Semidefinite and Second Order Cone Programming Seminar
Fall 2012
An Overview of the Paper
“Matrix Rank Minimization With Applications”

Instructor: Farid Alizadeh
Scribe: Ritwik Mitra

December 2, 2012

1 Overview

The dissertation by Maryam Fazel, “Matrix Rank Minimization with applications”, [3], focuses on minimizing ranks over matrices of convex sets which is genrally an NP-hard problem. The work provides a new heuristics for solving Rank Minimization Problem (RMP) for PSD matrices, which contrary to hitherto developed heuristics, do not require an initial point, are numerically efficient and provide a global lower bound on the optimal value. In addition it also provides a way of embedding a general RMP in a larger PSD problem.

2 Rank Minimization Problems

The general form of a RMP is as follows

$$\begin{array}{ll} \min & \operatorname{rank}(X) \\ \text{s.t.} & X \in \mathcal{C} \end{array} \tag{1}$$

where $X \in \mathfrak{R}^{m \times n}$ and $\mathcal{C}$ is a convex set. We can also enhance the minimization problem in (1) by employing a constraint on the set of matrices in $\mathcal{C}$ which help to specify the required accuracy/performance of the minimizer. In that sense $\operatorname{rank}(X)$ denotes the complexity of the model where as additional constraints impose accuracy requirements. This problem can be summarized as

$$\begin{array}{ll} \min & \operatorname{rank}(X) \\ \text{s.t.} & X \in \mathcal{C} \\ & f(X) \leq t \end{array} \tag{2}$$

Here the only requirement is that f, a measure of discrepancy, is convex. t is a tolerance level to be prespecified. In what follows, we provide a few examples of rank minimization problems to help set the ideas.

2.1 Examples of RMP

A myriad of problems can be formulated as a RMP. We provide a few examples of such cases.

2.1.1 Rank of A Covariance Matrix

Covarince matrix estimation problems arise in several areas of study. Statistics, Econometrics, Signal Processing to name a few. Let $\boldsymbol{z} \in \mathfrak{R}^n$ be a random vector ditributed according a probability law. We define $\Sigma = \mathbb{E}(\boldsymbol{z} - \mathbb{E}\boldsymbol{z})(\boldsymbol{z} - \mathbb{E}\boldsymbol{z})^T$ as the covariance matrix of the problem. A low rank of covariance matrix signifies that only a small number of linear combinations of variables in $\boldsymbol{z}$ are enough to explain the covariance of $\boldsymbol{z}$. This is essentially the so called "Factor Analysis" problem. We state

$$\begin{array}{ll} \min & \text{rank}(\Sigma) \\ \text{s.t.} & \|\Sigma - \widehat{\Sigma}\|_F \leq \epsilon \\ & \Sigma \succcurlyeq_{\mathcal{P}_{n\times n}} 0 \\ & \Sigma \in \mathcal{C} \end{array} \tag{3}$$

Here $\Sigma \in \mathfrak{R}^{n\times n}$ is the optimization variable. $\widehat{\Sigma}$ is an already measured covariance matrix and $\mathcal{C}$ might denote a a convex set of matrices that specify prior knowledge of the structure of Σ. $\|\|_F$ denotes a frobenius norm.

2.1.2 Reconstruction of Polygons From Moments

We consider a polygonal region $\mathcal{P}$ in the complex plane with verices $\boldsymbol{z} = (z_1, \cdots, z_m)$. The **complex moments**, c_k of $\mathcal{P}$ are given by

$$c_k = \int\int_{\mathcal{P}} z^k \, dxdy \tag{4}$$

Let us define τ_k as $\tau_k = \int\int_{\mathcal{P}} \frac{\partial^2}{\partial z^2}(z^k)\, dxdy$. The we clearly have

$$\tau_k = k(k-1)c_{k-2} \quad \text{with } \tau_0 = \tau_1 = 0 \tag{5}$$

A theorem by Davis is noted in this regard without proof. A statement of it can be found in Davis' original paper, [2], or in [5].

Theorem 1 (Davis) *Let $\boldsymbol{z} \in \mathbb{I}^m$ designate vertices of a polygon in $\mathcal{P}$ in the complex plane. Let f be analytic in the closure of $\mathcal{P}$. Let $f(\boldsymbol{z}) = (f(z_1), \cdots, f(z_m))$. Then there exist $\boldsymbol{a} \in \mathbb{I}^m$ such that*

$$\int\int_{\mathcal{P}} f^{''}(z) dxdy = \boldsymbol{a}^T f(\boldsymbol{z})$$

This immediately leads to $\tau_k = \mathbf{a}^T \mathbf{z}^k = \sum_{i=1}^{m} a_i z_i^k$. We define the Hankel Matrix $H_n \in \mathbb{I}^{n\times n}$ of τ_k's as

$$H_n = \begin{bmatrix} \tau_0 & \tau_1 & \cdots & \tau_{n-1} \\ \tau_1 & \tau_2 & \cdots & \tau_n \\ \tau_2 & \tau_3 & \cdots & \tau_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{n-1} & \tau_n & \cdots & \tau_{2n-2} \end{bmatrix} \tag{6}$$

Lemma 2 *For a polygon $\mathcal{P}$ in the complex plane of m vertices,* $\text{rank}(H_n) = m$

Proof: Let us define the Vandermonde Matrix $V_m \in \mathbb{I}^{m\times m}$ as

$$V_m = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_m \\ \vdots & \vdots & \ddots & \vdots \\ z_1^m & z_2^m & \cdots & z_m^m \end{bmatrix} \tag{7}$$

Since $\mathbf{z}$ are distinct $\text{rank}(V_m) = m$. If H_m is the leading block in H_n then $H_m = V_m \mathbf{diag}(\mathbf{a}) V_m^T$. Hence $\text{rank}(H_m) = \text{rank}(\mathbf{diag}(\mathbf{a})) = m$. Now if $n > m$, then $\text{rank}(H_n) = \text{rank}(V_n \mathbf{diag}(\mathbf{a}) V_n) \leq m$. Since rank of the leading block is m, we must have $\text{rank}(H_n) \geq m$. Thus proved. ■

We now suppose that we have $2n-1$ corrupted measurements $\mathbf{y}$ with $y_k = \tau_k + \nu_k$. Let ν_k's be gaussian 0 mean variable. Suppose m is unknown but $\leq n$. Goal is to estimate smallest possible $\mathcal{P}$ consisted with the data. Let $\widehat{H_n}$ be estimated from corrupted data $\mathbf{y}$. Consider the problem

$$\begin{array}{ll} \min & \text{rank}(H) \\ \text{s.t.} & \|H - \widehat{H}_n\|_F \leq \epsilon \\ & H \text{ is Hankel} \end{array} \tag{8}$$

Remark 3 *Without the restriction to Hankel matrices, this problem can be solved using SVD. Solution with restriction to Hankel matrices is rather hard.*

2.1.3 Cardinality Minimization Problem

Consider the linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}$ with $X \in \mathfrak{R}^{n\times p}$, $\boldsymbol{\beta} \in \mathfrak{R}^p$ and $\boldsymbol{\epsilon} \in \mathfrak{R}^n$ being a gaussian error vector. A very common problem in regression is to find the best possible fit to the data $\mathbf{y}$ by estimating $\boldsymbol{\beta}$. Variable selection means choosing such a $\boldsymbol{\beta}$ that is also smallest possible is size. We call this problem the cardinality minimization problem. It can be written as an RMP as

$$\begin{array}{ll} \min & \text{rank}(\mathbf{diag}(\boldsymbol{\beta})) \\ \text{s.t.} & \frac{1}{n}\|y - X\beta\|_2^2 \leq \varepsilon \end{array} \tag{9}$$

3 Semidefinite Embedding

If the matrix variable X in (1) is constrained to be in $\mathcal{P}_{n\times n}$, then we call the problem a PSD-RMP. Low-Rank Solutions for PSD-RMP have the advantage of always lying on the boundary of the cone. All the existing methods and also the heuristics developed in this work all apply only to PSD matrices. So in scenarios where X is non-PSD or even non-square, we need a tool to extend the methodologies developed here. This is the main motivation behind *Semidefinite Embedding Lemma*

Lemma 4 (Semidefinite Embedding) *Let* $X \in \mathfrak{R}^{m\times n}$ *be given. Then we have* $\mathrm{rank}(X) \le r$ *iff there exist matrices* $Y = Y^T \in \mathfrak{R}^{m\times m}$ *and* $Z = Z^T \in \mathfrak{R}^{n\times n}$ *such that*

$$\mathrm{rank}(Y) + \mathrm{rank}(Z) \le 2r \qquad \begin{bmatrix} Y & X \\ X^T & Z \end{bmatrix} \succcurlyeq 0$$

Proof: To prove our assertion first we make use of the following lemma. This is a general Schur Complement property. A prof of which can be found in [1]. We define Y^+ as the moore-penrose pseudo inverse of Y.

Lemma 5 *Consider marices* X, Y, Z *of appropriate dimensions with* Y *and* Z *symmetric. Then*

$$A = \begin{bmatrix} Y & X \\ X^T & Z \end{bmatrix} \succcurlyeq 0 \quad \text{iff} \quad \begin{array}{ll} i) & Y \succcurlyeq 0 \\ ii) & X^T(I - YY^+) = 0 \\ iii) & Z - X^T Y^+ Y \succcurlyeq 0 \end{array}$$

We now prove each direction separately.
$(\Rightarrow)$ Let $\mathrm{rank}(X) = r_0 \le r$. Consider the QR decomposition $X = LR$ with $L \in \mathfrak{R}^{m\times r_0}$ and $R \in \mathfrak{R}^{r_0\times n}$ with $\mathrm{rank}(L) = \mathrm{rank}(R) = r_0$. define $Y = LL^T$ and $Z = R^T R$. Then we have $\begin{bmatrix} LL^T & LR \\ R^T L^T & R^T R \end{bmatrix} \succcurlyeq 0$ since it has the square root $(L^T \ R)^T$
$(\Leftarrow)$ For proof in this direction we make use of Lemma 5. Without loss of generality we assume that $\mathrm{rank}(Y) \le \mathrm{rank}(Z)$. We note that $(I - YY^+)$ is a projection matrix into $\mathfrak{N}(Y)$ the nullspace of Y. Now (ii) implies $\mathfrak{N}(X^T) \supseteq \mathfrak{N}(Y)$ which implies $\mathrm{rank}(Y) \ge \mathrm{rank}(X^T) = \mathrm{rank}(X)$. ■
We now provide a few corollaries of the above lemma.

Corollary 6 *If* X *has a block diagonal structure, say* $X = \mathbf{diag}(X_i)_{i=1}^N$ *with* $X_i \in \mathfrak{R}^{m_i\times n_i}$ *then, without loss of generality we can assume that the slack variables* Y *and* Z *are each block diagonal say* $Y = \mathbf{diag}(Y_i)_{i=1}^N$ *with* $Y_i = Y_i^T \succcurlyeq 0 \in \mathfrak{R}^{m_i\times m_i}$ *and similarly for* Z *with* $Z_i \in \mathfrak{R}^{n_i\times n_i}$

Corollary 7 *If* X *is symmetric, then without loss generality we can take* $Y = Z$.

This embedding Lemma enables us to **equivalently** recode the RMP problem in (1) as

$$\begin{aligned} \min \quad & \frac{1}{2} \text{rank } \mathbf{diag}(Y, Z) \\ \text{s.t.} \quad & \begin{bmatrix} Y & X \\ X^T & Z \end{bmatrix} \succcurlyeq 0 \\ & X \in \mathcal{C} \end{aligned} \tag{10}$$

Equipped with these results we are now in a position to talk about the heuristic methodologies developed to tackle RMPs

4 The trace & log det Heuristic

In what follows we talk about two heuristic methods of solving RMPs. We go through them one by one.

4.1 The trace Heuristic

A heuristicaly known fact is that if $X \in \mathcal{P}_{n\times n}$, the PSD cone, then the problem in (1) can be replace by

$$\begin{aligned} \min \quad & \text{trace}(X) \\ \text{s.t.} \quad & X \in \mathcal{C} \\ & X \succcurlyeq_{\mathcal{P}_{n\times n}} 0 \end{aligned} \tag{11}$$

This is because of the fact that trace $X = \|\boldsymbol{\lambda}(X)\|_1$ where $\boldsymbol{\lambda}$ is the vector of eigen values of X. Also it is known that ℓ_1 minimization of a vector results in a sparse solution. Thus in this case (11) will result in a solution that is sparse in eigen values yielding in effect a low rank solution.

Remark 8 *If we define the ℓ_0-'norm' of a vector $\boldsymbol{\beta}$ as $\|\boldsymbol{\beta}\|_0 = \#\{j : \beta_j \neq 0\}$, then minimizing $\|\boldsymbol{\beta}\|_0$ is NP-hard because this norm-like function is neither a norm nor a convex function. It is a standard practice in regression to recode the problem by relaxing the ℓ_0 to ℓ_1 norm which leads to a convex problem.*

4.2 Nuclear Norm Heuristic

we begin by providing the definition of nuclear norm.

Definition 9 (Nuclear Norm) *Let $X \in \mathfrak{R}^{m\times n}$ be a matrix with singular values $\sigma_i(X)$ for $i = 1, \cdots, \min(m, n)$. The nuclear norm is defined as*

$$\|X\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(X) \tag{12}$$

4.2.1 Convex Envelope of Rank

Definition 10 (Convex Envelope) *The convex envelope of* $f : \mathcal{C} \to \Re$ *is defined as the largest convex function* g *such that* $g(x) \leq f(x)$ *for all* $x \in C$

We note that using Remark 8, rank(X) can be written as $\|\boldsymbol{\sigma}(X)\|_0$ which is non convex. As mentioned in said remark, a valid strategy could be to replace it by $\|\boldsymbol{\sigma}(X)\|_1 = \|X\|_*$, the nuclear norm. In what follows we show that *nuclear norm is the convex envelope of rank.* In other words it is the closest a convex function can get in approximating the rank of a matrix.

Let us denote the spectral norm of a matrix as $\|X\|_2$, the largest singular value of a matrix. Clearly $\|X\|_2 = \|\boldsymbol{\sigma}(X)\|_\infty$. Spectral norm is the dual of nuclear norm.

Theorem 11 (Convex Envelope Of Rank) *Consider the set*

$$\mathcal{S} = \{X \in \Re^{m \times n} : \ \|X\|_2 \leq 1\}$$

The Convex envelope of the function $\phi(X) = \text{rank}(X)$ *is given by* $\phi_{env}(X) = \|X\|_*$

We provide an outline of the proof without going nto the detail of the calculation.

Proof:(Outline Only)
The conjugate of a function f as $f^*(y) = \sup_{x \in C} \{\langle y, x\rangle - f(x)\}$. A basic result in convex analysis states that f^{**} is a convex envelope of f under certain easily verifiable conditions. We prove that conjugate of rank function ϕ is nuclear norm. Clearly $\phi^*(Y) = \sup_{\|X\|_2 \leq 1}\{\text{trace}(Y^T X) - \phi(X)\}$. Neumanns's Trace theorem, [4], states that

$$\text{trace}(Y^T X) \leq \sum_{i=1}^{\min(m,n)} \sigma_i(X)\sigma_i(Y)$$

Which leads us to write

$$\phi^*(Y) = \sup_{\|X\|_2 \leq 1} \left\{ \sum_{i=1}^{\min(m,n)} \sigma_i(X)\sigma_i(Y) - \text{rank}(X) \right\}$$

which is equivalent to

$$\phi^*(Y) = \max \left\{ 0, \sigma_1(Y) - 1, \cdots, \sum_{i=1}^{r} \sigma_i(Y) - r, \cdots, \sum_{i=1}^{\min(m,n)} \sigma_i(Y) - \min(m,n) \right\}$$

which in effect is equivalent to (Assuming r singular values of Y are > 1)

$$\phi^*(Y) = \sum_{i=1}^{\min(m,n)} (\sigma_i(Y) - 1)_+ = \sum_{i=1}^{r} \sigma_i(Y) - r \tag{13}$$

Now we find the conjugate of $\phi^*(Y)$ say $\phi^{**}(Z) = \sup_Y \left\{\text{trace}(Z^T Y) - \phi^*(Y)\right\}$. Application of Neumann's Trace Theorem yields

$$\phi^{**}(Z) = \sup_Y \left\{ \sum_{i=1}^{\min(m,n)} \sigma_i(Y)\sigma_i(Z) - \phi^*(Y) \right\}$$

If $\|Z\|_2 > 1$, then $\phi^{**} \to \infty$. If $\|Z\|_2 \leq 1 \& \|Y\|_2 \leq 1$ then $\phi^{**}(Z) = \|Z\|_*$. Finaly if $\|Z\|_2 \leq 1 \& \|Y\|_2 > 1$ then we can show that $\phi^{**}(Z) < \|Z\|_*$.
In essence we proved that over the set $\mathcal{S}$ defined before $\phi^{**}(Z) = \|Z\|_*$. Thus over this set $\|Z\|_*$ is the envelope of $\text{rank}(Z)$. ■

4.2.2 Applying Nuclear Norm

We showed in establishing the embedding lemma that the General RMP problem in (1) can be equivalently stated the PSD-RMP problem in (10). Clearly the trace heuristic mentioned at the begining of this section in (11) motivates yet another formulation.

$$\begin{aligned} \min \quad & \frac{1}{2}\,\text{trace}\,\mathbf{diag}(Y, Z) \\ \text{s.t.} \quad & \begin{bmatrix} Y & X \\ X^T & Z \end{bmatrix} \succcurlyeq 0 \\ & X \in \mathcal{C} \end{aligned} \tag{14}$$

We now proceed to provide an equivalence between (14) and a nuclear norm minimization problem. This equivalency is described in the following lemma.

Lemma 12 *Let $X \in \mathfrak{R}^{m\times n}$ and $t \in \text{Re}$. Then We can always find matrices $Y \in \mathfrak{R}^{m\times m}$ and $Z \in \mathfrak{R}^{n\times n}$ so that*

$$\|X\|_* \leq t \quad \textbf{\textit{iff}} \quad \begin{bmatrix} Y & X \\ X^T & Z \end{bmatrix} \succcurlyeq 0 \text{ and } \text{trace}(Y) + \text{trace}(Z) \leq 2t \tag{15}$$

Proof: ($\Leftarrow$) Let Y, Z satisfy (15). Consider the SVD of X as $X = UDV^T$. Let $\text{rank}(X) = r$. Then we can take $D \in \mathfrak{R}^{r\times r}$. Let us Consider the matrix $W = (U^T \;\; -V^T)^T$. then $WW^T \succcurlyeq 0$. Since the trace of product of two PSD matrices is non-negative we have $\text{trace}\left(WW^T \begin{bmatrix} Y & X \\ X^T & Z \end{bmatrix}\right) \geq 0$ which simplifies to

$$\text{trace}(UU^T Y) - \text{trace}(UV^T X^T) - \text{trace}(VU^T X) + \text{trace}(VV^T Z) \geq 0 \tag{16}$$

Let us extend the basis formed by columns of U with some more columns $\bar{U}$ so that $[U\ \bar{U}]$ is a full orthonormal basis and $UU^T + \bar{U}\bar{U}^T = I$. This yields $\text{trace}(U^T Y) \leq \text{trace}(Y)$. Similar considerations yield $\text{trace}(VV^T Z) \leq \text{trace}(Z)$. Also $\text{trace}(VU^T X) = \text{trace}(UV^T X^T) = \text{trace}(D) = \|X\|_*$. substituting these

values in (16) and using (15) we get $\|X\|_* \leq t$.
($\Rightarrow$) Let $\|X\|_* \leq t$. We find Y and Z as follows. Define

$$Y = UDU^T + \gamma I \quad \text{and} \quad Z = VDV^T + \gamma I \tag{17}$$

Then very clearly $\text{trace}(Y) + \text{trace}(Z) = 2\,\text{trace}(D) + \gamma(m+n)$. So if we choose γ as $\gamma = \frac{2(t - \|X\|_*)}{m+n}$ the we have $\text{trace}(Y) + \text{trace}(Z) = 2t$. Also it is easy to verify that

$$\begin{bmatrix} Y & X \\ X^T & Z \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix} D[U^T \; V^T] + \gamma I \succcurlyeq 0$$

■

Now we are ready to write the problem in (1) in terms of nuclear norm. First we rewrite (14) as

$$\begin{array}{ll} \min & t \\ \text{s.t.} & \text{trace}(Y) + \text{trace}(Z) \leq 2t \\ & \begin{bmatrix} Y & X \\ X^T & Z \end{bmatrix} \succcurlyeq 0 \\ & X \in \mathcal{C} \end{array} \tag{18}$$

It is then easily seen using Lemma 12 that we can write (18) as

$$\begin{array}{ll} \min & t \\ \text{s.t.} & \|X\|_* \leq t \\ X \in \mathcal{C} & \end{array} \tag{19}$$

4.3 The $\log\det$ Heuristic

In this heuristic $\log\det$ is used as a smooth surrogate for rank. This heuristic proposes an iterative linearization and minimization scheme for solving the problem of finding a local minima. Let $X \in \mathfrak{R}^{n\times n}$ and $X \succcurlyeq 0$. In this scenario we use another formulation of (1) as

$$\begin{array}{ll} \min & \log\det(X + \delta I) \\ \text{s.t.} & X \in \mathcal{C} \end{array} \tag{20}$$

Remark 13 *We note the following in connection to the above procedure*

- *We note that* $\log\det$ *is not a convex function but rather a concave one. However since it is smooth on* $\mathcal{P}_{n\times n}$, *we it can be minimized locally using any local minimization method*
- *Consider the Taylor expansion around* X_k

$$\log\det(X + \delta I) \approx \log\det(X + \delta I) + \text{trace}(X_k + \delta I)^{-1}(X - X_k) \tag{21}$$

We can find a new iterate value X_{k+1} *in terms of* X_k *by minimizing (21) instead of the concave function we started with.*

$$X_{k+1} = \arg\min_{X\in\mathcal{C}} \text{trace}(X_k + \delta I)^{-1} X \tag{22}$$

- *If we take* $\mathsf{X}_0 = \mathrm{I}$*, then it is clear that trace heuristic is the first step of the* log det *heuristic. Hence* X_1 *is the result of the* trace *heuristic and further iterations tend to reduce the rank even further yielding an even finer result*

5 Epilogue

The work by Maryam Fazel on matrix rank minimization problems provide a new albeit heuristic approach to solving RMPs. The work doesn't claim that the trace or log det heuristic yield lower rank solutions than other methods in all instances of RMPs. They, however, do come with a set of benefits.

Applicability This heuristic method applies to all possible instances of any general RMP

Initial Point There is no need for user-specified initial points. For log det heuristic the nuclear norm solution could work as a starting point for the log det methodology

Optimality The nuclear norm heuristic is optimal in the sense of being the convex envelope of the rank. It also provides a global lowerbound on the minimum rank if the feasible set is bounded

Fast Solution log det heuristic linearizes the concave problem making it covex at each iteration. Thus the convergence in this method is very fast and need very few iterations

References

[1] S. Boyd, L. El Ghaoul, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities In System And Control Theory*, volume 15. Society for Industrial Mathematics, 1987.

[2] P Davis. Plane regions determined by complex moments, 1977.

[3] M. Fazel. *Matrix Rank Minimization With Applications.* PhD thesis, PhD thesis, Stanford University, 2002.

[4] R.A. Horn and C.R. Johnson. *Matrix analysis.* Cambridge university press, 1990.

[5] P. Milanfar, M. Putinar, J. Varah, B. Gustafsson, and G. Golub. Shape reconstruction from moments: Theory, algorithms, and applications. In *Proceedings of SPIE*, volume 4116, pages 406–416, 2000.