W}\preceq\mathbb{I}, \begin{pmatrix}\boldsymbol{\theta} & \boldsymbol{\beta}\\ \boldsymbol{\beta}^\top & \boldsymbol{W}\end{pmatrix}\succeq \boldsymbol{0}.$$

We derive relaxation as worked example halfway through talk

Modeling Rank with Projection Matrices

Sparsity constraints can be modeled using binary variables

$$\|\boldsymbol{x}\|_0 \leq k \quad \Longleftrightarrow \quad \exists \boldsymbol{z} \in \mathcal{Z}_n : \boldsymbol{e}^\top \boldsymbol{z} \leq k, \boldsymbol{x} = \boldsymbol{z} \circ \boldsymbol{x},$$

Proof: Take $z_i = 1$ if $x_i \neq 0$, 0 otherwise

Rank constraints can be modeled using projection matrices

$$\mathrm{Rank}(\mathbf{X}) \leq k \iff \exists \mathbf{Y} \in \mathcal{Y}_n : \mathrm{tr}(\mathbf{Y}) \leq k, \ \mathbf{X} = \mathbf{Y}\mathbf{X}$$

where $\mathcal{Y}_n := \{\mathbf{P} \in S^n : \mathbf{P}^2 = \mathbf{P}\}$

*Proof: Take **Y** the orthogonal projection onto the span of **X***

Mixed-Projection Conic Optimization: A New Paradigm for Modeling Rank Constraints
D. Bertsimas, R. Cory-Wright, J. Pauphilet, Operations Research, 2021.

- Winner, 2020 INFORMS George Nicholson Best Paper Competition

Contributions

A New Perspective on Low-Rank Optimization
D. Bertsimas, **R. Cory-Wright**, J. Pauphilet, minor revision at Mathematical Programming, 2022.

Methodological: We propose a **simple preprocessing technique** which gives **strong & scalable** bounds for low-rank problems. Generalizes perspective reformulation technique from MIO

Theoretical: We invoke technique to **explicitly characterize** convex hulls of simple low-rank sets

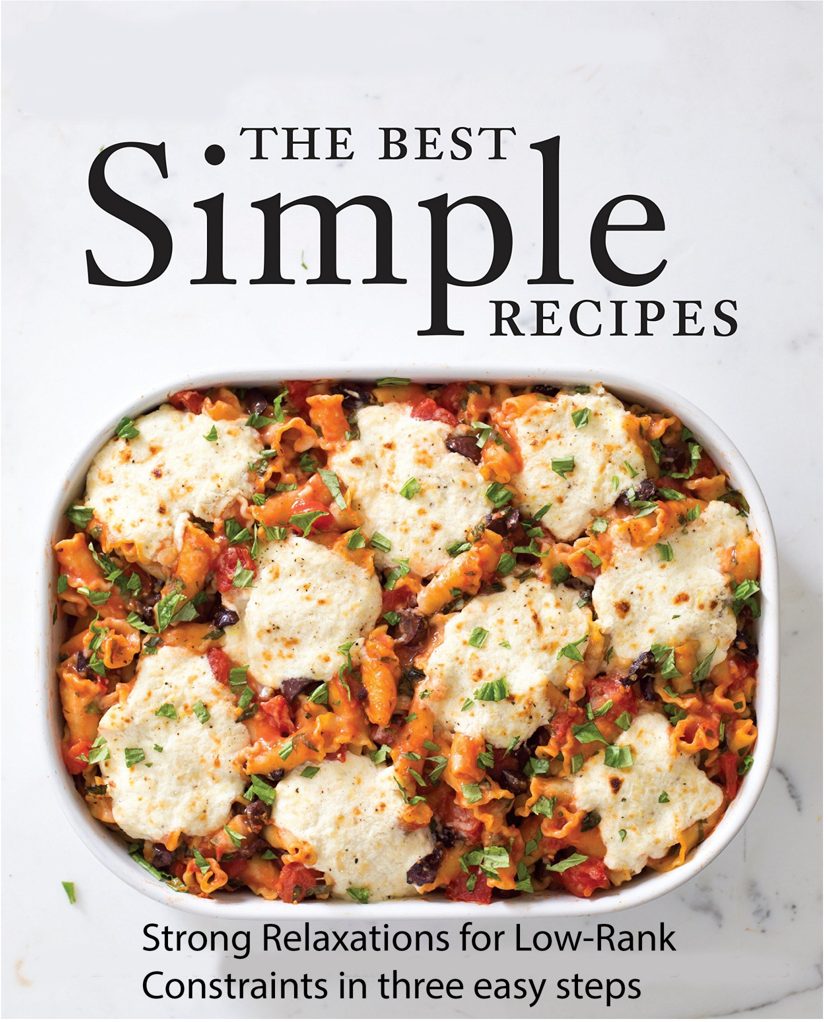
Algorithmic: We demonstrate technique's **efficacy** across diverse range of low-rank problems

Matrix Perspective Reformulation Technique: Recipe

1. Consider low-rank problem with regularization

2. Formulate as mixed-projection optimization problem

3. Take matrix perspective of regularizer



Matrix Perspective Reformulation Technique I: Regularization

Consider low-rank problem with spectral regularization

$$\min_{\boldsymbol{X} \in \mathcal{S}^n_+} \langle \boldsymbol{C}, \boldsymbol{X} \rangle + \boxed{\Omega(\boldsymbol{X})} + \mu \cdot \operatorname{Rank}(\boldsymbol{X}) \text{ s.t. } \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle = b_i \ \forall i \in [m], \ \boldsymbol{X} \in \mathcal{K}, \ \operatorname{Rank}(\boldsymbol{X}) \leq k,$$

Where:

- $\Omega(\boldsymbol{X}) := \sum_{i=1}^n \omega(\lambda_i(\boldsymbol{X})) = \operatorname{tr}(f(X))$ with ω univariate convex; f matrix convex generalization of ω
- Example: ridge regularization in regression
  - $\omega(\lambda) = \frac{1}{2\gamma}\lambda^2, \ \ \Omega(X) = \frac{1}{2\gamma}\sum_{i=1}^n \lambda_i(X)^2 = \frac{1}{2\gamma}\|X\|_F^2 = \frac{1}{2\gamma}\operatorname{tr}(X^T X)$

Matrix Perspective Reformulation Technique II: Formulation

Low-rank problem

$$\min_{\boldsymbol{X}\in\mathcal{S}^n_+} \langle \boldsymbol{C}, \boldsymbol{X}\rangle + \boxed{\Omega(\boldsymbol{X})} + \boxed{\mu \cdot \mathrm{Rank}(\boldsymbol{X})} \text{ s.t. } \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle = b_i \ \forall i \in [m], \ \boldsymbol{X} \in \mathcal{K}, \ \boxed{\mathrm{Rank}(\boldsymbol{X}) \leq k}$$

can be formulated as Mixed-Projection Optimization problem

$$\begin{aligned} \boxed{\min_{\boldsymbol{Y}\in\mathcal{Y}^k_n}} \min_{\boldsymbol{X}\in\mathcal{S}^n_+} \quad & \langle \boldsymbol{C}, \boldsymbol{X}\rangle + \boxed{\mu \cdot \mathrm{tr}(\boldsymbol{Y})} + \boxed{\mathrm{tr}(f(\boldsymbol{X}))} \\ \text{s.t.} \quad & \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle = b_i \quad \forall i \in [m], \ \boxed{\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{X},} \ \boldsymbol{X} \in \mathcal{K} \end{aligned}$$

where $\boldsymbol{Y}$ is a projection matrix

Matrix Perspective Reformulation Technique III: Reformulation

Mixed-Projection Conic Optimization problem

$$\min_{\boldsymbol{Y} \in \mathcal{Y}_n^k} \min_{\boldsymbol{X} \in \mathcal{S}_+^n} \quad \langle \boldsymbol{C}, \boldsymbol{X} \rangle + \mu \cdot \mathrm{tr}(\boldsymbol{Y}) + \boxed{\mathrm{tr}(f(\boldsymbol{X}))}$$
$$\text{s.t.} \quad \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle = b_i \quad \forall i \in [m], \boxed{\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{X},} \boldsymbol{X} \in \mathcal{K}.$$

Rewrite as equivalent problem which gives stronger relaxations

$$\min_{\boldsymbol{Y} \in \mathcal{Y}_n^k} \min_{\boldsymbol{X} \in \mathcal{S}_+^n} \quad \langle \boldsymbol{C}, \boldsymbol{X} \rangle + \mu \cdot \mathrm{tr}(\boldsymbol{Y}) + \boxed{\mathrm{tr}(g_f(\boldsymbol{X}, \boldsymbol{Y})) + (n - \mathrm{tr}(\boldsymbol{Y}))\omega(0)}$$
$$\text{s.t.} \quad \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle = b_i \quad \forall i \in [m], \ \boldsymbol{X} \in \mathcal{K},$$

where g_f, matrix perspective of f (Effros, 2009; Ebadian et al., 2011), is jointly convex in X,Y!

$$g_{f_\omega}(\boldsymbol{\beta}, \boldsymbol{P}) = \begin{cases} \boldsymbol{P}^{\frac{1}{2}} f_\omega \left(\boldsymbol{P}^{-\frac{1}{2}} \boldsymbol{\beta} \boldsymbol{P}^{-\frac{1}{2}} \right) \boldsymbol{P}^{\frac{1}{2}} & \text{if } \boxed{\mathrm{Span}(\boldsymbol{\beta}) \subseteq \mathrm{Span}(\boldsymbol{P})} \\ \infty & \text{otherwise} \end{cases}$$

Captures the bilinear constraint β=Pβ

Matrix Perspective Reformulation Technique IV: Relaxations

Mixed-Projection Conic Optimization relaxation very weak!

$$\min_{\boldsymbol{Y} \in \mathrm{Conv}(\mathcal{Y}_n^k)} \min_{\boldsymbol{X} \in \mathcal{S}_+^n} \quad \langle \boldsymbol{C}, \boldsymbol{X} \rangle + \mu \cdot \mathrm{tr}(\boldsymbol{Y}) + \mathrm{tr}(f(\boldsymbol{X}))$$

$$\text{s.t.} \quad \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle = b_i \quad \forall i \in [m], \ \boldsymbol{X} \in \mathcal{K},$$

Perspectified relaxation much stronger

$$\min_{\boldsymbol{Y} \in \mathcal{Y}_n^k} \min_{\boldsymbol{X} \in \mathcal{S}_+^n} \quad \langle \boldsymbol{C}, \boldsymbol{X} \rangle + \mu \cdot \mathrm{tr}(\boldsymbol{Y}) + \mathrm{tr}(g_f(\boldsymbol{X}, \boldsymbol{Y})) + (n - \mathrm{tr}(\boldsymbol{Y}))\omega(0)$$

Relax to convex hull

$$\text{s.t.} \quad \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle = b_i \quad \forall i \in [m], \ \boldsymbol{X} \in \mathcal{K},$$

Matrix Perspective Reformulation Technique IV: Relaxations

Mixed-Projection Conic Optimization relaxation very weak!

$$\min_{\boldsymbol{Y} \in \mathrm{Conv}(\mathcal{Y}_n^k)} \min_{\boldsymbol{X} \in \mathcal{S}_+^n} \quad \langle \boldsymbol{C}, \boldsymbol{X} \rangle + \mu \cdot \mathrm{tr}(\boldsymbol{Y}) + \mathrm{tr}(f(\boldsymbol{X}))$$
$$\text{s.t.} \quad \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle = b_i \quad \forall i \in [m], \ \boldsymbol{X} \in \mathcal{K},$$

Perspectified relaxation much stronger

$$\min_{\boldsymbol{Y} \in \mathrm{Conv}(\mathcal{Y}_n^k)} \min_{\boldsymbol{X} \in \mathcal{S}_+^n} \quad \langle \boldsymbol{C}, \boldsymbol{X} \rangle + \mu \cdot \mathrm{tr}(\boldsymbol{Y}) + \mathrm{tr}(g_f(\boldsymbol{X}, \boldsymbol{Y}) + (n - \mathrm{tr}(\boldsymbol{Y}))\omega(0)$$
$$\text{s.t.} \quad \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle = b_i \quad \forall i \in [m], \ \boldsymbol{X} \in \mathcal{K},$$

Questions on the recipe?

Matrix Perspective Reformulation: Worked Example

Reduced Rank Regression: Fit interpretable model using small number of singular values

Step 1: Consider problem with spectral regularization:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p \times n}} \quad \frac{1}{2m}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \boxed{\frac{1}{2\gamma}\|\boldsymbol{\beta}\|_F^2} + \mu \cdot \mathrm{Rank}(\boldsymbol{\beta})$$

Where $\Omega(X) = \frac{1}{2\gamma}\sum_{i=1}^{n} \lambda_i(\beta)^2 = \frac{1}{2\gamma}||\beta||_F^2$

Matrix Perspective Reformulation: Worked Example

Reduced Rank Regression: Fit **interpretable** model using small number of singular values

Step 2: Formulate as Mixed-Projection problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p \times n}, \boldsymbol{W} \in \mathcal{Y}_n^n} \frac{1}{2m}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_F^2 + \mu \cdot \mathrm{tr}(\boldsymbol{W}), \boldsymbol{W} = \boldsymbol{\beta}\boldsymbol{W}$$

Where $\mathcal{Y}_n := \{\mathbf{P} \in S^n : \mathbf{P}^2 = \mathbf{P}\}$ is set of $n \times n$ projection matrices

Matrix Perspective Reformulation: Worked Example

Reduced Rank Regression: Fit interpretable model using small number of singular values

Step 3: Reformulate by taking matrix perspective

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p\times n}, \boldsymbol{W}\in\mathcal{S}^n_+, \boldsymbol{\theta}\in S^p_+} \quad \frac{1}{2m}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \boxed{\frac{1}{2\gamma}\mathrm{tr}(\boldsymbol{\theta})} + \mu\cdot\mathrm{tr}(\boldsymbol{W}) \quad \text{s.t.} \quad \boldsymbol{W}\preceq\mathbb{I}, \boxed{\begin{pmatrix}\boldsymbol{\theta} & \boldsymbol{\beta}\\ \boldsymbol{\beta}^\top & \boldsymbol{W}\end{pmatrix}\succeq\mathbf{0}}$$

Questions on the worked example?

Theoretical Contribution: Convex Hulls of Low-Rank Sets

Bertsimas, Cory-Wright, and Pauphilet (21+): Theorem 2

Let T denote epigraph of spectral function under rank constraints:

$$\mathcal{T} = \{\boldsymbol{X} \in \mathcal{S}^n_+ : \mathrm{tr}(f(\boldsymbol{X})) + \mu \cdot \mathrm{Rank}(\boldsymbol{X}) \leq t, \mathrm{Rank}(\boldsymbol{X}) \leq k\}$$

$\omega(\cdot)$ scalar convex function such that $tr(f(X)) = \sum_{i=1}^{n} \omega(\lambda_i(X))$ for matrix convex f

Then, extended formulation of convex hull of T given by:

$$\mathcal{T}^c = \{(\boldsymbol{X}, \boldsymbol{Y}) \in \mathcal{S}^n_+ \times \mathrm{Conv}(\mathcal{Y}^k_n) : \mathrm{tr}(g_f(\boldsymbol{X}, \boldsymbol{Y})) + \mu \cdot \mathrm{tr}(\boldsymbol{Y}) + (n - \mathrm{tr}(\boldsymbol{Y}))\omega(0) \leq t\}$$

Where:

- g_f matrix perspective of f
- $\mathrm{Conv}(\mathcal{Y}^k_n) = \{\boldsymbol{Y} \in S^n_+ : \boldsymbol{Y} \preceq \mathbb{I}, \mathrm{tr}(\boldsymbol{Y}) \leq k\}$ is convex hull of rank-k projection matrices.

Matrix perspective reformulation gives convex hull of simple low-rank sets

Application: Proof SVD is Convex Opt in Lifted Space

Eckart-Mirsky-Young Theorem

The following "non-convex" optimization problem is exactly solvable via a top-k SVD

$$\min_{\boldsymbol{X} \in \mathbb{R}^{n \times m}} \quad \|\boldsymbol{X} - \boldsymbol{A}\|_F^2 : \ \mathrm{Rank}(\boldsymbol{X}) \leq k$$

Bertsimas, C., Pauphilet (2021b) pp16

The following two optimization problems attain the same optimal value:

$$\min_{\boldsymbol{X} \in \mathbb{R}^{n \times m}} \quad \|\boldsymbol{X} - \boldsymbol{A}\|_F^2 : \ \mathrm{Rank}(\boldsymbol{X}) \leq k$$

$$\min_{\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta}} \quad \frac{1}{2}\mathrm{tr}(\boldsymbol{\theta}) - \langle \boldsymbol{A}, \boldsymbol{X} \rangle + \frac{1}{2}\|\boldsymbol{A}\|_F^2 \ \text{s.t.} \ \boldsymbol{Y} \preceq \mathbb{I}, \ \mathrm{tr}(\boldsymbol{Y}) \leq k, \begin{pmatrix} \boldsymbol{\theta} & \boldsymbol{X} \\ \boldsymbol{X}^\top & \boldsymbol{Y} \end{pmatrix} \succeq \boldsymbol{0}.$$

Suggests that if Y*, solution to relaxation, is not proj matrix then we should round via top-k SVD

Approximate Solutions via Greedily Rounding Relaxation

Consider Y* solution to relaxation.

If Y* already projection matrix, relaxation tight, otherwise:

1. Greedily round Y* via top-k SVD -> obtain Y

2. Solve for X under constraint $X = YX$

Conclusion: If f(Y) Lipschitz continuous, greedy near optimal in theory and practice.

Application I: Reduced Rank Regression

Formulation

$$\min_{\beta \in \mathbb{R}^{p \times n}} \quad \frac{1}{2m}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_F^2 + \mu \cdot \mathrm{Rank}(\boldsymbol{\beta}),$$

Decision variables/Problem data

β: Low-rank coefficient matrix

Y: Matrix of outputs

X: Matrix of inputs

Portfolio Selection: Predict Weekly Log-Returns of Each Security in S&P 500

- Given many factors as inputs, e.g., gas prices, supply chain bottlenecks

- To avoid overfitting, restrict complexity of models, regularize.

Reminder: Rank Regression and Relaxations

Reduced Rank Regression: Fit **interpretable** model using small number of singular values

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p\times n}} \quad \frac{1}{2m}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_F^2+\frac{1}{2\gamma}\|\boldsymbol{\beta}\|_F^2+\mu\cdot\mathrm{Rank}(\boldsymbol{\beta})$$

Bertsimas, C., and Pauphilet (2021) Equation (6)

The following matrix perspective relaxation is a valid relaxation for reduced rank regression:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p\times n},\boldsymbol{W}\in\mathcal{S}_+^n,\boldsymbol{\theta}\in S_+^p} \quad \frac{1}{2m}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_F^2+\frac{1}{2\gamma}\mathrm{tr}(\boldsymbol{\theta})+\mu\cdot\mathrm{tr}(\boldsymbol{W}) \quad \text{s.t.} \quad \boldsymbol{W}\preceq\mathbb{I}, \begin{pmatrix}\boldsymbol{\theta} & \boldsymbol{\beta}\\ \boldsymbol{\beta}^\top & \boldsymbol{W}\end{pmatrix}\succeq\mathbf{0}.$$

We refer to this relaxation as the “Matrix Perspective” relaxation

An Even Stronger Relaxation

(Dong, Chen and Linderoth, 2015): In sparse linear regression, apply perspective relaxation to "natural" separable regularizer, plus "extra" diagonal term extracted from matrix $X^T X$. Gives stronger relaxations!

Saddle-Point Rank Relaxation (new): Use same approach in low-rank case

Bertsimas, C., and Pauphilet (2021) Equation (7)

The following matrix perspective relaxation is a valid relaxation for reduced rank regression:

$$\min_{\boldsymbol{\theta}\in\mathcal{S}^n_+,\boldsymbol{\beta}\in\mathbb{R}^{p\times n},\boldsymbol{B}\in\mathcal{S}^n_+,\boldsymbol{W}\in\mathcal{S}^n_+} \quad \frac{1}{2m}\|\boldsymbol{Y}\|_F^2 - \frac{1}{m}\langle \boldsymbol{Y}, \boldsymbol{X}\boldsymbol{\beta}\rangle + \frac{1}{2}\langle \boldsymbol{B}, \frac{1}{\gamma}\mathbb{I} + \frac{1}{m}\boldsymbol{X}^\top\boldsymbol{X}\rangle + \mu\cdot\mathrm{tr}(\boldsymbol{W})$$

$$\text{s.t.} \quad \begin{pmatrix} \boldsymbol{B} & \boldsymbol{\beta} \\ \boldsymbol{\beta} & \boldsymbol{W} \end{pmatrix} \succeq \mathbf{0}, \boldsymbol{W} \preceq \mathbb{I}.$$

We refer to this relaxation as the "DCL" relaxation

Application I: Reduced Rank Regression

Example:

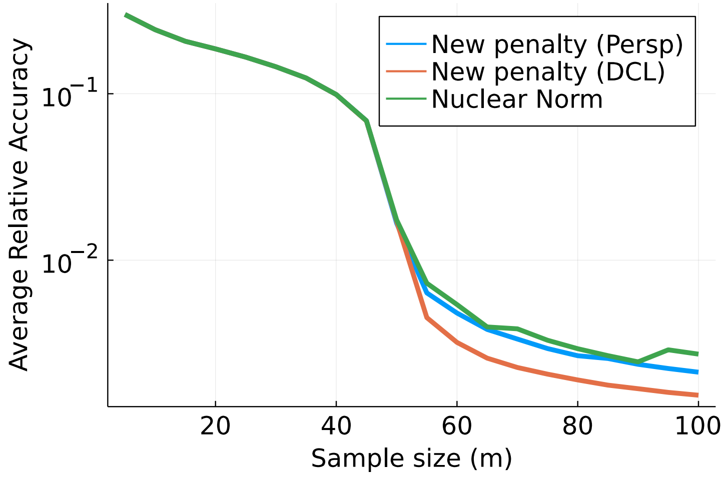
Recover rank-10 50 x m matrix:

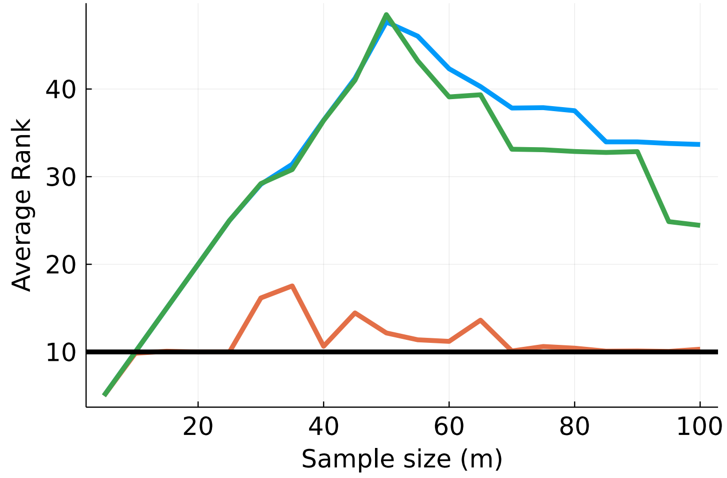
- Vary m, measure MSE, rank from relaxations
- Compare against nuclear norm

- Matrix perspective dominates nuclear norm
- DCL more accurate than matrix perspective or NN, recovers true rank

- DCL w. Mosek solves for 300x300 matrices on Macbook Pro in minutes, NN takes hours for 150x150.

Code available on GitHub:
ryancorywright/MatrixPerspectiveSoftware





Application II: Matrix Completion

Formulation:

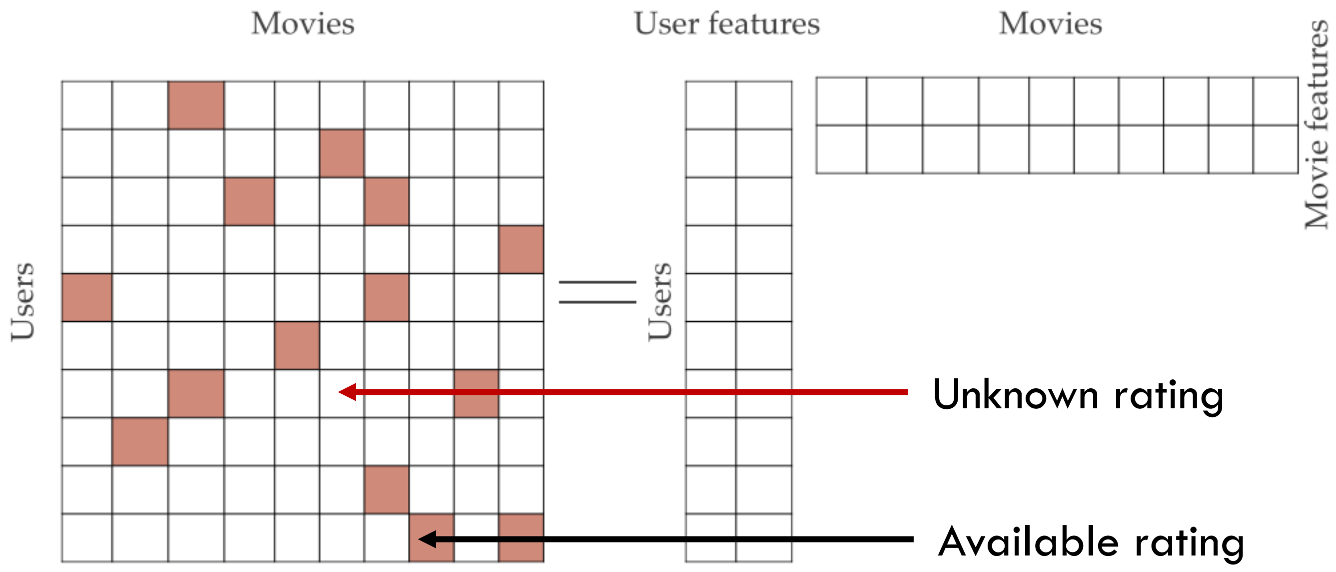
$$\min_{\boldsymbol{X}\in\mathbb{R}^{n\times p}} \quad \frac{1}{2}\sum_{(i,j)\in\mathcal{I}} (X_{i,j} - A_{i,j})^2 \quad \text{s.t.} \quad \text{Rank}(\boldsymbol{X}) \leq k.$$

Decision variables/Problem data

$X_{i,j}$: Predicted rating movie j by user i
$A_{i,j}$: Reported rating movie j by user i

Movie Recommendation:

- Given user movie ratings, predict ratings for unseen movies.
- To make problem tractable, assume ratings depend on k factors (lead actor, lead actress, director, genre, year, ..)



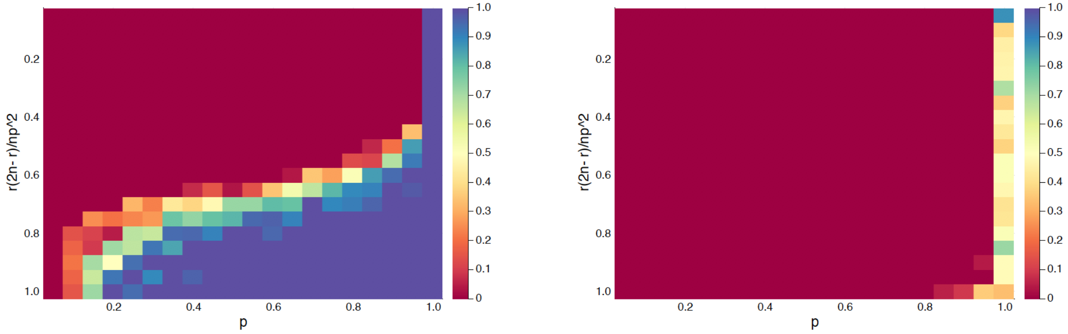
Application II: Matrix Completion

Example:

Recover low-rank **100x100** matrix:
- Vary rank, proportion entries sampled
- Measure % time recover matrix to 1% MSE (more purple=better)

- Nuclear norm *by far* worst approach

- New penalty better, new penalty with rounding much better



(a) New Penalty (b) Nuclear Norm

Sum of singular values

Application II: Matrix Completion

Example:

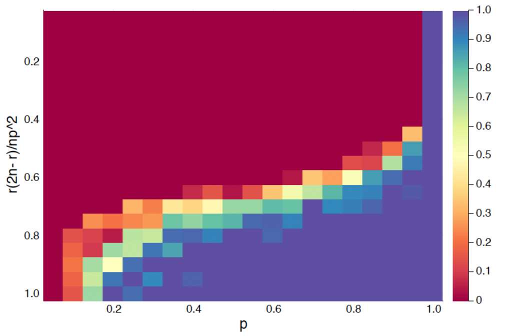
Recover low-rank **100x100** matrix:

- Vary rank, proportion entries sampled
- Measure % time recover matrix to 1% MSE (more purple=better)

- Nuclear norm *by far* worst approach

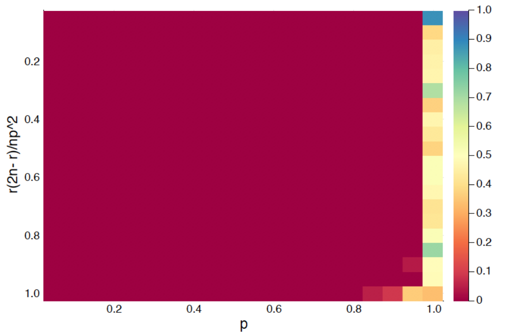
- New penalty better, new penalty with rounding much better

- Code available on GitHub
[ryancorywright/MixedProjectionSoftware](https://github.com/ryancorywright/MixedProjectionSoftware)



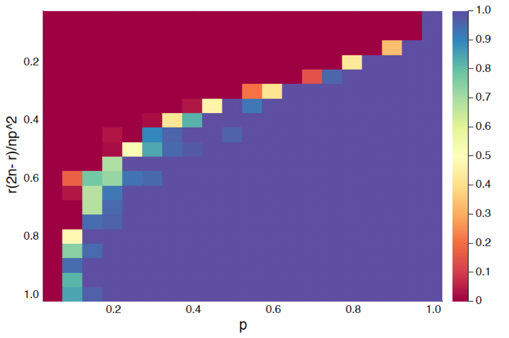
(a) New Penalty

Avg MSE: 0.161



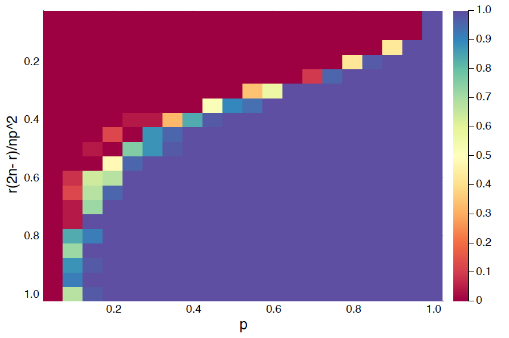
(b) Nuclear Norm

Avg MSE: 0.181



(c) SVD+Local Improvement

Avg MSE: 0.147



(d) New Penalty+Local Improvement

Avg MSE: 0.054

In practice, new penalty is viable and *often* more accurate

Conclusion

Matrix perspective is natural generalization of perspective reformulation

- Exploit separability of eigenvalues to obtain "embarrassingly tight" formulation.
- Leads to relaxations which outperform state-of-the-art for central problems in OR/ML.
- Suggests this is a very general story, often useful to think about problems this way.

Two future directions:

1. Writing a book → *Integer and Matrix Optimization: A Nonlinear Approach*
2. Branch-and-bound → perspective relax eventually lead to B&B which solves sparse regression at scale. Similar approach for matrix completion in progress

Thank you for listening!
Lingering questions?
Email r.cory-wright@imperial.ac.uk

Selected References I

A Unified Approach to Mixed-Integer Optimization Problems With Logical Constraints
D. Bertsimas, **R. Cory-Wright**, J. Pauphilet, SIAM Journal on Optimization **31**(3): 2340-2367, 2021.
- **Winner, 2019 INFORMS Computing Society Best Student Paper Competition**

Mixed-Projection Conic Optimization: A New Paradigm for Modeling Rank Constraints
D. Bertsimas, **R. Cory-Wright**, J. Pauphilet, Operations Research, accepted, 2021.
- **Winner, 2020 INFORMS George Nicholson Best Paper Competition**

A New Perspective on Low-Rank Optimization
D. Bertsimas, **R. Cory-Wright**, J. Pauphilet, major revision at Mathematical Programming, 2022.

A Nonlinear Programming Algorithm for Solving Semidefinite Programs via Low-Rank Factorization
S. Burer, R. Monteiro, Mathematical Programming **95** 329-357, 2003

Trace Inequalities and Quantum Entropy: An Introductory Course
E. Carlen, Entropy and the Quantum, 529:73-140, 2010.

Regularization vs. Relaxation: A Convexification Perspective of Statistical Variable Selection
H. Dong, K. Chen, J. Linderoth, submitted to Mathematical Programming, 2015.

Selected References II

Perspectives of Matrix Convex Functions
A. Ebadian, I. Nikoufar, M. E. Gordji, Proc. Natl. Acad. Sci **108**(18), 7313-7314, 2011.

A Matrix Convexity Approach to Some Celebrated Quantum Inequalities
E. G. Effros, Proc. Natl. Acad. Sci **106**(4), 1006-1008, 2009.

Semidefinite Approximations of the Matrix Logarithm
H. Fawzi, J. Saunderson, P. Parrilo. Foundations of Computational Mathematics **19**(2): 259-296, 2019

Perspective Cuts for a Class of Convex 0-1 Mixed Integer Programs
A. Frangioni, C. Gentile. Mathematical Programming **106**:225-236 (2006)

Perspective Reformulations of Mixed Integer Nonlinear Programs With Indicator Variables
O. Günlük, J. Linderoth. Mathematical Programming **124**:183-205 (2010)

Mixed-Integer Convex Representability
M. Lubin, I. Zadik, J.P. Vielma, Mathematics of Operations Research, 2021

Guaranteed Minimum-Rank Solutions of Linear Matrix Inequalities via Nuclear Norm Minimization
B. Recht, M. Fazel, P. Parrilo. SIAM Review **52**(3):471-501 (2010)

What does MPCO (not) generalize from MIO?

MIO captures notions of

- Finiteness: $z \in \{0, 1\}$
- Algebraicity: $z^2 - z = 0$

While MPCO captures notions of algebraicity ($Y^2 = Y$) but NOT finiteness-uncountably infinitely many Y

Therefore [what follows is conjecture]

- Results from MIO which depend on algebraic arguments (perspective reformulation, taking convex hulls)
- Or where enumeration argument can be replaced with coverage argument (branch-and-bound/cut)

Generalize from MIO. But..

- Results in MIO which depend on discreteness (e.g., MIR cuts) probably do not

Therefore, QCQP cuts (split cuts, PSD cuts) can be used by MPCO, but MIO cuts (Knapsack/flow cover) cannot

Remark: determining whether MIO result due to finiteness is non-trivial