Prediction from low-rank missing data

Elad Hazan Roi Livni Yishay Mansour

Princeton U Hebrew U Tel-Aviv U

& Microsoft Research (all of us)

Recommendation systems

18,000 movies

480,000 users

x	1	1	x	...	x
x	x	x	5	...	x
x	x	3	x	...	x
x	4	3	x	...	2
...	x	x	x	...	x
x	5	x	1	...	x
x	x	3	3	...	x
x	1	x	x	...	2

Predicting from low-rank missing data

Gender? Annual income?
Will buy “Halo4”?
Likes cats or dogs?

18,000 movies

480,000 users

x	1	1	x	...	x
x	x	x	5	...	x
x	x	3	x	...	x
x	4	3	x	...	2
...	x	x	x	...	x
x	5	x	1	...	x
x	x	3	3	...	x
x	1	x	x	...	2

1
0
0
1
1
1
0
1

Formally: predicting w. low-rank missing data

Unknown distribution on vectors/rows x'_i in $\{0,1\}^n$, missing data x_i in $\{*,0,1\}^n$ (observed), X has rank k, training data y in {0,1}, every row has >= k observed entries

Find: efficient machine M: $\{*,0,1\}^n \rightarrow R$
s.t. with poly(δ,ε,k,n) samples, with probability 1-δ:

$$E_i[(M(x_i) - y_i)^2] - \min_{\|w\| \leq 1} E_i[(w^\top x_i - y_i)^2] \leq \epsilon$$

Kernel version:

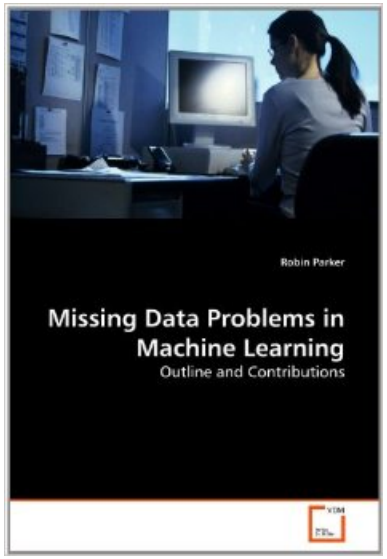
$$E_i[(M(x_i) - y_i)^2] - \min_{\|w\| \leq 1} E_i[(w^\top \phi(x_i) - y_i)^2] \leq \epsilon$$

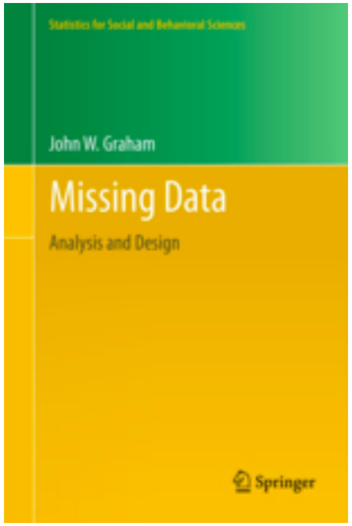
Difficulties

- Missing data (usually MOST data is missing)
- Structure in missing data (low rank)
- NP-hard (low-rank reconstruction is a special case)
- Can we use a non-proper approach? (distributional assumptions, convex relaxations for reconstruction)

Missing data (statistics & ML)

Statistics books: i.i.d missing entries.
recovery from (large) constant percentage (MCAR,MAR)
Or generative model for missing-ness (MNAR)
very different from what we need...





approach 1: Completion & prediction

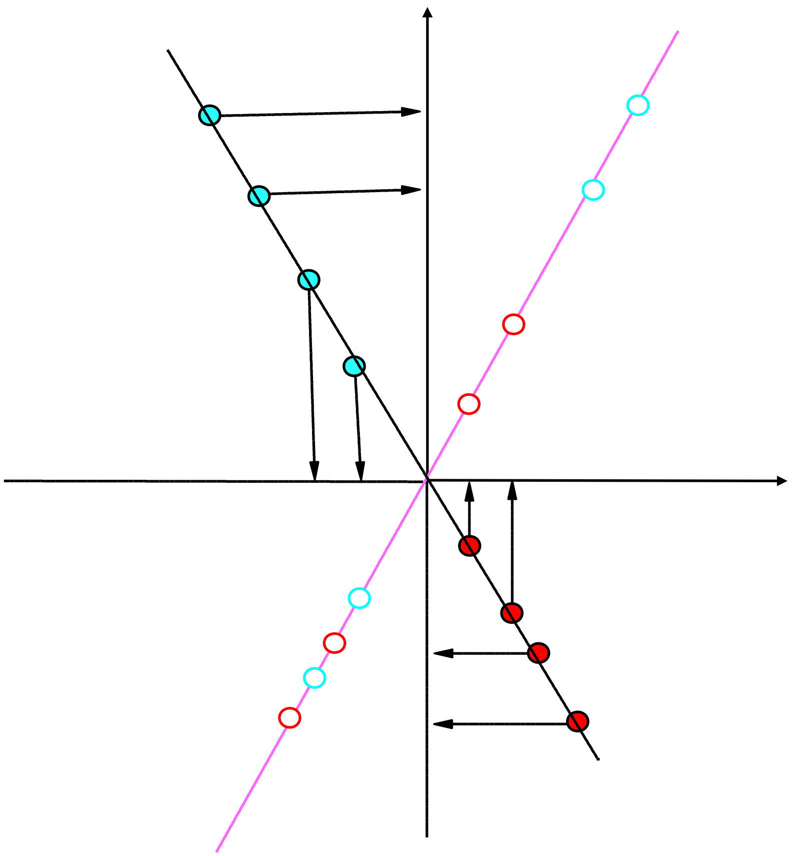
[Goldberg, Zhu, Recht, Xu, Nowak '10]

Method: add predictions y as another column in X, use matrix completion to reconstruct & predict.

Can we use approach 1? Completion & prediction

[Goldberg, Zhu, Recht, Xu, Nowak '10]

reconstruction is not sufficient nor necessary!!



Can we use approach 1? Completion & prediction

[Goldberg, Zhu, Recht, Xu, Nowak '10]

1	*	1	*
*	-1	*	-1
1	-1	1	-1
1	*	1	*

1
-1
-1
??

1	-1	1	-1
1	-1	1	-1
1	-1	1	-1
1	-1	1	-1

1
-1
-1
-1

Both are rank-2 completions

1	1	1	1
-1	-1	-1	-1
1	-1	1	-1
1	1	1	1

1
-1
-1
1

Can we use approach 1?
Completion & prediction

[Goldberg, Zhu, Recht, Xu, Nowak ‘10]

K				K			
0	1	0	1	*	*	*	*
1	0	1	0	*	*	*	*
0	0	0	0	*	*	*	*
1	1	1	1	*	*	*	*
*	*	*	*	1	1	1	1
*	*	*	*	1	0	1	0
*	*	*	*	0	1	0	1
*	*	*	*	1	0	1	0

Gender? Annual income?
Will buy “Halo4”?

1
0
0
1
1
1
0
1

There is a recoverable k-dim subspace!!

Our results (approach 2)

- Agnostic learning – compete with the best linear predictor that knows all the data, assuming it is rank k (or close)
- Provable
- Efficient (theoretically & practically)
- Significantly improves prediction over standard datasets (Netflix, Jester, ….)
- Generalizes to kernel (non-linear) prediction

Our results (approach 2)
Formally:

Unknown distribution on rows x'_i in $\{0,1\}^n$, missing data x_i in $\{*,0,1\}^n$ (observed), X' has rank k, training data y in $\{0,1\}$, every row has >= k observed entries

We build efficient machine $M: \{*,0,1\}^n \rightarrow R$
s.t. with poly($\log \delta, k, n^{\log(1/\varepsilon)}$) samples, with probability $1-\delta$:

$$E_i[(M(x_i) - y_i)^2] - \min_{\|w\| \leq 1} E_i[(w^\top x_i - y_i)^2] \leq \epsilon$$

Extends to arbitrary kernels, # samples increases w. degree (polynomial kernels)

Warm up: agnostic, non-proper & useless (inefficient)

- Data matrix = X of size m *n (X' is full matrix, X with hidden entries)
  rank = k
  every row has k visible entries

- "Optimal predictor" = subspace + linear predictor (SVM)
  - B = basis , k * n matrix
  - w = predictor, vector in R^k

- Given x = row in X, unknown label y predict according to:

$$B\alpha = x$$

$$\hat{y} = \alpha^\top w$$

Warm up: inefficient, agnostic

- Given x = row in X, unknown label y predict according to:

$$B\alpha = x$$

$$\hat{y} = \alpha^\top w$$

Inefficiently: learn B, w (bounded sample complexity/regret – compact sets)

(distributional world – bounded fat-shattering dimension)

Learning a hidden subspace

Learning a hidden subspace is hidden-clique hard! [Berthet & Rigollet '13], any hope for efficient algorithms?

Hardness applies only for proper learning!!

Efficient agnostic algorithm

- Let s be the set of k coordinates that are visible in a certain x. Then:

$$\begin{aligned} B\alpha &= x \\ \hat{y} &= \alpha^\top w \end{aligned} \quad \Longleftrightarrow \quad \hat{y} = (B_s^{-1} x_s)^\top w$$

Where B_s and x_s are the submatrix (vector) corresponding to the coordinates s.

“2 operations” – subset of s rows & inverse

Step 1: “rid of inverse”

Replace inverse by polynomial (need condition on the eigenvalues):

$$w^\top B_s^{-1} x_s = w^\top \left[\sum_{j=1}^{\infty} (I_s - B_s)^j \right] x_s$$

Let C = I – B, and up to precision independent of k,n:

$$w^\top B_s^{-1} x_s = w^\top \left[\sum_{j=1}^{q} C_s^j \right] x_s + O(\frac{1}{q})$$

Thus, consider (non-proper) hypothesis class:

$$g_{C,w}(x_s) = w^\top \left[\sum_{j=1}^{q} C_s^j \right] x_s$$

Step 2: “rid of column selection”

Observation:

$$g_{C,w}(x_s) = \sum_{\ell \subseteq s \; |\ell| \leq q} w_{\ell_1} C_{\ell_1,\ell_2} \times \ldots \times C_{\ell_{|\ell|-1},\ell_{|\ell|}} \cdot x_{\ell_{|\ell|}}$$

(polynomial in C,w multiplied by coefficients of x)

Thus, there is a kernel mapping, and vector
v=v(C,w) such that

$$g_{C,w}(x_s) = v^\top \Phi(x_s)$$

$$v = v(C,w) \in \mathcal{R}^{n^q}$$

Observation 3

Kernel inner products take the form:

$$\phi(x_s^{(1)}) \cdot \phi(x_t^{(2)}) = \frac{|s \cap t|^q - 1}{|s \cap t| - 1} \sum_{k \in s \cap t} x_k^{(1)} x_k^{(2)}$$

Inner product $\phi\ (x_s)*\phi(x_t)$ –computed in time n*q

Algorithm

Kernel function

$$\phi(x_s^{(1)}) \cdot \phi(x_t^{(2)}) = \frac{|s \cap t|^q - 1}{|s \cap t| - 1} \sum_{k \in s \cap t} x_k^{(1)} x_k^{(2)}$$

Algorithm: SVM kernel with this particular kernel.

Guarantee – agnostic, non-proper, as good as best subspace embedding.

Nearly same algorithm for all degree q!

λ - regularity

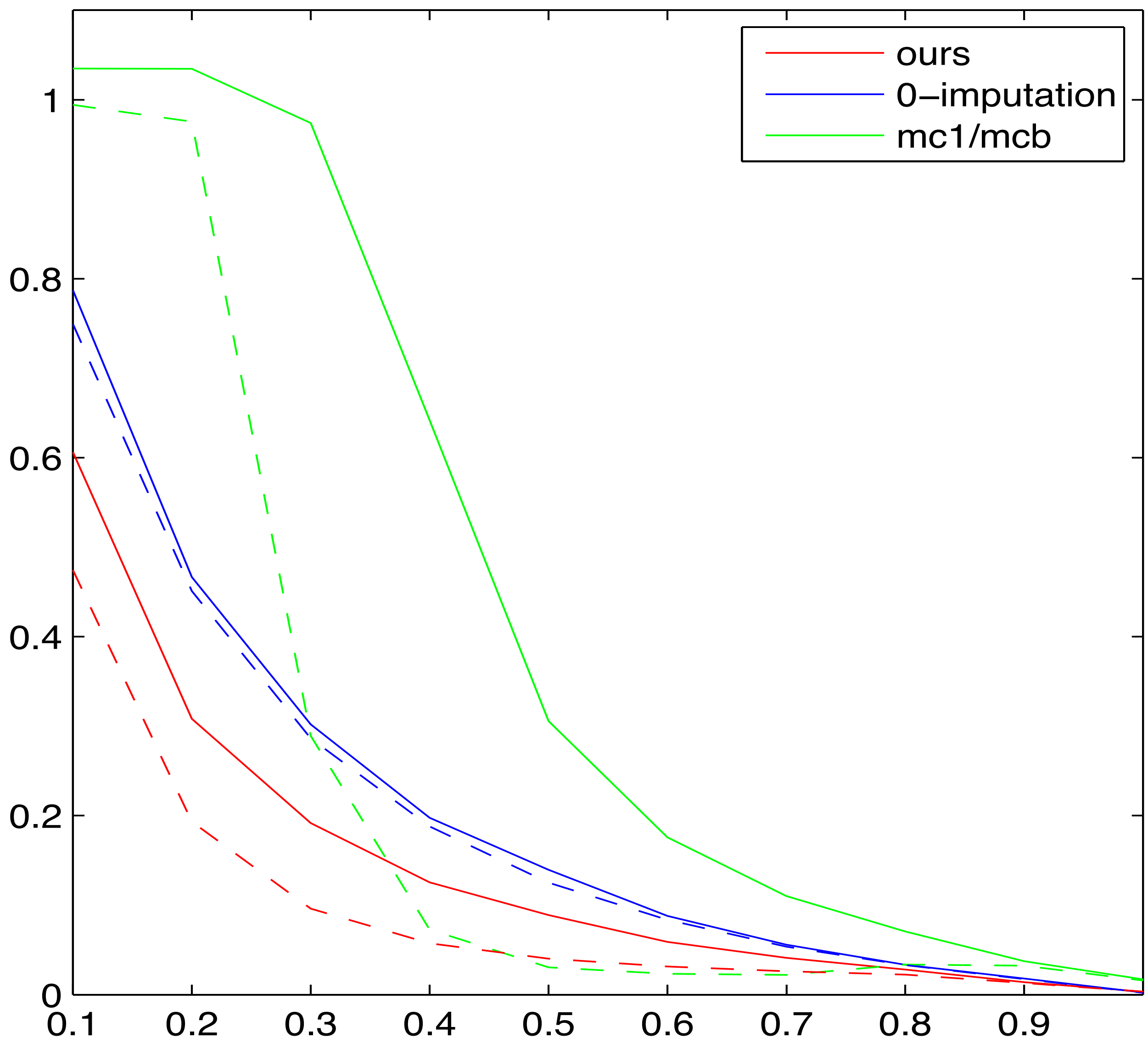
To apply the Taylor series – eigenvalues need to be in unit circle.

Reduces to an assumption on appearance of missing data. This is provably necessary.

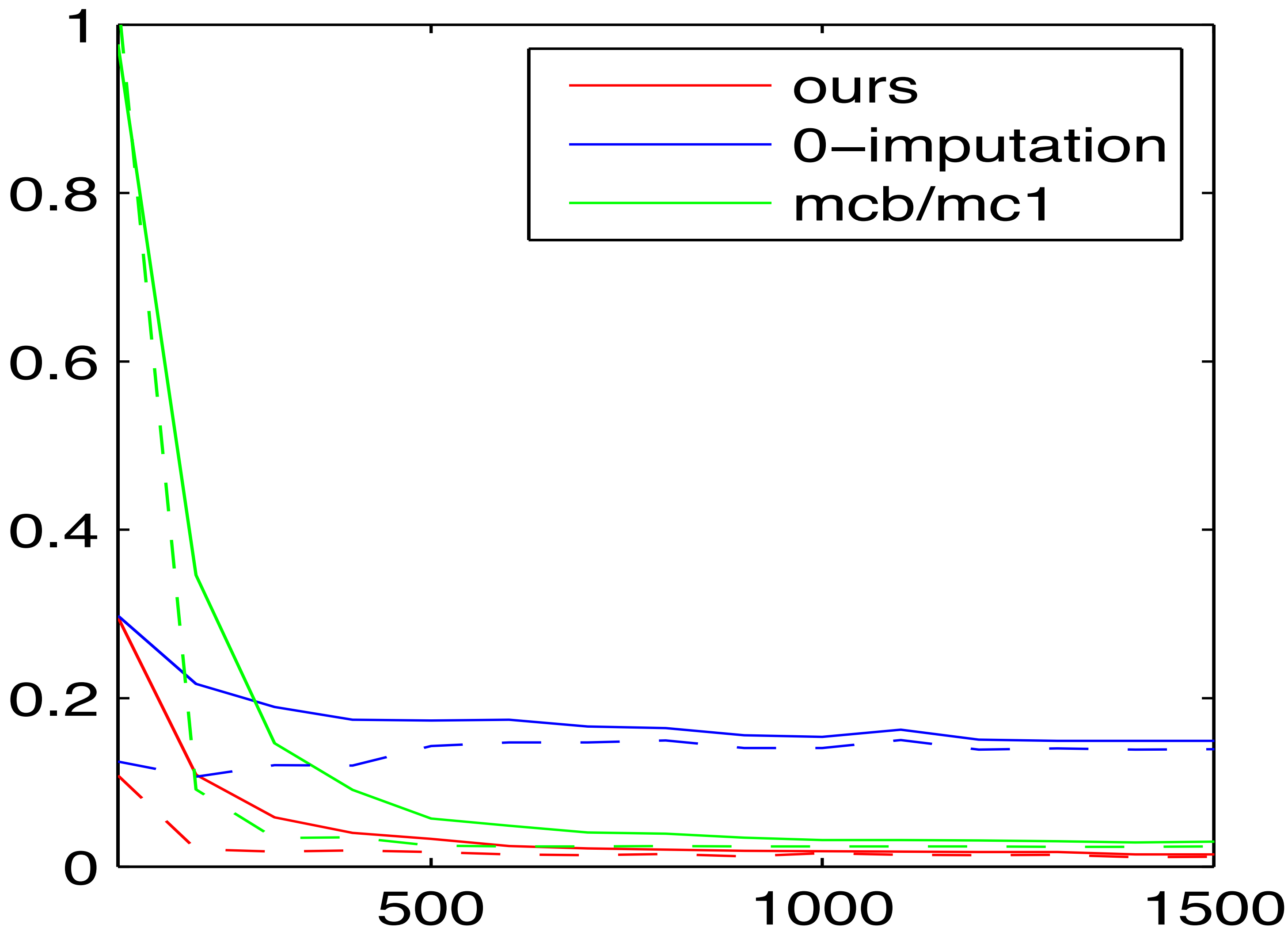
Regret bound (sample complexity) depend on this parameter – which is provably a constant independent of the rank/problem dimensions.

Running time – independent of this parameter.

Preliminary benchmarks
MAR data



Preliminary benchmarks
NMAR data (blocks)



Preliminary benchmarks real data

	Karma	0-svm	Mcb0	Mcb1	Geom
mamographic	0.17	0.17	0.17	0.18	0.17
bands	0.24	0.34	0.41	0.40	0.35
hepatitis	0.23	0.17	0.23	0.21	0.22
wisconsin	0.03	0.03	0.03	0.04	0.04
Horses	0.35	0.36	0.55	0.37	0.36
Movielens (age)	0.16	0.22	0.25	0.25	NaN

Summary

Prediction from recommendation data:

- Reconstruction+relaxation approach doomed to fail
- Non-proper agnostic learning gives provable guarantees, efficient algorithm
- Benchmarks are promising
- Non-reconstructive approach fo[illegible] types of missing data? Fully-polynomial alg?
- When does reconstruction fail and agnostic/non-proper learning work?