An incomplete survey of matrix completion

Benjamin Recht
Department of Computer Sciences
University of Wisconsin-Madison
Abstract Setup: Matrix Completion

M =

- How do you fill in the missing data?

M_{ij} known for black cells
M_{ij} unknown for white cells

Rows index movies
Columns index users

M = L * R*

k x n
k x r
r x n

kn entries
r(k+n) entries
Low-rank Matrix Completion

- **PROBLEM:** Find the matrix of lowest rank has the specified entries

\[
\begin{align*}
\text{minimize} & \quad \text{rank}(X) \\
\text{subject to} & \quad X_{ij} = M_{ij} \quad \forall (i, j) \in \Omega
\end{align*}
\]

- **When is this problem easy?**
  - Which algorithms?
  - Which sampling sets?
  - Which low-rank matrices?
Which algorithm?

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

\[
\Phi(X) = y \\
\Phi : \mathbb{R}^{k×n} \rightarrow \mathbb{R}^m
\]

\[
\text{minimize } \quad \text{rank}(X) \\
\text{subject to } \quad \Phi(X) = y
\]

• **NP-HARD:**
  – Reduce to MAXCUT
  – Hard to approximate
  – Exact algorithms are awful
Which Algorithm?

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

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

- Nuclear norm is the “numerical rank” in numerical analysis
- The “trace heuristic” from controls if $X$ is p.s.d.
First theory result

\[ \Phi(X) = y \quad \Phi : \mathbb{R}^{k \times n} \rightarrow \mathbb{R}^m \]

- If \( m > c_0 r(k+n-r) \log(kn) \), the heuristic succeeds for most maps \( \Phi \).

Recht, Fazel, and Parrilo. 2007.

- Number of measurements \( c_0 \ r(k+n-r) \ \log(kn) \)

**Approach:** Show that a random \( \Phi \) is nearly an isometry on the manifold of low-rank matrices.

- Stable to noise in measurement vector \( y \) and returns as good an answer as a truncated SVD of the true \( X \).
If we can choose the samples...

- Generically, first $r$ rows and $r$ columns are sufficient:

  $$M = \begin{bmatrix} A & B \\ C & CA^{-1}B \end{bmatrix}$$

- [FKV98, DKM03, etc.]: sample proportional to norms of columns. Low-rank matrix approximations.

If we can’t choose the samples...

- Most sets with more than $2r n^\beta \log(n)$ entries have at least one entry for every row and column, the row-column graph is connected.

- [AM04]: random sampling sufficient to obtain an additive error approximation to $M$. 

Bounds for Matrix Completion

- Suppose $\mathbf{X}$ is $k \times n$ ($k \leq n$) has rank $r$ and has row and column spaces with coherence bounded above by $\mu$. Then the nuclear norm heuristic recovers $\mathbf{X}$ from most subsets of entries $\Omega$ with cardinality at least

$$|\Omega| \geq C\mu n^{6/5} r \log(n)$$

_Candès and Recht. 2008_

special case extensions:

$$|\Omega| \geq C\mu^2 n r \log^6(n)$$

[CT09] _stronger assumptions_

$$|\Omega| > C'n\log(n)$$

[KMO09] _rank = o(1), $\sigma_1/\sigma_r$ bounded_

$$|\Omega| \geq 32\mu r(n + k) \log^2(2n)$$

[GLFBE09,G09,R09]
Incoherence is necessary

\[ X = \begin{pmatrix} \end{pmatrix} (= e_1 e_1^*) \]

- Any subset of entries that misses the (1,1) component tells you nothing!

\[ X = \begin{pmatrix} \end{pmatrix} (= e_1 v^*) \]

- Still need to see the entire first row
- Want each entry to provide nearly the same amount of information
• **RNLA**: matrix is an oracle, summarize the matrix, lift the sketch to the full matrix

• **MC**: entries are provided in advance, solve a convex program subject to agreeing with the observations

<table>
<thead>
<tr>
<th></th>
<th>RNLA</th>
<th>MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Get to choose your samples</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Assumptions on matrix</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Optimal Rates</td>
<td>YES</td>
<td>Õ(YES)</td>
</tr>
<tr>
<td>Robustness</td>
<td>A lot</td>
<td>Some</td>
</tr>
<tr>
<td>Fast algorithms</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>

- If you can control your sampling set at all *do it*.
- When do we ever get uniform samples in reality?
- Why bother with matrix completion at all?
- Perhaps there’s something something bigger going on.
Linear Inverse Problems

• Find me a solution of

\[ y = \Phi x \]

• \( \Phi \) m x n, m<n

• Of the infinite collection of solutions, which one should we pick?

• Leverage structure:

- [Sparsity]
- [Rank]
- [Smoothness]
- [Symmetry]

• How do we design algorithms to solve underdetermined systems problems with priors?
Sparsity

- 1-sparse vectors of Euclidean norm 1

- Convex hull is the unit ball of the $l_1$ norm
  \[
  \{ x : \| x \|_1 \leq 1 \}
  \]

\[
\| x \|_1 = \sum_{i=1}^{n} |x_i|
\]
minimize $\|x\|_1$
subject to $\Phi x = y$

Compressed Sensing: Candes, Romberg, Tao, Donoho, Tanner, Etc...
• 2x2 matrices
• plotted in 3d

\[
\begin{bmatrix}
x & y \\
y & z \\
\end{bmatrix}
\]

rank 1
\[x^2 + z^2 + 2y^2 = 1\]

Convex hull:
\[
\{X : \|X\|_* \leq 1\}
\]

\[
\|X\|_* = \sum_i \sigma_i(X)
\]
• 2x2 matrices
• plotted in 3d

\[
\begin{bmatrix}
x & y \\
y & z
\end{bmatrix} \overset{\ast}{\leq} 1
\]

\[
\|X\|_\ast = \sum_i \sigma_i(X)
\]

Nuclear Norm Heuristic

Fazel 2002.
R, Fazel, and Parillo 2007
Rank Minimization/Matrix Completion
Integer Programming

- Integer solutions: all components of $x$ are $\pm 1$

- Convex hull is the unit ball of the $l_1$ norm
  \[ \{ x : \| x \|_\infty \leq 1 \} \]

  \[ \| x \|_\infty = \max_i |x_i| \]
minimize $\|x\|_\infty$
subject to $\Phi x = y$

Donoho and Tanner 2008
Mangasarian and Recht. 2009.
Parsimonious Models

- Search for best linear combination of fewest atoms
- “rank” = fewest atoms needed to describe the model

\[ x = \sum_{k=1}^{r} w_k \alpha_k \]

\[ \|x\|_A \equiv \inf_{(w,\alpha)} \sum_{k=1}^{r} |w_k| \]
Union of Subspaces

- X has structured sparsity: linear combination of elements from a set of subspaces \{U_g\}.
- Atomic set: unit norm vectors living in one of the U_g

Permutations and Rankings

- X a sum of a few permutation matrices
- Examples: Multiobject Tracking, Ranked elections, BCS
- Convex hull of permutation matrices: doubly stochastic matrices.
Moments: convex hull of \([1, t, t^2, t^3, t^4, \ldots]\), \(t \in T\), some basic set.

System Identification, Image Processing, Numerical Integration, Statistical Inference

Solve with semidefinite programming

Cut-matrices: sums of rank-one sign matrices.

Collaborative Filtering, Clustering in Genetic Networks, Combinatorial Approximation Algorithms

Approximate with semidefinite programming

Low-rank Tensors: sums of rank-one tensors

Computer Vision, Image Processing, Hyperspectral Imaging, Neuroscience

Approximate with alternating least-squares
Atomic Norms

• Given a basic set of atoms, $\mathcal{A}$, define the function
  \[ \|x\|_{\mathcal{A}} = \inf \{ t > 0 : x \in t\text{conv}(\mathcal{A}) \} \]

• When $\mathcal{A}$ is centrosymmetric, we get a norm
  \[ \|x\|_{\mathcal{A}} = \inf \{ \sum_{a \in \mathcal{A}} |c_a| : x = \sum_{a \in \mathcal{A}} c_a a \} \]

\textbf{IDEA:} \minimize \|z\|_{\mathcal{A}} \quad \text{subject to} \quad \Phi z = y

• When does this work?
• How do we solve the optimization problem?
Tangent Cones

- Set of directions that decrease the norm from $x$ form a cone:
  $$\mathcal{T}_A(x) = \{ d : \| x + \alpha d \|_A \leq \| x \|_A \text{ for some } \alpha > 0 \}$$

- $x$ is the unique minimizer if the intersection of this cone with the null space of $\Phi$ equals $\{0\}$

$$y = \Phi z$$

minimize $\| z \|_A$

subject to $\Phi z = y$

$$\{ z : \| z \|_A \leq \| x \|_A \}$$
Mean Width

\[ S_C(d) = \sup_{x \in C} d'x \]

Support Function: \( S_C(d) = \sup_{x \in C} d'x \)

\( S_C(d) + S_C(-d) \)
measures width of \( C \) when projected onto span of \( d \).

mean width: \( w(C) = \int_{S^{n-1}} S_C(u) du \)
• When does a random subspace, $U$ in $\mathbb{R}^n$, intersect a convex cone $C$ at the origin?

• **Gordon (1988):** with high probability if

$$\text{codim}(U) \geq n \ w(C \cap S^{n-1})^2$$

where $w(C \cap S^{n-1}) = \int_{S^{n-1}} S_C(u) du$ is the mean width.

• **Corollary:** For inverse problems, if $\Phi$ is a random Gaussian matrix with $m$ rows, need

$$m \geq n \ w(\mathcal{T}_A(x) \cap S^{n-1})^2$$

for exact recovery of $x$. 
**Duality**

\[
\begin{align*}
w(C \cap S^n) &= \mathbb{E}_u \left[ \max_{v \in C, \|v\| = 1} \langle v, u \rangle \right] \\
&\leq \mathbb{E}_u \left[ \max_{v \in C, \|v\| \leq 1} \langle v, u \rangle \right] \\
&= \mathbb{E}_u \left[ \min_{w \in C^*} \|u - w\| \right]
\end{align*}
\]

- \(C^*\) is the polar cone.
- \(C^* = \{w : \langle w, z \rangle \leq 0 \ \forall z \in C\}\)

\[
\mathcal{T}_A(x)^* = \mathcal{N}_A(x)
\]

- \(\mathcal{N}_A(x)\) is the *normal cone*. Equal to the cone induced by the subdifferential of the atomic norm at \(x\).
Rates

• Hypercube: \[ m \geq n/2 \]

• Sparse Vectors, n vector, sparsity s
  \[ m \geq 2s \log \left( \frac{n}{s} \right) + \frac{5s}{4} \]

• Block sparse, M groups (possibly overlapping), maximum group size B, k active groups
  \[ m \geq k \left( \sqrt{2 \log (M - k)} + \sqrt{B} \right)^2 + kB \]

• Low-rank matrices: \( n_1 \times n_2 \), \( n_1 < n_2 \), rank \( r \)
  \[ m \geq 3r(n_1 + n_2 - r) \]
General Cones

• **Theorem:** Let $C$ be a nonempty cone with polar cone $C^*$. Suppose $C^*$ subtends normalized solid angle $\mu$. Then

$$w(C) \leq 3\sqrt{\log \left( \frac{4}{\mu} \right)}$$

• **Corollary:** For a vertex-transitive (i.e., “symmetric”) polytope with $p$ vertices, $O(\log p)$ Gaussian measurements are sufficient to recover a vertex via convex optimization.

• For $n \times n$ permutation matrix: $m = O(n \log n)$
• For $n \times n$ cut matrix: $m = O(n)$
Atomic Norm Minimization

**IDEA:**

\[
\begin{align*}
\text{minimize} & \quad \|z\|_{\mathcal{A}} \\
\text{subject to} & \quad \Phi z = y
\end{align*}
\]

- Generalizes existing, powerful methods
- Rigorous formula for developing new analysis algorithms
- Precise, tight bounds on number of measurements needed for model recovery
- One algorithm prototype for all data-mining applications
Extensions

• Width Calculations for more general structures (clustering matrices, spectrum estimation, system identification)

• Recovery bounds for structured measurement matrices (application specific)

• Understanding of the loss due to convex relaxation and norm approximation

• Scaling generalized shrinkage algorithms to massive data sets
Acknowledgements

• See:

  http://pages.cs.wisc.edu/~brecht/publications.html

  for all references

• Results developed in collaboration with Emmanuel Candes, Venkat Chandrasekaran, Maryam Fazel, Babak Hassibi, Pablo Parrilo, Weiyu Xu, and Alan Willsky.