

# Unsupervised cross-validation

for SVD, NMF,  $k$ -means

Art B. Owen

Patrick O. Perry

Department of Statistics

Stanford University

# Statistics as usual

	Variable 1	...	Variable C
Case 1			
⋮			
Case R			

- 1) Variables are named entities:
  - E.g. pressure, volume, income ...
  - They persist
- 2) Cases are anonymous replicates
  - Sampled IID from some  $F$
  - Of no inherent interest
  - We'd rather just know  $F$

Under statistic as usual ...

... we only care about cases because they show relationships among variables.

# Variables by variables

Rating	Viewer 1	Viewer 2	Viewer 3	...	Viewer C
Movie 1	4	4	1	...	4
Movie 2	5	5	NA	...	NA
Movie 3	3	3	NA	...	2
⋮	⋮	⋮	⋮	⋮	⋮
Movie R	NA	5	3	...	4

Sometimes specific rows and columns are both of persistent interest:

IPs            ×    books            →    purchases

terms           ×    documents        →    counts

candidate      ×    interviewer      →    rating

nodes           ×    more nodes        →    labeled edges

# Triples

	Movie	Viewer	Rating
Case 1	1	1	4
Case 2	1	2	4
Case 3	2	1	5
⋮	⋮	⋮	⋮
Case N	R	C	4

- Now cases are anonymous
- We don't store the NAs
- 2 categorical variables with lots of levels
- Not independent:
  - Cases 1 & 2 share a movie
  - Cases 1 & 3 share a viewer

How should we bootstrap and cross-validate data like this?

Should we resample cases? leave out cases?

# Sample reuse as usual

## Cross validation

- 1) Pairs  $(X_i, Y_i)$  are IID from  $F$
- 2) Leave out some pairs
- 3) Fit  $\hat{Y} = f(X)$  from retained pairs
- 4) Predict held out  $Y$ 's

## Bootstrap

- 1) Data  $X_i$  are IID from  $F$  (unknown)
- 2) Estimate  $F$  by  $\hat{F}$  (known)
- 3) Sample  $X_i^*$  from  $\hat{F}$
- 4)  $X_i$  are to  $F$  as  $X_i^*$  are to  $\hat{F}$

... in a nutshell

# What we want

## Bootstrap

For complicated models, shake the data around, see what is stable and what is not.

## Cross-validation

Leave out some data, predict by the rest.  
Tricky to predict a row if we left it out.

We should get the **same** answer for  $X$  and  $X'$

## The challenge

- 1) Rows are not IID
- 2) Neither are columns
- 3) And the matrix is not replicated

# The answer

	Bootstrap	Cross-validation
By Elements	Wrong	OK (awkward)
By Rows and Cols	OK (approx)	Good

# Outer product models

Data are

$$X_{ij}, \text{ for } i = 1, \dots, m \text{ and } j = 1, \dots, n$$

Usual ANOVA

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

Fisher & MacKenzie, 1923

$$X_{ij} = \mu + \alpha_i + \beta_j + du_i v_j + \varepsilon_{ij}$$

Focus on multiplicative part



$$X_{ij} \doteq du_i v_j$$

Add more factors

$$X_{ij} \doteq \sum_{\ell=1}^k d_{\ell} u_{i\ell} v_{j\ell}$$

# Outer product models

Expressions:

$$X_{ij} \doteq \sum_{\ell=1}^k \sigma_{\ell} u_{i\ell} v_{j\ell}$$

$$X \doteq \sum_{\ell=1}^k \sigma_{\ell} u_{\ell} v_{\ell}' \quad u_{\ell} \in \mathbb{R}^m, v_{\ell} \in \mathbb{R}^n$$

$$X \doteq U \Sigma V'$$

$$X \doteq LR \quad L \in \mathbb{R}^{m \times k}, R \in \mathbb{R}^{k \times n}$$

Examples:

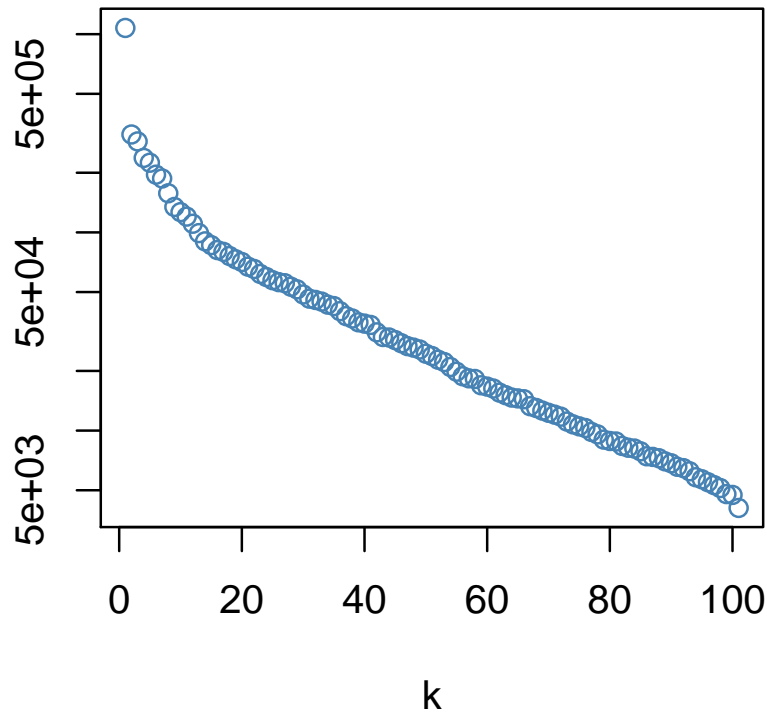
- 1) Factor analysis
- 2) Principal components
- 3) Singular value decomp (SVD)
- 4) Nonnegative matrix factorization (NMF)
- 5) Semi-discrete decomp

Problem: how to pick  $k$ ?

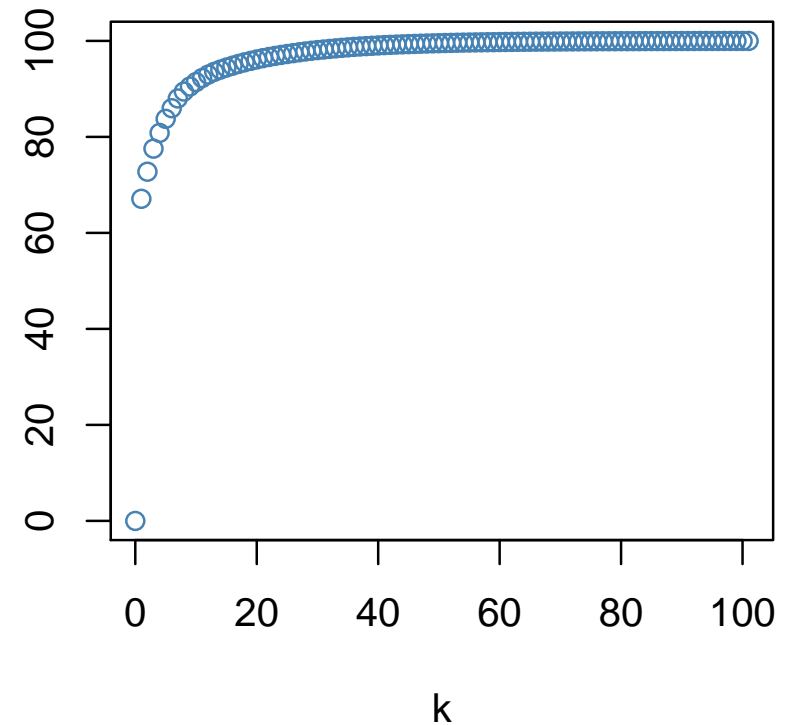
# Picking $k$

Novartis:  $n = 101$  tissues  $m = 12600$  genes, expression

**Singular values Novartis data**



**% Squared error explained**



- Bigger  $k \implies$  smaller error (might overfit)
- Not always an obvious gap

# Classical solution

- 1) Fit SVD with  $k$  and with  $k + 1$
- 2) Find sum squares  $SS(k)$
- 3) Count parameters  $d(k)$
- 4) Prefer  $k + 1$  if  $F = \frac{SS(k) - SS(k+1)}{d(k+1) - d(k)}$  is large

## Problems

- 1) No good way to count df
- 2) ... and it doesn't work dos Dias & Krzanowski 2003
  - i) get 5% vs 66% for  $k = 0$  vs  $k = 1$
  - ii) then too conservative for larger  $k$

# Cross validating the SVD

Hold out one value, say  $X_{11}$

Eastment & Krzanowski 1982

- 1) Do SVD leaving out **row 1** ... take  $v$ 's
- 2) Do SVD leaving out **col 1** ... take  $u$ 's
- 3) Pool the  $\sigma$ 's (geometric mean)
- 4) Reassemble
- 5) Peek at  $X_{11}$  to pick sign  $\pm u_k v'_k$

Still get bigger  $k \implies$  better fit

Besse & Ferré 1993

Wold 1978 holds out scattered elements. We don't investigate that.

Holmes-Junca 1985 compares E-K CV to bootstrap.

Gabriel 2002

- 1) 
$$X = \begin{pmatrix} X_{11} & X_{1\ 2:n} \\ X_{2:m\ 1} & X_{2:m\ 2:n} \end{pmatrix}$$
- 2) Fit  $X_{2:m\ 2:n} \doteq U \Sigma_{1:k} V'$
- 3)  $\hat{X}_{11} = X_{12:n} (V \Sigma_{1:k}^+ U') X_{2:m,1}$
- 4) Seems to work (in crop science)

But **why** does it work?

# Generalize to $r \times s$ blocks

$$X = \left( \begin{array}{c|c} A & B \\ \hline C & D \end{array} \right)$$

- Take  $X \in \mathbb{R}^{m \times n}$
- Leave out  $A \in \mathbb{R}^{r \times s}$
- Fit SVD to  $D \in \mathbb{R}^{m-r \times n-s}$
- Cut at  $k$  terms  $\hat{D}^{(k)}$
- $\hat{A} = B(\hat{D}^{(k)})^+ C$  (pseudo  $BD^{-1}C$ )
- Repeat for  $(m/r) \times (n/s)$  blocks
- Sum squared err  $\|\hat{A} - A\|^2$

# Toy example

$$X = \begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix}$$

$$A - BD^+C = 1 - 26^{-1}3 = 0$$

Separate row and col deleted SVDs fail

$$\begin{aligned} \begin{pmatrix} 2 \\ 6 \end{pmatrix} &= \sqrt{40} \begin{pmatrix} \frac{1}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} \end{pmatrix} \\ \begin{pmatrix} 3 & 6 \end{pmatrix} &= \sqrt{45} \begin{pmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{pmatrix} \\ \sqrt{\sqrt{45}\sqrt{40}} \begin{pmatrix} \frac{1}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{pmatrix} &= \begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix} \times \frac{3\sqrt{2}}{\sqrt{5}} \end{aligned}$$

# Self-consistency lemma O & Perry

Suppose that  $X = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$  has rank  $k$  and **so does**  $D$ . Then

$$A = BD^+C = B(\hat{D}^{(k)})^+C$$

where  $D^+$  is the Moore-Penrose generalized inverse.

This justifies treating  $A - B(\hat{D}^{(k)})^+C$  as a residual from rank  $k$ .

We could also replace  $B, C$  by  $B^{(k)}, C^{(k)}$

(but it doesn't seem to work as well)

## Idea of proof

$$X = \begin{pmatrix} A & B \\ C & D \end{pmatrix} = U\Sigma V' = \begin{pmatrix} U_1\Sigma V_1' & U_1\Sigma V_2' \\ U_2\Sigma V_2' & U_2\Sigma V_2' \end{pmatrix}$$

- It would be immediate if  $D = U_2\Sigma V_2'$  were an SVD. [then  $D^+$  would be  $V_2\Sigma^+U_2'$ ]
- But  $U_2$  and  $V_2$  are not orthogonal matrices.
- We get  $\dots D^+ = V_2(V_2'V_2)^{-1}\Sigma^+(U_2'U_2)^{-1}U_2'$
- Then  $U_1\Sigma V_2' \left( V_2(V_2'V_2)^{-1}\Sigma^+(U_2'U_2)^{-1}U_2' \right) U_2\Sigma V_1' = U_1\Sigma V_1' = A$

# Exceptions: $\text{rank}(D) < \text{rank}(X)$

**Spike**  $X = \left( \begin{array}{c|cccc} 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right)$   $A = 1$  but  $\hat{A} = 0$

**Stripe**  $X = \left( \begin{array}{c|cccc} 1 & 1 & 1 & 1 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right)$   $A = 1$  but  $\hat{A} = 0$

**Arrow**  $X = \left( \begin{array}{c|cccc} 1 & 1 & 1 & 1 \\ \hline 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{array} \right)$   $A = 1$  but  $\hat{A} = 0$

# Upshot

- Any feature with fewer rows or columns than we hold out could be lost
- This could bring robustness versus noise
- Or make us miss sparse features
- Almost the same thing.

Fix (if you want it)

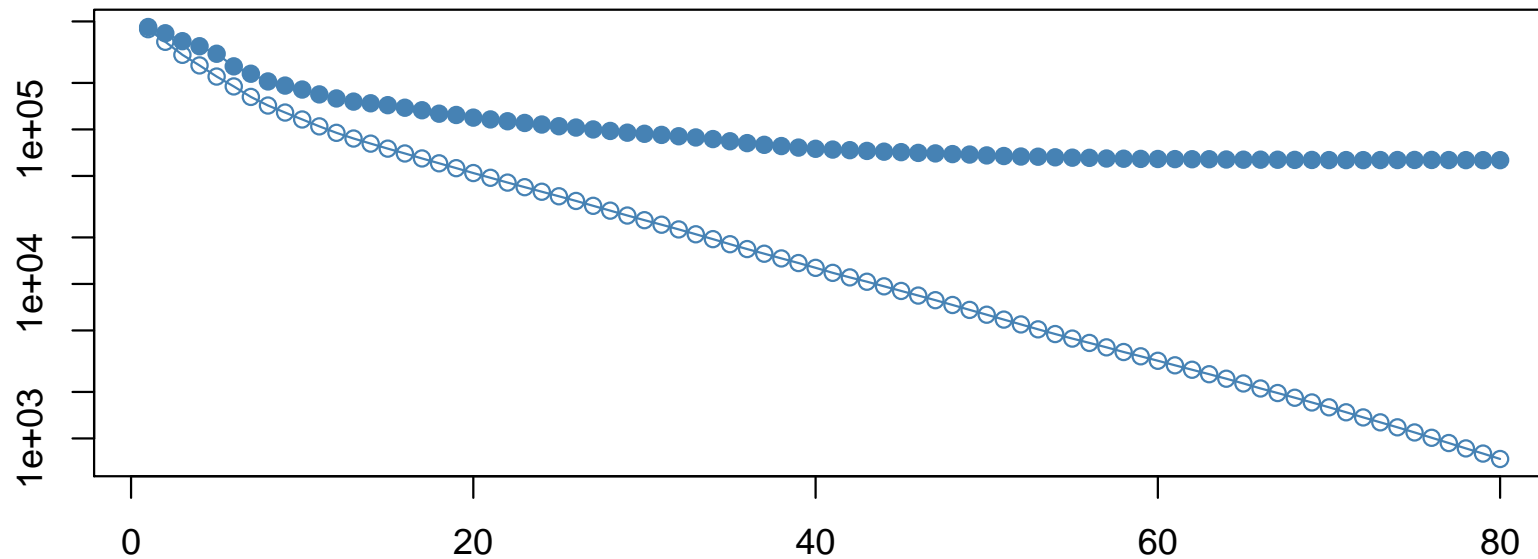
$$\tilde{X} = \mathcal{O}_L X \mathcal{O}_R$$

$$\mathcal{O}_L \in \mathbb{R}^{m \times m} \quad \text{random orthogonal}$$

$$\mathcal{O}_R \in \mathbb{R}^{n \times n} \quad \text{random orthogonal}$$

$X$  and  $\tilde{X}$  have same singular values

# Novartis results



- Solid = cross-validated squared error
- Open = naive squared error

Min is at fairly large  $k$  . . . but only a few percent better than for  $k = 20$

# First steps: rank 1 vs rank 0

$$X = \kappa uv' + Z$$

$$u \in \mathbb{R}^m \quad u'u = 1$$

$$v \in \mathbb{R}^n \quad v'v = 1$$

$$Z \sim \mathcal{N}(0, I_m \otimes I_n)$$

- True rank is 0 or 1
- depending on  $\kappa \geq 0$
- We try ranks  $k = 0$  and  $k = 1$

## Limits

$$n \rightarrow \infty$$

$$m \rightarrow \infty$$

$$m/n \rightarrow c \in (0, \infty)$$

## References

Onatski 2007

Johnstone 2001

Muirhead 1982

# If true rank is 0

Hold out  $r \times s$  matrix  $A$

Fit rank  $k = 0$

$$\begin{aligned} & \mathbb{E} \left( \|A - B(\hat{D}^{(0)})^+ C\|^2 \right) \\ &= \mathbb{E}(\|A\|^2) = rs \end{aligned}$$

Fit rank  $k = 1$

$$\begin{aligned} & \mathbb{E} \left( \|A - B(\hat{D}^{(1)})^+ C\|^2 \right) \\ &= \mathbb{E}(\|A\|^2) + \mathbb{E} \left( \|B(\hat{D}^{(1)})^+ C\|^2 \right) \end{aligned}$$

The true rank has an advantage . . . let's measure it

# The difference

$$\widehat{D}^{(1)} = \sigma_1 uv' \quad (\widehat{D}^{(1)})^+ = \sigma_1^{-1} vu'$$

$\sigma_1$  is largest singular value of  $\mathcal{N}(0, I_{m-r} \otimes I_{n-s})$

Use  $\mathbb{E}(\mathbb{E}(\dots | D))$

$$\mathbb{E}(\|B(\widehat{D}^{(1)})^+ C\|^2 | D) = \sigma_1^{-2} \mathbb{E}(\|Bvu' C\|^2 | D) = rs\sigma_1^{-2}$$

$$\mathbb{E}(\sigma_1^{-2}) \approx (\sqrt{m-r} + \sqrt{n-s})^{-2} \quad \text{Johnstone}$$

Summing over all  $mn/(rs)$  holdouts

$$\frac{1}{mn} \mathbb{E}(\text{CVSS}(0)) = 1 \quad \frac{1}{mn} \mathbb{E}(\text{CVSS}(1)) \approx 1 + \frac{1}{(\sqrt{m-r} + \sqrt{n-s})^2}$$

Larger holdouts  $r/m$  &  $s/n$  protect more against overfitting

# If true rank is 1

$$X = \kappa uv' + Z$$

If we fit  $k = 0$ :

$$\text{then } \mathbb{E}(\text{CVSS}(0)) = \kappa^2 + mn$$

Fixed  $u$  and Gaussian  $v \rightarrow$

spiked covariance of **Johnstone**

General  $u$  &  $v$  more complicated

## Signal strength

$$\kappa^2 \propto mn$$

strong factors

**Bai 2003**

$$\kappa^2 \sim \delta \sqrt{mn} \quad \delta > 1$$

weak factors

**Onatski 2007**

$$\kappa^2 < \sqrt{mn}$$

invisible factors

We work with weak factors (large finite  $\delta$ )

# Partition of $X$

$$X = \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \kappa \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}' + \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{pmatrix}$$

$$D = \kappa u_2 v_2' + Z_{22} = \tilde{\kappa} \frac{u_2 v_2'}{|u_2| |v_2|} + Z_{22} \quad \tilde{\kappa} = \kappa |u_2| |v_2|$$

$$\hat{D}^{(1)} = \hat{\kappa} \hat{u}_2 \hat{v}_2'$$

## Steps

- We watch  $\hat{\kappa}/\tilde{\kappa}$  get close to 1
- Also  $\hat{u}_2' u_2 / |\hat{u}_2|$  and  $\hat{v}_2' v_2 / |v_2|$
- Keeping  $|u_2|^2$  and  $|v_2|^2 > \eta > 0$  so  $\tilde{\kappa}/\kappa$  is near 1

assume we don't hold out most of the signal

easy for Gaussian  $u$  and  $v$

(incoherence)

# Summary

$\mathbb{E}(\text{CVSS}(k)) - mn$	True $k = 0$	True $k = 1$
Fitted $k = 0$	0	$\delta\sqrt{mn}$
Fitted $k = 1$	$\frac{\sqrt{mn}}{\sqrt{c+1}/\sqrt{c+2}}$	$\sqrt{mn}(\delta(1 - \eta^{-1})^2 + \sqrt{c} + 1/\sqrt{c} + 1/\delta)$

## Assumptions for the table

$$r = o(m) \quad s = o(n) \quad \frac{m}{n} \sim c \quad \kappa = \delta\sqrt{mn} \quad \eta \geq 1/2$$

Retain at least  $\eta$  of  $u$  and  $v$  in  $D$

## What counters large hold outs:

They make  $|u_2|$  and  $|v_2|$  small, diminishing  $\delta(1 - \eta^{-1})^2$

# Later theory

Thesis of [Perry \(2009\)](#) works on spiked model

$$X \stackrel{d}{=} \sqrt{n} U D V^T + E$$

$$E \sim \mathcal{N}(0, I \otimes I), \quad \text{where}$$

$$U^T U = V^T V = I_k$$

$$D = \text{diag}(d_1, \dots, d_k)$$

Suppose best rank  $r$  minimizes  $\|\hat{X}^{(r)} - \sqrt{n} U D V^T\|_F^2$

There are two thresholds at work. When  $d_k > \theta_1$  it becomes detectable.

When  $d_k > \theta_2 > \theta_1$  it becomes advantageous.

In between . . . corresponding  $u_k$  and  $v_k$  not well enough determined

Also: Asymptotically optimal holdout is 52% of rows and 52% of columns or any block of  $\approx 0.52^2 mn$  elements.

# Simulation

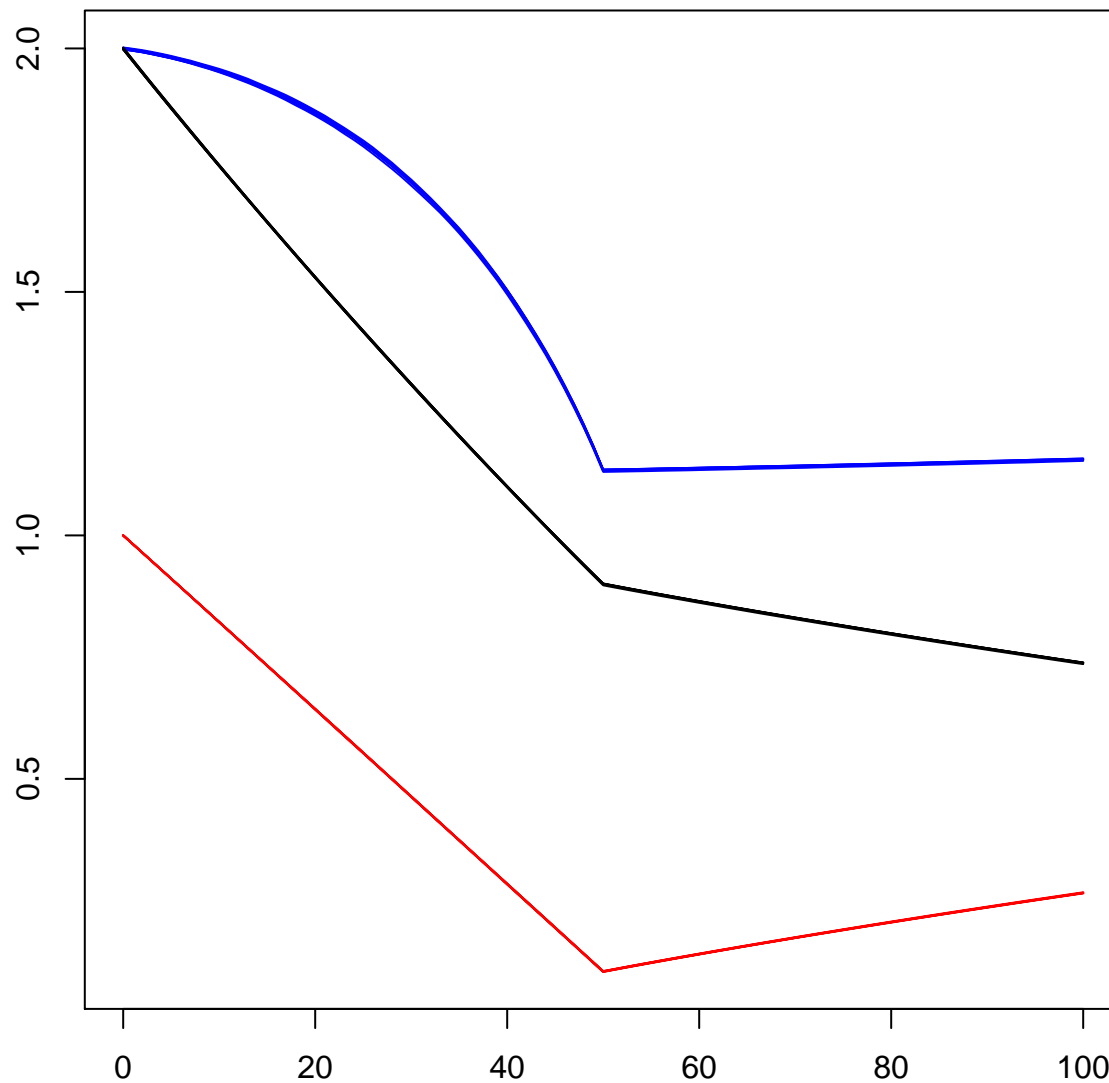
- $X = \mu + Z$   $\mu$  is *known* signal  $Z$  is noise
- Fit  $X \doteq \hat{X}_k$  truncated SVD  $k = 1, \dots, K$
- Estimate  $k$  by  $\hat{k}$  somehow
- Keep score via  $\|\hat{X}_{\hat{k}} - \mu\|^2$

## Details

- 1)  $Z \sim \mathcal{N}(0, I \otimes I)$
- 2)  $\|\mu\|^2 / \mathbb{E}\|Z\|^2 \in \{1, 0.1, 0.01\}$  Hi Med Low signal
- 3)  $\mu$  has singular values  $\propto (1, 1, 1, \dots, 1, 0, 0, \dots)$  **Binary** singular values
- 4) OR  $\mu$  has singular values  $\propto (1, 1/2, 1/4, 1/8, \dots)$  **Geometric** singular values
- 5)  $\mu = U\Sigma V'$  with 'uniform'  $U$  and  $V$  (from a random Gaussian matrix)

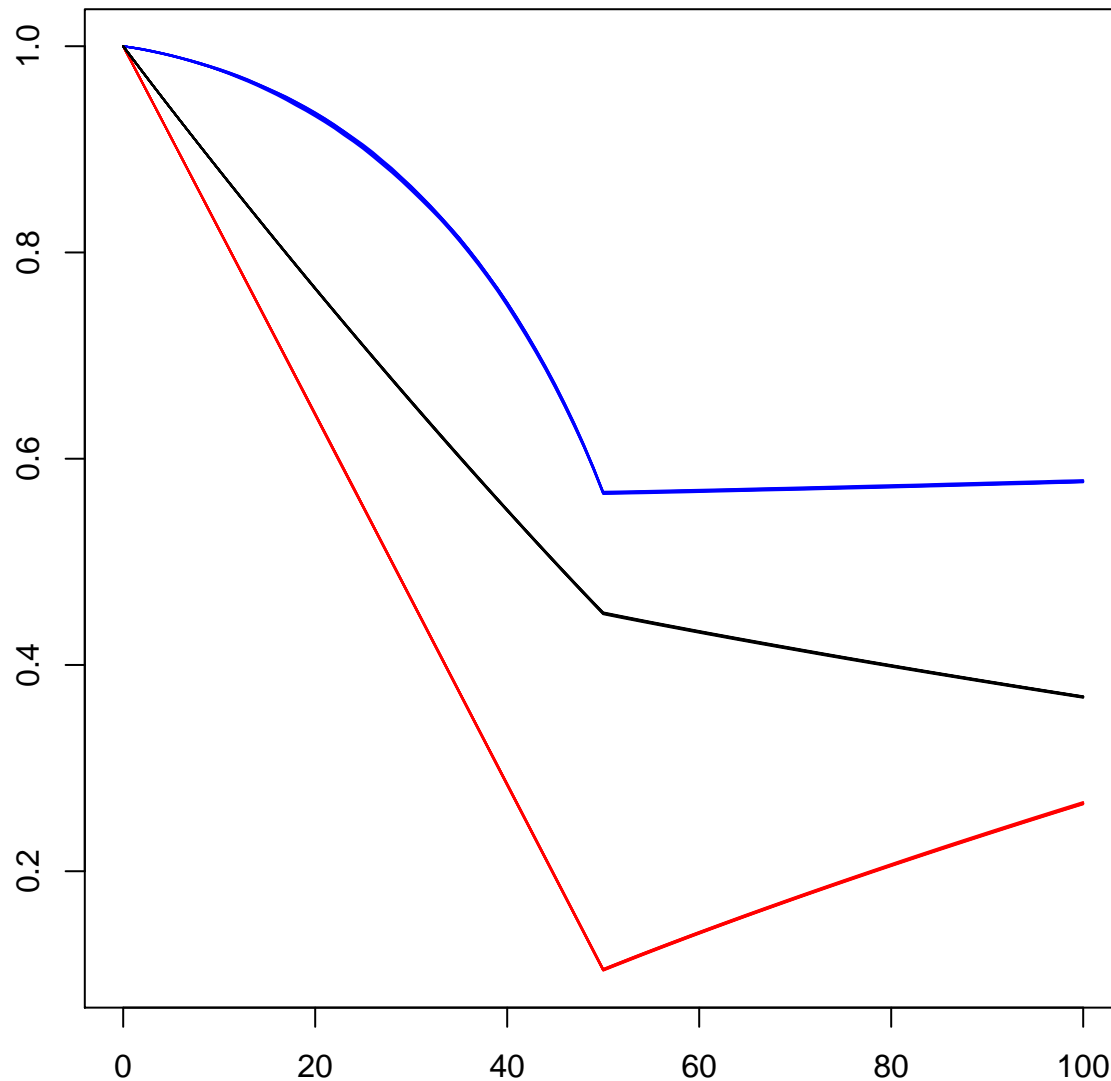
Simulation sizes  $X \in \mathbb{R}^{40 \times 50}$  or  $X \in \mathbb{R}^{1000 \times 1000}$

# Results for large simulation



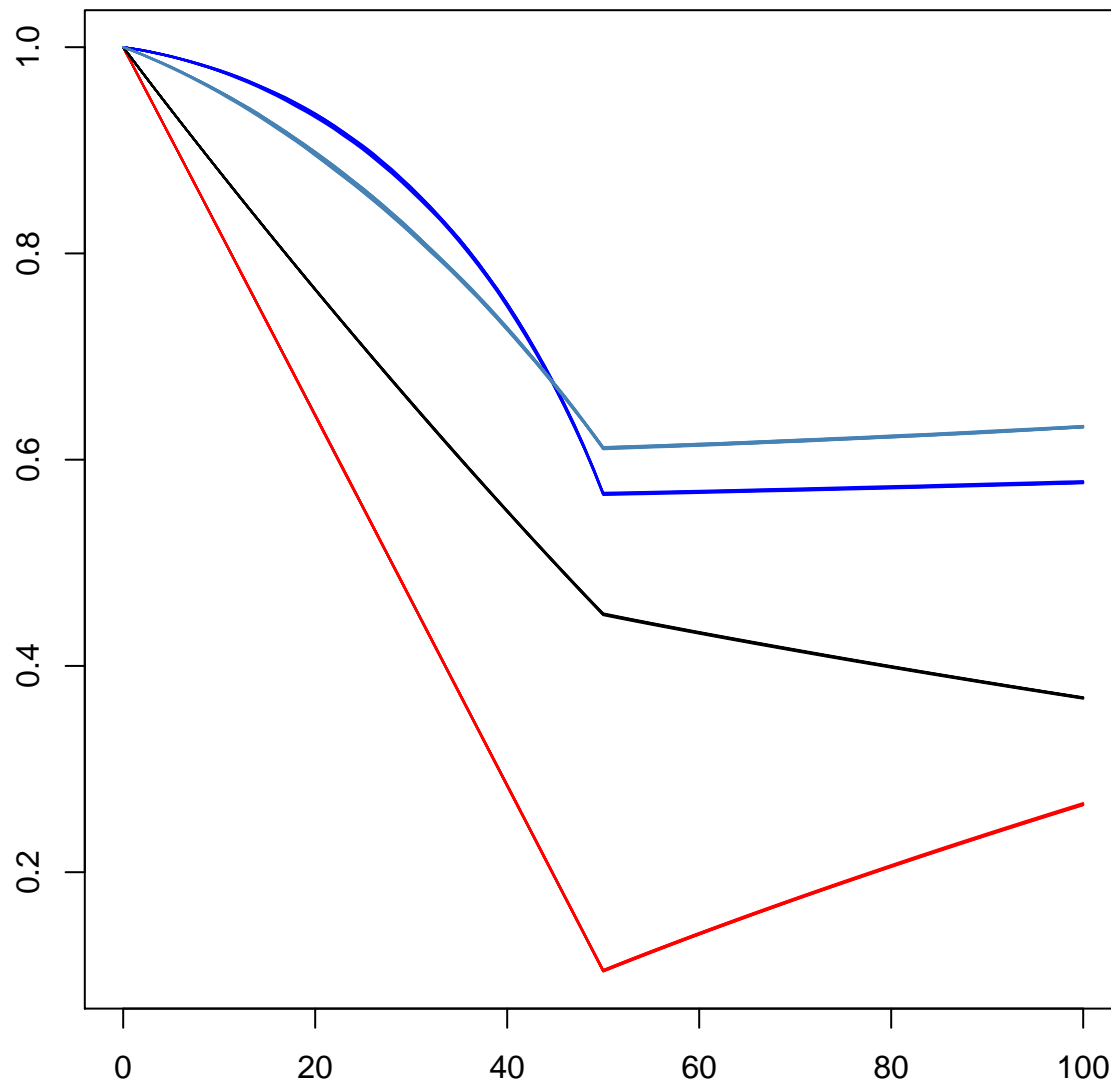
- Binary s.v.: 50 1s and 950 0s
- High signal  $\|\mu\|^2 = E\|Z\|^2$
- $x$ -axis = rank
- $y$ -axis = MSE
- True loss in red
- Naive loss in black
- Bi-CV loss in blue
- Using  $200 \times 200$  holdouts
- 10 replicates are shown

# Results: MSE relative to rank 0



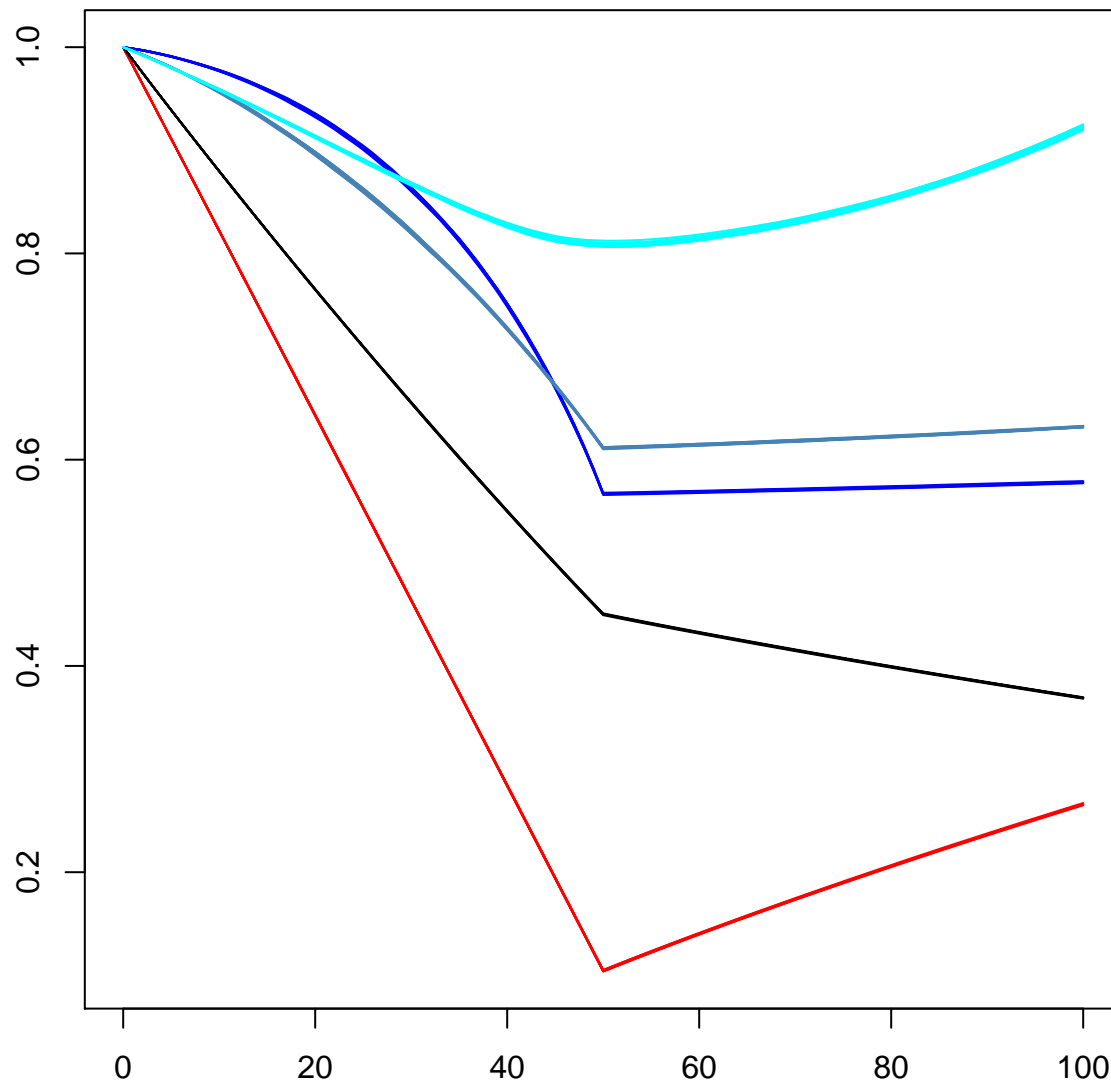
- Binary s.v.: 50 1s and 950 0s
- High signal  $\|\mu\|^2 = E\|Z\|^2$
- $x$ -axis = rank
- $y$ -axis = MSE / MSE at rank 0
- True loss in red
- Naive loss in black
- Bi-CV loss in blue
- Using  $200 \times 200$  holdouts
- 10 replicates are shown

# Relative MSE, hold out 1/5 and 1/2



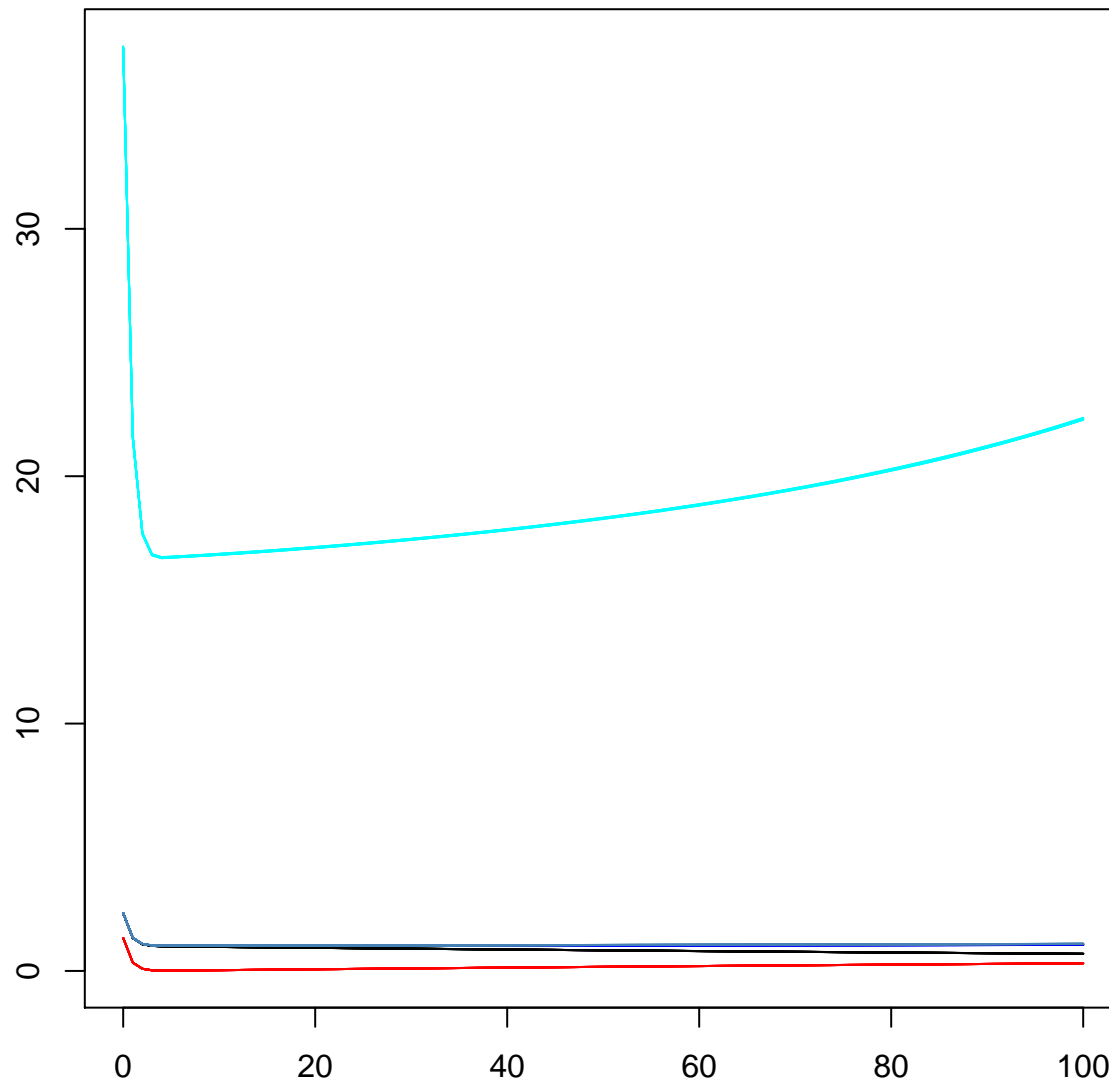
- Binary s.v.: 50 1s and 950 0s
- High signal  $\|\mu\|^2 = E\|Z\|^2$
- $x$ -axis = rank
- $y$ -axis = MSE / MSE at rank 0
- True loss in red
- Naive loss in black
- Bi-CV loss in blue
- Using  $200 \times 200$  holdouts
- 10 replicates are shown
- $500 \times 500$  holdouts in pale blue

# Relative MSE, hold out 1/5, 1/2, 4/5



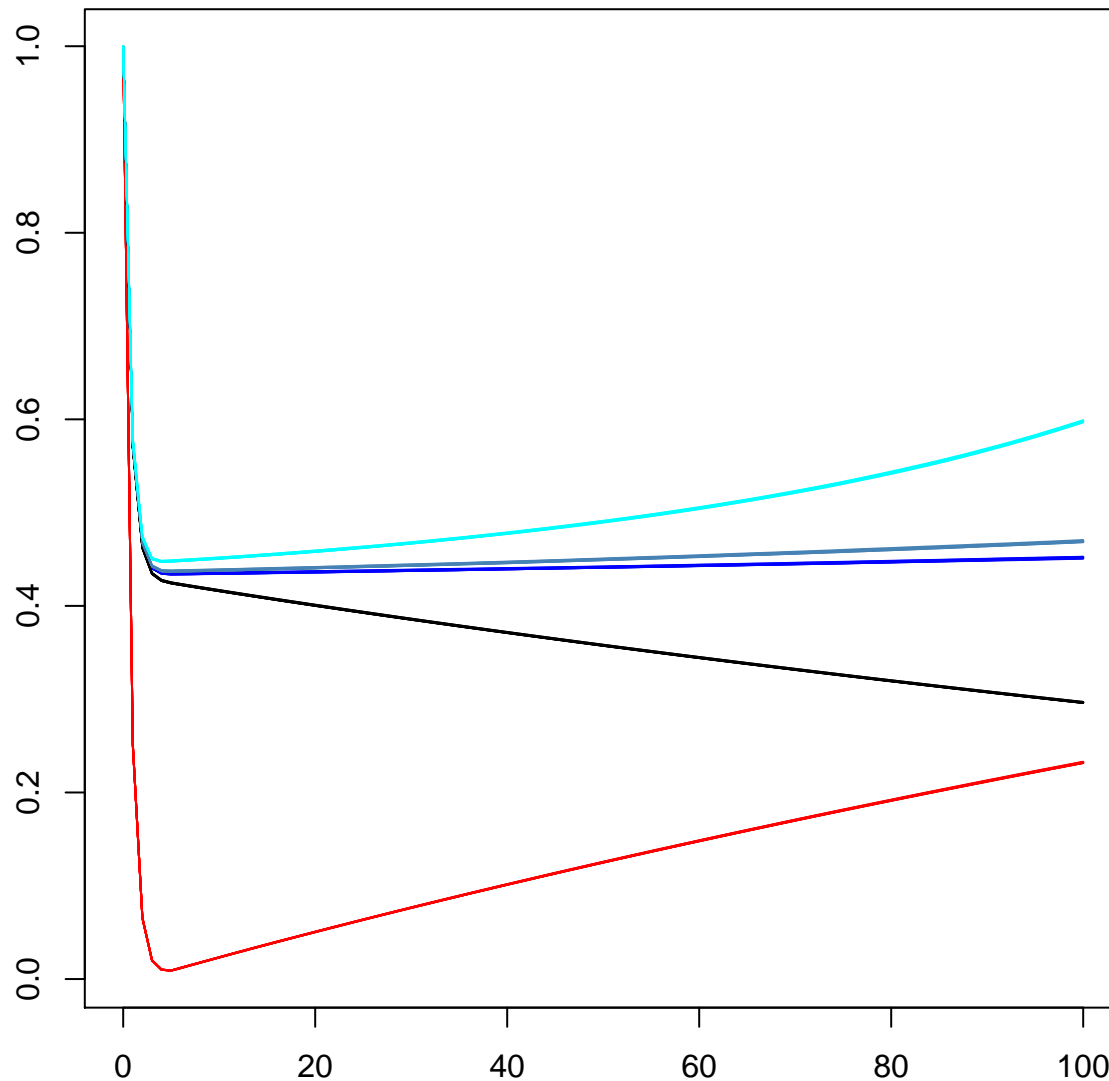
- Binary s.v.: 50 1s and 950 0s
- High signal  $\|\mu\|^2 = E\|Z\|^2$
- $x$ -axis = rank
- $y$ -axis = MSE / MSE at rank 0
- True loss in red
- Naive loss in black
- $200 \times 200$  holdouts in blue
- $500 \times 500$  holdouts in pale blue
- $800 \times 800$  holdouts in cyan
- 10 replicates are shown

# Geometric singular values



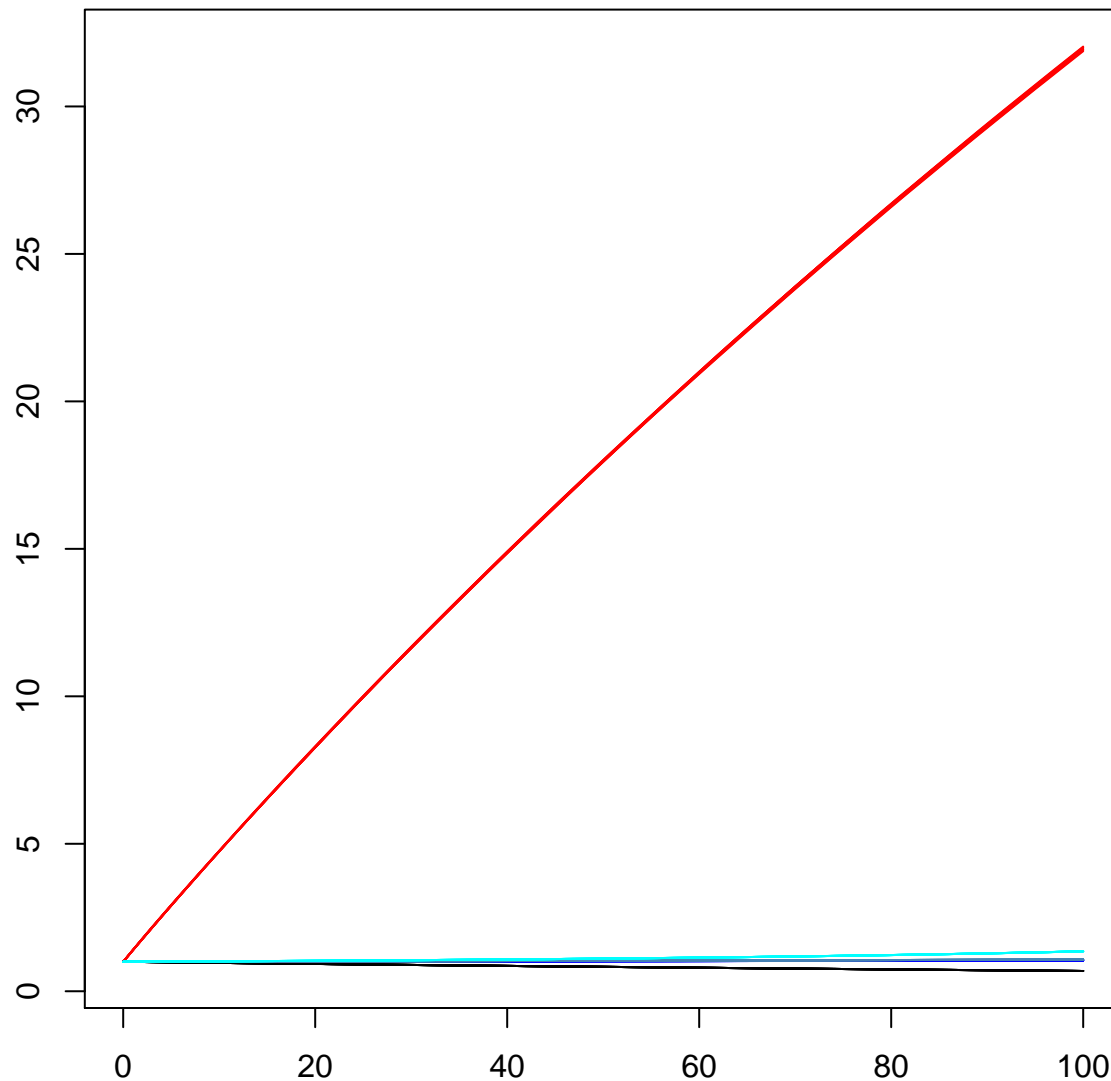
- S.v.  $\propto (1, 1/2, 1/4, 1/8, \dots)$
- High signal  $\|\mu\|^2 = E\|Z\|^2$
- $x$ -axis = rank
- $y$ -axis = MSE (not relative)
- True loss in red
- Naive loss in black
- $200 \times 200$  holdouts in blue
- $500 \times 500$  holdouts in pale blue
- $800 \times 800$  holdouts in cyan
- 10 replicates are shown

# Geometric sv, scaled mse



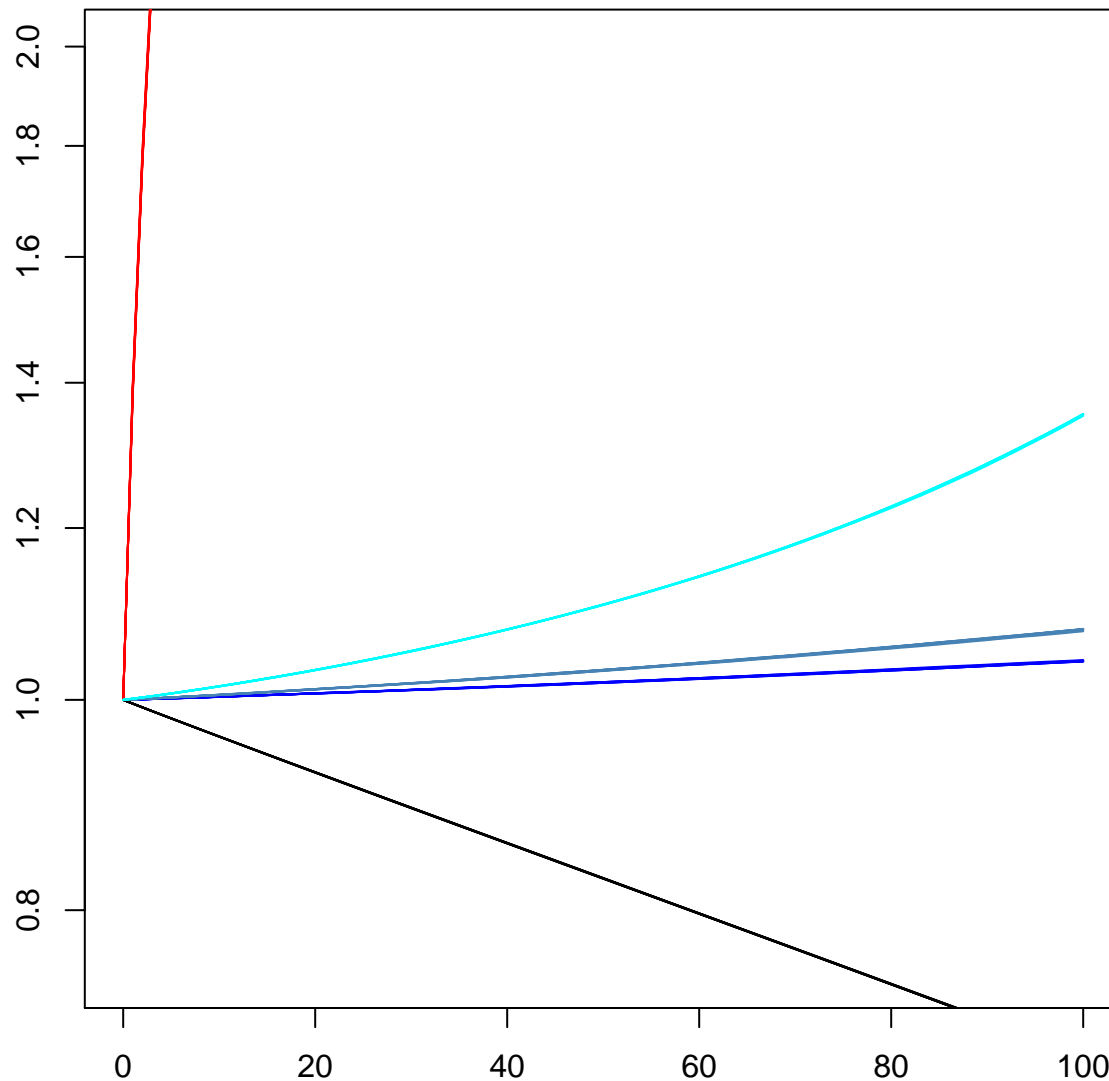
- S.v.  $\propto (1, 1/2, 1/4, 1/8, \dots)$
- High signal  $\|\mu\|^2 = E\|Z\|^2$
- $x$ -axis = rank
- $y$ -axis = MSE / MSE at rank 0
- True loss in red
- Naive loss in black
- $200 \times 200$  holdouts in blue
- $500 \times 500$  holdouts in pale blue
- $800 \times 800$  holdouts in cyan
- 10 replicates are shown

# Binary sv low signal



- 50 nonzero sv's, 950 are 0
- Low signal  $\|\mu\|^2 = 0.01E\|Z\|^2$
- $x$ -axis = rank
- $y$ -axis = MSE / MSE at rank 0
- True loss in red
- Naive loss in black
- $200 \times 200$  holdouts in blue
- $500 \times 500$  holdouts in pale blue
- $800 \times 800$  holdouts in cyan
- 10 replicates are shown

# Binary sv low signal . . . zoom



- 50 nonzero sv's, 950 are 0
- Low signal  $\|\mu\|^2 = 0.01E\|Z\|^2$
- $x$ -axis = rank
- $y$ -axis = MSE / MSE at rank 0
- True loss in red
- Naive loss in black
- $200 \times 200$  holdouts in blue
- $500 \times 500$  holdouts in pale blue
- $800 \times 800$  holdouts in cyan
- 10 replicates are shown

# Regret estimates

$$k^* = \arg \min_{0 \leq k \leq K} \|\hat{X}^{(k)} - \mu\|^2 \quad \text{Best } k$$

$$\hat{k} = \arg \min_{0 \leq k \leq K} \text{CV-MSE}(k) \quad \text{Est. } k$$

$$\text{Regret ratio} = \frac{\|\hat{X}^{(\hat{k})} - \mu\|^2}{\|\hat{X}^{(k^*)} - \mu\|^2}$$

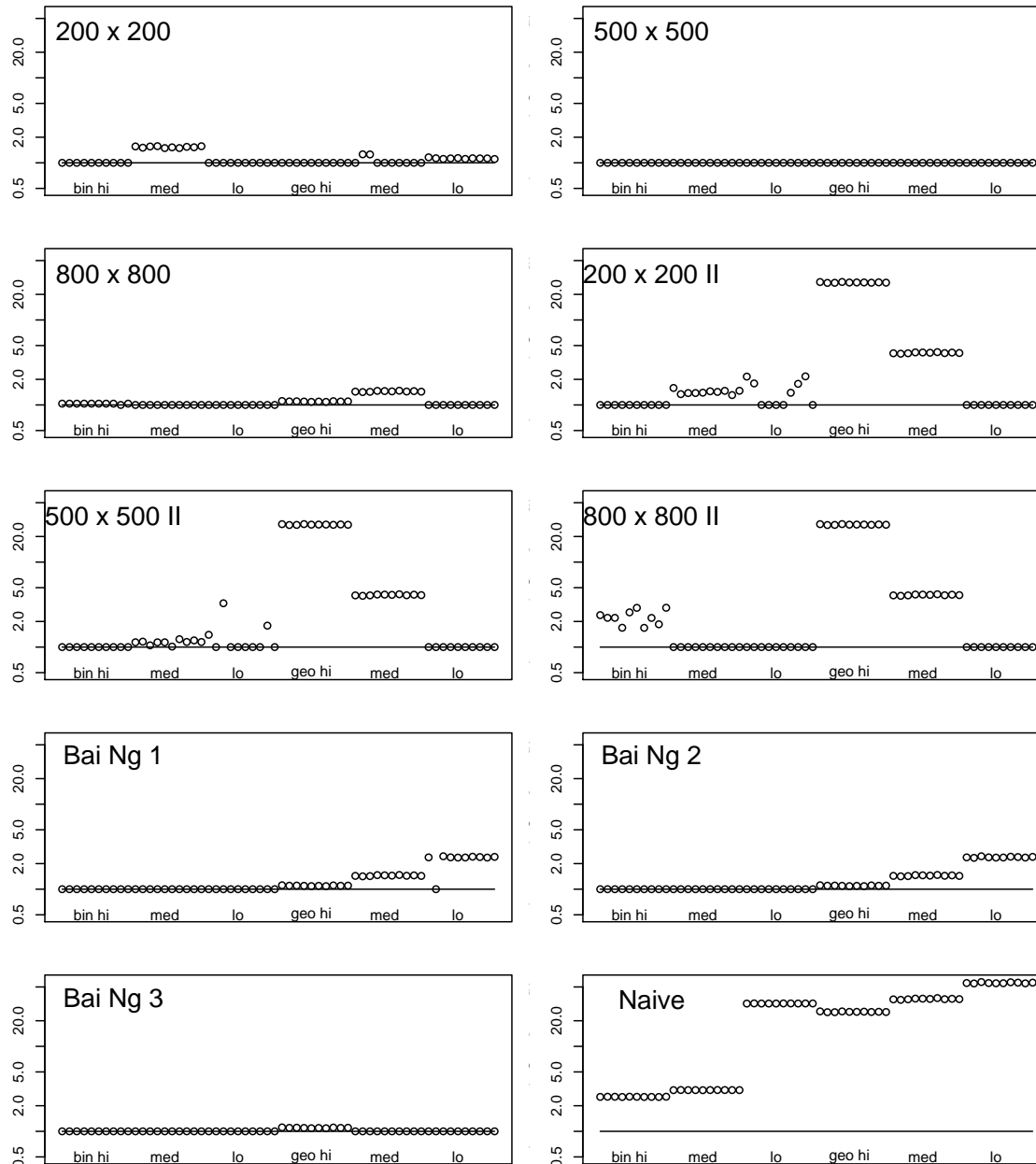
Bai and Ng: BIC style penalties

For  $X \in \mathbb{R}^{m \times n}$ , let  $c = c_{m,n} = \sqrt{\min(m, n)}$

- 1)  $\log(\|\hat{X}^{(k)} - X\|^2) + k \frac{m+n}{mn} \log \frac{mn}{m+n}$
- 2)  $\log(\|\hat{X}^{(k)} - X\|^2) + k \frac{m+n}{mn} \log c^2$
- 3)  $\log(\|\hat{X}^{(k)} - X\|^2) + k \frac{m+n}{mn} \frac{\log c^2}{c^2}$

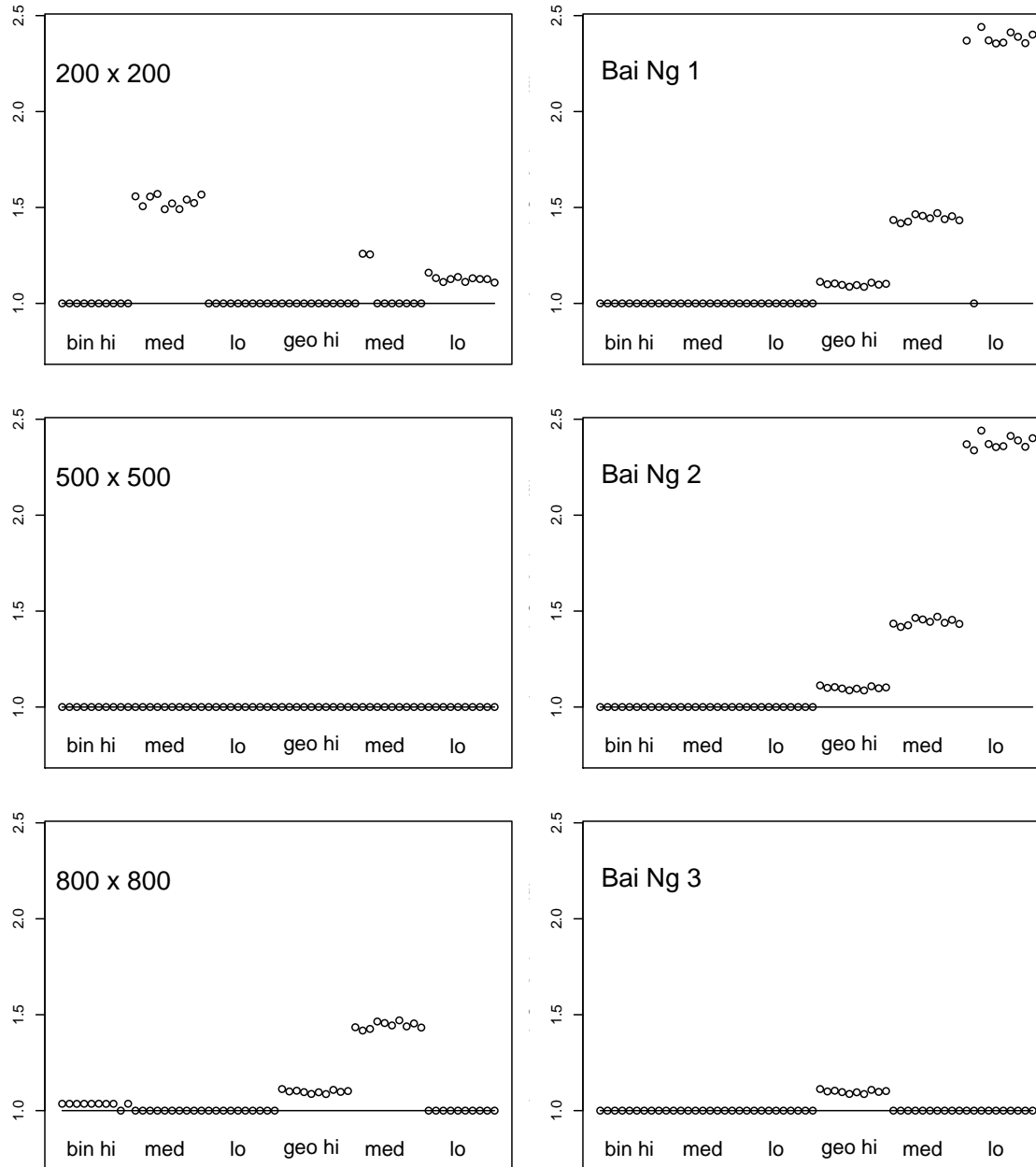
Their goal: estimate correct  $k$

# Regret ratios



- hold out 200x200 500x500 800x800
- residual style *I* or *II*
- $I: A - B(\hat{D}^{(k)})^+ C$
- $II: A - \hat{B}^{(k)} (\hat{D}^{(k)})^+ \hat{C}^{(k)}$
- Bai & Ng 1,2,3
- Naive estimate
- Hold out 1/2 by 1/2 matched oracle
- Bai & Ng matched on binary case
- Style *II* works badly

# Regret ratios, zoom



- hold out 200x200 500x500 800x800
- residual style I only
- Bai & Ng 1,2,3

# Cross validating the NMF

Lee & Seung (1999)

$$X \doteq WH$$

$$W \in [0, \infty)^{m \times k}$$

$$H \in [0, \infty)^{k \times n}$$

- Non-negative entries.
- Often sparse.
- Interpretation as parts.

Let  $D \doteq W_D H_D$

Then  $\hat{A} = B(W_D H_D)^+ C$

We don't always get

$$(YZ)^+ = Z^+ Y^+.$$

But by **MacDuffee's** theorem

$$(W_D H_D)^+ = H_D^+ W_D^+$$

when both have rank  $k$

# NMF criterion

- $(W_D H_D)^+ = H_D^+ W_D^+$
- but  $H_D^+$  and  $W_D^+$  may have negative elements
- we can reimpose nonnegativity (see article)

Different methods for different outer product models.

EG  $k$ -means factors differently on left and right.

# NMF Example

## Original Data:

Cornell classic 3 corpus

3893 abstracts from 3 sources

medical, info retrieval, aeronautical

193,848 total words

4463 distinct word stems

$X_{ij} = \#(\text{word } i \text{ in doc } j)$

$X_{ij}$  small integers

## Synthetic Data:

Fit rank 1 NMF to each journal

Merge them into  $WH$

$W \in [0, \infty)^{m \times 3}$   $H \in [0, \infty)^{3 \times n}$

Sample  $X \sim \text{Poisson}(\epsilon + WH)$

$\epsilon = \text{average value of } WH$

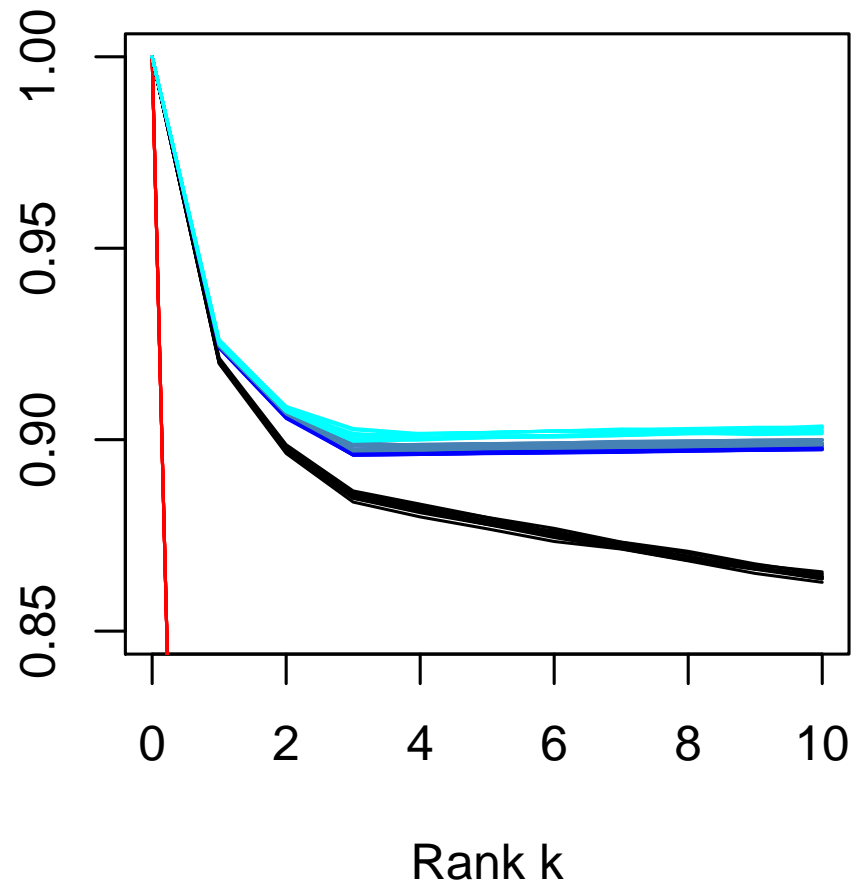
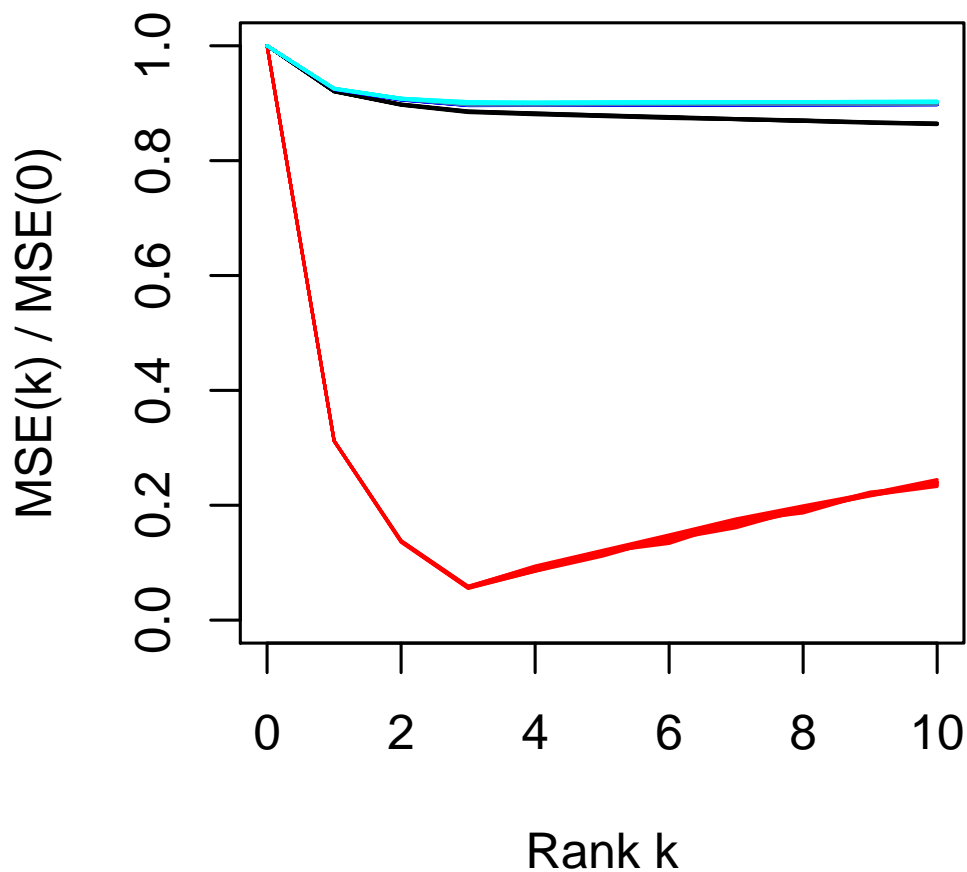
signal/noise  $\approx 1$

Best number of terms should be 3 by construction

or 4 ... since  $\text{rank}(\epsilon + WH) = 4$

# True and estimated loss

## BCV relative errors versus rank



True loss

Naive loss

2x2, 3x3, 5x5 BCV loss

10 repeats

2x2 and 3x3 got rank 3 10 times

5x5 got rank 4

# Other outer product models

Semidiscrete decomposition  $X \doteq U\Sigma V'$   $U \in \{-1, 0, 1\}^{m \times k}$   $V \in \{-1, 0, 1\}^{k \times n}$

$k$  means  $X \doteq U\Sigma V'$   $U \in \{0, 1\}^{m \times k}$   $V \in \mathbb{R}^{k \times n}$

Plaid  $X_{ij} = \sum_{\ell} \rho_{i\ell} \kappa_{j\ell} (\mu_{\ell} + \alpha_{i\ell} + \beta_{j\ell})$   $\rho_{i\ell}, \kappa_{j\ell} \in \{0, 1\}$

also can be cross-validated

(it remains to see how well)

# Cross validation conclusion

- 1) We can leave out blocks  $r \times s$
- 2) Leaving out  $1 \times 1$  may be slow but is not incorrect.  
It might give  $\hat{k}$  too high  
Leaving out scattered point sets should be ok.  
But it requires missing data methods.

# Acknowledgments

- 1) Joint work with [Patrick O. Perry](#), Stanford Statistics → Harvard E.E. (9/09).
- 2) Thanks to [Alexei Onatski](#) for sharing pre-publication versions of his random matrix work.
- 3) Thanks to [Statistica Sinica](#) for sponsoring the session.
- 4) And the organizers [Peter Hall](#), [Jing-Shiang Hwang](#), [Kung-Yee Liang](#).
- 5) Supported by DMS-0604939 from the [U.S. NSF](#)