Seven Lemmas on Nonlinear Models for Matrix Completion You Won't Believe

(Number Six Will Blow Your Mind!)

Rebecca Willett, University of Wisconsin-Madison

SIAM Annual Meeting 2017

Nonlinearities in recommender systems

						
	10	9	8	6	4	1
	10	10	?	10	9	1

Low-rank matrix models predict Roummel's rating as a weighted sum of other users' ratings.

Nonlinear models can yield more accurate predictions of human preferences

General setup with missing data

- We have s points in $\mathbb{R}^n$:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_s \end{bmatrix} \in \mathbb{R}^{n \times s}$$

- We only observe m of the n entries in each $\boldsymbol{x}_i$; let $\boldsymbol{\Omega}$ indicate the locations of the observed entries and $\mathcal{P}_{\boldsymbol{\Omega}}(\cdot)$ be the projection onto this set.
- The incomplete version of $\boldsymbol{X}$ (with missing entries) is $\boldsymbol{X}_0$

General setup with missing data

- We have s points in $\mathbb{R}^n$:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_s \end{bmatrix} \in \mathbb{R}^{n \times s}$$

- We only observe m of the n entries in each $\boldsymbol{x}_i$; let $\boldsymbol{\Omega}$ indicate the locations of the observed entries and $\mathcal{P}_{\boldsymbol{\Omega}}(\cdot)$ be the projection onto this set.
- The incomplete version of $\boldsymbol{X}$ (with missing entries) is $\boldsymbol{X}_0$
- With low-rank matrix completion, we might set

$$\hat{\boldsymbol{X}} = \arg\min_{\boldsymbol{X}} \operatorname{rank}(\boldsymbol{X}) \text{ subject to } \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}_0)$$

General setup with missing data

- We have s points in $\mathbb{R}^n$:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \ldots & \boldsymbol{x}_s \end{bmatrix} \in \mathbb{R}^{n \times s}$$

- We only observe m of the n entries in each $\boldsymbol{x}_i$; let $\boldsymbol{\Omega}$ indicate the locations of the observed entries and $\mathcal{P}_{\boldsymbol{\Omega}}(\cdot)$ be the projection onto this set.
- The incomplete version of $\boldsymbol{X}$ (with missing entries) is $\boldsymbol{X}_0$
- With low-rank matrix completion, we might set

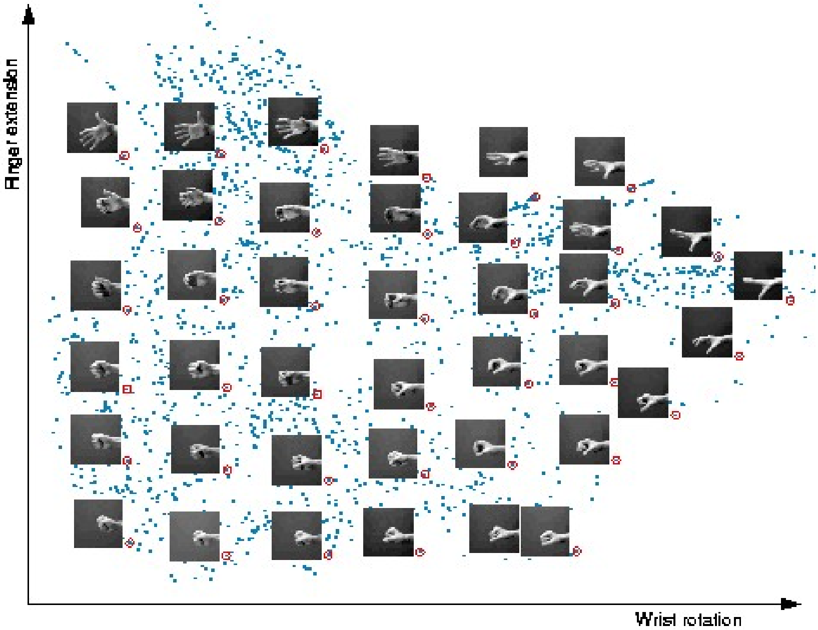
$$\hat{\boldsymbol{X}} = \arg\min_{\boldsymbol{X}} \operatorname{rank}(\boldsymbol{X}) \text{ subject to } \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}_0)$$

$$\hat{\boldsymbol{X}} = \arg\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_* \text{ subject to } \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}_0)$$

or

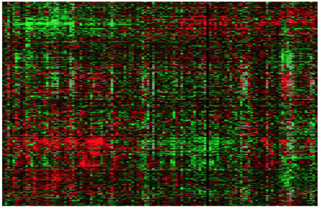
$$(\hat{\boldsymbol{U}}, \hat{\boldsymbol{V}}) = \arg\min_{\substack{\boldsymbol{U} \in \mathbb{R}^{n \times r} : \|\boldsymbol{U}\|_F \leq 1, \\ \boldsymbol{V} \in \mathbb{R}^{s \times r}}} \|\boldsymbol{X}_0 - \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{U}\boldsymbol{V}^\top)\|_F^2$$

Nonlinear representations of images

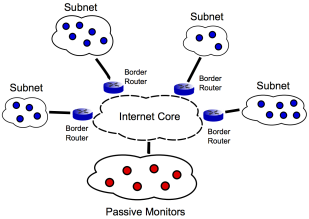


Nonlinearities abound

Computer Vision



Genomics



Network Topology Inference

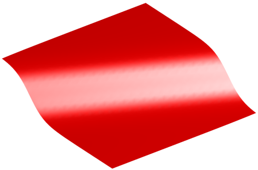
Can we extend the successes of low-rank matrix completion to non-linear structures?

We currently lack a unified, systematic framework for learning nonlinear models with missing data

How much missing data can be tolerated?

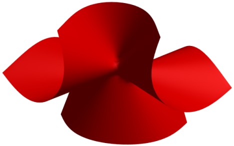
Efficient optimization algorithms?

Today: Three nonlinear models



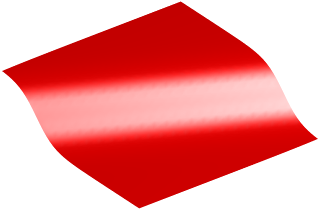
Single Index Models

Unions of Subspaces



Algebraic Varieties

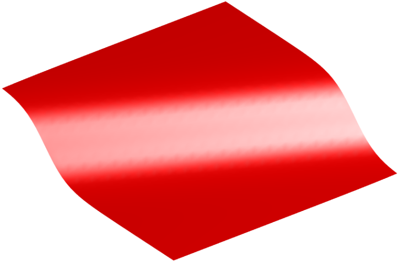
Matrix completion via single index models



Ravi Ganti

Laura Balzano

Single index models[1]



$$\boldsymbol{Z} \in \mathbb{R}^{n \times s} \quad \text{is a latent low-rank matrix}$$

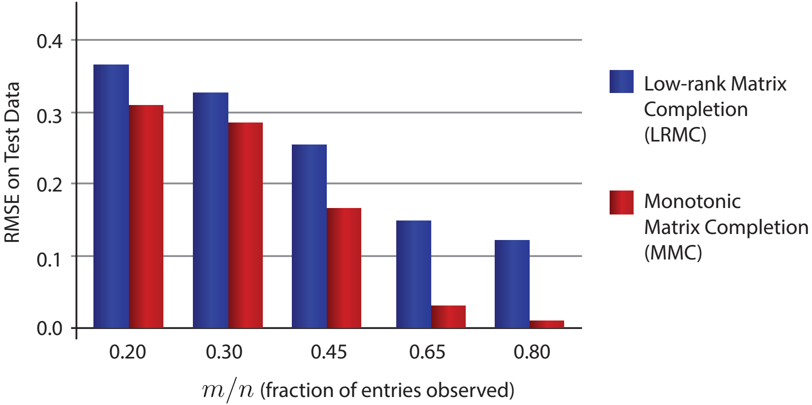
$$\boldsymbol{X} = g(\boldsymbol{Z}) \in \mathbb{R}^{n \times s} \quad \text{is a monotonic nonlinear transformation}$$

$$\boldsymbol{X}_{i,j} = g(\boldsymbol{Z}_{i,j}) \quad \text{of each element of } \boldsymbol{Z}$$

$$(\hat{g}, \hat{\boldsymbol{Z}}) = \underset{\substack{(g \text{ monotonic}, \\ \boldsymbol{Z} \text{ rank}-r)}}{\arg\min} \|\mathcal{P}_{\boldsymbol{\Omega}}(X_0 - g(\boldsymbol{Z}))\|_F^2$$

1 [Ichimura, 1993, Horowitz and Härdle, 1996, Kalai and Sastry, 2009, Kakade et al., 2011, Ganti et al., 2015]

Monotonic matrix completion in action (synthetic data)



$$n = 30, \ s = 20, \ r = 5, \ g(z) = (1 + e^{-z})^{-1}$$

Monotonic matrix completion in action (real data)

Dataset	Dimensions	Effective rank	Low-rank matrix completion	Monotonic matrix completion
PaperReco	3426×50	47	0.4026	0.2965
Jester-3	24938×100	66	6.8728	5.2348
ML-100k	1682×943	391	3.3101	1.1533
Cameraman	1536×512	393	0.0754	0.06885

RMSE of different methods on real datasets.
Roughly 10% of the entries were observed in each case.

Monotonic matrix completion theory[2]

Lemma 1: We can bound the MSE of the output of the MMC algorithm $(\hat{\boldsymbol{Z}}, \hat{g})$ as a function of

- how much data is missing,
- the data dimension,
- the number of samples, and
- the underlying subspace rank

as long as

$$\|\boldsymbol{X} - \boldsymbol{Z}\| \preceq \sqrt{n}$$

i.e., as long as the true g is not "too nonlinear".

2 [Ganti et al., 2015]

Monotonic matrix completion theory[2]

Lemma 1: We can bound the MSE of the output of the MMC algorithm $(\hat{\boldsymbol{Z}}, \hat{g})$ as a function of

- how much data is missing,
- the data dimension,
- the number of samples, and
- the underlying subspace rank

as long as

$$\|\boldsymbol{X} - \boldsymbol{Z}\| \preceq \sqrt{n}$$

i.e., as long as the true g is not "too nonlinear".

Challenge: need more flexibility than single index models provide

[2] [Ganti et al., 2015]

Matrix completion for unions of subspaces

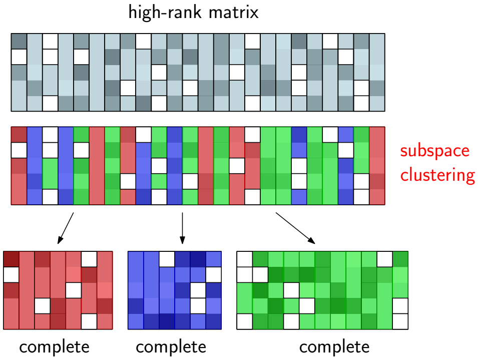
Daniel Pimentel

Roummel Marcia

Laura Balzano

Robert Nowak

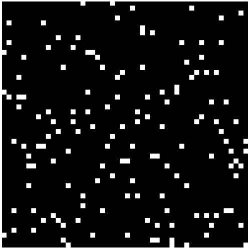
Unions of subspaces



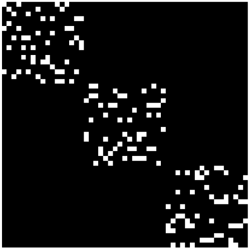
Clustering followed by low-rank matrix completion [3]

- Sparse subspace clustering (SSC):

$$c_i = \underset{c:\langle c, e_i \rangle = 0}{\arg\min} \ \|c\|_1 + \lambda \|\mathcal{P}_{\Omega_i}(x_i - X_{0,\setminus i} c)\|_2^2$$



c_i's



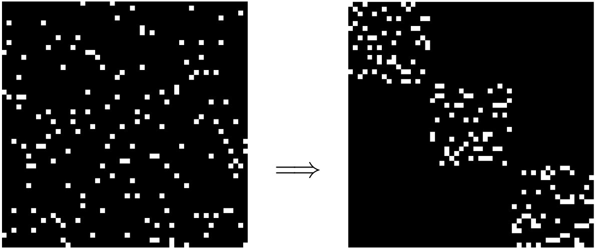
sorted c_i's

[3] [Elhamifar and Vidal, 2013, Yang et al., 2015]

Clustering followed by low-rank matrix completion [3]

- Sparse subspace clustering (SSC):

$$\boldsymbol{c}_i = \underset{\boldsymbol{c}:\langle \boldsymbol{c}, \boldsymbol{e}_i \rangle = 0}{\arg\min} \; \|\boldsymbol{c}\|_1 + \lambda \|\mathcal{P}_{\boldsymbol{\Omega}_i}(\boldsymbol{x}_i - \boldsymbol{X}_{0,\setminus i}\boldsymbol{c})\|_2^2$$



$\boldsymbol{c}_i$'s $\Longrightarrow$ sorted $\boldsymbol{c}_i$'s

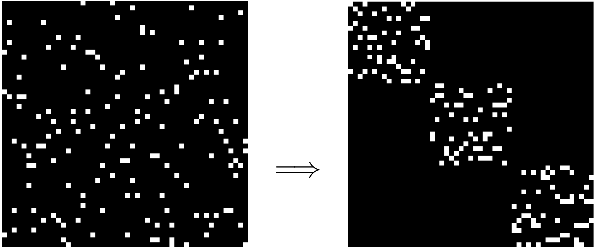
- spectral clustering on the $\boldsymbol{c}_i$'s
- low-rank matrix completion on each cluster

[3] [Elhamifar and Vidal, 2013, Yang et al., 2015]

Clustering followed by low-rank matrix completion [3]

- Sparse subspace clustering (SSC):

$$c_i = \underset{c:\langle c, e_i \rangle = 0}{\arg\min} \; \|c\|_1 + \lambda \|\mathcal{P}_{\Omega_i}(x_i - X_{0,\setminus i} c)\|_2^2$$



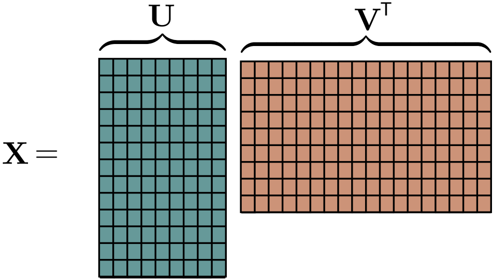
c_i's sorted c_i's

- spectral clustering on the c_i's
- low-rank matrix completion on each cluster

Does not allow improved clustering based on completed estimate

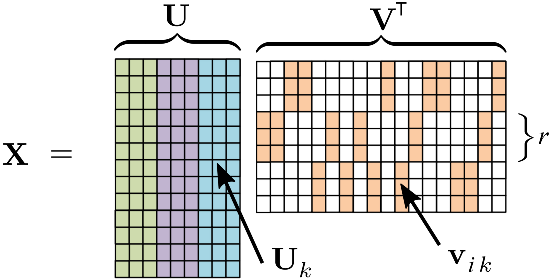
[3] [Elhamifar and Vidal, 2013, Yang et al., 2015]

Group sparse matrix factorization[4]



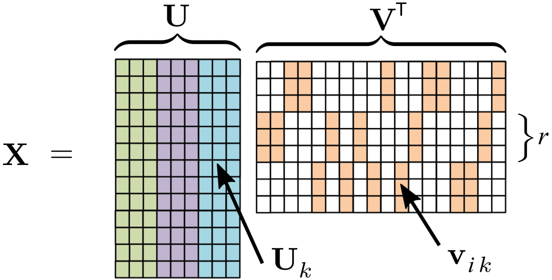
4 [Pimentel-Alarcon et al., 2016]

Group sparse matrix factorization[4]



[4] [Pimentel-Alarcon et al., 2016]

Group sparse matrix factorization[4]

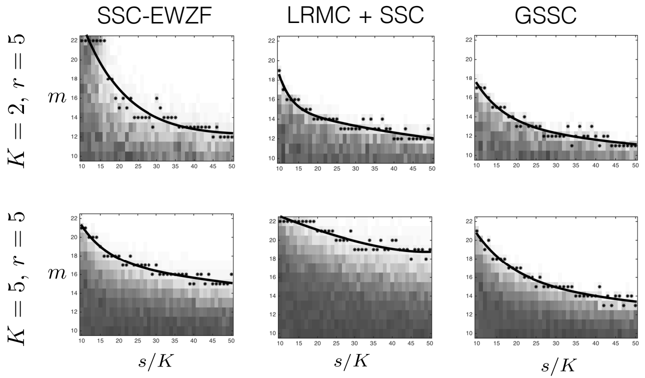


$$(\hat{\boldsymbol{U}}, \hat{\boldsymbol{V}}) = \underset{\boldsymbol{U}:\|\boldsymbol{U}\|_F \leq 1, \boldsymbol{V}}{\arg\min} \|\boldsymbol{X}_0 - \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{U}\boldsymbol{V}^T)\|_F^2 + \lambda \sum_{i=1}^{s} \sum_{k=1}^{K} \|\boldsymbol{v}_{i,k}\|_2$$

Lemma 2: Accumulation point exists and is a critical point of the objective function.

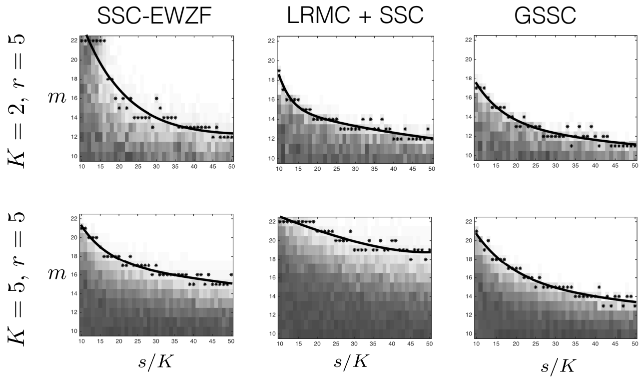
[4] [Pimentel-Alarcon et al., 2016]

GSSC Results



Proportion of correctly classified points as a function of s/K (number of columns per subspace) and m (number of observed entries per column). White represents 100% accuracy. $n = 25$.

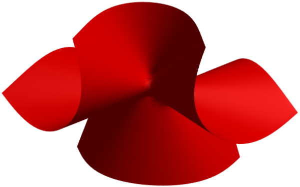
GSSC Results



Proportion of correctly classified points as a function of s/K (number of columns per subspace) and m (number of observed entries per column). White represents 100% accuracy. $n = 25$.

Challenge: accuracy depends heavily on quality of initial clustering

Matrix completion for algebraic varieties



Greg Ongie

Laura Balzano

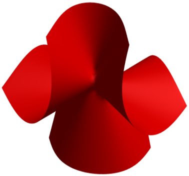
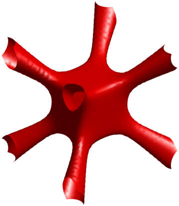
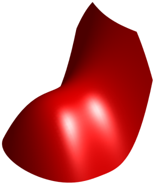
Robert Nowak

Algebraic Varieties

An algebraic variety is the solution set of a system of polynomial equations:

$$V = \{\boldsymbol{x} \in \mathbb{R}^n : p_1(\boldsymbol{x}) = \cdots = p_K(\boldsymbol{x}) = 0\}$$

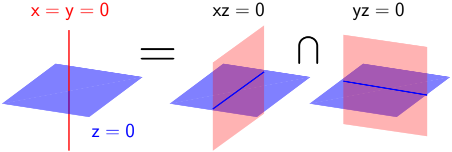
for some polynomials $p_1, ..., p_K$ in variables $\boldsymbol{x} = (x_1, ..., x_n)$.

A union of subspaces is a variety[5]

Example: Union of line and plane

$$\boldsymbol{U} = \{z = 0\},$$
$$\boldsymbol{V} = \{x = 0, y = 0\},$$
$$\boldsymbol{U} \cup \boldsymbol{V} = \underbrace{\{xz = 0, yz = 0\}}_{\text{system of quadratic eqns}}$$

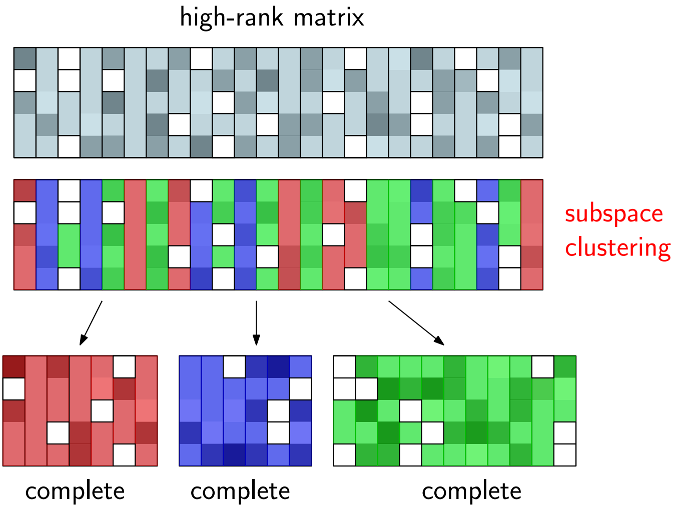


Lemma 3: If $\boldsymbol{U}_1, ..., \boldsymbol{U}_K$ are subspaces, then

$$\cup_{k=1}^{K} \boldsymbol{U}_k = \{\boldsymbol{x} : \underbrace{\ell_1(\boldsymbol{x}) \cdots \ell_K(\boldsymbol{x})}_{\text{product of linear forms}} = 0, \; \ell_k \text{ linear}, \ell_k \text{ vanishes on } \boldsymbol{U}_k\}$$

[5] Algebraic Subspace Clustering/Generalized PCA [Vidal et al., 2016]

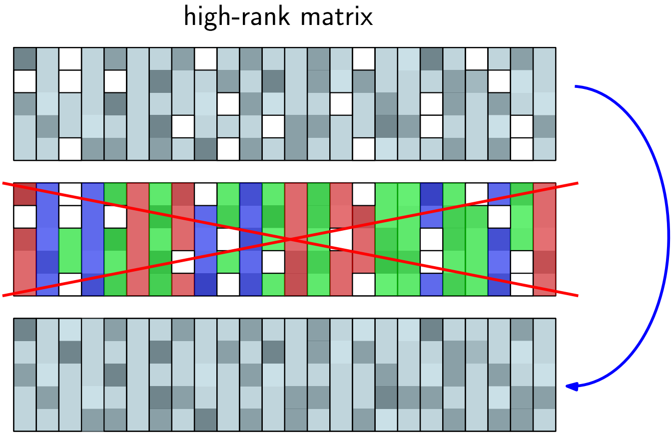
Matrix completion under a union of subspaces model[6]



Clustering is difficult with missing data.

[6] [Eriksson et al., 2012, Yang et al., 2015, Pimentel-Alarcón et al., 2016]

Matrix completion under a union of subspaces model[6]



Variety formulations bypass clustering.

6 [Eriksson et al., 2012, Yang et al., 2015, Pimentel-Alarcón et al., 2016]

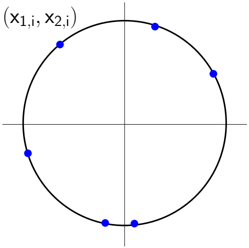
Veronese mappings

Key observation: Data belonging to a variety are rank deficient under a Veronese embedding.

- Consider matrix of points in $\mathbb{R}^2$ draw from a quadratic curve:

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

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



Veronese mappings

Key observation: Data belonging to a variety are rank deficient under a Veronese embedding.

- Consider matrix of points in $\mathbb{R}^2$ draw from a quadratic curve:

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

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

- Map each point to all monomials with degree ≤ 2:

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

- $\boldsymbol{X}$ is full rank, but $\boldsymbol{Y}$ is rank deficient:
  $\boldsymbol{c}^T\boldsymbol{Y} = \boldsymbol{0}$ with $\boldsymbol{c} = (c_0, ..., c_5)^T \implies \operatorname{rank}(\boldsymbol{Y}) \leq 5.$

Veronese embeddings

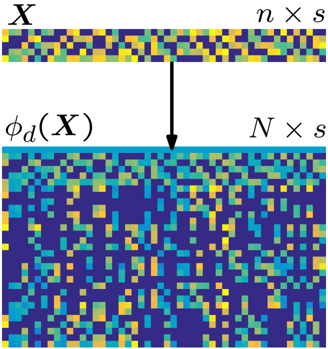
- For $\boldsymbol{x} = [x_1, ..., x_n]^T \in \mathbb{R}^n$ define

$$\phi_d(\boldsymbol{x}) := \underbrace{(x_1^{\alpha_1} \cdots x_n^{\alpha_n})_{|\alpha| \leq d}}_{\text{all degree } \leq d \text{ monomials}} \in \mathbb{R}^N$$

  for $N = \binom{n+d}{d}$

- For a matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s] \in \mathbb{R}^{n \times s}$,

$$\phi_d(\boldsymbol{X}) := [\phi_d(\boldsymbol{x}_1), ..., \phi_d(\boldsymbol{x}_s)] \in \mathbb{R}^{N \times s}$$



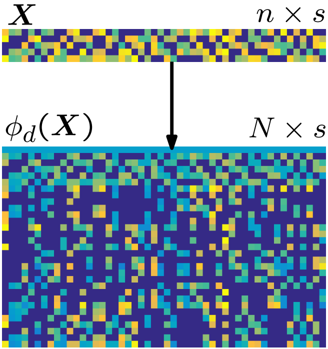
Veronese embeddings

- For $\boldsymbol{x} = [x_1, ..., x_n]^T \in \mathbb{R}^n$ define

$$\phi_d(\boldsymbol{x}) := \underbrace{(x_1^{\alpha_1} \cdots x_n^{\alpha_n})_{|\alpha| \le d}}_{\text{all degree } \le d \text{ monomials}} \in \mathbb{R}^N$$

  for $N = \binom{n+d}{d}$
- For a matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s] \in \mathbb{R}^{n \times s}$,

$$\phi_d(\boldsymbol{X}) := [\phi_d(\boldsymbol{x}_1), ..., \phi_d(\boldsymbol{x}_s)] \in \mathbb{R}^{N \times s}$$



Lemma 4: $\phi_d(\boldsymbol{X})$ is rank deficient if and only if columns of $\boldsymbol{X}$ lie on a variety generated by polynomials of degree $\le d$:

$$\boldsymbol{C}^T \phi_d(\boldsymbol{X}) = \boldsymbol{0}$$

Restatement of Main Objective

Main objective:

Complete a partially observed matrix $\boldsymbol{X}$ under the assumption that the columns of $\boldsymbol{X}$ lie on a variety?

$$\Updownarrow$$

Complete a partially observed matrix $\boldsymbol{X}$ under the assumption that $\phi_d(\boldsymbol{X})$ is low-rank

Restatement of Main Objective

Main objective:

Complete a partially observed matrix $\boldsymbol{X}$ under the assumption that the columns of $\boldsymbol{X}$ lie on a variety?

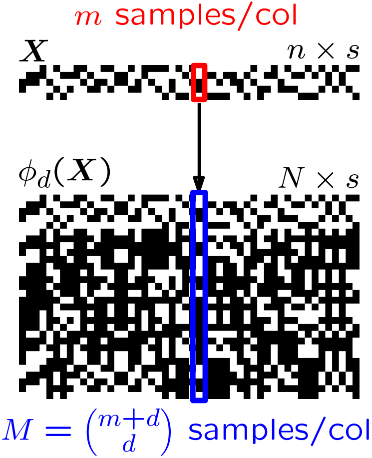
$\Updownarrow$

Complete a partially observed matrix $\boldsymbol{X}$ under the assumption that $\phi_d(\boldsymbol{X})$ is low-rank

Optimization formulation:

$$\min_{\boldsymbol{X}} \operatorname{rank} \phi_d(\boldsymbol{X}) \text{ subject to } \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}_0)$$

When could this work?

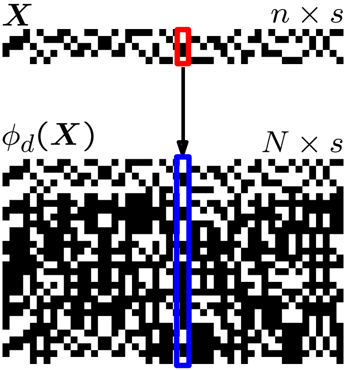


When could this work?

Degrees of freedom (DoF):

of a $N \times s$ rank-R matrix $= R(N + s - R)$

m samples/col



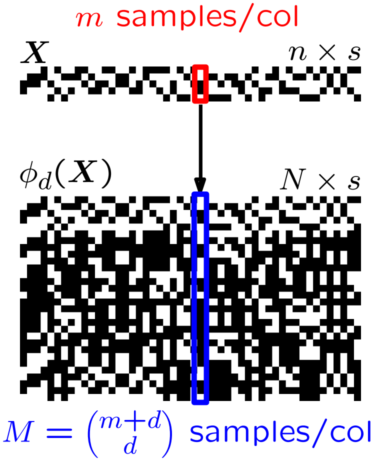
$M = \binom{m+d}{d}$ samples/col

When could this work?

Degrees of freedom (DoF):

of a $N \times s$ rank-R matrix $= R(N + s - R)$

of a $N \times s$ rank-R Veronese embedding matrix $= R(n + s - R)$

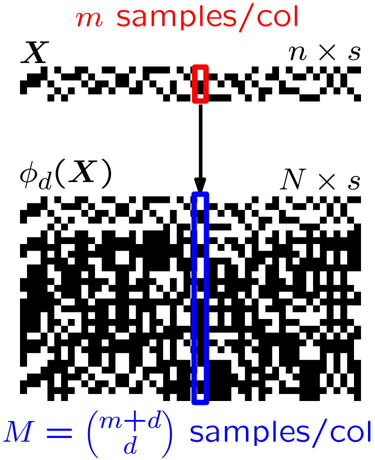


When could this work?

Degrees of freedom (DoF):

of a $N \times s$ rank-R matrix $= R(N + s - R)$

of a $N \times s$ rank-R Veronese embedding matrix $= R(n + s - R)$



Lemma 5: (Predicted minimal sampling rate)

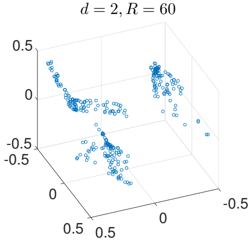
$$Ms \geq R(n + s - R)$$

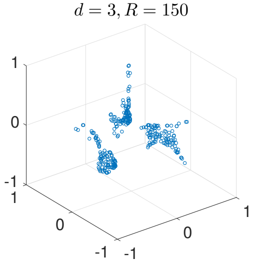
if

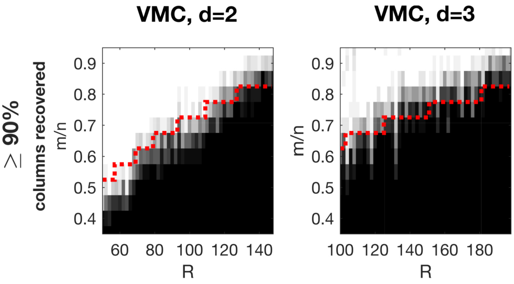
$$m \geq n \left(\frac{R}{N} \right)^{\frac{1}{d}}, \text{ for } s \gg R$$

Phase transitions - Parametric Curves/Surfaces

Example Datasets:







ambient dimension	$n = 20$
datapoints	$s = 300$
embedding space rank	R
samples per column	m/n

Unions of Subspaces

Recall that a union of subspaces is a variety.

Lemma 6: If the columns of $\boldsymbol{X} \in \mathbb{R}^{n \times s}$ belong to a union of K subspaces, each with dimension at most r, then

$$R = \operatorname{rank} \phi_d(\boldsymbol{X}) \leq K \binom{r+d}{d}.$$

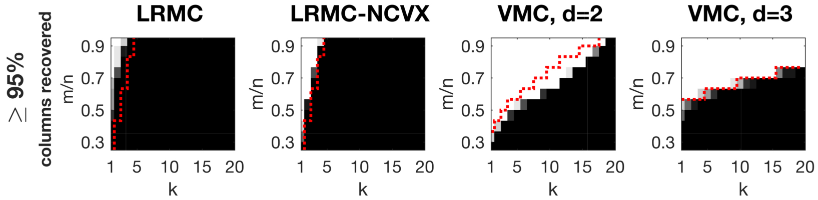
Then the minimal number of observed entries per column is

$$m \geq n \left(\frac{R}{N}\right)^{\frac{1}{d}} \approx K^{1/d} r$$

- To perform low-rank matrix completion in $\boldsymbol{X}$, we'd need $m \approx Kr$
- Bigger d isn't always better, as we need $s = O(Kr^d)$

Phase transitions - Union of Subspaces

Predicted sampling rate: $m/n = O(K^{1/d} r)$



Randomly generate UoS data:	
ambient dimension	$n = 15$
subspace dimension	$r = 3$
number of subspaces	$K = 1, ..., 20$
samples per column	m/n

Schatten-p quasi-norm minimization

- Relaxed formulation:

$$\min_{\boldsymbol{X}} \|\phi_d(\boldsymbol{X})\|_{\mathcal{S}_p}^p \text{ subject to } \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}_0)$$

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

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

with $\sigma_i(\boldsymbol{Y})$ denoting the i^{th} singular value of $\boldsymbol{Y}$.

- For $p = 1$ we recover the nuclear norm; for $p < 1$ penalty is non-convex.
- We call this optimization formulation variety-based matrix completion (VMC).

Iterative Reweighted Least Squares (IRLS) Algorithm[7]

- **Example**: Low-rank matrix completion via nuclear norm minimization

$$\min_{\boldsymbol{Y}} \|\boldsymbol{Y}\|_* \text{ subject to } \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{Y}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{Y}_0),$$

- Basic IRLS approach

$$\|\boldsymbol{Y}\|_* = \mathsf{tr}\left(\boldsymbol{Y}^T\boldsymbol{Y}\right)^{\frac{1}{2}} = \mathsf{tr}\left(\boldsymbol{Y}^T\boldsymbol{Y}\right)\underbrace{\left(\boldsymbol{Y}^T\boldsymbol{Y}\right)^{-\frac{1}{2}}}_{\boldsymbol{W}}$$

[7] [Fornasier et al., 2011, Mohan and Fazel, 2012]

Iterative Reweighted Least Squares (IRLS) Algorithm[7]

- **Example**: Low-rank matrix completion via nuclear norm minimization

$$\min_{\boldsymbol{Y}} \|\boldsymbol{Y}\|_* \text{ subject to } \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{Y}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{Y}_0),$$

- Basic IRLS approach

$$\|\boldsymbol{Y}\|_* = \operatorname{tr}(\boldsymbol{Y}^T\boldsymbol{Y})^{\frac{1}{2}} = \operatorname{tr}(\boldsymbol{Y}^T\boldsymbol{Y})\underbrace{(\boldsymbol{Y}^T\boldsymbol{Y})^{-\frac{1}{2}}}_{\boldsymbol{W}}$$

while not converged **do**
$\quad \boldsymbol{W} \leftarrow (\boldsymbol{Y}^T\boldsymbol{Y})^{-\frac{1}{2}}$
$\quad \boldsymbol{Y} \leftarrow \arg\min_{\boldsymbol{Y}} \operatorname{tr}(\boldsymbol{Y}^T\boldsymbol{Y})\boldsymbol{W} \text{ subject to } \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{Y}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{Y}_0)$
end while

7 [Fornasier et al., 2011, Mohan and Fazel, 2012]

IRLS for Variety Completion

IRLS for low-rank matrix completion

while not converged **do**
$\boldsymbol{W} \leftarrow (\boldsymbol{Y}^T\boldsymbol{Y})^{\frac{p}{2}-1}$
$\boldsymbol{Y} \leftarrow \arg\min_{\boldsymbol{Y}} \operatorname{tr}(\boldsymbol{Y}^T\boldsymbol{Y})\boldsymbol{W}$ subject to $\mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{Y}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{Y}_0)$
end while

IRLS for variety-based matrix completion

while not converged **do**
$\boldsymbol{W} \leftarrow (\phi_d(\boldsymbol{X})^T\phi_d(\boldsymbol{X}))^{\frac{p}{2}-1}$
$\boldsymbol{X} \leftarrow \arg\min_{\boldsymbol{X}} \operatorname{tr}\phi_d(\boldsymbol{X})^T\phi_d(\boldsymbol{X})\boldsymbol{W}$ subject to $\mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}_0)$
end while

Challenge: embedding space dimension $N = \binom{n+d}{d} = O(n^d)$ is large.

The Kernel Trick[8]

Efficiently compute inner-products with polynomial kernel:

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

For matrices $\boldsymbol{X}, \boldsymbol{Y}$:

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

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.

Substantially reduces working dimension:
$k_d(\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{R}^{s \times s}$ vs. $\boldsymbol{X} \in \mathbb{R}^{N \times s}$.

8 [Muller et al., 2001]

IRLS for variety-based matrix completion

while not converged **do**

$\boldsymbol{W} \leftarrow (\phi_d(\boldsymbol{X})^T \phi_d(\boldsymbol{X}))^{\frac{p}{2}-1}$

$\boldsymbol{X} \leftarrow \arg\min_{\boldsymbol{X}} \mathsf{tr}\, \phi_d(\boldsymbol{X})^T \phi_d(\boldsymbol{X}) \boldsymbol{W}$ subject to $\mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}_0)$

end while

Kernelized IRLS for variety-based matrix completion

while not converged **do**

$\boldsymbol{W} \leftarrow k_d(\boldsymbol{X}, \boldsymbol{X})^{\frac{p}{2}-1}$

$\boldsymbol{X} \leftarrow \arg\min_{\boldsymbol{X}} \operatorname{tr} k_d(\boldsymbol{X}, \boldsymbol{X})\boldsymbol{W}$ subject to $\mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}_0)$

end while

Kernelized IRLS for variety-based matrix completion

while not converged **do**

$\boldsymbol{W} \leftarrow k_d(\boldsymbol{X}, \boldsymbol{X})^{\frac{p}{2}-1}$

$\boldsymbol{X} \leftarrow \arg\min_{\boldsymbol{X}} \operatorname{tr} k_d(\boldsymbol{X}, \boldsymbol{X})\boldsymbol{W} \text{ subject to } \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}_0)$

end while

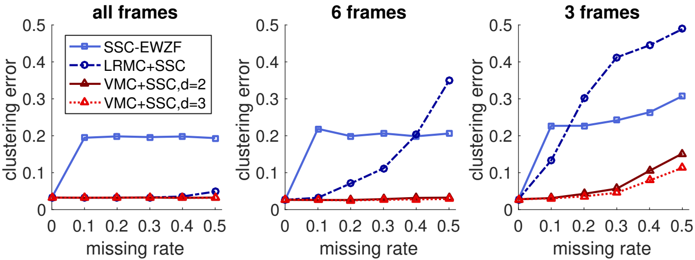
Lemma 7: Every limit point of the iterates generated by the kernelized IRLS algorithm is a stationary point of the ϵ-smoothed Schatten-p norm objective function

$$\min_{\boldsymbol{X}} \operatorname{tr}(k_d(\boldsymbol{X}, \boldsymbol{X}) + \epsilon \boldsymbol{I})^{\frac{p}{2}} \text{ s.t. } \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}_0)$$

Subspace clustering with missing data

Bootstrap into a subspace clustering algorithm with missing data (VMC+SSC)

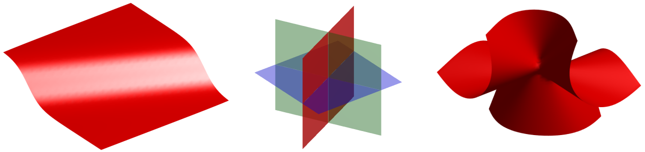
1. Fill in missing data with VMC
2. Sparse Subspace Clustering (SSC)[9]



Motion segmentation on Hopkins 155 dataset

9 [Elhamifar and Vidal, 2009]

Nonlinear models for matrix completion



- Nonlinearities appear throughout in real-world data but are ignored by low-rank matrix completion – SAD!
- Leveraging nonlinear models improves missing data inference – TERRIFIC!
- Variety-based models offer TREMENDOUS flexibility without clustering

Nonlinear models for matrix completion

- Nonlinearities appear throughout in real-world data but are ignored by low-rank matrix completion – SAD!
- Leveraging nonlinear models improves missing data inference – TERRIFIC!
- Variety-based models offer TREMENDOUS flexibility without clustering
- Open questions: Are convex formulations possible? Or stronger guarantees for non-convex formulations? Will Roummel like Wonder Woman?

Thank you

More details:

https://arxiv.org/abs/1703.09631

https://arxiv.org/abs/1512.08787

http://ieeexplore.ieee.org/document/7551734/

References I

Elhamifar, E. and Vidal, R. (2009).
Sparse subspace clustering.
In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2790–2797. IEEE.

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

Eriksson, B., Balzano, L., and Nowak, R. D. (2012).
High-rank matrix completion.
In *AISTATS*, pages 373–381.

Fornasier, M., Rauhut, H., and Ward, R. (2011).
Low-rank matrix recovery via iteratively reweighted least squares minimization.
SIAM Journal on Optimization, 21(4):1614–1640.

References II

Ganti, R. S., Balzano, L., and Willett, R. (2015).
Matrix completion under monotonic single index models.
In *Advances in Neural Information Processing Systems*, pages 1873–1881.

Horowitz, J. L. and Härdle, W. (1996).
Direct semiparametric estimation of single-index models with discrete covariates.
Journal of the American Statistical Association, 91(436):1632–1640.

Ichimura, H. (1993).
Semiparametric least squares (sls) and weighted sls estimation of single-index models.
Journal of Econometrics, 58(1-2):71–120.

Kakade, S. M., Kanade, V., Shamir, O., and Kalai, A. (2011).
Efficient learning of generalized linear and single index models with isotonic regression.
In *Advances in Neural Information Processing Systems*, pages 927–935.

References III

Kalai, A. T. and Sastry, R. (2009).
The isotron algorithm: High-dimensional isotonic regression.
In *COLT*.

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

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

Pimentel-Alarcon, D., Balzano, L., Marcia, R., Nowak, R., and Willett, R. (2016).
Group-sparse subspace clustering with missing data.
In *IEEE Statistical Signal Processing Workshop*.

References IV

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

Vidal, R., Ma, Y., and Sastry, S. (2016).
Generalized Principal Component Analysis.
Springer New York.

Yang, C., Robinson, D., and Vidal, R. (2015).
Sparse subspace clustering with missing entries.
In *Proceedings of The 32nd International Conference on Machine Learning*, pages 2463–2472.