Semidefinite and Second Order Cone Programming Seminar Fall 2012 Low Rank Matrix Completion

Marta Cavaleiro

12/03/2012

1 Overview

Imagine we have an $n_1 \times n_2$ matrix from which we only get to see a small number of the entries. Is it possible from the available entries to guess the many entries that are missing? In general it is an impossible task because the unknown entries could be anything. However, if one knows that the matrix is low rank and makes a few reasonable assumptions, then the matrix can indeed be reconstructed and often from a surprisingly low number of entries. This field of research, matrix completion, was started with the results in [1] and [2]. There, it was shown, that under some conditions, recovering a rank$-r$ matrix from randomly selected matrix elements, can be done efficiently by minimizing the nuclear norm of the matrix, which can be converted in a semi-definite program.

In this work we review in an intuitive way the main results of two seminal papers and some of the well-known applications of the matrix completion problem.

2 Some Applications of Matrix Completion

Suppose we are interested in recovering a data matrix M of size $n_1 \times n_2$ and only get to know m of its entries, and m is much smaller than the total number of the entries $n_1 n_2$. The problem is clearly impossible if we do not impose some assumptions. It turns out that in many applications we know that the data represented in the matrix has low dimension, so one natural assumption is to consider that the matrix we want to find has low rank.

Next we present two popular applications of low rank matrix completion.

- **Collaborative filtering & the *Netflix* problem**

  Collaborative filtering problem consists in making predictions about the interests of a user by collecting preference information from many users. The *Netflix* problem is a well-known example of such problem. In this case the goal is to predict customers ratings of unwatched movies given the information about their own preferences and others users, so that *Netflix*, the movie-rental service, can do recommendations.

  Consider the matrix where each row correspond to each *Netflix* customer and each column to a movie, and each entry i, j of the matrix corresponds to the rating given by user i to movie j. The size of this matrix is obviously very large and, since each user rates only a few movies, there are many entries of the matrix that are missing, and that *Netflix* is interested in predicting. Now, it turns out that only few factors determine a user's preference in movies (e.g. genre, lead actor/actresses, director, year, etc.), that is, there is a relatively small number of "types" of people with respect to movie preferences. Also, customers who agreed on movies ratings in the past will be likely to agree in the future. For these reasons it is a natural assumption to consider that the *Netflix* matrix is low rank.

- **Positioning from local distances**

  Consider the problem of trying to find the positions of a large number of points, $x_1, ..., x_n \in \mathbb{R}^d$, that is its coordinates relatively to each other, from information about the pairwise distances between them. If all the pairwise distances are known exactly, then the shape of the network can actually be recovered via a technique called *Multidimensional Scaling*. But, a more interesting and practical case is when many of the distances are unknown. This problem can be considered as a matrix completion one, where we have as many rows and columns as points, and the entry (i, j) corresponds to the square of the distance between the points i and j. It turns out that the matrix of the squares of the distances between the points has a fixed maximum rank depending on the dimension of the space in which the points are embedded. To see this consider x_i^k the k-th coordinate of x_i. Since

  $$\|x_i - x_j\|^2 = \|x_i\|^2 - 2x_i x_j^T + \|x_j\|^2,$$

  the matrix of the squares of the pairwise distances, M, can be written as

  $$M = \begin{bmatrix} \|x_1\|^2 & -2x_1^1 & \cdots & -2x_1^d & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ \|x_n\|^2 & -2x_n^1 & \cdots & -2x_n^d & 1 \end{bmatrix} \begin{bmatrix} 1 & \cdots & 1 \\ x_1^1 & \cdots & x_n^1 \\ \vdots & & \vdots \\ x_1^d & \cdots & x_n^d \\ \|x_1\|^2 & \cdots & \|x_n\|^2 \end{bmatrix}.$$

  So, M can be written as the product of a matrix with $d + 2$ columns and another with $d + 2$ rows. Then, the rank of M is at most the minimum

between the ranks of those two matrices, so it is bounded by $d+2$, which is much smaller than n.

3 Rank Minimization Approach

Throughout these notes we will assume that the matrix we want to recover, M, is of size $n_1 \times n_2$ and has rank r such that $r \ll \min(n_1, n_2)$. The set of available entries is $\Lambda \subset \{1, ..., n_1\} \times \{1, ..., n_2\}$. Now, observe the following:

Claim 1 *A matrix of size $n_1 \times n_2$ and rank r has $(n_1 + n_2 - r)r$ number of degrees of freedom.*

Proof: Consider the Singular Value Decomposition of M:

$$M = \sum_{1 \leq k \leq r} \sigma_k \mathbf{u}_k \mathbf{v}_k^T.$$

Since $\mathbf{u}_1$ is a unit vector, then it only has $n_2 - 1$ degrees of freedom. Now, $\mathbf{u}_2$ besides being unit as well, also must be orthogonal to $\mathbf{u}_1$, and therefore has only $n_2 - 2$ degrees of freedom. And so on, for the first r $\mathbf{u}_k$ singular vectors. The same applies to the singular vectors $\mathbf{v}_k$. The r non-zero singular values constitute r more degrees of freedom. So the total number of degrees of freedom is:

$$(n_2 - 1) + ... + (n_2 - r) + (n_1 - 1) + ... + (n_1 - r) + r =$$
$$n_1 r - \frac{r}{2}(r+1) + n_2 r - \frac{r}{2}(r+1) + r =$$
$$n_1 r + n_2 r - r^2.$$

■

If M is low rank, we have a small number of degrees of freedom, so a natural question is: do we really need to see everything in the matrix to get to know it, when the number of degrees of freedom is much smaller than the size of the matrix? In general the answer is yes, even if the number of observed entries is large sometimes such task is impossible. Also, note that if the number of observed entries is less than the degrees of freedom then clearly no matter which entries are available, there can be an infinite number of matrices of rank at most r with the exact same entries.

Thus, if the measurements are sufficiently many and somehow 'in the right positions', one might hope that there is only one low rank matrix that has those entries. If this was true, a common sense approach would be to solve the optimization problem:

$$\begin{array}{ll} \min & \operatorname{rank}(X) \\ \text{s.t.} & X_{ij} = M_{ij}, \quad (i,j) \in \Lambda \end{array} \tag{1}$$

If there was only one low rank matrix fitting the data then this would recover M. However, solving this problem is not practical since it is known to be NP-hard. Section 5 will introduce an alternative approach that turns out to be efficient under some assumptions. Bur firstly, we give some intuition in the next section that concern the requirements that a matrix need to obey in order to completion from few entries be possible.

4 The Coherence Property

Since we know that we cannot recover all matrices, so what kind of requirements does a matrix need to obey in order to be recovered by only a small number of its entries?

Consider M of rank 1 and of the form $\boldsymbol{x}\boldsymbol{y}^T$. If we do not have samples from a given row, say the i-th one, then one could never guess the value of the first component x_i, by any method whatsoever, since no information about x_i is observed. In the case of the *Netflix* problem that would mean that there is a user that did not rate any movie, so trying to guess his/her preferences is impossible. The same naturally holds for the existence of an unobserved column. So this shows that at least one needs one observation per row and per column. This leads to the assumption:

Assumption: *The observed entries of the matrix are selected uniformly at random.*

This way we are considering what happens for most sampling sets since this assumption makes it unlikely that cases like there are rows or columns totally unobserved happen.

Let M be the following matrix:

$$M = \boldsymbol{e}_1\boldsymbol{x}^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_{n_2} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

Clearly, this matrix cannot be recovered from a sampling of its entries unless the sample is close to exhaustive. The reason is that for most sampling sets we would never be able to see all the entries of the first row, which cannot be recovered in any other way.

More generally, consider a row (column) that has no relationship with the other rows (columns) in the sense that it is orthogonal to them, so it occupies its own separate component of the singular value decomposition of M. Such a row (column) is then impossible to complete exactly without sampling the entire row (column). Thus, to get exact matrix completion from a small fraction of entries, one needs some geometric assumption on the singular vectors, which spreads them out across all coordinates in a roughly even manner, as opposed to being concentrated on just a few values.

Such informal considerations led the authors of [8] to introduce the incoherence assumption, that somehow quantifies how close to the standard basis the vectors of a subspace are.

Definition 1 (Coherence parameter, [1]) *Let U be a subspace of $\mathbb{R}^d$ of dimension r, and P_U be the orthogonal projection onto U. The coherence of U (with respect to the standard basis, $\{e_1, ..., e_d\}$, is defined to be*

$$\mu(U) = \frac{d}{r} \max_{1 \leq i \leq n} \|P_U e_i\|^2. \tag{2}$$

Observation 2 *For any subspace U the coherence parameter is always such that:*

$$1 \leq \mu(U) \leq \frac{d}{r}.$$

Note that $\mu(U) \leq \frac{d}{r}$ since if one of the standard basis vectors is in U then the norm of its projection is 1. An intuition behind why $\mu(U) \geq 1$ is that the smallest value for $\mu(U)$ will correspond to a space U where all vectors of the standard basis are "equally close" to U. So, for instance, when U is the subspace (of $\mathbb{R}^d$) generated by the vector $x = \frac{1}{\sqrt{d}}[1, 1, ..., 1]^d$ then, for any $i = 1, ..., d$:

$$P_U e_i = xx^T e_i = \frac{1}{d}[1, 1, ..., 1]^d,$$

so $\|P_U e_i\|^2 = \frac{1}{d} = \frac{r}{d}$ for all $i = 1, ..., d$.

Consider the row space of M, U, and the column space V. Thus, from our intuition we are interested in matrices whose subspaces U and V have small coherence. Hence, with that in mind, consider the following definition:

Definition 3 (μ_0-Incoherence) *Given a $n_1 \times n_2$ matrix M of rank r, we say that M is μ_0-incoherent if:*

$$\max(\mu(U), \mu(V)) \leq \mu_0, \quad \text{for some} \quad \mu_0.$$

Thus we are interested in matrices with low coherence.

Observe that the above definition is equivalent to:

$$\|P_U e_i\|^2 \leq \frac{\mu_0 r}{n_1}, \text{ for all } i = 1, ..., n_1,$$

$$\|P_V e_j\|^2 \leq \frac{\mu_0 r}{n_2}, \text{ for all } j = 1, ..., n_2.$$

And, naturally, $\mu_0 \geq 1$.

The following theorem reflects the importance of the coherence factor and imposes a bound on the number of observed entries so that it can be fully recovered. We present a simplified version to avoid unnecessary technicalties.

Theorem 4 (Candés and Tao, 2009, [2]) *Suppose we want to recover a matrix* M *of size* $n_1 \times n_2$ *and rank* r *from the set of samples* Λ *taken uniformly at random. Suppose* M *is* μ*-incoherent and consider* $n = \min\{n_1, n_2\}$ *and* $0 < \delta < 1$*. Then, if*

$$|\Lambda| \leq \mu n r \log\left(\frac{n}{\delta}\right),$$

there are infinitely many μ*-incoherent matrices* $X \neq M$ *of rank at most* r *such that* $X_{ij} = M_{ij}$*, for* $(i,j) \in \Lambda$*, with probability at least* δ*.*

Recall that the number of degrees of freedom is about $2nr$ (a very rough upper bound), so the theorem implies that to recover an arbitrary rank-r and μ-incoherent matrix with a decent probability by any method, the minimum number of samples must be about the number of degrees of freedom times $\mu \log n$. Since $\mu \geq 1$, then at least about $nr \log n$ samples are really needed.

5 Exact solution through Nuclear Norm Minimization

As we have seen, the rank minimization approach (1) is not a possible way of solving our problem. An alternative approach is minimizing the nuclear norm of the matrix which is a convex function. The nuclear norm of a matrix X is defined as

$$\|X\|_* = \sum_{k=1}^{n} \sigma_k(X),$$

that is, the sum of its singular values. In the case where X is positive semidefinite (PSD) the nuclear norm is exactly the trace of the matrix. Now let $\lambda(X)$ be the vector of the eigenvalues of X, and observe that

$$\|\lambda(X)\|_0 = \operatorname{rank}(X),$$

$$\|\lambda(X)\|_1 = \operatorname{trace}(X).$$

It is well-known that to obtain a sparse vector from an underdetermined linear system, minimizing its ℓ_1-norm is an effective heuristic that tends to find sparse solutions. This is the intuition behind the use of the nuclear norm. This heuristic was studied in [3] where it was shown that the nuclear norm approach can be used for any matrix (not necessarily PSD or even square).

Thus, we will solve (3) instead of (1):

$$\begin{array}{ll} \min & \|X\|_* \\ \text{s.t.} & X_{ij} = M_{ij}, \quad (i,j) \in \Lambda \end{array} \tag{3}$$

The first results on the minimum number of needed observed entries of M so that it can be recovered by minimizing the nuclear norm, were presented

in [1] using the μ-incoherence property. There, it was proved that at least $O(n^{1.2} r \log n)$ samples were needed to recover the matrix with high probability, for $n = \min\{n_1, n_2\}$. Later, using a stronger version of incoherence, Candès and Tao in [2] improved the bound above. The following definition allows us to state their main theorem.

Definition 5 (Strong Incoherence Property, [2]) *We say that* M *obeys the strong incoherence property with parameter* μ *if:*

$$\left| \langle e_i, P_U e_j \rangle - \frac{r}{n_1} 1_{i=j} \right| \leq \mu \frac{\sqrt{r}}{n_1}, \quad \textit{for all } (i,j) \in \{1, ..., n_1\}^2;$$

$$\left| \langle e_i, P_V e_j \rangle - \frac{r}{n_2} 1_{i=j} \right| \leq \mu \frac{\sqrt{r}}{n_2}, \quad \textit{for all } (i,j) \in \{1, ..., n_2\}^2;$$

and, for $E := \sum_{1 \leq k \leq r} u_i v_i^T$ *the following holds*

$$|E_{ij}| \leq \mu \frac{\sqrt{r}}{\sqrt{n_1 n_2}} \quad \textit{for all } (i,j) \in \{1, ..., n_1\} \times \{1, ..., n_2\}$$

This property is related to, but slightly different from, the incoherence property. Also it can be proved that $\mu \geq 1$.

Theorem 6 (Candès and Tao, 2009, [2]) *Let* M *be a* $n_1 \times n_2$ *matrix of rank* r *obeying the strong incoherence property with parameter* μ*. Consider* $n = \min\{n_1, n_2\}$*. Suppose we observe the entries* $\Lambda \subset \{1, ..., n_1\} \times \{1, ..., n_2\}$ *sampled uniformly at random. Then there is a constant* $C > 0$ *such that if*

$$|\Lambda| \geq C \mu^2 n r \log^6 n, \tag{4}$$

then M *is the unique solution of problem (3) with probability at least* $1 - n^{-3}$*.*

In other words, with high probability and given condition (4) on the number of observed entries, nuclear norm minimization recovers all the entries of M with no error.

For an overview of the main ideas behind the proof we refer the reader to [6] while the proof can be found in [2].

6 Final Remarks

Theorem 6 states that if a matrix is strongly incoherent and the cardinality of the sampled set is about the number of degrees of freedom times a few logarithmic factors, then nuclear norm minimization is exact, a result quite surprising. Similar results were proved later improving the bound of Theorem 6. To the best of our knowledge, the best bound was proved in [4] using a slight variation of coherence. There, for the same probability as in Theorem 6, the bound (4)

was improved to $\mathcal{O}(\mu n r \log^2 n)$, where μ corresponds to definition of coherence adopted in the paper.

A word on what kind of matrices satisfy the strong incoherence property with a small value of μ is missing. We refer two examples shown in [2]. For that consider again $n = \min(n_1, n_2)$ and the SVD decomposition of M:

$$M = \sum_{1 \le k \le r} \sigma_k \mathbf{u}_k \boldsymbol{\nu}_k^T.$$

- Suppose that the singular vectors of M, obey the following :

  $$\|u_k\|_\infty \le \sqrt{\alpha / n_1} \quad \text{and} \quad \|\nu_k\|_\infty \le \sqrt{\alpha / n_2}$$

  with $\alpha = \mathcal{O}(1)$, then M obeys the strong incoherence property with $\mu = \mathcal{O}(\sqrt{\log n})$.

- Assume that the matrices $[\mathbf{u}_1, \ldots, \mathbf{u}_r]$ and $[\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_r]$ are independent random orthogonal matrices, then with high probability, M obeys the strong incoherence property with $\mu = \mathcal{O}(\log n)$.

We finally mention that there are other approaches that also use coherence like definitions to solve the low rank matrix completion other than by minimizing the nuclear norm. As an example we refer the reader to [5] where they focused particularly in the positioning from local distances problem.

References

[1] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. of Comput. Math.*, 2008.

[2] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 2009.

[3] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Elec. Eng. Dept, Stanford University,, 2002.

[4] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, 2011.

[5] Sewoong Oh. *Matrix Completion: Fundamental Limits and Efficient Algorithms*. PhD thesis, Elec. Eng. Dept, Stanford University, 2010.

[6] T. Tao. The power of convex relaxation: near-optimal matrix completion. `http://terrytao.wordpress.com/2009/03/10/the-power-of-convex-relaxation-near-optimal-matrix-completion/`, 2009. [Online; accessed December, 1st, 2012].