The convex algebraic geometry of rank minimization

Pablo A. Parrilo

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology

International Symposium on Mathematical Programming
August 2009 - Chicago
First, my coauthors...

Ben Recht  
(UW-Madison)

Maryam Fazel  
(U. Washington)

Venkat Chandrasekaran  
(MIT)

Sujay Sanghavi  
(UT Austin)

Alan Willsky  
(MIT)
Rank minimization

**PROBLEM:** Find the *lowest rank* matrix in a given convex set.

- E.g., given an affine subspace of matrices, find one of lowest possible rank

- In general, NP-hard (e.g., reduction from max-cut, sparsest vector, etc.).

*Many* applications...
Application 1: Quadratic optimization

- Boolean quadratic minimization (e.g., max-cut)

\[
\min_{x_i \in \{-1,1\}^n} x^T Q x
\]

- Relax using the substitution \( X := xx^T \):

\[
\min_{X \succeq 0, X_{ii} = 1} \text{Tr } Q X
\]

- If solution \( X \) has rank 1, then we solved the original problem!
Application 2: Matrix Completion

\[ M = \begin{bmatrix}
M_{ij} & M_{ij} & M_{ij} \\
M_{ij} & M_{ij} & M_{ij} \\
M_{ij} & M_{ij} & M_{ij}
\end{bmatrix} \]

- Partially specified matrix, known pattern
- Often, random sampling of entries
- Applications:
  - Partially specified covariances (PSD case)
  - Collaborative prediction (e.g., Rennie-Srebro 05, Netflix problem)
Application 3: Sum of squares

\[ P(x) = \sum_{i=1}^{r} q_i^2(x) \]

- Number of squares equal to the rank of the Gram matrix
- How to compute a SOS representation with the minimum number of squares?
- Rank minimization with SDP constraints
Application 4: System Identification

\[ \text{noise} \]
\[ \omega(t) \rightarrow P \rightarrow h \ast \omega(T) \]

- Measure response at time \( T \), with random input
- Response at time \( T \) is linear in \( h \).

\[
\text{hank}(h) := \begin{bmatrix}
  h(0) & h(1) & \cdots & h(N) \\
  h(1) & h(2) & \cdots & h(N+1) \\
  \vdots & \vdots & \ddots & \vdots \\
  h(N) & h(N+1) & \cdots & h(2N)
\end{bmatrix}.
\]

- “Complexity” of \( P \approx \text{rank}(\text{hank}(h)) \)
Application 5: Video inpainting

(Ding-Sznaier-Camps, ICCV 07)

• Given video frames with missing portions

• Reconstruct / interpolate the missing data

• “Simple” dynamics <-> Low rank (multi) Hankel
Application 5: Video inpainting (cont)

(Ding-Sznaier-Camps, ICCV 07)
Overview

Rank optimization is ubiquitous in optimization, communications, and control. Difficult, even under linear constraints.

This talk:
• Underlying convex geometry and nuclear norm
• Rank minimization can be efficiently solved (under certain conditions!)
• Links with sparsity and “compressed sensing”
• Combined sparsity + rank
Outline

- Applications
- Mathematical formulation
- Rank vs. sparsity
- What is the underlying geometry?
- Convex hulls of varieties
- Geometry and nuclear norm

- Sparsity + Rank
- Algorithms

Rank \( r \) matrices
**PROBLEM:** Find the matrix of smallest rank that satisfies the underdetermined linear system:

\[ \mathcal{A}(X) = b \quad \mathcal{A} : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^p \]

- Given an affine subspace of matrices, find one of lowest rank
Rank can be complicated...

What does a rank constraint look like?

(section of)
3x3 matrices, rank 2

(section of)
7x7 matrices, rank 6
How to solve this?

Many methods have been proposed (after all, it’s an NLP!)

- Newton-like local methods
- Manifold optimization
- Alternating projections
- Augmented Lagrangian
- Etc...

Sometimes (often?) work very well.

But, very hard to prove results on global performance.
Geometry of rank varieties

What is the structure of the set of matrices of fixed rank?

\[ \mathcal{P}(k) \triangleq \{ M \in \mathbb{R}^{n \times n} \mid \text{rank}(M) \leq k \} . \]

An \textit{algebraic variety} (solution set of polynomial equations), defined by the vanishing of all \((k+1) \times (k+1)\) minors of \(M\).

Its dimension is \(k \times (2n-k)\), and is nonsingular, except on those matrices of rank less than or equal to \(k-1\).

At smooth points \(M = U \Sigma V^T\) well-defined tangent space:

\[ T(M) = \{ UX^T + YV^T \mid X, Y \in \mathbb{R}^{n \times k} \} . \]
Nuclear norm

- **Sum of the singular values** of a matrix

\[ \|X\|_* := \sum_{i=1}^{r} \sigma_i(X), \quad \sigma_i(X) := \sqrt{\lambda_i(X^T X)} \]

- Unitarily invariant

\[ \|UXV\|_* = \|X\|_* \]

- Also known as *Schatten 1-norm, Ky-Fan r-norm, trace norm,* etc...
Why is nuclear norm relevant?

• Bad nonconvex problem -&gt; Convexify!

• Nuclear norm is “best” convex approximation of rank
Comparison with sparsity

Consider sparsity minimization
- Geometric interpretation
- Take “sparsity 1” variety
- Intersect with unit ball
- Take convex hull

L1 ball! (crosspolytope)
Nuclear norm

• Same idea!

• Take “rank 1” variety

• Intersect with unit ball

• Take convex hull

\[ \text{conv}\{uv^T : u \in \mathbb{R}^n, v \in \mathbb{R}^m, \|u\| = 1, \|v\| = 1\} \]

Nuclear ball!
Convex hulls of algebraic varieties

- Systematic methods to produce (exact or approximate) SDP representations of convex hulls

- Based on sums of squares (Shor, Nesterov, Lasserre, P., Nie, Helton)

- Parallels/generalizations from combinatorial optimization (*theta bodies*, e.g., Gouveia-Laurent-P.-Thomas 09)
Nuclear norm and SDP

• The nuclear norm is SDP-representable!

\[ \|X\|_* := \sum_{i=1}^{r} \sigma_i(X), \]

• Semidefinite programming characterization:

\[
\begin{align*}
\max_Y & \quad \text{Tr}(X'Y) \\
\text{s.t.} & \quad \begin{bmatrix} I_m & Y \\
Y' & I_n \end{bmatrix} \succeq 0.
\end{align*}
\]

\[
\begin{align*}
\min_{W_1, W_2} & \quad \frac{1}{2}(\text{Tr}(W_1) + \text{Tr}(W_2)) \\
\text{s.t.} & \quad \begin{bmatrix} W_1 & X \\
X' & W_2 \end{bmatrix} \succeq 0.
\end{align*}
\]
A convex heuristic

**PROBLEM:** Find the matrix of lowest rank that satisfies the underdetermined linear system

\[ \mathcal{A}(X) = b \quad \mathcal{A} : \mathbb{R}^{n \times m} \to \mathbb{R}^p \]

• Convex optimization **heuristic**
  – Minimize *nuclear norm* (sum of singular values) of \(X\)
  – This is a *convex* function of the matrix \(X\)
  – Equivalent to a SDP problem

\[
\begin{align*}
\text{minimize} & \quad \|X\|_* = \sum_{i=1}^m \sigma_i(X) \\
\text{subject to} & \quad \mathcal{A}(X) = b
\end{align*}
\]
Nuclear norm heuristic

Affine Rank Minimization:
\[
\begin{align*}
\text{minimize} & \quad \text{rank}(X) \\
\text{subject to} & \quad A(X) = b
\end{align*}
\]

Relaxation:
\[
\begin{align*}
\text{minimize} & \quad \|X\|_* = \sum_{i=1}^{m} \sigma_i(X) \\
\text{subject to} & \quad A(X) = b
\end{align*}
\]

• Proposed in Maryam Fazel’s PhD thesis (2002).
• Nuclear norm is the convex envelope of rank
• Convex, can be solved efficiently
• Seems to work well in practice
Nice, but will it work?

**Affine Rank Minimization:**

\[
\begin{align*}
\text{minimize} & \quad \text{rank}(X) \\
\text{subject to} & \quad \mathcal{A}(X) = b
\end{align*}
\]

**Relaxation:**

\[
\begin{align*}
\text{minimize} & \quad \|X\|_* = \sum_{i=1}^{m} \sigma_i(X) \\
\text{subject to} & \quad \mathcal{A}(X) = b
\end{align*}
\]

Let’s see...
Numerical experiments

• Test matrix arbitrarily chosen ;)

• Rank 5 matrix, 46x81 pixels
• Generate random equations, Gaussian coeffs.
• Nuclear norm minimization via SDP (SeDuMi)
Phase transition
How to explain this?

Apparently, under certain conditions, nuclear norm minimization “works”.

How to formalize this?
How to \textit{prove} that relaxation will work?

**Affine Rank Minimization:**
\[
\begin{align*}
\text{minimize} & \quad \text{rank}(X) \\
\text{subject to} & \quad A(X) = b
\end{align*}
\]

**Relaxation:**
\[
\begin{align*}
\text{minimize} & \quad \|X\|_* = \sum_{i=1}^{m} \sigma_i(X) \\
\text{subject to} & \quad A(X) = b
\end{align*}
\]

**General recipe:**

- Find a \textit{deterministic} condition that ensures success
- Sometimes, condition may be hard to check
- If so, invoke \textit{randomness} of problem data, to show condition \textit{holds with high probability (concentration of measure)}
Compressed sensing - Overview

- Compressed sensing is a new paradigm for data acquisition/estimation.
- Influential recent work includes Donoho/Tanner, Candès/Romberg/Tao, and Baraniuk, among others.

Compressed sensing relies on sparsity (on some domain, e.g., Fourier, wavelet, etc).
- "Few" random measurements + smart decoding
- Many applications, particularly MRI, geophysics, radar, etc.
Nice, but will it work?

**Affine Rank Minimization:**

\[
\begin{align*}
\text{minimize} & \quad \text{rank}(X) \\
\text{subject to} & \quad A(X) = b
\end{align*}
\]

**Relaxation:**

\[
\begin{align*}
\text{minimize} & \quad \|X\|_* = \sum_{i=1}^{m} \sigma_i(X) \\
\text{subject to} & \quad A(X) = b
\end{align*}
\]

- Use a “restricted isometry property” (RIP)
- Then, this heuristic *provably* works.
- For “random” operators, RIP holds with overwhelming probability
Restricted Isometry Property (RIP)

• Let $\mathcal{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^p$ be a linear map. For every positive integer $r \leq m$, define the $r$-restricted isometry constant to be the smallest number $\delta_r(\mathcal{A})$ such that

\[
(1 - \delta_r(\mathcal{A})) \|X\|_F \leq \|\mathcal{A}(X)\| \leq (1 + \delta_r(\mathcal{A})) \|X\|_F
\]

holds for all matrices $X$ of rank at most $r$.

• Similar to RIP condition for sparsity studied by Candès and Tao (2004).
RIP ⇒ Heuristic Succeeds

**Theorem:** Let $X_0$ be a matrix of rank $r$. Let $X_*$ be the solution of $A(X) = A(X_0)$ of smallest nuclear norm. Suppose that $r \geq 1$ is such that $\delta_{5r}(A) < 1/10$. Then $X_* = X_0$.

- Deterministic condition on $A$
- Checking RIP can be hard
- However, “random” $A$ will have RIP with high probability

*Independent of $m, n, r, p$*
“Random” ⇒ RIP holds

**Theorem:** Fix $0 < \delta < 1$. If $\mathcal{A}$ is “random”, then for every $r$, there exist a constant $c_0$ depending only on $\delta$ such that $\delta_r(\mathcal{A}) \leq \delta$ whenever $p \geq c_0 \cdot r(2n-r) \log(n^2)$ with probability exponentially close to 1.

- Number of measurements $c_0 \cdot r(2n-r) \log(n^2)$

- Typical scaling for this type of result.
Comparison

**Compressed Sensing**
- cardinality
- disjoint support implies cardinality of sum is sum of cardinalities
- $\ell_1$ norm
  - dual norm: $\ell_\infty$
  - best convex approximation of cardinality on unit ball of $\ell_\infty$ norm
  - Optimized via LP
- RIP: $\forall$ vectors of cardinality at most $s$
  \[ 1 - \delta_s \leq \frac{\|Ax\|}{\|x\|} \leq 1 + \delta_s \]

**Low-rank Recovery**
- rank
- “disjoint” row and column spaces implies rank of sum is sum of ranks
- nuclear norm
  - dual norm: operator norm
  - best convex approximation of rank on unit ball of operator norm
  - Optimized via SDP
- RIP: $\forall$ matrices of rank at most $r$
  \[ 1 - \delta_r \leq \frac{\|AX\|}{\|X\|_F} \leq 1 + \delta_r \]
Secant varieties

• What is common to the two cases? Can this be further extended?

• A natural notion: secant varieties

• Generalize notions of rank to other objects (e.g., tensors, nonnegative matrices, etc.)

• However, technical difficulties
  – In general, these varieties may not be closed
  – In general, associated norms are not polytime computable
Nuclear norm minimization: Algorithms

Convex, nondifferentiable, special structure.

• Many possible approaches:
  – Interior-point methods for SDP
  – Projected subgradient
  – Low-rank parametrization (e.g., Burer-Monteiro)

• Much exciting recent work:
  – Liu-Vandenberghe (interior point, exploits structure)
  – Cai-Candès-Shen (singular value thresholding)
  – Ma-Goldfarb-Chen (fixed point and Bregman)
  – Toh-Yun (proximal gradient)
  – Lee-Bresler (atomic decomposition)
  – Liu-Sun-Toh (proximal point)
Low-rank completion problem

- Recent work of Candès-Recht, Candès-Tao, Keshavan-Montanari-Oh, etc.
- These provide theoretical guarantees of recovery, either using nuclear norm, or alternative algorithms
- E.g., Candès-Recht:
  \[ p \geq C r n^{(6/5)} \log(n) \]
- Additional considerations: noise robustness, etc.
What if sparsity pattern is not known?

Given Composite matrix

Unknown Sparse Matrix

Unknown support, values

Unknown Low-rank Matrix

Unknown rank, eigenvectors

Task: given $C$, recover $A^*$ and $B^*$
A probabilistic model given by a graph.

- Then, the inverse covariance $\Sigma^{-1}$ is sparse.

No longer true if graphical model has *hidden variables*.

But, it is the sum of a sparse and a low-rank matrix (Schur complement)

$$\hat{K}_o = \Sigma_o^{-1} = K_o - K_{o,h}K_h^{-1}K_{h,o}.$$
Application: Matrix rigidity

**Smallest** number of changes to reduce rank [Valiant 77]

Rigidity of a matrix NP-hard to compute

Rigidity bounds have a number of applications
  – Complexity of linear transforms [Valiant 77]
  – Communication complexity [Lokam 95]
  – Cryptography
Identifiability issues

Problem can be *ill-posed*. Need to ensure that terms cannot be simultaneously sparse and low-rank.

Define two geometric quantities, related to tangent spaces to sparse and low-rank varieties:

\[
\mu_1(M) = \max_{S \in \Omega(M)} \frac{||S||}{||S||_{\infty}}
\]

\[
\mu_2(M) = \max_{S \in T(M)} \frac{||S||_{\infty}}{||S||_{T(M)}}
\]
Tangent space identifiability

Necessary and sufficient condition for exact decomposition
Tangent spaces must intersect transversally

\[ \Omega(A^*) \cap T(B^*) = \{0\} \]

Sufficient condition for transverse intersection

\[ \mu_1(A^*) \mu_2(B^*) < 1 \quad \Rightarrow \quad \Omega(A^*) \cap T(B^*) = \{0\} \]
Back to decomposition

If \( C = A^* + B^* \), where \( A^* \) is sparse and \( B^* \) is low-rank,

\[
\min_{A, B} \quad \gamma \| A \|_0 + \text{rank}(B)
\]

s.t. \( A + B = C \)

• **combinatorial**, NP-hard in general

• **not known** how to choose \( \gamma \)

• when does this exactly recover \((A^*, B^*)\)?
Natural convex relaxation

\[ \|A\|_0 \quad \rightarrow \quad \|A\|_1 = \sum_{i,j} |a_{i,j}| \]

\[ \text{rank}(B) = \|\sigma(B)\|_0 \quad \rightarrow \quad \|B\|_* = \sum_i \sigma_i(B) \]

Propose:

\[(\hat{A}, \hat{B}) = \arg \min_{A,B} \quad \gamma \|A\|_1 + \|B\|_* \quad \text{s.t. } A + B = C \]

Convex program (in fact, an SDP)
Matrix decomposition

Theorem: For any $A^*$ and $B^*$

$$\mu_1(A^*)\mu_2(B^*) < \frac{1}{8} \implies (\hat{A}, \hat{B}) = (A^*, B^*) \text{ is unique optimum of convex program for a range of } \gamma$$

Essentially a refinement of tangent space transversality conditions

<table>
<thead>
<tr>
<th>Transverse intersection:</th>
<th>$\mu_1(A^<em>)\mu_2(B^</em>) &lt; 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convex recovery:</td>
<td>$\mu_1(A^<em>)\mu_2(B^</em>) &lt; \frac{1}{8}$</td>
</tr>
</tbody>
</table>

Under “natural” random assumptions, condition holds w.h.p.
Summary

• Sparsity, rank, and beyond
  – Many applications
  – Common geometric formulation: secant varieties
  – Convex hulls of these varieties give “good” proxies for optimization
  – Algebraic and geometric aspects

• Theoretical challenges
  – Efficient descriptions
  – Sharper, verifiable conditions for recovery
  – Other formulations (e.g., Ames-Vavasis on planted cliques)
  – Finite fields?

• Algorithmic issues
  – Reliable, large scale methods
Thank you!

Want to know more? Details below, and in references therein:


Thanks for your attention!