Pushing the Limits of Affine Rank Minimization by Adapting Probabilistic PCA

Bo Xin ¹ David Wipf ²

¹Peking University
²Microsoft Research, Beijing

July 8, 2015
Outline

1. Introduction
2. Bayesian Affine Rank Minimization
3. Experimental Results
4. Conclusions
Section 1

1. Introduction

2. Bayesian Affine Rank Minimization

3. Experimental Results

4. Conclusions
General problem

\[
\min_{\mathbf{X} \in \mathbb{R}^{n \times m}} \text{rank}(\mathbf{X}) \\
\text{s.t. } \mathcal{A}(\mathbf{X}) = \mathbf{b} \quad \mathcal{A} : \mathbb{R}^{n \times m} \to \mathbb{R}^{p}
\]

(1)

Special case: Matrix completion

\[
\min_{\mathbf{X} \in \mathbb{R}^{n \times m}} \text{rank}(\mathbf{X}) \\
\text{s.t. } \mathbf{X}_{ij} = (\mathbf{X}_0)_{ij}, \quad (i, j) \in \Omega, \quad |\Omega| = p.
\]

(2)

Problems

- NP-hard
Common Solutions

The target

\[
\min_{X \in \mathbb{R}^{n \times m}} \text{rank}(X) \quad \text{s.t.} \quad \mathcal{A}(X) = b \quad \mathcal{A} : \mathbb{R}^{n \times m} \to \mathbb{R}^{p}
\]  

(3)

common surrogates

\[
\min_{X \in \mathbb{R}^{n \times m}} \sum_i f(\sigma_i[X]) \quad \text{s.t.} \quad \mathcal{A}(X) = b \quad \mathcal{A} : \mathbb{R}^{n \times m} \to \mathbb{R}^{p}
\]  

(4)

Special cases

- \( f(z) = I[z \neq 0] \rightarrow \text{matrix rank}. \)
- \( f(z) = z \rightarrow \text{(commonly applied) convex nuclear norm}. \)
- \( f(z) = \log(z) \text{ or } f(z) = z^q \text{ with } q \leq 1 \rightarrow \text{non-convex}. \)
Figure: Plots of different surrogates for matrix rank in a 1D feasible subspace. Here the convex nuclear norm does not retain the correct global minimum. In contrast, although the non-convex $\sum_i \log (\sigma_i [X]^2 + \gamma)$ penalty exhibits the correct minimum when $\gamma$ is sufficiently small, it also contains spurious minima.

Key issues
- Nuclear norm: strong assumptions on measurement process ($\mathcal{A}$) and the underlying matrix ($X_0$).
- Non-convex ones: convergence to local minimum (including tuning parameters).
As a Consequence

**Figure:** Plots from Lu et al. (2014). The performance of a matrix completion task.

**Remarks**

- Degree of freedom: $r(m + n) - r^2$. Theoretical limit: $p = r(m + n) - r^2$.
- Specifically, here $m = n = 150$, $p = 0.5 \times 150^2 = 11250$; $r_{\text{possible}} = 43$.
- Nuclear norm failed at $r = 24$; current best non-convex surrogate failed at $r = 33$. 
Key contributions

- A deceptively simple and parameter-free algorithm.
- Strong empirical performance against the theoretical limit, which has never been demonstrated previously.
- Theoretical inquiry for such substantial performance gains.
Section 2

1 Introduction

2 Bayesian Affine Rank Minimization

3 Experimental Results

4 Conclusions
Basic Model

View it from a Bayesian perspective

\[
p(b|X; A, \lambda) \propto \exp \left[ -\frac{1}{2\lambda} \|A(X) - b\|^2 \right], \tag{5}\]

\[
p(X; \Psi, \nu) = \prod_i \mathcal{N}(x_i; 0, \nu_i \Psi) \propto \exp \left[ x^\top \bar{\Psi}^{-1} x \right], \tag{6}\]

\[
\hat{x} = \text{vex}[\hat{X}] = \bar{\Psi}A^\top \left( \lambda I + A\bar{\Psi}A^\top \right)^{-1} b. \tag{7}\]

Choose \(\Psi\) by maximizing the likelihood marginalized over \(X\)

\[
\max_{\Psi \in H^+, \nu \geq 0} \int p(b|X; A, \lambda)p(X; \Psi, \nu) dX, \tag{8}\]

After a \(-2\log\) transformation, this leads to a new objective

\[
\mathcal{L}(\Psi, \nu) = b^\top \Sigma_b^{-1} b + \log |\Sigma_b|, \tag{9}\]

\[
\Sigma_b = A\bar{\Psi}A^\top + \lambda I \quad \text{and} \quad \bar{\Psi} = \text{diag}[\nu] \otimes \Psi.
\]
Construct upper bounds for both terms

\[ b^\top \Sigma_b^{-1} b \leq \frac{1}{\lambda} \| b - Ax \|_2^2 + x^\top \bar{\Psi}^{-1} x \]  

(10)

\[ \log |\Sigma_b| \equiv m \log |\Psi| + \log \left| \lambda A^\top A + \bar{\Psi}^{-1} \right| \]

\[ \leq m \log |\Psi| + \text{tr} \left[ \Psi^{-1} \nabla_{\Psi^{-1}} \right] + C, \]  

(11)

\[ \nabla_{\Psi^{-1}} = \sum_{i=1}^{m} \Psi - \Psi A_i^\top \left( A\bar{\Psi} A^\top + \lambda I \right)^{-1} A_i \Psi, \]  

(12)

Check the equality conditions

\[ \hat{x} = \text{vex}[\hat{X}] = \bar{\Psi} A^\top \left( \lambda I + A\bar{\Psi} A^\top \right)^{-1} b. \]  

(13)

\[ \Psi^{opt} = \arg \min_X \text{tr} \left[ \Psi^{-1} \left( XX^\top + \nabla_{\Psi^{-1}} \right) \right] + m \log |\Psi| \]

\[ = \frac{1}{m} \left[ \hat{X} \hat{X}^\top + \nabla_{\Psi^{-1}} \right]. \]  

(14)
Symmetric improvements

\[ \tilde{\Psi} = \frac{1}{2} (\Psi_r \otimes I + I \otimes \Psi_c). \]  

(15)

The corresponding updating rules

\[ \hat{x} = \text{vec}[\hat{X}] = \frac{1}{2} (\tilde{\Psi}_r + \tilde{\Psi}_c) A^\top \left[ \lambda I + A \frac{1}{2} (\tilde{\Psi}_r + \tilde{\Psi}_c) A^\top \right]^{-1} b. \]  

(16)

\[ \nabla_{\Psi_r^{-1}} = \sum_{i=1}^{m} \Psi_r - \Psi_r A_{ri}^\top \left( A \tilde{\Psi}_r A^\top + \lambda I \right)^{-1} A_{ri} \Psi_r, \]  

(17)

\[ \nabla_{\Psi_c^{-1}} = \sum_{i=1}^{n} \Psi_c - \Psi_c A_{ci}^\top \left( A \tilde{\Psi}_c A^\top + \lambda I \right)^{-1} A_{ci} \Psi_c, \]  

(18)

\[ \Psi_r^{opt} = \frac{1}{n} \left[ \hat{X}^\top \hat{X} + \nabla_{\Psi_r^{-1}} \right], \quad \Psi_c^{opt} = \frac{1}{m} \left[ \hat{X} \hat{X}^\top + \nabla_{\Psi_c^{-1}} \right]. \]  

(19)
Lemma (Global optimal always retained)

Define $r$ as the smallest rank of any feasible solution to $b = A \text{vec}[X]$, where $A \in \mathbb{R}^{p \times nm}$ satisfies $\text{spark}[A] = p + 1$. Then if $r < p/m$, any global minimizer $\{\Psi^*, \nu^*\}$ of (9) in the limit $\lambda \to 0$ is such that $x^* = \bar{\Psi}^* A^\top \left( A \bar{\Psi}^* A^\top \right)^\dagger b$ is feasible and $\text{rank}[X^*] = r$ with $\text{vec}[X^*] = x^*$.

Lemma (Scale Invariance)

Let $\tilde{A} = AD$, where $D = \text{diag}[\alpha_1 \Gamma, \ldots, \alpha_m \Gamma]$ is a block-diagonal matrix with invertible blocks $\Gamma \in \mathbb{R}^{n \times n}$ of unit norm scaled with coefficients $\alpha_i > 0$. Then iff $\{\Psi^*, \nu^*\}$ is a minimizer (global or local) to (9) in the limit $\lambda \to 0$, then $\{\Gamma^{-1} \Psi^*, \text{diag}[\alpha]^{-1} \nu^*\}$ is a minimizer when $\tilde{A}$ replaces $A$. The corresponding estimates of $X$ are likewise in one-to-one correspondence.

Theorem (Exclusive cases of local minimum smoothed away)

Let $b = A \text{vec}[X]$, where $A$ is block diagonal, with blocks $A_i \in \mathbb{R}^{p_i \times n}$. Moreover, assume $p_i > 1$ for all $i$ and that $\bigcap_i \text{null}[A_i] = \emptyset$. Then if $\min_X \text{rank}[X] = 1$ in the feasible region, any minimizer $\{\Psi^*, \nu^*\}$ of (9) (global or local) in the limit $\lambda \to 0$ is such that $x^* = \bar{\Psi}^* A^\top \left( A \bar{\Psi}^* A^\top \right)^\dagger b$ is feasible and $\text{rank}[X^*] = 1$ with $\text{vec}[X^*] = x^*$. Furthermore, no cost function in the form of (4) can satisfy the same result. In particular, there can always exist local and/or global minima with rank greater than one.
Illustration

Figure: Plots of different surrogates for matrix rank in a 1D feasible subspace. The cost function of BARM smoothes away local minimum while simultaneously retaining the correct global optima.
Section 3

1. Introduction

2. Bayesian Affine Rank Minimization

3. Experimental Results

4. Conclusions
Matrix Completion

Task

- \( X_0 = M_L M_R \), with \( M_L \in \mathbb{R}^{n \times r} \) and \( M_R \in \mathbb{R}^{r \times m} \) \((n = m = 150)\) as iid \( \mathcal{N}(0, 1) \).
- 50% of all entries are then hidden uniformly at random.
- FoS denotes Frequency of Success, where the relative error \( REL = \frac{\|X_0 - \hat{X}\|_F}{\|X_0\|_F} \leq 10^{-3} \).

Figure: Reproduction of the task in Lu et al. (2014).

Remarks

- We significantly outperforms the state-of-the-art and reaches the theoretical limit.
Comparisons with Rank-Aware Algorithms

Task

- $X_0 = M_L M_R$, with $M_L \in \mathbb{R}^{n \times r}$ and $M_R \in \mathbb{R}^{r \times m}$ ($n = m = 150$) as iid $\mathcal{N}(0, 1)$.
- $[U, S, V] = X_0$; $s = \text{diag}[S]$; $\forall i, s_i = s_i \ast \left( \frac{1}{i^{0.8}} \right)$; $X_0 = U\text{diag}[s] V^T$.
- 50% of all entries are then hidden uniformly at random.

![Graphs showing comparisons with rank-aware algorithms. BARM has no knowledge of the true rank.](image)

(a) $X$ iid Gaussian

(b) $X$ has decaying singular values

**Figure:** Comparisons with rank-aware algorithms. BARM has no knowledge of the true rank.
General Problems

**Task**

- \( X_0 = M_L M_R \), with \( M_L \in \mathbb{R}^{n \times r} \) and \( M_R \in \mathbb{R}^{r \times m} \) \((n = m = 100)\) as iid \( \mathcal{N}(0, 1) \).
- Uncorrelated \( A \): iid \( \mathcal{N}(0, 1) \), \( p \times n^2 \) matrix \((p=1000)\);
- Correlated \( A \): \( \sum_{i=1}^{p} i^{-1/2} u_i v_i^\top \), where \( u_i \in \mathbb{R}^p \) and \( v_i \in \mathbb{R}^{n^2} \) are iid \( \mathcal{N}(0, 1) \) vectors.

![Graphs showing FoS and REL for different ranks and methods: (a) \( A \) uncorrelated, (b) \( A \) correlated.]

**Remarks**

- Over a wide battery of empirical tests, our algorithm is consistently able to reach the theoretical limits.
Pushing the Limit and Delve into Failure Cases

Figure: Singular value averages of failure cases. (at $p = (m + n)r - r^2$).

Remarks

- Solutions of correct minimal rank are obtained even though $\hat{X} \neq X_0$.
- Define rank success as when $\sigma_r[\hat{X}] / \sigma_{r+1}[\hat{X}] > 10^3$, where $r$ is the rank of the true low-rank $X_0$. FoRS denotes Frequency of Rank Success.
### Table: Further matrix completion comparisons of BARM with IRLS0 by pushing the limits.

<table>
<thead>
<tr>
<th>Problem</th>
<th>IRLSO</th>
<th>BARM</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>n(=m)</td>
<td>r</td>
</tr>
<tr>
<td>0.9</td>
<td>100</td>
<td>14</td>
</tr>
<tr>
<td>0.95</td>
<td>100</td>
<td>14</td>
</tr>
<tr>
<td>0.99</td>
<td>100</td>
<td>14</td>
</tr>
</tbody>
</table>

### Remarks
- BARM failures are converted to successes under the FoRS metric.
- The other algorithms display almost identical behavior under either metric.
Application-Low-rank Image Rectification

Task
- Construct a first-order Taylor series approximation around the current rectified image estimate
- Reduced to rank minimization under general affine constraints.

Figure: Image rectification comparisons using a checkboard image. *Top:* Original image with observed region (red box) and estimated transformation (green box). *Bottom:* Rectified image estimates.

(a) Nuclear norm (easy)  (b) BARM (easy)  (c) Nuclear norm (hard)  (d) BARM (hard)
Section 4

1. Introduction

2. Bayesian Affine Rank Minimization

3. Experimental Results

4. Conclusions
Discussions

- Model justification based on technical considerations rather than the legitimacy of priors.

- Computational complexity: worst case scale linearly in the elements of $X$ and quadratically in the number of observations.

- More challenging test would be interesting and could potentially show the robustness of BARM.

- Extension to tensor analysis would be interesting.
Conclusions and Discussions

Take home messages

- A deceptively simple and parameter-free algorithm, with very strong empirical performance and theoretical inquiries.
- Code: [http://idm.pku.edu.cn/staff/boxin/](http://idm.pku.edu.cn/staff/boxin/)
Thank you!


