Inferring Polynomial Relationships from Partial Data Using Matrix Rank Minimization

Christopher Gadzinski, me@cgad.ski

University of Coimbra

Low Rank Matrix Completion and Company

Suppose $M = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ is a matrix of datapoints, some of whose coordinates are unknown.

$$M = \begin{bmatrix} m_{1,1} & m_{1,2} & ? & m_{1,4} & m_{1,5} & \ldots \\ ? & m_{2,2} & ? & m_{2,4} & ? & \ldots \\ m_{3,1} & ? & m_{3,3} & m_{3,4} & ? & \ldots \\ ? & m_{4,2} & m_{4,3} & ? & m_{4,5} & \ldots \end{bmatrix}$$

How can we infer latent relationships in this kind of **partially observed** data? This problem has applications in the design of *recommender systems*. For example: how can we infer patterns in the way that users rate movies when we have a small set of reviews from each user?

If our data is expected to be low-rank, it's natural to recover M as the matrix of *minimal rank* that agrees with the observed entries. This is the problem of **low rank matrix completion** (LRMC):

LRMC:

$$\begin{array}{rl} \text{minimize} & \operatorname{rank} M \\ \text{subject to} & m_{i,j} = c_{i,j},\ (i,j) \in \Omega \end{array}$$

The problem is essentially to determine the column space V of M from the information that V incides in certain affine subspaces. We call this a **hitting subspace problem**. Conversely, a general hitting subspace problem can be posed as a matrix rank minimization problem if we allow arbitrary affine constraints on each column. We call this **column-affine rank minmization** (CARM). This pair of equivalent problems generalizes the classic notion of LRMC.

Hitting subspace problem:

$$\begin{array}{rl} \text{minimize} & \dim V \\ \text{subject to} & V \cap W_i \neq \emptyset,\ i = 1 \ldots n \end{array}$$

CARM:

$$\begin{array}{rl} \text{minimize} & \operatorname{rank}([\mathbf{x}_1, \ldots \mathbf{x}_n]) \\ \text{subject to} & \pi_i(\mathbf{x}_i) = \tilde{\mathbf{x}}_i,\ i = 1 \ldots n \end{array}$$

In [2], Candès and Recht showed that LRMC is often formally equivalent to a certain semidefinite program. This observation leads to some practical algorithms for LRMC, like the singular value thresholding algorithm (SVT) from [1]. In turn, SVT can be easily generalized to solve the more general CARM problem.

How Much Data is Enough?

When is CARM well-posed, information-theoretically? There are two perspectives to consider:

Are enough degrees of freedom cut? When the r-dimensional column space of M is fixed, each additional column contributes an additional r degrees of freedom, so, a new column with $k > r$ affine constraints should impose $k - r$ extra constraints on V. Naively, we expect that completion of a rank r matrix should be possible when at least

$$\left\lceil \frac{r(m-r)}{k-r} \right\rceil$$

datapoints are given with k observed coordinates each.

Is V identifiable from the projections π_i? Suppose we observe arbitrarily many datapoints under each projection $\pi_1, \ldots, \pi_l$. Generally, the most we can learn about V is summarized by the images $\pi_i(V)$. In fact, for generic subspaces V, LRMC will succeed on a sufficiently large dataset projected under the maps π_i if and only if

$$V = \bigcap_i \pi_i^{-1}(\pi_i(V)).$$

In this case, we say that V is *identifiable* under the projections π_i.

LRMC in Polynomial Feature Space

What if our data is drawn from a nonlinear *algebraic variety*? Can we generalize LRMC to exploit higher degree polynomial relationships?

Let $v \colon \mathbb{R}^m \to (\mathbb{R}^m)^{\otimes p}$ be the Veronese map, sending a point x to its p-fold symmetric tensor power. If the columns $\mathbf{x}_i$ of $M = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ are drawn from a variety $\mathcal{V}$ that is cut out by degree-p polynomials, then the matrix $v(M) = [v(\mathbf{x}_1), \ldots, v(\mathbf{x}_n)]$ will be rank-deficient. So, the natural way to infer M from the column-wise affine constraints $\pi_i(\mathbf{x}_i) = \tilde{\mathbf{x}}_i$ is to solve the following "lifted" version of LRMC.

Nonlinear LRMC:

$$\begin{array}{rl} \text{minimize} & \operatorname{rank}[v(\mathbf{x}_1), \ldots, v(\mathbf{x}_n)] \\ \text{subject to} & \pi_i(\mathbf{x}_i) = \tilde{\mathbf{x}}_i,\ i = 1 \ldots n. \end{array}$$

However, *affine* constraints on the columns of M give rise to implicit *nonlinear* constraints on the columns of $v(M)$. **How can we solve this new class of rank minimization problems?** In [3], Ongie et al. suggested that $v(M)$ could be inferred (in some cases) **as a solution to a LRMC problem.** In our work, we suggest an improved way to infer it **as a solution to a hitting subspace problem.**

Given a coordinate projection $\pi \colon \mathbb{R}^m \to \mathbb{R}^k$ onto a subset S of coordinates, let $\pi^{\otimes p}$ denote the naturally associated map from $(\mathbb{R}^m)^{\otimes p}$ to $(\mathbb{R}^k)^{\otimes p}$ that projects onto the coordinates corresponding to monomials supported on S. When $v \colon \mathbb{R}^k \to (\mathbb{R}^k)^{\otimes p}$ is the Veronese map for $\mathbb{R}^k$, $\pi^{\otimes p}$ makes the following diagram commute.

$$\begin{array}{ccc} \mathbb{R}^m & \xrightarrow{v} & (\mathbb{R}^m)^{\otimes p} \\ \downarrow{\scriptstyle \pi} & & \downarrow{\scriptstyle \pi^{\otimes p}} \\ \mathbb{R}^k & \xrightarrow{v} & (\mathbb{R}^k)^{\otimes p} \end{array}$$

If $\pi(\mathbf{x}) = \tilde{\mathbf{x}}$, then $\pi^{\otimes p}(v(\mathbf{x})) = v(\pi(\mathbf{x})) = v(\tilde{\mathbf{x}})$, so $\pi^{\otimes p}(\mathbf{y}) = v(\tilde{\mathbf{x}})$ is a relaxation of the constraint that $\mathbf{y} \in v(\pi^{-1}(\tilde{\mathbf{x}}))$. On the other hand, a stronger condition is to stipulate that y should lie in the linear span of $v(\pi^{-1}(\tilde{\mathbf{x}}))$. These two approaches give the following rank-minimization problems for $v(M)$.

Tensorized LRMC:

$$\begin{array}{rl} \text{minimize} & \operatorname{rank}[\mathbf{y}_1, \ldots, \mathbf{y}_n] \\ \text{subject to} & y_i \in (\mathbb{R}^m)^{\otimes p}, \\ & \pi_i^{\otimes p}(y_i) = v(\tilde{x}_i),\ i = 1 \ldots n \end{array}$$

Tensorized hitting subspace problem:

$$\begin{array}{rl} \text{minimize} & \operatorname{rank}[\mathbf{y}_1, \ldots, \mathbf{y}_n] \\ \text{subject to} & \mathbf{y} \in \langle v(\pi_i^{-1}(\tilde{\mathbf{x}}_i)) \rangle \setminus 0,\ i = 1 \ldots n \end{array}$$

Ongie et al. showed that *tensorized LRMC* is well-posed on sufficiently large datasets drawn from certain *unions of subspaces*. However, our new *tensorized hitting subspace problem* may let us recover $v(M)$ more frequently, because it imposes more constraints on each column of $v(M)$.

Consider the case $p = 2$. Let $\mathbf{x} = (x_1, \ldots, x_m)$ be a column of M, and suppose the coordinates $x_1, \ldots, x_k$ are known. What constraints can we impose on the corresponding column of $v(M)$,

$$\mathbf{y} = v(x) = (y_{i,j})_{1 \le i \le j \le n} = (x_i x_j)_{1 \le i \le j \le n}?$$

In the LRMC approach, we simply use the $\binom{k+1}{2}$ equations $y_{i,j} = x_i x_j$ for $1 \le i \le j \le k$. However, the tensorized hitting subspace problem leads us to consider, e.g., the $(k-1)(m-k)$ new equations

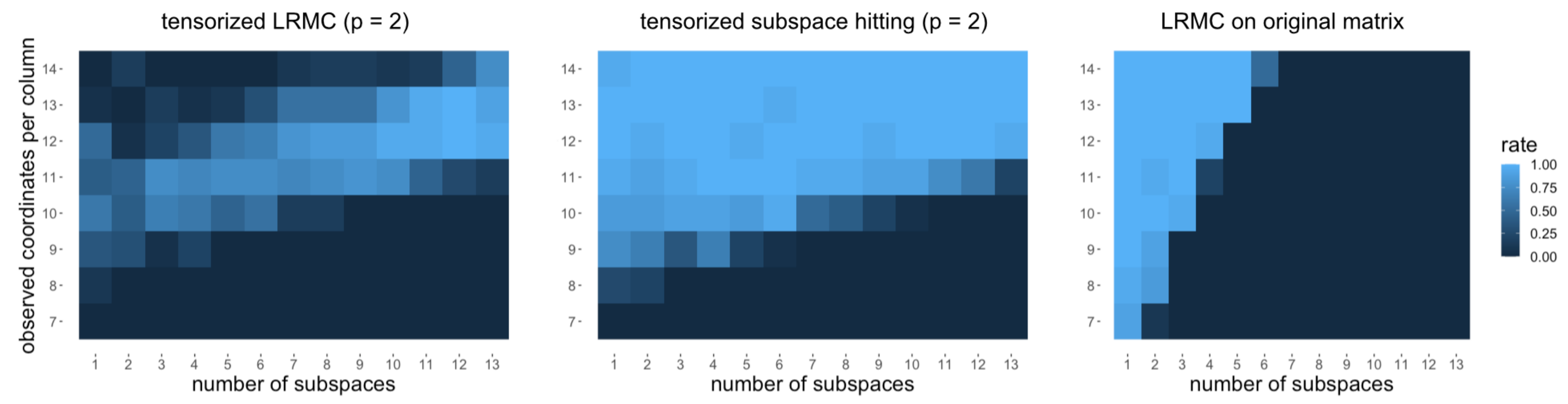
$$y_{1,i} = \frac{x_1}{x_j} y_{j,i}$$

for $k < i \le n$ and $2 \le j \le k$. Especially when $k \ll m$, this is a big improvement!

An Experiment on Unions of Subspaces

Choose K 2-dimensional subspaces $S_1, \ldots, S_K$ in $\mathbb{R}^{15}$. Generate a dataset of $50K$ points, with 50 points drawn from each subspace. Permute the dataset so that the clusters of points drawn from each subspace are unknown, and retain only m coordinates from each datapoint. **Can we recover the missing coordinates?**

We compare the performance of regular LRMC against algorithms derived from the tensorized LRMC problem and the tensorized hitting subspace problem. For a given value of K and m, we solve 20 pseudorandom problem instances and report the fraction of times that our process imputes the missing coordinates of the original matrix to within a modest tolerance. For the LRMC / CARM subproblems, we use naive implementation of singular value thresholding, limited to 200 iterations.



When $K > 7$, the matrix M we are completing is no longer rank-deficient, so applying LRMC to the original matrix completion problem cannot possibly succeed. Meanwhile, as already noted by Ongie et al., lifting the problem into the tensorized domain does let us impute M in this high-rank situation. We extend on this observation: at least with the rank-minimization algorithm we have used, using the *tensorized hitting subspace problem* lets us impute M even more reliably.

Future Directions

It can be shown that the success of tensorized LRMC is contingent on some special properties of the data. Specifically, if no more than k coordinates of each data point are observed, Ongie's method will not succeed at completing data drawn from an algebraic variety $\mathcal{V}$ unless the space of pth degree homogeneous polynomials in the vanishing ideal of $\mathcal{V}$ is generated by polynomials supported on no more than k coordinates. However, while degree-of-freedom reasoning suggests that the tensorized hitting subspace problem will recover $v(\mathcal{V})$ in more situations, little is currently known about its precise limitations.

References and Acknowledgements

[1] Jian-Feng Cai, Emmanuel J. Candes, and Zuowei Shen.
A singular value thresholding algorithm for matrix completion, 2008.

[2] Emmanuel J. Candes and Benjamin Recht.
Exact matrix completion via convex optimization, 2008.

[3] Greg Ongie, Daniel Pimentel-Alarcón, Laura Balzano, Rebecca Willett, and Robert D. Nowak.
Tensor methods for nonlinear matrix completion, 2020.

The work presented in this paper was partially carried out in the scope of the MobiWise project: From mobile sensing to mobility advising (P2020 SAICTPAC/0011/2015), co-financed by COMPETE 2020, Portugal 2020 - Operational Program for Competitiveness and Internationalization (POCI), European Union's ERDF (European Regional Development Fund), and the Portuguese Foundation for Science and Technology (FCT).