Algebraic Variety Models for High-Rank Matrix Completion

Greg Ongie [1] **Rebecca Willett** [2] **Robert D. Nowak** [2] **Laura Balzano** [1]

Abstract

We consider a generalization of low-rank matrix completion to the case where the data belongs to an algebraic variety, *i.e.,* each data point is a solution to a system of polynomial equations. In this case the original matrix is possibly high-rank, but it becomes low-rank after mapping each column to a higher dimensional space of monomial features. Many well-studied extensions of linear models, including affine subspaces and their union, can be described by a variety model, as well as a rich class of nonlinear quadratic and higher degree curves and surfaces. We study the sampling requirements for matrix completion under a variety model with a focus on a union of affine subspaces. We also propose an efficient matrix completion algorithm that minimizes a convex or non-convex surrogate of the rank of the matrix of monomial features, using the well-known "kernel trick" to avoid working directly with the high-dimensional monomial matrix. We show the proposed algorithm is able to recover synthetically generated data up to the predicted sampling complexity bounds, and outperforms standard low rank matrix completion and subspace clustering algorithms in experiments with real data.

1. Introduction

Work in the last decade on matrix completion has shown that it is possible to leverage linear structure in order to interpolate missing values in a low-rank matrix (Candes & Recht, 2012). The high-level idea of this work is that if the data defining the matrix belongs to a structure having fewer degrees of freedom than the entire dataset, that structure provides redundancy that can be leveraged to complete the matrix. The assumption that the matrix is low-rank is equivalent to assuming the data lies on (or near) a low-dimensional linear subspace.

It is of great interest to generalize matrix completion to exploit low-complexity *nonlinear* structures in the data. Several avenues have been explored in the literature, from generic manifold learning (Lee et al., 2013), to unions of subspaces (Eriksson et al., 2012; Elhamifar & Vidal, 2013), to low-rank matrices perturbed by a nonlinear monotonic function (Ganti et al., 2015; Song et al., 2016). In each case missing data has been considered, but there lacks a clear, unifying framework for these ideas.

In this work we study the problem of completing a matrix whose columns belong to an *algebraic variety*, *i.e.,* the set of solutions to a system of polynomial equations (Cox et al., 2015). This is a strict generalization of the linear (or affine) subspace model, which can be written as the set of points satisfying a system of linear equations. Unions of subspaces and unions of affine spaces are also algebraic varieties. Plus, a much richer class of non-linear curves, surfaces, and their unions, are captured by a variety model.

The matrix completion problem using a variety model can be formalized as follows. Let $\boldsymbol{X} = \left[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s\right] \in \mathbb{R}^{n \times s}$ be a matrix of s data points where each column $\boldsymbol{x}_i \in \mathbb{R}^n$. Define $\phi_d : \mathbb{R}^n \to \mathbb{R}^N$ as the mapping that sends the vector $\boldsymbol{x} = (x_1, ..., x_n)$ to the vector of all monomials in $x_1, ..., x_n$ of degree at most d, and let $\phi_d(\boldsymbol{X})$ denote the matrix that results after applying ϕ_d to each column of $\boldsymbol{X}$, which we call the *lifted matrix*. We will show the lifted matrix is rank deficient if and only if the columns of $\boldsymbol{X}$ belong to an algebraic variety. This motivates the following matrix completion approach:

$$\min_{\hat{\boldsymbol{X}}} \ \operatorname{rank} \phi_d(\hat{\boldsymbol{X}}) \ \text{ such that } \ \mathcal{P}_\Omega(\hat{\boldsymbol{X}}) = \mathcal{P}_\Omega(\boldsymbol{X}) \qquad (1)$$

where $\mathcal{P}_\Omega(\cdot)$ represents a projection that restricts to some observation set $\Omega \subset \{1, \ldots, n\} \times \{1, \ldots, s\}$. The rank of $\phi_d(\hat{\boldsymbol{X}})$ depends on the choice of the polynomial degree d and the underlying "complexity" of the variety, in a sense we will make precise. Figure 1 shows three examples of datasets that have low-rank in the lifted space for different polynomial degrees d.

In this work we investigate the factors that influence the sampling complexity of varieties as well as algorithms for

[1] Department of EECS, University of Michigan, Ann Arbor, Michigan, USA [2] Department of ECE, University of Wisconsin, Madison, Wisconsin, USA. Correspondence to: Greg Ongie <gongie@umich.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

completion. The challenges are (a) to characterize varieties having low-rank (and therefore few degrees of freedom) in the lifted space, *i.e.,* determine when $\phi_d(\boldsymbol{X})$ is low-rank, (b) devise efficient algorithms for solving (1) that can exploit these few degrees of freedom in a matrix completion setting, and (c) determine the trade-offs relative to existing matrix completion approaches. This work contributes considerable progress towards these goals.

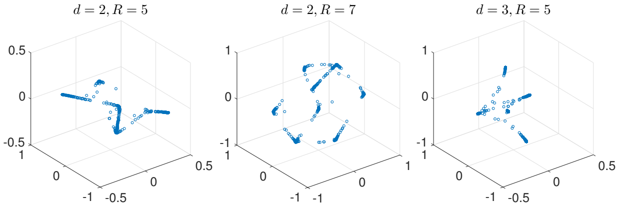


Figure 1. Data belonging to algebraic varieties in $\mathbb{R}^3$. The original data is full rank, but a nonlinear embedding of the matrix to a feature space consisting of monomials of degree at most d is low-rank with rank R, indicating the data has few degrees of freedom.

Our main contributions are as follows. We identify bounds on the rank of a matrix $\phi_d(\boldsymbol{X})$ when the columns of the data matrix $\boldsymbol{X}$ belong to an algebraic variety. We study how many entries of such a matrix should be observed in order to recover the full matrix from an incomplete sample. We show as a case study that monomial representations produce low-rank representations of unions of subspaces, and we characterize the rank. The standard union of subspace representation as a discrete collection of individual subspaces is inherently non-smooth in nature, whereas the algebraic variety allows for a purely continuous parameterization. This leads to a general algorithm for completion of a data matrix whose columns belong to a variety. The algorithm's performance is showcased on data simulated as a union of subspaces, a union of low-dimensional parametric surfaces, and real data from a motion segmentation dataset and a motion capture dataset. The simulations show that the performance of our algorithm matches our predictions and outperforms other methods. In addition, the analysis of the degrees of freedom associated with the proposed representations introduces several new research avenues at the intersection of nonlinear algebraic geometry and random matrix theory.

1.1. Related work

There has been a great deal of research activity on matrix completion problems since (Candes & Recht, 2012), where the authors showed that one can recover an incomplete matrix from few entries using a convex relaxation of the rank minimization optimization problem. For example, it is now known that only $O(rn)$ entries are necessary and sufficient (Pimentel-Alarcón et al., 2016b) for almost every $n \times n$ rank r matrix as long as the measurement pattern satisfies certain deterministic conditions. However, these methods and theory are restricted to low-rank linear models. A great deal of real data exhibit nonlinear structure, and so it is of interest to generalize this approach. Work in that direction has dealt with union of subspaces models (Eriksson et al., 2012; Yang et al., 2015; Elhamifar, 2016; Pimentel-Alarcón et al., 2016a; Pimentel-Alarcon & Nowak, 2016), locally linear approximations (Lee et al., 2013), as well as low-rank models perturbed by an arbitrary nonlinear link function (Ganti et al., 2015; Song et al., 2016; Rao et al., 2017). In this paper we instead seek a more general model that captures both linear and nonlinear structure. The variety model has as instances low-rank subspaces and their union as well as quadratic and higher degree curves and surfaces.

Work on kernel PCA (*cf.,* (Sanguinetti & Lawrence, 2006; Nguyen & Torre, 2009)) leverage similar geometry to ours. In *Kernel Spectral Curvature Clustering* (Chen et al., 2009), the authors similarly consider clustering of data points via subspace clustering in a lifted space using kernels. These works are algorithmic in nature, with promising numerical experiments, but do not systematically consider missing data or analyze relative degrees of freedom.

This paper also has close ties to algebraic subspace clustering (ASC) (Vidal et al., 2003; 2005; 2016; Tsakiris & Vidal, 2015), also known as generalized PCA. Similar to our approach, the ASC framework models unions of subspaces as an algebraic variety, and makes use of monomial liftings of the data to identify the subspaces. Characterizations of the rank of data belonging to union of subspaces under the monomial lifting are used in the ASC framework (Vidal et al., 2016) based on results in (Derksen, 2007). The difference of the results in (Derksen, 2007) and those in Prop. 1 is that ours hold for monomial liftings of all degrees d, not just $d \geq k$, where k is the number of subspaces. Also, the main focus of ASC is to recover unions of subspaces or unions of affine spaces, whereas we consider data belonging to a more general class of algebraic varieties. Finally, the ASC framework has not been adapted to the case of missing data, which is the main focus of this work.

2. Variety Models

As a toy example to illustrate our approach, consider a matrix

$$\boldsymbol{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,6} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,6} \end{pmatrix} \in \mathbb{R}^{2\times 6}$$

whose six columns satisfy the quadratic equation

$$c_0 + c_1\, x_{1,i} + c_2\, x_{2,i} + c_3\, x_{1,i}^2 + c_4\, x_{1,i}x_{2,i} + c_5\, x_{2,i}^2 = 0 \quad (2)$$

for $i = 1, \ldots, 6$ and some unknown constants $c_0, \ldots, c_5$ that are not all zero. Generically, $\boldsymbol{X}$ will be full rank. However, suppose we vertically expand each column of the matrix to make a 6×6 matrix

$$\boldsymbol{Y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{1,1} & x_{1,2} & \cdots & x_{1,6} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,6} \\ x_{1,1}^2 & x_{1,2}^2 & \cdots & x_{1,6}^2 \\ x_{1,1}x_{2,1} & x_{1,2}x_{2,2} & \cdots & x_{1,6}x_{2,6} \\ x_{2,1}^2 & x_{2,2}^2 & \cdots & x_{2,6}^2 \end{pmatrix} \in \mathbb{R}^{6\times 6},$$

i.e., we augment each column of $\boldsymbol{X}$ with a 1 and with the quadratic monomials $x_{1,i}^2$, $x_{1,i}x_{2,i}$, $x_{2,i}^2$. This allows us to re-express the polynomial equation (2) as the matrix-vector product $\boldsymbol{Y}^T\boldsymbol{c} = \boldsymbol{0}$ where $\boldsymbol{c} = (c_0, c_1, .., c_5)^T$. In other words, $\boldsymbol{Y}$ is rank deficient. Suppose, for example, that we are missing entry $x_{1,1}$ of $\boldsymbol{X}$. Since $\boldsymbol{X}$ is full rank, there is no way to uniquely complete the missing entry by leveraging linear structure alone. Instead, we ask: Can we complete $x_{1,1}$ using the linear structure present in $\boldsymbol{Y}$? Due to the missing entry $x_{1,1}$, the first column of $\boldsymbol{Y}$ will having the following pattern of missing entries: $(1, -, x_{2,1}, -, -, x_{2,1}^2)^T$. However, assuming the five complete columns in $\boldsymbol{Y}$ are linearly independent, we can uniquely determine the nullspace vector $\boldsymbol{c}$ up to a scalar multiple. Then from (2) we have

$$c_3\, x_{1,1}^2 + (c_1 + c_4\, x_{2,1})x_{1,1} = -c_0 - c_2\, x_{2,1} - c_5\, x_{2,1}^2.$$

In general, this equation will yield at most two possibilities for $x_{1,1}$. Moreover, there are conditions where we can *uniquely* recover $x_{1,1}$, namely when $c_3 = 0$ and $c_1 + c_4\, x_{2,1} \neq 0$.

This example shows that even without a priori knowledge of the particular polynomial equation satisfied by the data, it is possible to uniquely recover missing entries in the original matrix by leveraging induced linear structure in the matrix of expanded monomials. We now show how to considerably generalize this example to the case of data belonging to an arbitrary algebraic variety.

2.1. Formulation

Let $\boldsymbol{X} = \left[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s\right] \in \mathbb{R}^{n\times s}$ be a matrix of s data points where each column $\boldsymbol{x}_i \in \mathbb{R}^n$. Define $\phi_d : \mathbb{R}^n \to \mathbb{R}^N$ as the mapping that sends the vector $\boldsymbol{x} = (x_1, ..., x_n)$ to the vector of all monomials in $x_1, ..., x_n$ of degree at most d:

$$\phi_d(\boldsymbol{x}) = (\boldsymbol{x}^{\boldsymbol{\alpha}})_{|\boldsymbol{\alpha}|\leq d} \in \mathbb{R}^N$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)$ is a multi-index of non-negative integers, with $\boldsymbol{x}^{\boldsymbol{\alpha}} := x_1^{\alpha_1} \cdots x_n^{\alpha_n}$, and $|\boldsymbol{\alpha}| := \alpha_1 + \cdots + \alpha_n$. In the context of kernel methods in machine learning, the map ϕ_d is often called a polynomial feature map (Muller et al., 2001). Borrowing this terminology, we call $\phi_d(\boldsymbol{x})$ a *feature vector*, the entries of $\phi_d(\boldsymbol{x})$ *features*, and the range of ϕ_d *feature space*. Note that the number of features is given by $N = N(n,d) = \binom{n+d}{n} = \binom{n+d}{d}$, the number of unique monomials in n variables of degree at most d. When $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_s]$ is an $n \times s$ matrix, we use $\phi_d(\boldsymbol{X})$ to denote the $N \times s$ matrix $[\phi_d(\boldsymbol{x}_1), ..., \phi_d(\boldsymbol{x}_s)]$.

The problem we consider is this: can we complete a partially observed matrix $\boldsymbol{X}$ under the assumption that $\phi_d(\boldsymbol{X})$ is low-rank? This can be posed as the optimization problem given above in (1). We give a practical algorithm for solving a relaxation of (1) in Section 4. Similar to previous work cited above on using polynomial feature maps, our method leverages the *kernel trick* for efficient computations. The success of this optimization and its relaxations will depend on many factors, but clearly the rank of $\phi_d(\boldsymbol{X})$ and the number of sampled entries will play an important role. The number of samples, rank, and dimensions all grow in the mapping to feature space, but they grow at different rates depending on the underlying geometry; it is not immediately obvious what conditions on the geometry and sampling rates impact our ability to determine the missing entries. In the remainder of this section, we show how to relate the rank of $\phi_d(\boldsymbol{X})$ to the underlying variety, and we study the sampling requirements necessary for the completion of the matrix in feature space.

2.2. Rank properties

To better understand what determines the rank of the matrix $\phi_d(\boldsymbol{X})$, we introduce some additional notation and concepts from algebraic geometry. Let $\mathbb{R}[\boldsymbol{x}]$ denote the space of all polynomials with real coefficients in n variables $\boldsymbol{x} = (x_1, ..., x_n)$. We model a collection of data as belonging to a *real (affine) algebraic variety* (Cox et al., 2015), which is defined as the common zero set of a system of polynomials $P \subset \mathbb{R}[\boldsymbol{x}]$:

$$V(P) = \{\boldsymbol{x} \in \mathbb{R}^n : f(\boldsymbol{x}) = 0 \text{ for all } f \in P\}.$$

Suppose the variety $V(P)$ is defined by the finite set of polynomials $P = \{f_1, ..., f_q\}$, where each f_i has degree at most d. Let $\boldsymbol{C} \in \mathbb{R}^{N\times q}$ be the matrix whose columns are given by the vectorized coefficients $(c_{\boldsymbol{\alpha},i})_{|\boldsymbol{\alpha}|\leq d}$ of the polynomials $f_i(x)$, $i = 1, ..., q$ in P. Then the columns of $\boldsymbol{X}$ belong to the variety $V(P)$ if and only if $\phi_d(\boldsymbol{X})^T\boldsymbol{C} = \boldsymbol{0}$. In particular, assuming the columns of $\boldsymbol{C}$ are linearly independent, this shows that $\phi_d(\boldsymbol{X})$ has rank $\leq \min(N - q, s)$. In particular, when the number of data points $s > N - q$, then $\phi_d(\boldsymbol{X})$ is rank deficient.

However, the exact rank of $\phi_d(\boldsymbol{X})$ could be much smaller than $\min(N - q, s)$, especially when the degree d is large. This is because the coefficients $\boldsymbol{c}$ of *any polynomial* that vanishes at every column of $\boldsymbol{X}$ satisfies $\phi_d(\boldsymbol{X})^T\boldsymbol{c} = \boldsymbol{0}$. We will find it useful to identify this space of coefficients with a finite dimensional vector space of polynomials. Let $\mathbb{R}_d[\boldsymbol{x}]$ be the space of all polynomials in n real variables of degree at most d. We define the *vanishing ideal of degree* d corresponding to a set $\mathcal{X} \subset \mathbb{R}^n$, denoted by $\mathcal{I}_d(\mathcal{X})$, to be subspace of polynomials belonging to $\mathbb{R}_d[\boldsymbol{x}]$ that vanish at all points in $\mathcal{X}$:

$$\mathcal{I}_d(\mathcal{X}) := \{f \in \mathbb{R}_d[\boldsymbol{x}] : f(\boldsymbol{x}) = 0 \text{ for all } \boldsymbol{x} \in \mathcal{X}\}.$$

We also define the *non-vanishing ideal of degree* d corresponding to X, denoted by $\mathcal{S}_d(\mathcal{X})$, to be the orthogonal complement of $\mathcal{I}_d(\mathcal{X})$ in $\mathbb{R}_d[\boldsymbol{x}]$:

$$\mathcal{S}_d(\mathcal{X}) := \{g \in \mathbb{R}_d[\boldsymbol{x}] : \langle f, g\rangle = 0 \text{ for all } f \in \mathcal{I}_d(\mathcal{X})\},$$

where the inner product $\langle f, g\rangle$ of polynomials $f, g \in \mathbb{R}_d[\boldsymbol{x}]$ is defined as the inner product of their coefficient vectors. Hence, the rank R of $\phi_d(\boldsymbol{X})$ can expressed in terms of the dimension of non-vanishing ideal of degree d corresponding to $\mathcal{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_s\}$, the set of all columns of $\boldsymbol{X}$. Specifically, we have rank $\phi_d(\boldsymbol{X}) = \min(R, s)$ where

$$R = \dim \mathcal{S}_d(\mathcal{X}) = N - \dim \mathcal{I}_d(\mathcal{X})\ .$$

In general the dimension of the space $\mathcal{I}_d(\mathcal{X})$ or $\mathcal{S}_d(\mathcal{X})$ is difficult to determine when $\mathcal{X}$ is an arbitrary set of points. However, if we assume $\mathcal{X}$ is a subset of a variety V, then $\mathcal{I}_d(V) \subseteq \mathcal{I}_d(\mathcal{X})$ and hence

$$\text{rank } \phi_d(\boldsymbol{X}) \leq \dim \mathcal{S}_d(V).$$

In certain cases dim $\mathcal{S}_d(V)$ can be computed exactly or bounded using properties of the polynomials defining V. For example, it is possible to compute the dimension of $\mathcal{S}_d(V)$ directly from a *Gröbner basis* for the vanishing ideal associated with V (Cox et al., 2015). In Section 3 we show how to bound the dimension of $\mathcal{S}_d(V)$ in the case where V is a union of subspaces.

2.3. Sampling rate

Informally, the *degrees of freedom* of a class of objects is the minimum number of free variables needed to describe an element in that class uniquely. For example, a $n \times s$ rank r matrix has $r(n + s - r)$ degrees of freedom: nr parameters to describe r linearly independent columns making up a basis of the column space, and $r(s - r)$ parameters to describe the remaining $s - r$ columns in terms of this basis. It is impossible to uniquely complete a matrix in this class if we sample fewer than this many entries.

We can make a similar argument to specify the minimum number of samples needed to uniquely complete a matrix that is low-rank when mapped to feature space. First, we characterize how missing entries of the data matrix translate to missing entries in feature space. For simplicity, we will assume a sampling model where we sample a fixed number of entries m from each column of the original data matrix. Let $\boldsymbol{x} \in \mathbb{R}^n$ represent a single column of the data matrix, and $\Omega \subset \{1, ..., n\}$ with $m = |\Omega|$ denote the indices of the sampled entries of $\boldsymbol{x}$. The pattern of revealed entries in $\phi_d(\boldsymbol{x})$ corresponds to the set of multi-indices:

$$\{\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n) : |\boldsymbol{\alpha}| \leq d,\ \alpha_i = 0 \text{ for all } i \in \Omega^c\},$$

which has the same cardinality as the set of all monomials of degree at most d in m variables, *i.e.,* $\binom{m+d}{d}$. If we call this quantity M, then the ratio of revealed entries in $\phi_d(x)$ to the feature space dimension is

$$\frac{M}{N} = \frac{\binom{m+d}{d}}{\binom{n+d}{d}} = \frac{(m+d)(m+d-1)\cdots(m+1)}{(n+d)(n+d-1)\cdots(n+1)},$$

which is on the order of $(\frac{m}{n})^d$ for small d. More precisely, we have the bounds

$$\left(\frac{m}{n}\right)^d \leq \frac{M}{N} \leq \left(\frac{m+d}{n}\right)^d. \tag{3}$$

In total, observing m entries per column of the data matrix translates to M entries per column in feature space. Suppose the $N \times s$ lifted matrix $\phi_d(\boldsymbol{X})$ is rank R. By the preceding discussion, we need least $R(N + s - R)$ entries of the feature space matrix $\phi_d(\boldsymbol{X})$ to complete it uniquely among the class of all $N \times s$ matrices of rank R. Hence, at minimum we need to satisfy

$$Ms \geq R(N + s - R). \tag{4}$$

Let m_0 denote the minimal value of m such that $M = \binom{m+d}{d}$ achieves the bound (4), and set $M_0 = \binom{m_0+d}{d}$. Dividing (4) through by the feature space dimension N and s gives

$$\frac{M_0}{N} \geq \left(\frac{R}{N}\right)\left(\frac{N+s-R}{s}\right) = \left(\frac{R}{s} + \frac{R}{N}\left(1 - \frac{R}{s}\right)\right), \tag{5}$$

and so from (3) we see we can guarantee this bound with

$$\rho_0 := \frac{m_0}{n} \geq \left(\frac{R}{s} + \frac{R}{N}\left(1 - \frac{R}{s}\right)\right)^{\frac{1}{d}}, \tag{6}$$

and this in fact will result in tight satisfaction of (5) because $(M_0/N)^{\frac{1}{d}} \approx m_0/n$ for small d and large n.

At one extreme where the matrix $\phi_d(\boldsymbol{X})$ is full rank, then $R/s = 1$ or $R/N = 1$ and according to (6) we need $\rho_0 \approx 1$, *i.e.,* full sampling of every data column. At the other extreme where instead we have many more data points than the feature space rank, $R/s \ll 1$, then (6) gives the asymptotic bound $\rho_0 \approx (R/N)^{\frac{1}{d}}$.

The above discussion bounds the degrees of freedom of a matrix that is rank-R in feature space. Of course, the proposed variety model has potentially fewer degrees of freedom than this, because additionally the columns of the lifted matrix are constrained to lie in the image of the feature map. We use the above bound only as a rule of thumb for sampling requirements on our matrix. Furthermore, we note that sample complexities for standard matrix completion often require that locations are observed uniformly at random, whereas in our problem the locations of observations in the lifted space will necessarily be structured. However, there is recent work that shows matrix completion can suceed without these assumptions (Pimentel-Alarcón et al., 2016b; Chen et al., 2014) that gives reason to believe random samples in the original space may allow completion in the lifted space, and our empirical results in Section 5 support this rationale.

3. Case Study: Union of Affine Subspaces

A union of affine subspaces can be modeled as an algebraic variety. For example, with $(x, y, z) \in \mathbb{R}^3$, the union of the plane $z = 1$ and the line $x = y$ is the zero-set of the *quadratic* polynomial $(z-1)(x-y)$. In general, if $\mathcal{A}_1, \mathcal{A}_2 \subset \mathbb{R}^n$ are affine spaces of dimension r_1 and r_2, respectively, then we can write $\mathcal{A}_1 = \{\boldsymbol{x} : f_i(\boldsymbol{x}) = 0 \text{ for } i = 1, ..., n-r_1\}$ and $\mathcal{A}_2 = \{\boldsymbol{x} : g_i(\boldsymbol{x}) = 0 \text{ for } i = 1, ..., n-r_2\}$ where the f_i and g_i are affine functions. The union $\mathcal{A} \cup \mathcal{B}$ can be expressed as the common zero set of all possible products of the f_i and g_i, *i.e.,* $\mathcal{A}_1 \cup \mathcal{A}_2$ is the common zero set of a system of $(n-r_1)(n-r_2)$ quadratic equations. Similarly, a union of k affine subspaces of dimensions $r_1, ..., r_k$ is a variety described by a system of $\prod_{i=1}^{k}(n-r_i)$ polynomial equations of degree k.

In this section we establish bounds on the feature space rank for data belonging to a union of affine subspaces. We will make use of the following lemma that shows the dimension of a vanishing ideal is fixed under an affine change of variables:

Lemma 1. *Let $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ be an affine change of variables,* i.e., *$\mathcal{T}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$, where $\boldsymbol{b} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is invertible. Then for any $S \subset \mathbb{R}^n$,*

$$\dim \mathcal{I}_d(S) = \dim \mathcal{I}_d(\mathcal{T}(S)). \tag{7}$$

We omit the proof for brevity, but the result is elementary and relies on the fact the degree of a polynomial is unchanged under an affine change of variables. Our next result establishes a bound on the feature space rank for a single affine subspace:

Proposition 1. *If the columns of a matrix $\boldsymbol{X}^{n \times s}$ belong to an affine subspace of dimension at most r, then*

$$\textit{rank}\, \phi_d(\boldsymbol{X}) \leq \binom{r+d}{d}, \ \textit{for all} \ \ d \geq 1. \tag{8}$$

Proof. By Lemma 1, $\dim \mathcal{I}_d(\mathcal{A})$ is preserved under an affine transformation of $\mathcal{A}$. Note that we can always find an affine change of variables $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{c}$ with invertible $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{c} \in \mathbb{R}^n$ such that in the coordinates $\boldsymbol{y} = (y_1, ..., y_n)$ the variety $\mathcal{A}$ becomes

$$\mathcal{A} = \{(y_1, \ldots, y_r, 0, \ldots, 0) : y_1, ..., y_r \in \mathbb{R}\}. \tag{9}$$

For any polynomial $f(\boldsymbol{y}) = \sum_{|\boldsymbol{\alpha}| \leq d} c_{\boldsymbol{\alpha}} \boldsymbol{y}^{\boldsymbol{\alpha}}$, the only monomial terms in $f(\boldsymbol{y})$ that do not vanish on $\mathcal{A}$ are those of the form $y_1^{\alpha_1} \cdots y_r^{\alpha_r}$. Furthermore, any polynomial in just these monomials that vanishes on all of $\mathcal{A}$ must be the zero polynomial, since the $y_1, ..., y_r$ are free variables. Hence,

$$\mathcal{S}_d(\mathcal{A}) = \text{span}\{y_1^{\alpha_1} \cdots y_r^{\alpha_r} : \alpha_1 + \cdots + \alpha_r \leq d\} \tag{10}$$

i.e., the non-vanishing ideal coincides with the space of polynomials in r variables of degree at most d, which has dimension $\binom{r+d}{d}$, proving the claim. □

We note that for s sufficiently large, the bound in (8) becomes an equality, provided the data points are distributed generically within the affine subspace, meaning they are not the solution to additional non-trivial polynomial equations of degree at most d.

Proposition 1 shows that points belonging to a single affine subspace of dimension r are mapped to a linear subspace of dimension $\binom{r+d}{d}$ under ϕ_d. Therefore, if the columns of a data matrix are drawn from a union of k affine subspaces of dimension r, their image under ϕ_d will belong to a union of k linear subspaces each of dimension at most $\binom{r+d}{d}$. The linear span of this union has dimension at most $k\binom{r+d}{d}$, which yields the following result:

Proposition 2. *If the columns of a matrix $\boldsymbol{X}^{n \times s}$ belong to a union of k affine subspaces each of dimension at most r, then*

$$\textit{rank}\, \phi_d(\boldsymbol{X}) \leq k\binom{r+d}{d}, \ \textit{for all} \ d \geq 1. \tag{11}$$

In some cases the bound (11) is (nearly) tight. For example, if the data lies on the union of two r-dimensional affine subspaces $\mathcal{A}$ and $\mathcal{B}$ that are mutually orthogonal, one can show[1] rank $\phi_d(\boldsymbol{X}) = 2\binom{r+d}{d} - 1$. Empirically, we observe that the bound in (11) is order-optimal with respect to k, r, and d. In this case, the feature space rank to dimension ratio is $R/N = O(k\left(\frac{r}{n}\right)^d)$. Recall that the minimum sampling rate is approximately $(R/N)^{\frac{1}{d}}$ for $s \gg R$. Hence the mininum number of samples per column m should be

$$m \approx O(k^{\frac{1}{d}} r). \tag{12}$$

This rate is favorable to low-rank matrix completion approaches, which need $m = O(kr)$ for a union of k subspaces having dimension r. At first glance, this bound suggests it is always better to take the degree d as large as possible. However, this is only true for sufficiently large s.

[1] The rank is one less than the bound in (11) because $\mathcal{S}_d(\mathcal{A}) \cap \mathcal{S}_d(\mathcal{B})$ has dimension one, coinciding with the space of constant polynomials.

To take advantage of the improved sampling rate implied by (12), according to (6) we need the number of data vectors per subspace to be $O(r^d)$. In other words, our model is able to accommodate more subspaces with larger d but at the expense of requiring exponentially more data points per subspace. Note that if the number of data points is sufficiently large, we could take $d = \log k$ and require only $m \approx O(r)$ observed entries per column. In this case, for moderately sized k (*e.g.,* $k \leq 20$) we should choose $d = 2$ or 3. In fact, we find that for these values of d we get excellent empirical results for the recovery of union of subspaces data, as shown in Section 5.

4. Algorithm

There are several existing matrix completion algorithms that could potentially be adapted to solve a relaxation of the rank minimization problem (1), such as singular value thresholding (Cai et al., 2010), or alternating minimization (Jain et al., 2013). However, these approaches do not easily lend themselves to "kernelized" implementations, *i.e.,* ones that do not require forming the high-dimensional lifted matrix $\phi_d(X)$ explicitly, but instead make use of the efficiently computable *kernel function* for polynomial feature maps [2]

$$k_d(\boldsymbol{x}, \boldsymbol{y}) := \phi_d(\boldsymbol{x})^T \phi_d(\boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y} + 1)^d. \tag{13}$$

For matrices $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_s], \boldsymbol{Y} = [\boldsymbol{y}_1, ..., \boldsymbol{y}_s] \in \mathbb{R}^{n \times s}$, we use $k_d(\boldsymbol{X}, \boldsymbol{Y})$ to denote the matrix whose (i,j)-th entry is $k_d(\boldsymbol{x}_i, \boldsymbol{y}_j)$, or equivalently,

$$k_d(\boldsymbol{X}, \boldsymbol{Y}) := \phi_d(\boldsymbol{X})^T \phi_d(\boldsymbol{Y}) = (\boldsymbol{X}^T \boldsymbol{Y} + \mathbf{1})^{\odot d}, \tag{14}$$

where $\mathbf{1} \in \mathbb{R}^{s \times s}$ is the matrix of all ones, and $(\cdot)^{\odot d}$ denotes the entrywise d-th power of a matrix. A kernelized implentation is critical for even modest sizes of d, since the number of rows of the lifted matrix scales exponentially with d.

One class of algorithm that kernelizes very naturally is the iterative reweighted least squares (IRLS) approach of (Fornasier et al., 2011; Mohan & Fazel, 2012) for low-rank matrix completion. The algorithm also has the advantage of being able to accommodate the non-convex Schatten-p relaxation of the rank penalty, in addition to the convex nuclear norm relaxation. Specifically, we use an IRLS approach to solve the following variety-based matrix completion (VMC) optimization problem:

$$\min_{\boldsymbol{X}} \|\phi_d(\boldsymbol{X})\|_{\mathcal{S}_p}^p \;\text{ s.t. }\; \mathcal{P}_\Omega(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{X}_0), \tag{VMC}$$

where $\|\boldsymbol{Y}\|_{\mathcal{S}_p}$ is the Schatten-p quasi-norm defined as

$$\|\boldsymbol{Y}\|_{\mathcal{S}_p} := \left(\textstyle\sum_i \sigma_i(\boldsymbol{Y})^p\right)^{\frac{1}{p}}, \;\; 0 < p \leq 1 \tag{15}$$

with $\sigma_i(\boldsymbol{Y})$ denoting the i^{th} singular value of $\boldsymbol{Y}$. Algorithm 1 gives the pseudo-code of the proposed IRLS algorithm for solving (VMC), which we derive below.

Algorithm 1 Kernelized IRLS to solve (VMC).

```
Require: Initialize $X = X_0$, $\gamma = \gamma_0$. Choose $\eta$,$\gamma_{\text{min}}$.
  while not converged do
    Step 1: Inverse power of kernel matrix
    $K \leftarrow k_d(X, X)$
    $(V, S) = \texttt{eig}(K)$.
    $W \leftarrow V(S + \gamma I)^{\frac{p}{2}-1} V^T$

    Step 2: Projected gradient descent step
    $\tau \leftarrow \gamma^{1-\frac{p}{2}}$
    $X \leftarrow X - \tau X(W \odot k_{d-1}(X, X))$
    $X \leftarrow \mathcal{P}_\Omega(X_0) + \mathcal{P}_{\Omega^c}(X)$
    $\gamma \leftarrow \max\{\gamma/\eta, \gamma_{\text{min}}\}$
  end while
```

First, consider the simpler problem of minimizing the Schatten-p norm of a matrix variable $\boldsymbol{Y}$ belonging to a constraint set $\mathcal{C}$. The main idea behind the IRLS approach is re-express the Schatten-p quasi-norm as

$$\|\boldsymbol{Y}\|_{\mathcal{S}_p}^p = \text{tr}[(\boldsymbol{Y}^T\boldsymbol{Y})^{\frac{p}{2}}] = \text{tr}[(\boldsymbol{Y}^T\boldsymbol{Y})\boldsymbol{W}], \tag{16}$$

where $\boldsymbol{W} := (\boldsymbol{Y}^T\boldsymbol{Y})^{\frac{p}{2}-1}$. Note if $\boldsymbol{W}$ is treated as constant, then (16) is a smooth, quadratic function of $\boldsymbol{Y}$. This motivates the following iterative approach:

$$\begin{aligned} \boldsymbol{W}_n &= (\boldsymbol{Y}_n^T \boldsymbol{Y}_n + \gamma_n)^{\frac{p}{2}-1} \\ \boldsymbol{Y}_{n+1} &= \underset{\boldsymbol{Y} \in \mathcal{C}}{\arg\min} \;\text{tr}[(\boldsymbol{Y}^T\boldsymbol{Y})\boldsymbol{W}_n]. \end{aligned}$$

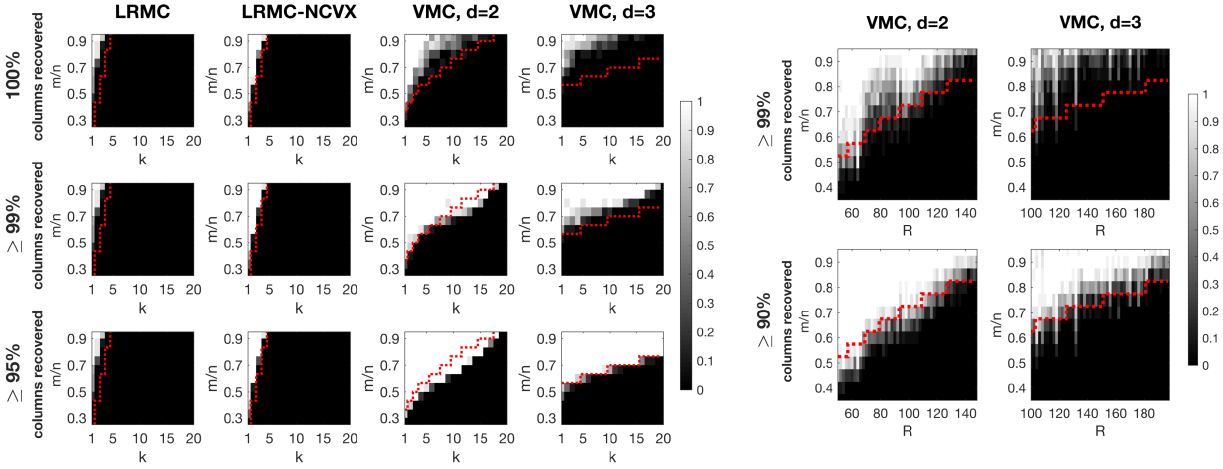
Here γ_n is a sequence of smoothing parameters satisfying $\gamma_n \to \gamma_{\text{min}}$ as $n \to \infty$, where γ_{min} is close to zero, which is included to improve numerical stability and avoid local minima; this is equivalent to minimizing a smooth approximation of the Schatten-p cost (Mohan & Fazel, 2012).

Making the substitution $\boldsymbol{Y} = \phi_d(\boldsymbol{X})$ in the above derivation, gives the following approach for solving (VMC):

$$\begin{aligned} \boldsymbol{W}_n &= (k(\boldsymbol{X}_n, \boldsymbol{X}_n) + \gamma_n I)^{\frac{p}{2}-1} \\ \boldsymbol{X}_{n+1} &= \underset{\boldsymbol{X}}{\arg\min} \;\text{tr}[k(\boldsymbol{X},\boldsymbol{X})\boldsymbol{W}_n] \text{ s.t. } \mathcal{P}_\Omega(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{X}_0) \end{aligned}$$

Rather than finding the exact minimum in the $\boldsymbol{X}$ update, which could be costly, following the approach in (Mohan & Fazel, 2012), we instead take a single projected gradient descent step to update $\boldsymbol{X}$. A straightforward calculation shows that the gradient of

[2] Strictly speaking, k_d is not kernel associated with the polynomial feature map ϕ_d as defined in (2.1). Instead, it is the kernel of the related map $\tilde{\phi}_d(\boldsymbol{x}) := \{\sqrt{c_{\boldsymbol{\alpha}}}\boldsymbol{x}^{\boldsymbol{\alpha}} : |\boldsymbol{\alpha}| \leq d\}$ where $c_{\boldsymbol{\alpha}}$ are appropriately chosen multinomial coefficients.



(a) Union of Subspaces

(b) Parametric Data

Figure 2. Phase transitions for matrix completion of synthetic variety data. In (a) we simulate data belonging to a union of k subspaces for varying k. In (b) we simulate data belonging union of few parametric curves and surfaces having known feature space rank R. We randomly undersample each column of the data matrix at the rate m/n. The grayscale values 0–1 indicate the fraction of random trials where the columns of the data matrix were successfully recovered up to the specified percentage (white is success, black is failure). In all figures the red dashed line indicates the predicted minimal sampling rate $\rho_0 = m_0/n$ determined by (4).

the objective $F(\boldsymbol{X}) = \text{tr}[k(\boldsymbol{X},\boldsymbol{X})\boldsymbol{W}]$ is given by $\nabla F(\boldsymbol{X}) = \boldsymbol{X}(\boldsymbol{W} \odot k_{d-1}(\boldsymbol{X},\boldsymbol{X}))$, where $\odot$ denotes an entry-wise product. Hence a projected gradient step with step-size $\tau_n > 0$ is given by

$$\tilde{\boldsymbol{X}}_n = \boldsymbol{X}_n - \tau_n \boldsymbol{X}_n(\boldsymbol{W}_n \odot k_{d-1}(\boldsymbol{X}_n,\boldsymbol{X}_n))$$
$$\boldsymbol{X}_n = \mathcal{P}_\Omega(\boldsymbol{X}_0) + \mathcal{P}_{\Omega^c}(\tilde{\boldsymbol{X}}_n).$$

Similar to (Mohan & Fazel, 2012), one can show that every limit point of the above iterates converges to a stationary point of a smoothed Schatten-p cost for appropriate choices of step-sizes τ_n. Heuristics are given in (Mohan & Fazel, 2012) for updating the smoothing parameter γ_n, which we adopt as well. Specifically, we set $\gamma_n = \gamma_0/\eta^n$, where γ_0 and η are user-defined parameters, and update $\tau_n = \gamma_n^{1-p/2}$. The appropriate choice of γ_0 and η will depend on the scaling and spectral properties of the data. Empirically, we find that setting $\gamma_0 = (0.1)^d\lambda_{max}$, where λ_{max} is the largest eigenvalue of the kernel matrix obtained from the initialization, and $\eta = 1.01$ work well in a variety of settings. For all our experiments in Section 5 we fix $p = 1/2$, which was found to give the best matrix recovery results for synthetic data. We also use a zero-filled initialization $\boldsymbol{X}_0$ in all cases.

5. Numerical Experiments

5.1. Empirical validation of sampling bounds

In Figure 2 we report the results of two experiments to validate the predicted minimum sampling rate ρ_0 in (4) on synthetic variety data. In the first experiment we generated $n \times s$ data matrices whose columns belong to a union of k subspaces each of dimension r (with $n = 15$, $s = 100k$, $r = 3$). In the second experiment we generated data matrices of size 20×300 whose columns belong to a union of randomly generated parametric surfaces of low dimension, where we sorted each dataset by its empirically determined feature space rank R. For both experiments, we undersampled each column of the matrix taking m entries uniformly at random at various values of k and R, and then attempted to recover the missing entries using our proposed IRLS algorithm for VMC (Algorithm 1 with $p = 1/2$) for $d = 2, 3$. For the union of subspaces data, we also compare with low-rank matrix completion in the original matrix domain via nuclear norm minimization (LRMC) and non-convex Schatten-$1/2$ minimization (LRMC-NCVX), implemented using Algorithm 1 with a linear kernel ($d = 1$ in (13)). We said a column was successfully recovered if $\|\boldsymbol{x} - \boldsymbol{x}_0\|/\|\boldsymbol{x}_0\| \leq 10^{-5}$, where $\boldsymbol{x}$ is the recovered column and $\boldsymbol{x}_0$ is the original column. For each pair of parameters (m, k) or (m, R) we perform 10 random trials to determine the probability of successful recovery.

Consistent with our theory, VMC is successful at recovering most of the data columns above the predicted minimum sampling rate, substantially extending the range of recovery over LRMC. While VMC often fails to recover 100% of the columns near the predicted rate, in fact a large proportion of the columns (%99–%90) are still successfully completed. Sometimes the recovery dips below the predicted rate (*e.g.,* VMC, $d = 2$ in Fig. 2(a) and VMC, $d = 3$ in Fig. 2(b)). However, since the predicted rate relies on what is likely an over-estimate of the true degrees of freedom, it is not surprising that the VMC algorithm occasionally succeeds below this rate, too.

5.2. Motion segmentation of real data

In Figure 3 we apply VMC to the problem of motion segmentation (Kanatani, 2001) with missing data using the Hopkins 155 dataset (Tron & Vidal, 2007). This data consists of several feature points tracked across frames of the video. We reproduce the experimental setting in (Yang et al., 2015), and simulate high-rank data by undersampling frames of the dataset. We simulate missing trajectories by sampling uniformly at random from the feature points across all frames. To obtain a clustering we first completed the missing entries using VMC and then ran the sparse subspace clustering (SSC) algorithm (Elhamifar & Vidal, 2009) on the result, calling this VMC+SSC. A similar approach of standard LRMC followed by SSC (LRMC+SSC) provides a consistent baseline for subspace clustering with missing data (Yang et al., 2015; Elhamifar, 2016). We also compare against SSC with entrywise zerofill (SSC-EWZF) (Yang et al., 2015). We find the VMC+SSC approach gives similar or lower clustering error than LRMC+SCC for low missing rates. Likewise, VMC+SSC also substantially outperforms SSC-EWZF for high missing rates. Unlike SSC-EWZF and the other algorithms in (Yang et al., 2015), VMC+SSC also succeeds in setting where the data is low-rank (i.e., when all frames are retained). This is because the performance of VMC is similar to standard LRMC in the low-rank setting.

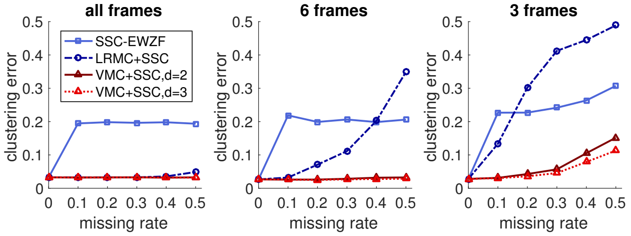


Figure 3. Subspace clustering error on Hopkins 155 dataset for varying rates of missing data and undersampling of frames.

5.3. Completion of motion capture data

In Figure 4 we demonstrate VMC for completing time-series trajectories from motion capture sensors using a dataset from the CMU Mocap database[3] (subject 56, trial 6). Empirically, this dataset has been shown to be locally low-rank over the time frames corresponding to each separate activity, and can be modeled as a union of subspaces (Elhamifar, 2016). The data had measurements from $n = 62$ sensors at $s = 6784$ time instants. We randomly undersampled the columns of this matrix and attempt to complete the data using VMC, LRMC, and LRMC-NCVX and measure the resulting *completion error*: $\|\boldsymbol{X} - \boldsymbol{X}_0\|_F / \|\boldsymbol{X}_0\|_F$, where $\boldsymbol{X}$ is the recovered matrix and $\boldsymbol{X}_0$ is the original matrix. Similar to results on synthetic data, we find the VMC approach outperforms LRMC-NCVX for appropriately chosen degree d. In particular, VMC with $d = 2, 3$ perform similar for small missing rates, but VMC $d = 2$ gives lower completion error over $d = 3$ for large missing rates, consistent with the results in Figure 2.

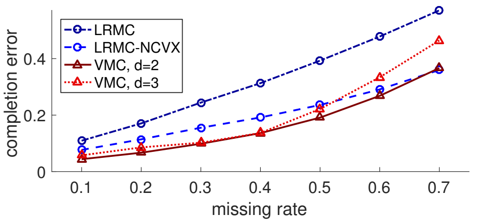


Figure 4. Completion error on CMU Mocap dataset using the proposed VMC approach compared with convex and non-convex LRMC algorithms.

6. Conclusion

We introduce a matrix completion approach that generalizes low-rank matrix completion to a much wider class of variety models, including data belonging to a union of subspaces. We present a hypothesized sampling complexity bound for the completion of a matrix whose columns belong to an algebraic variety. A surprising result of our analysis that that a union of k affine subspaces of dimension r should be recoverable from $O(rk^{1/d})$ measurements per column, provided we have $O(r^d)$ data points (columns) per subspace, where d is the degree of the feature space map. In particular, if we choose $d = \log k$, then we need only $O(r)$ measurements per column as long as we have $O(r^{\log k})$ columns per subspace. We additionally introduce an efficient algorithm based on an iterative reweighted least squares approach that realizes these hypothesized bounds on synthetic data, and reaches state-of-the-art performance on for matrix completion on several real high-rank datasets.

Our algorithm can easily accommodate other smooth kernels, including the popular Gaussian RBF kernel (Muller et al., 2001). A similar optimization formulation to ours was presented in the recent pre-print (Garg et al., 2016) using Gaussian RBF kernels in place of polynomial kernels, showing good empirical results in a matrix completion context. However, analysis of the sample complexity in this case is complicated by the fact that a feature space representation for Gaussian RBF kernel is necessarily infinite-dimensional. Understanding the sample requirements in this case would be an interesting avenue for future work.

Acknowledgements

For this work, Balzano and Ongie were supported in part by ARO grant W911NF-14-1-0634. Willett and Nowak were supported in part by NSF IIS-1447449, NSF CCF-0353079, and NIH 1 U54 AI117924-01, and Nowak also by AFOSR FA9550-13-1-0138.

[3] http://mocap.cs.cmu.edu

References

Cai, Jian-Feng, Candès, Emmanuel J, and Shen, Zuowei. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Candes, Emmanuel and Recht, Benjamin. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

Chen, Guangliang, Atev, Stefan, and Lerman, Gilad. Kernel spectral curvature clustering (kscc). In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pp. 765–772. IEEE, 2009.

Chen, Yudong, Bhojanapalli, Srinadh, Sanghavi, Sujay, and Ward, Rachel. Coherent matrix completion. In *Proceedings of The 31st International Conference on Machine Learning*, pp. 674–682, 2014.

Cox, David A., Little, John, and O'Shea, Donal. *Ideals, Varieties, and Algorithms*. Springer International Publishing, 2015.

Derksen, Harm. Hilbert series of subspace arrangements. *Journal of pure and applied algebra*, 209(1):91–98, 2007.

Elhamifar, Ehsan. High-rank matrix completion and clustering under self-expressive models. In *Advances in Neural Information Processing Systems*, pp. 73–81, 2016.

Elhamifar, Ehsan and Vidal, René. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2790–2797. IEEE, 2009.

Elhamifar, Ehsan and Vidal, René. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35 (11):2765–2781, 2013.

Eriksson, Brian, Balzano, Laura, and Nowak, Robert D. High-rank matrix completion. In *AISTATS*, pp. 373–381, 2012.

Fornasier, Massimo, Rauhut, Holger, and Ward, Rachel. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM Journal on Optimization*, 21(4):1614–1640, oct 2011. doi: 10.1137/100811404.

Ganti, Ravi Sastry, Balzano, Laura, and Willett, Rebecca. Matrix completion under monotonic single index models. In *Advances in Neural Information Processing Systems*, pp. 1873–1881, 2015.

Garg, Ravi, Eriksson, Anders, and Reid, Ian. Non-linear dimensionality regularizer for solving inverse problems. *arXiv preprint arXiv:1603.05015*, 2016.

Jain, Prateek, Netrapalli, Praneeth, and Sanghavi, Sujay. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674. ACM, 2013.

Kanatani, Ken-ichi. Motion segmentation by subspace separation and model selection. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pp. 586–591. IEEE, 2001.

Lee, Joonseok, Kim, Seungyeon, Lebanon, Guy, and Singer, Yoram. Local low-rank matrix approximation. *ICML (2)*, 28:82–90, 2013.

Mohan, Karthik and Fazel, Maryam. Iterative reweighted algorithms for matrix rank minimization. *The Journal of Machine Learning Research*, 13(1):3441–3473, 2012.

Muller, K-R, Mika, Sebastian, Ratsch, Gunnar, Tsuda, Koji, and Scholkopf, Bernhard. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.

Nguyen, Minh H and Torre, Fernando. Robust kernel principal component analysis. In *Advances in Neural Information Processing Systems*, pp. 1185–1192, 2009.

Pimentel-Alarcón, D, Balzano, L, Marcia, R, Nowak, R, and Willett, R. Group-sparse subspace clustering with missing data. In *Statistical Signal Processing Workshop (SSP), 2016 IEEE*, pp. 1–5. IEEE, 2016a.

Pimentel-Alarcon, Daniel and Nowak, Robert. The information-theoretic requirements of subspace clustering with missing data. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 802–810, 2016.

Pimentel-Alarcón, Daniel L, Boston, Nigel, and Nowak, Robert D. A characterization of deterministic sampling patterns for low-rank matrix completion. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):623–636, 2016b.

Rao, Nikhil, Ganti, Ravi, Balzano, Laura, Willett, Rebecca, and Nowak, Robert. On learning high dimensional structured single index models. In *Proceedings of the 31st AAAI conference on artificial intelligence*, 2017.

Sanguinetti, Guido and Lawrence, Neil D. Missing data in kernel PCA. In *European Conference on Machine Learning*, pp. 751–758. Springer, 2006.

Song, Dogyoon, Lee, Christina E, Li, Yihua, and Shah, Devavrat. Blind regression: Nonparametric regression for latent variable models via collaborative filtering. In *Advances in Neural Information Processing Systems*, pp. 2155–2163, 2016.

Tron, Roberto and Vidal, René. A benchmark for the comparison of 3-d motion segmentation algorithms. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8. IEEE, 2007.

Tsakiris, Manolis C and Vidal, René. Algebraic clustering of affine subspaces. *arXiv preprint arXiv:1509.06729*, 2015.

Vidal, René, Soatto, Stefano, Ma, Yi, and Sastry, Shankar. An algebraic geometric approach to the identification of a class of linear hybrid systems. In *Decision and Control, 2003. Proceedings. 42nd IEEE Conference on*, volume 1, pp. 167–172. IEEE, 2003.

Vidal, René, Ma, Yi, and Sastry, Shankar. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (12):1945–1959, 2005.

Vidal, René, Ma, Yi, and Sastry, Shankar. *Generalized Principal Component Analysis*. Springer New York, 2016.

Yang, Congyuan, Robinson, Daniel, and Vidal, René. Sparse subspace clustering with missing entries. In *Proceedings of The 32nd International Conference on Machine Learning*, pp. 2463–2472, 2015.