

Low-rank matrix completion

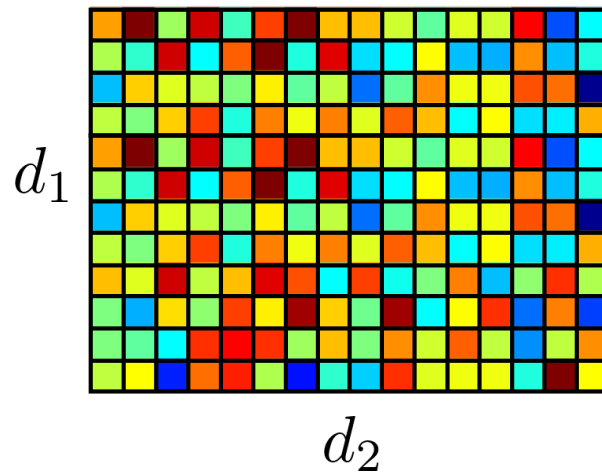
An overview and some recent advances...

Mark A. Davenport

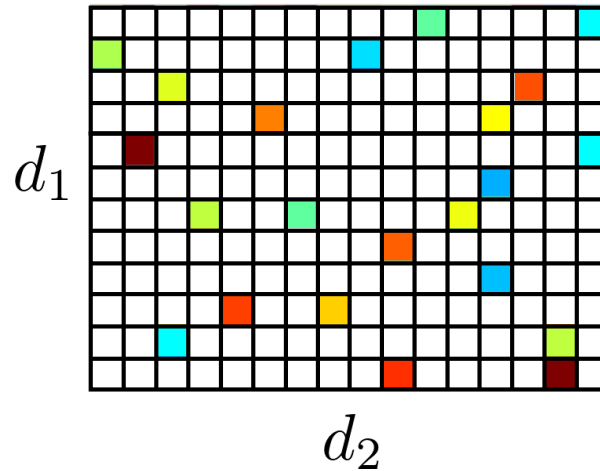


Georgia Institute
of **Tech**nology

Matrix completion

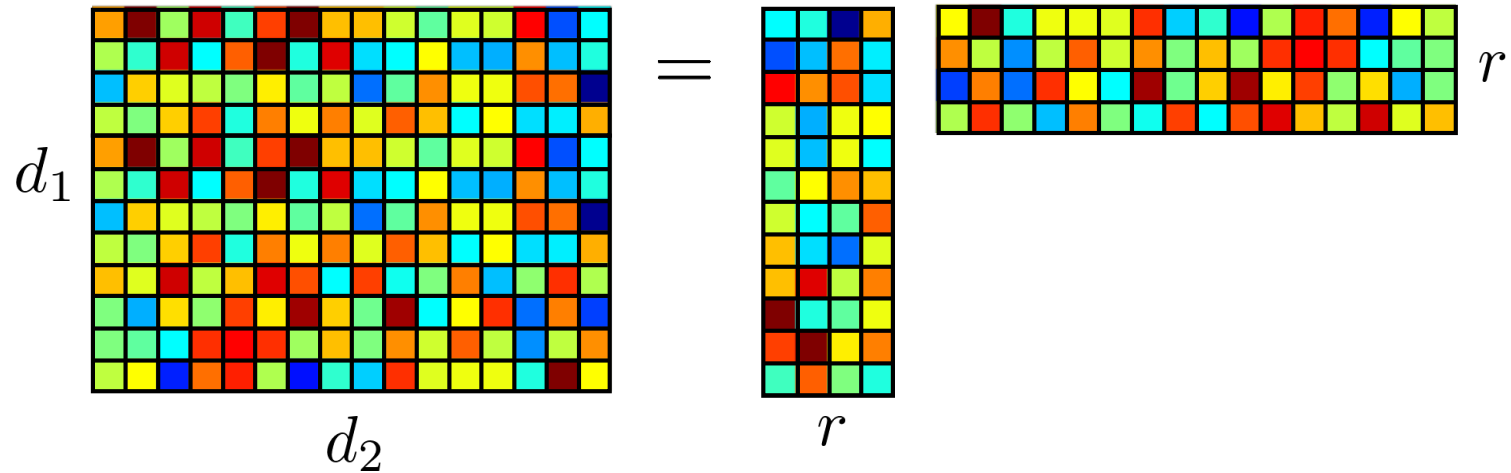


Matrix completion



- When is it possible to recover the original matrix?
- How can we do this efficiently?
- How many samples will we need?

Low-rank matrices



Singular value decomposition:

$$M = U\Sigma V^T \quad \longrightarrow \quad \approx r(d_1 + d_2) \ll d_1 d_2$$

degrees of freedom

Applications

- Recommendation systems
- Analysis of student response data
- Recovery of incomplete survey data
- Analysis of voting data
- Localization/multidimensional scaling
- Undersampled ensembles of correlated signals
- Blind deconvolution
- Phase recovery
- Quantum state tomography
- ...

Application: Recommendation systems

The “Netflix Problem”

$$M(i, j) = \text{how much user } i \text{ likes movie } j$$

Rank 1 model: u_i = how much user i likes romantic movies

v_j = amount of romance in movie j

$$M(i, j) = u_i v_j$$

Rank 2 model: w_i = how much user i likes zombie movies

x_j = amount of zombies in movie j

$$M(i, j) = u_i v_j + w_i x_j$$

LOVE
MEANS
NEVER
HAVING
TO SAY
YOU'RE
UNDEAD



WARM BODIES

PG-13 PARENTS STRONGLY CAUTIONED
SOME MATERIAL MAY BE INAPPROPRIATE FOR CHILDREN UNDER 13
ZOMBIE VIOLENCE AND SOME LANGUAGE

FEBRUARY 1 #WARMBODIES

Quebec



Low-rank matrix recovery

We will assume we are given

- a $d_1 \times d_2$ matrix M with rank r
- a set of m linear measurements: $y = \mathcal{A}(M)$

Examples of \mathcal{A} include:

- samples of M on the set Ω : $y = M_\Omega$
- general linear measurements: $y_i = \langle A_i, M \rangle = \text{trace}(A_i^T M)$

How can we recover M ?

$$\underset{X}{\text{minimize}} \quad \|y - \mathcal{A}(X)\|_2^2$$

$$\text{subject to } \text{rank}(X) = r$$

Algorithms for low-rank recovery

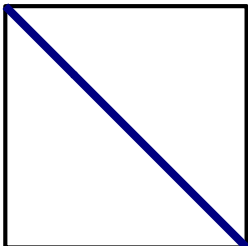
Let's start with something easier!

$$\begin{aligned} & \underset{X}{\text{minimize}} \quad \|Y - X\|_F^2 \\ & \text{subject to} \quad \text{rank}(X) = r \end{aligned}$$

Still a non-convex problem, but has an easy solution...

Eckart-Young Theorem

Compute SVD $Y = U\Sigma V^T$ and truncate: $\widehat{M} = U\bar{\Sigma}V^T$

$\bar{\Sigma} =$  } keep only r largest singular values

Hard and soft thresholding

An equivalent formulation is to solve

$$\underset{X}{\text{minimize}} \quad \|Y - X\|_F^2 + \lambda \cdot \text{rank}(X)$$

Solution given by **hard thresholding**: $\bar{\sigma}_k = \begin{cases} \sigma_k, & \sigma_k \geq \sqrt{\lambda} \\ 0, & \sigma_k < \sqrt{\lambda} \end{cases}$

There is also a **convex** optimization problem with a very similar solution

$$\underset{X}{\text{minimize}} \quad \|Y - X\|_F^2 + \lambda \cdot \|X\|_* \leftarrow \text{ **nuclear norm** }$$

Solution given by **soft thresholding**: $\bar{\sigma}_k = \begin{cases} \sigma_k - \lambda, & \sigma_k \geq \lambda \\ 0, & \sigma_k < \lambda \end{cases}$

Nuclear norm minimization

When we are given indirect observations $y = \mathcal{A}(M)$, we can no longer efficiently solve

$$\underset{X}{\text{minimize}} \quad \|y - \mathcal{A}(X)\|_2^2 + \lambda \cdot \text{rank}(X)$$

However, we *can* still efficiently solve problems like

$$\underset{X}{\text{minimize}} \quad \|y - \mathcal{A}(X)\|_2^2 + \lambda \cdot \|X\|_* \qquad \underset{X}{\text{minimize}} \quad \|y - \mathcal{A}(X)\|_2^2$$

subject to $\|X\|_* \leq \tau$

Can be solved via convex optimization techniques:

- proximal algorithms
- iterative thresholding

Burer-Monteiro Heuristic

Since we expect the solution to be low-rank, we can save on memory by optimizing over $X = LR^T$ where

- L is $d_1 \times r$
- R is $d_2 \times r$

In this case, one can show that $\|X\|_* = \min_{L,R} \frac{1}{2} (\|L\|_F^2 + \|R\|_2^2)$

Thus, we can alternatively solve

$$\underset{L,R}{\text{minimize}} \quad \|y - \mathcal{A}(LR^T)\|_2^2 + \frac{\lambda}{2} \|L\|_F^2 + \frac{\lambda}{2} \|R\|_F^2$$

This is a non-convex problem, but there are conditions under which we can ensure that ***any local minimum is actually a global minimum***

Alternating minimization

Another very common strategy in practice is to solve the factorized version of the problem

$$\underset{L,R}{\text{minimize}} \quad \|y - \mathcal{A}(LR^T)\|_2^2 + \frac{\lambda}{2}\|L\|_F^2 + \frac{\lambda}{2}\|R\|_F^2$$

by alternately fixing L (or R) and solving for R (or L)

With one factor fixed, solving for the other reduces to a simple least squares problem

Again, this is a non-convex problem, but there is now a growing body of theoretical guarantees for this approach

Recovery from Gaussian observations

To provide any theoretical guarantees, we need to make more concrete assumptions on \mathcal{A}

To see what might be possible, suppose that $y_i = \langle A_i, M \rangle$ where A_i has i.i.d. Gaussian entries

When $m \gtrsim r(d_1 + d_2)$, a Gaussian \mathcal{A} acts as an approximate isometry on the set of low-rank matrices

In this case, one can show that nuclear norm minimization will recover M **exactly** (with high probability)!

Very sharp bounds can be established by appealing to Gaussian widths: $m \geq 3r(d_1 + d_2 - r) + 1$

Matrix completion

Theoretical results for matrix completion are a bit more delicate

Not all low-rank matrices can be completed!

Sparse matrices cannot be uniquely determined by simply observing a few entries

To avoid such problematic cases, we typically restrict our attention to matrices M which are ***incoherent***

Define $\mu(U) := \frac{d}{r} \max_{1 \leq i \leq d} \|\mathcal{P}_U e_i\|_2^2$

We will assume that $M = U\Sigma V^T$ satisfy $\mu(U), \mu(V) \leq \mu_0$

Matrix completion

With high probability, nuclear norm minimization will result in exact recovery provided that we observe at least $m \gtrsim \mu_0 r (d_1 + d_2) \log^2(d_1 + d_2)$ entries selected at random

Note that we need at least $m \gtrsim (d_1 + d_2) \log(d_1 + d_2)$ just to ensure that we sample each column/row at least once

Similar guarantees are possible for other algorithms (e.g., alternating minimization)

The (simplest) proofs involve the use of concentration inequalities for sums of random matrices to construct (approximate) “dual certificates”

Matrix completion in practice

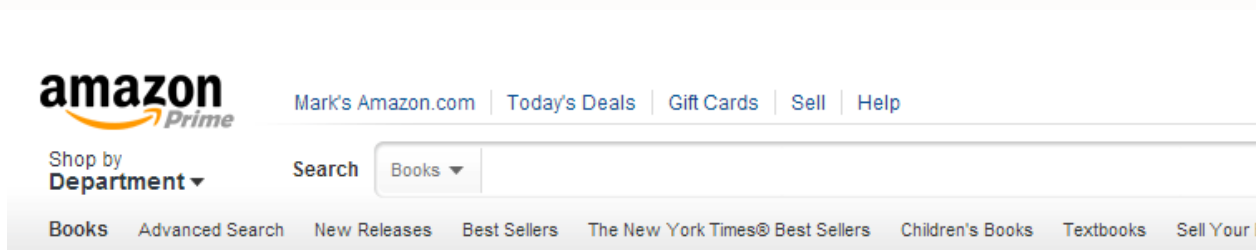
- **Quantization**

- Netflix: Ratings are integers between 1 and 5
- Response data: Correct/Incorrect, Yes/No, Agree/Disagree

- **Dynamics**

- Netflix: Changing tastes over time
- Student response data: Students are (hopefully!) learning

The trouble with quantization



amazon Prime

Mark's Amazon.com | Today's Deals | Gift Cards | Sell | Help

Shop by Department

Search Books

Books | Advanced Search | New Releases | Best Sellers | The New York Times® Best Sellers | Children's Books | Textbooks | Sell Your

Customer Reviews

Linear Operator Theory in Engineering and Science (Applied Mathematical Sciences)



Average Customer Review
★★★★★ (4 customer reviews)

Share your thoughts with other customers

Create your own review

★★★★★ i would give 6 stars!, February 26, 2012

By [Chris](#) See all my reviews

Amazon Verified Purchase (What's this?)

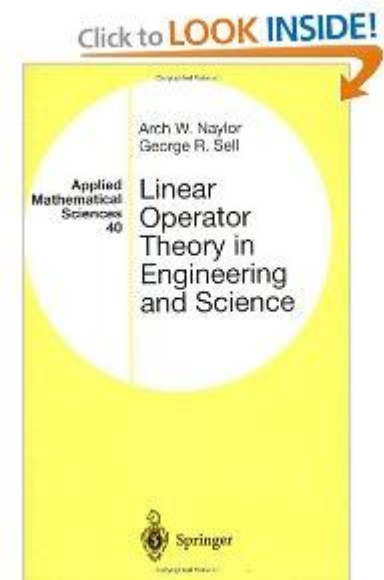
This review is from: **Linear Operator Theory in Engineering and Science (Applied Mathematical Sciences) (Paperback)**

I'm doing a PhD in econometrics and I need to apply operator theories in constructing a linear or nonlinear operator to help explain individual economic behaviour. This book contains numerous useful ideas and applications with exercises thoroughly designed; one of the questions in the exercise gave me an idea of creating a matrix for describing a nonlinear operator. That question asks for a matrix that describes a second order differential operator and that gave me an idea that Taylor series approximation can be used to linearise a nonlinear operator and hence a nonlinear operator may also be described by a matrix.

Help other customers find the most helpful reviews

Was this review helpful to you?

[Report abuse](#) | [Permalink](#)



1-bit matrix completion

Extreme case of quantization:

$$y = \text{sign}(M_{\Omega})$$

Claim: Recovering M from y is impossible!

$$M = \begin{bmatrix} \lambda & \lambda & \lambda & \lambda \\ \lambda & \lambda & \lambda & \lambda \\ \lambda & \lambda & \lambda & \lambda \\ \lambda & \lambda & \lambda & \lambda \end{bmatrix}$$

No matter how many samples we obtain, all we can learn is whether $\lambda > 0$ or $\lambda < 0$

Is there any hope?

If we consider a noisy version of the problem, recovery becomes feasible!

$$y = \text{sign}(M_{\Omega} + Z_{\Omega})$$

$$M + Z = \begin{bmatrix} \lambda + Z_{1,1} & \lambda + Z_{1,2} & \lambda + Z_{1,3} & \lambda + Z_{1,4} \\ \lambda + Z_{2,1} & \lambda + Z_{2,2} & \lambda + Z_{2,3} & \lambda + Z_{2,4} \\ \lambda + Z_{3,1} & \lambda + Z_{3,2} & \lambda + Z_{3,3} & \lambda + Z_{3,4} \\ \lambda + Z_{4,1} & \lambda + Z_{4,2} & \lambda + Z_{4,3} & \lambda + Z_{4,4} \end{bmatrix}$$

Fraction of positive/negative observations tells us something about λ

Reminiscent of “dithering” and stochastic resonance

Maximum likelihood estimation

Log-likelihood function:

$$F(X) = \sum_{(i,j) \in \Omega_+} \log(f(X_{i,j})) + \sum_{(i,j) \in \Omega_-} \log(1 - f(X_{i,j}))$$



CDF of the pre-quantization noise

$$\underset{X}{\text{maximize}} \quad F(X)$$

$$\text{subject to} \quad \frac{1}{d\alpha} \|X\|_* \leq \sqrt{r}$$

$$\|X\|_\infty \leq \alpha$$

Gaussian model

Theorem

Assume that $\frac{1}{d\alpha} \|M\|_* \leq \sqrt{r}$ and $\|M\|_\infty \leq \alpha$. We observe $y = \text{sign}(M_\Omega + Z_\Omega)$ where Z has iid $\mathcal{N}(0, \sigma^2)$ entries. If Ω is chosen at random with $\mathbb{E}|\Omega| = m > (d_1 + d_2) \log(d_1 + d_2)$, then with high probability

$$\frac{1}{d_1 d_2} \|\widehat{M} - M\|_F^2 \leq C \left(\frac{\alpha}{\sigma} + 1 \right) e^{\alpha^2/2\sigma^2} \sigma \alpha \sqrt{\frac{r(d_1 + d_2)}{m}}$$

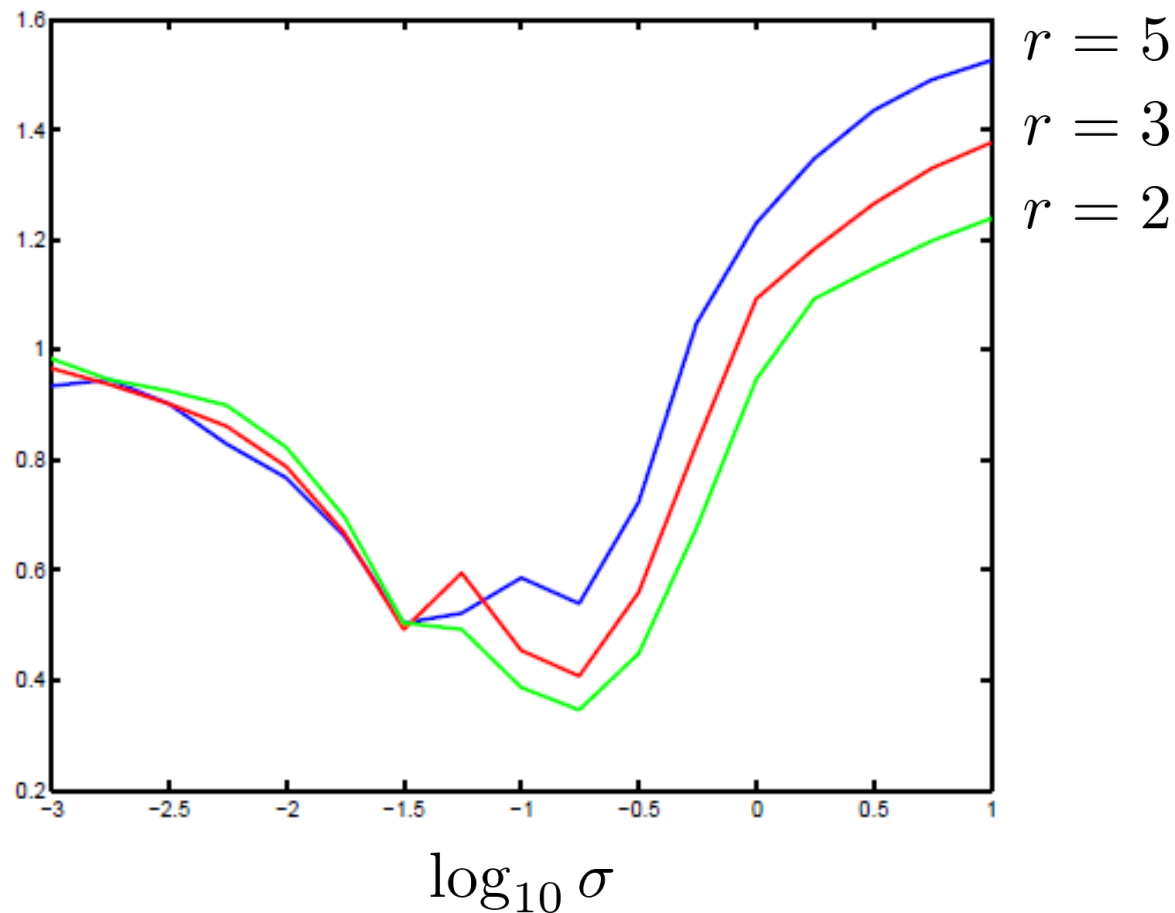
For any fixed α , optimal bound is achieved by $\sigma \approx 1.3\alpha$, in which case the bound reduces to

$$\frac{1}{d_1 d_2} \|\widehat{M} - M\|_F^2 \leq 3.1 C \alpha^2 \sqrt{\frac{r(d_1 + d_2)}{m}}$$

Results in practice

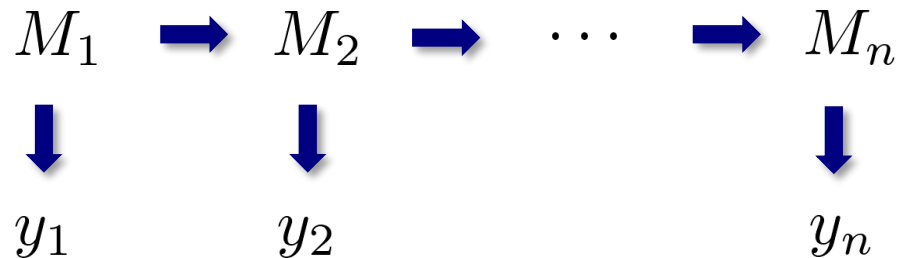
$d = 500$ $m = 0.15d^2$

$$\frac{\|\widehat{M} - M\|_F}{\|M\|_F}$$



Dynamic matrix completion

Suppose that the underlying matrix is changing over time



Goal: Recover M_n

Assumptions: $M_i = UV_i^T$

$$V_i = V_{i-1} + \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, \sigma_1^2)$$

$$y_i = [M_i + z_i]_{\Omega_i} \quad z_i \sim \mathcal{N}(0, \sigma_2^2)$$

Locally weighted matrix smoothing

$$\begin{aligned} & \underset{X}{\text{maximize}} && \frac{1}{2} \sum_{i=1}^n w_i \|y_i - X_{\Omega_i}\|_2^2 \\ & \text{subject to} && \text{rank}(X) \leq r \\ & && \|X\|_{\infty} \leq \alpha \end{aligned}$$

Can give theoretical upper bound that depends on the weights and on $\kappa = \frac{\sigma_1^2}{\sigma_2^2}$

Optimizing this bound gives us a formula for the weights:

$$w_i \propto \frac{1}{1+(n-i)\kappa}$$

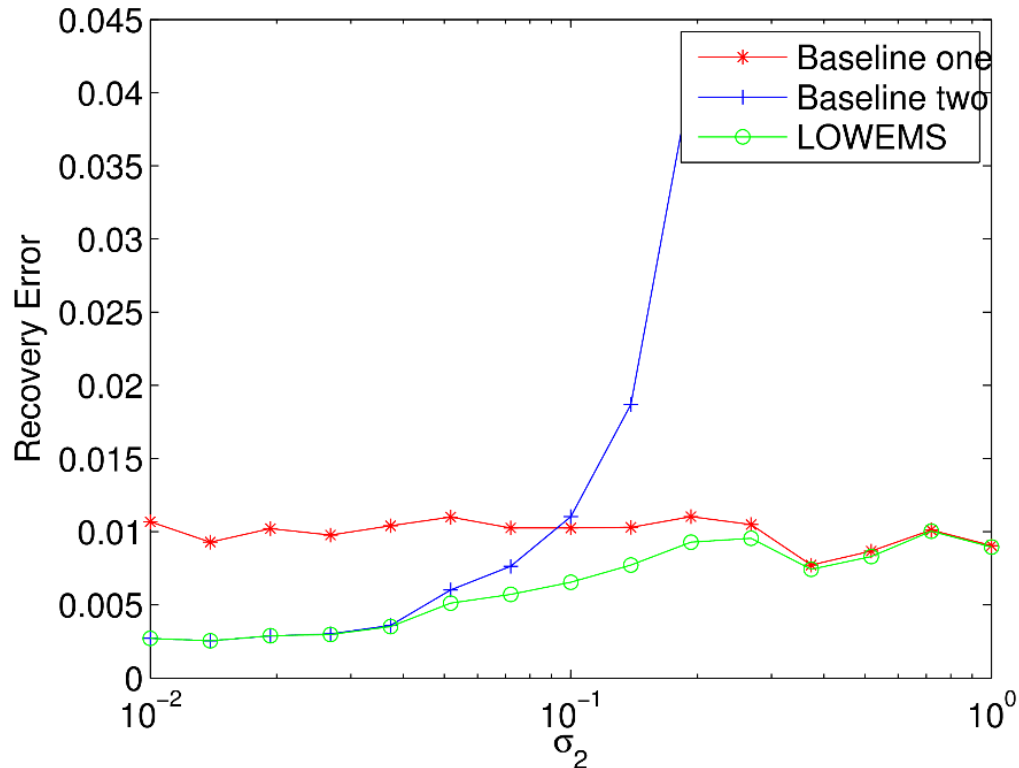
Synthetic simulation

Baseline one: $\kappa = \frac{\sigma_1^2}{\sigma_2^2} \rightarrow \infty$

(Recovery using only y_n)

Baseline two: $\kappa = \frac{\sigma_1^2}{\sigma_2^2} \rightarrow 0$

(Recovery using equal weighting of y_1, \dots, y_n)



Ongoing work

- Promising results on real data
 - Netflix
 - educational data
- Provable guarantees for better/practical algorithms?
- Improved theory for the quantized case?
- More realistic dynamic models?
 - can we learn the dynamics from the data?

Thank You!

For more details and references, see [arxiv:1601.06422](https://arxiv.org/abs/1601.06422)