# Bi-cross-validation for factor analysis

Art B. Owen, Stanford University

and

Jingshu Wang, Stanford University

# Summary

- Factor analysis is over $100$ years old Spearman (1904)
- It is widely used (economics, finance, psychology, engineering $\cdots$)
- It remains problematic to choose the number of factors $k$
- Best methods justified by simulation (e.g., parallel analysis)
- Recent random matrix theory (RMT) applies to homoscedastic case

## Our contribution

1) We hold out some rows and some columns
2) Fit a $k$-factor model to held-in data
3) Predict held-out data to choose $k$
4) Justify via simulations
5) RMT informs the simulation design
6) Proposed method dominates recent proposals

# Factor analysis model

$$Y = X + \Sigma^{1/2}E \in \mathbb{R}^{N\times n}$$

$$X = LR^{\mathsf{T}}, \qquad \text{or } UDV^{\mathsf{T}} \text{ (svd)}$$

$$L \in \mathbb{R}^{N\times k}$$

$$R \in \mathbb{R}^{n\times k}$$

$$E \sim \mathcal{N}(0, I_N \otimes I_n)$$

$$\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_N^2)$$

## For example

$Y$ has expression of $N$ genes for $n$ subjects.

Should not model $n \to \infty$ with $N$ fixed.

# Our criterion

$$Y = X + \Sigma^{1/2} E$$

There is a true $X$. Choosing $k$ yields $\widehat{X} = \widehat{X}_k$, and

**The best $k$ minimizes**

$$\|X - \widehat{X}_k\|_F^2 = \sum_{ij} (\widehat{X}_{k,ij} - X_{ij})^2$$

**True rank**

The true rank of $X$ is not necessarily the best $k$.

The true # factors may not even be a meaningful quantity.

There could be more than $\min(n, N)$ real phenomena.

Maybe $k = \infty$.

# Principal components

Use $y \in \mathbb{R}^N$ for one observation, i.e,. one column of $Y$.

## Principal components

$$\mathrm{Var}(y) = P\Lambda P^\mathsf{T} = \sum_j \lambda_j p_j p_j^\mathsf{T} \qquad \text{truncate to } k \text{ terms}$$

## Factor analysis

$$Y = X + \Sigma E = LR^\mathsf{T} + \Sigma E$$

$$\mathrm{Var}(y) = LL^\mathsf{T} + \Sigma \qquad \text{if rows of } R \sim \mathcal{N}(0, I_k)$$

Sometimes the principal components solution is used for factor analysis or as a starting point.

Sample $\widehat{\mathrm{Var}}(y)$ not necessarily low rank plus diagonal.

# Choosing $k$ in factor analysis

- try several and use 'WOW' criterion* Johnson and Wichern (1988)
- plot sorted eigenvalues of $\widehat{\mathrm{Var}}(y)$ and cut at an 'elbow' $\cdots$ scree plot
- standardize so that $\widehat{\mathrm{Var}}(y)$ is a correlation matrix. Keep any $\lambda_j > 1$ Kaiser (1960)
- hypothesis tests $H_{0,k}$:rank $= k$ vs $H_{A,k}$:rank $> k$
- parallel analysis Horn (1965) and Buja & Eyoboglu (1992) more later

## Estimating $L$ and $R$ or just $X = LR^{\mathsf{T}}$

This is tricky even for fixed $k$.

We want to recover $X$ and the likelihood from $Y = X + \Sigma E$ is unbounded. (as $\sigma_i \to 0$)

Other approaches fit $\widehat{\mathrm{Var}}(y) \doteq$ rank $k +$ diagonal

*as in "Wow, I understand these factors."

# Classical asymptotics

Fix dimension $N$ of $y$.

Let number $n$ of observations $\to \infty$.

Work out asymptotically valid hypothesis tests.

Not convincing when $N$ and $n$ are comparable, or $N \gg n$

## Also

Why approach an estimation problem via testing?

# Random matrix theory

Applies more directly to principal components

Survey by Paul & Aue (2014)

Suppose that $Y \sim \mathcal{N}(0, I_N \otimes I_n)$

All true eigenvalues of $\mathrm{Var}(y)$ are 1. Biggest is biased up.

## Marcenko-Pastur (1967)

Let $n, N \to \infty$ with $N/n \to \gamma \in (0,1)$ $\qquad \widehat{\mathrm{Var}}(y) = n^{-1}YY^{\mathsf{T}}$

$\max \mathrm{eig}\, \widehat{\mathrm{Var}} \to M \equiv (1+\sqrt{\gamma})^2$ & $\min \mathrm{eig}\, \widehat{\mathrm{Var}} \to m \equiv (1-\sqrt{\gamma})^2$

Limiting density is

$$f_\gamma(\lambda) = \frac{\sqrt{(M-\lambda)(\lambda-m)}}{2\pi\gamma\lambda}, \qquad m \leqslant \lambda \leqslant M$$

## For $\gamma > 1$

Get a lump at $0$ and the $f_{1/\gamma}$ density

# Tracy-Widom

Marcenko-Pastur describes the 'bulk spectrum'.

Tracy & Widom (1994, 1996) find the distribution of max eigenvalue for (related) Wigner matrix.

Johnstone (2001) finds distn of $\max \mathrm{eig}\,(n\widehat{\mathrm{Var}})$

Again $N/n \to \gamma \in (0, 1]$

## The limit

$$\frac{\lambda_1 - \mu_{n,N}}{\sigma_{n,N}} \xrightarrow{\mathrm{d}} W_1 \qquad \text{Tracy-Widom density}$$

$$\mu_{n,N} = (\sqrt{n-1} + \sqrt{N})^2 \qquad \text{center}$$

$$\sigma_{n,N} = (\sqrt{n-1} + \sqrt{N})\left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{N}}\right)^{1/3} \qquad \text{scale}$$

Then $\sigma_{n,N}/\mu_{n,N} = O(n^{-2/3})$

# Spiked covariance

$$\mathrm{Var}(y) = \sum_{j=1}^{k} d_j u_j u_j^{\mathsf{T}} + \sigma^2 I_N, \qquad u_j \text{ orthonormal}$$

$$d_1 \geqslant d_2 \geqslant \cdots \geqslant d_k > 0, \qquad k \text{ fixed}$$

## Theorem for $\sigma^2 = 1, \quad N/n \to \gamma$

$\mathrm{Var}(y) \in \mathbb{R}^{N\times N}$ has eigenvalues

$$\ell_1 \geqslant \ell_2 \geqslant \ldots \geqslant \ell_k > 1 = \ell_{k+1} = \cdots = \ell_{\min(n,N)}$$

$\widehat{\mathrm{Var}}(y)$ has eigenvalues $\hat{\ell}_j$. Then

$$\hat{\ell}_j \xrightarrow{\text{a.s.}} \begin{cases} (1+\sqrt{\gamma})^2, & \ell_j \leqslant 1+\sqrt{\gamma} \\ \ell_j\Big(1+\frac{\gamma}{\ell_j - 1}\Big), & \text{else.} \end{cases}$$

**Consequence** $d_j \leqslant \sqrt{\gamma}$ undetectable (via svd)

Don't need Gaussianity.

# Parallel analysis

The biggest eigenvalue is biased up.

Horn (1965) simulates Gaussian data with no factors.

He tracks the largest singular values

Takes $k \geqslant 1$ if observed $d_1$ attains a high quantile

Goes on to the second largest . . .

## Permutations

Buja & Eyoboglu (1992) adapt to non Gaussian data.

They permute all $N$ of the variables

Uniform over $n!$ permutations of each.

## Upshot

They're wise to the bias. Horn's article came out **before** Marcenko-Pastur.

Tricky consequences from using for example $95$'th quantile:

One huge factor might make it hard to detect some other real ones.

Also $\cdots$ why $0.95$?

# Bi-cross-validation

$$Y = X + E.$$

Seek $k$ so that $\widehat{Y}_k$ (truncated SVD) is close to underlying $X$

Derived for $E$ with all variances equal.

We will adapt it to $\Sigma^{1/2}E$.

## Literature

O & Perry (2009) Ann. Stat.

Perry (2009) Thesis

# Bi-cross-validation

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

- Take $Y \in \mathbb{R}^{m \times n}$
- Hold 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 $\|\hat{A} - A\|^2$

Randomize the order of the rows and columns.

# Toy example

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

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

# Self-consistency lemma

O & Perry (2009)

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

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

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

This justifies treating $A - B(\widehat{D}^{(k)})^{+}C$ 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)

Truncated SVD, also non-negative matrix factorization

(unpublished: $k$-means)

# Perry's 2009 thesis

Recall the $\sqrt{\gamma}$ spike detection threshold.

If $d_j u_j u_j$ is barely detectable then $\hat{d}_j \hat{u}_j \hat{u}_j^\mathsf{T}$ is a bad estimate of $d_j u_j u_j^\mathsf{T}$

It makes $\widehat{X}$ worse (better to use $j-1$ terms).

## The second threshold

Perry found a second threshold. Above that, including $\hat{d}_j \hat{u}_j \hat{u}_j^\mathsf{T}$ helps.

Also found by Gavish & Donoho (2014)

$$\frac{1+\gamma}{2} + \sqrt{\Bigl(\frac{1+\gamma}{2}\Bigr)^2 + 3\gamma}$$

# Optimal holdout size

$Y = \sqrt{n} U D V^{\mathsf{T}} + \sigma E$ Fixed $D = \mathrm{diag}(d_1, \dots, d_{k_0})$.

$N/n \to \gamma \in (0, \infty)$ $U^{\mathsf{T}} U = V^{\mathsf{T}} V = I_{k_0}$

Hold out $n_0$ of $n$ columns and $N_0$ of $N$ rows

Oracle picks $k_*$ BCV gives $\widehat{\mathrm{PE}}_k$ we take $\hat{k} = \arg\min_k \widehat{\mathrm{PE}}_k$

## Result

Let

$$\bar{\gamma} = \Bigl(\frac{\gamma^{1/2} + \gamma^{-1/2}}{2}\Bigr)^2 \quad \text{and} \quad \rho = \frac{n - n_0}{n} \times \frac{N - N_0}{N}$$

Then $k_* - \arg\min_k E(\widehat{\mathrm{PE}}_k) \to 0$ if

$$\sqrt{\rho} = \frac{\sqrt{2}}{\sqrt{\bar{\gamma}} + \sqrt{\bar{\gamma} + 3}}$$

For example: for $n = N$ hold out about 53% of the rows and cols

NB: held out aspect ratio does **not have to** match data's

Perry (2009)

# Unequal variances

$$Y = X + \Sigma^{1/2}E = LR^{\mathsf{T}} + \Sigma^{1/2}E$$

$$\begin{aligned}
\begin{pmatrix} Y_{00} & Y_{01} \\ Y_{10} & Y_{11} \end{pmatrix} &= \begin{pmatrix} L_0 \\ L_1 \end{pmatrix} \begin{pmatrix} R_0 \\ R_1 \end{pmatrix}^{\mathsf{T}} + \begin{pmatrix} \Sigma_0^{1/2} & 0 \\ 0 & \Sigma_1^{1/2} \end{pmatrix} \begin{pmatrix} E_{00} & E_{01} \\ E_{10} & E_{11} \end{pmatrix} \\
&= \begin{pmatrix} L_0 R_0^{\mathsf{T}} & L_0 R_1^{\mathsf{T}} \\ L_1 R_0^{\mathsf{T}} & L_1 R_1^{\mathsf{T}} \end{pmatrix} + \begin{pmatrix} \Sigma_0^{1/2} E_{00} & \Sigma_0^{1/2} E_{01} \\ \Sigma_1^{1/2} E_{10} & \Sigma_1^{1/2} E_{11} \end{pmatrix}
\end{aligned}$$

Held-in data (lower right corner)

$$Y_{11} = X_{11} + \Sigma_{11}^{1/2} E_{11} = L_1 R_1^{\mathsf{T}} + \Sigma_{11}^{1/2} E_{11}$$

# Hold-out estimation

From held in data

$$Y_{11} = L_1 R_1^\mathsf{T} + \Sigma_{11}^{1/2} E_{11}$$

we get $\hat{L}_1$, $\hat{R}_1$, and $\hat{\Sigma}_1$ (how to do that later). But we need $\hat{X}_{00} = \hat{L}_0 \hat{R}_0^\mathsf{T}$

## Least squares to get $\hat{L}_0$

$$Y_{01}^\mathsf{T} \doteq \hat{R}_1 \hat{L}_0^\mathsf{T} + E_{01}^\mathsf{T} \Sigma_0^{1/2}$$

$$\hat{L}_0 = Y_{01} \hat{R}_1 (\hat{R}_1^\mathsf{T} \hat{R}_1)^{-1}$$

## Weighted least squares to get $\hat{R}_0$

$$Y_{10} \doteq \hat{L}_1 \hat{R}_0^\mathsf{T} + \Sigma_1^{1/2} E_{10}$$

$$\hat{R}_0^\mathsf{T} = (\hat{L}_1^\mathsf{T} \hat{\Sigma}_1^{-1} \hat{L}_1)^{-1} \hat{L}_1^\mathsf{T} \hat{\Sigma}_1^{-1} Y_{10}$$

## Held-out estimate

$$\hat{X}_{00} = \hat{L}_0 \hat{R}_0^\mathsf{T}$$

$\hat{X}_{00}$ is unique even though factorization $\hat{X}_{11} = \hat{L}_1 \hat{R}_1^\mathsf{T}$ is not.

# Solving the hold-in problem

$$Y = LR^{\mathsf{T}} + \Sigma^{1/2}E, \quad L \in \mathbb{R}^{N\times k}, R \in \mathbb{R}^{n\times k}$$

Likelihood

$$\log \mathcal{L}(X,\Sigma) = -\frac{Nn}{2}\log(2\pi) - \frac{n}{2}\log\det\Sigma + \mathrm{tr}\Bigl[-\frac{1}{2}\Sigma^{-1}(Y-X)(Y-X)^{\mathsf{T}}\Bigr]$$

Max over $\Sigma$ given $\hat{X}$

$$\hat{\Sigma} = n^{-1}\mathrm{diag}[(Y-\hat{X})(Y-\hat{X})^{\mathsf{T}}]$$

Max over rank $k$ $X$ given $\hat{\Sigma}$

$$\widetilde{Y} = \hat{\Sigma}^{-1/2}X, \quad \hat{X} = \hat{\Sigma}^{1/2} \times \text{truncated SVD of } \widetilde{Y}$$

Except that

It doesn't work. The likelihood is unbounded. Diverges as some $\sigma_i \downarrow 0$

# Making it work

## Constrain $\sigma_i \geqslant \epsilon > 0$

Then you'll get back some $\hat{\sigma}_i = \epsilon$. So choice of $\epsilon$ is hard.

## Place a prior on $\sigma_i$

SImilar story. Delicate to make the prior strong enough but not too strong. Also: it didn't perform well even using the true prior.

## Place a regularizing penalty on $\Sigma$

Then we would need a cross-validation step in the middle of our cross-validation procedure.

## Regularization by early stopping

Do $m$ iterations and then stop. Extensive simulations show that $m = 3$ is nearly as good as an oracle's $m$. Also it is faster than trying to iterate to convergence.

This seemed to be the biggest problem we faced.

# The algorithm

Use Perry's recommended hold-out fraction

Make the held out matrix nearly square

Predict the held out values

Repeat the cross-validation

Retain the winning $k$

# Extensive simulations

We generated a battery of test cases of different sizes, aspect ratios, and mixtures of factor strengths.

### Factor strengths

$$X = \sqrt{n} U D V^{\mathsf{T}}$$

- Undetectable. $d_j < \sqrt{\gamma}$.
- Harmful. Above the detection threshold but below the second threshold.
- Helpful. Above both thresholds.
- Strong. A factor explains a fixed $\%$ of variance of the data matrix. $d_i \propto \sqrt{n}$.

### Weak factors have fixed $d_i$ as $n \to \infty$

They explain $O(1/n)$ of the total variance.
Strong factors are prevalent, but sometimes less interesting since we may already know about them.

# Methods

Comparing to recent proposals in econometrics.

1) Parallel analysis via permutations Buja & Eyuboglu (1992)

2) Eigenvalue difference method Onatski (2010)
$\hat{k} = \max\{k \leqslant k_{\max} \mid \lambda_k^2 - \lambda_{k+1}^2 \geqslant \delta\}$

3) Eigenvalue ratio method Ahn & Horenstein (2013)
$\hat{k} = \arg\max\{0 \leqslant k \leqslant k_{\max} \mid \lambda_k^2/\lambda_{k+1}^2\}$
$\lambda_0^2 \equiv \sum_{i\geqslant 1} \lambda_i^2/\log(\min(n,N))$

4) Information criterion Bai & Ng (2002)
$\arg\min_k \log(V(k)) + k\left(\frac{n+N}{nN}\right)\log\left(\frac{nN}{n+N}\right)$
$V(k) = \|Y - \widehat{Y}_k\|_F^2/(nN)$

5) Weak factor estimator of Nadakuditi & Edelman (2008)

## Also

The **true** rank, and an **oracle** that always picks the best rank for each data set.

# Noise levels

$$Y = X + \Sigma^{1/2}E \qquad E \sim \mathcal{N}(0, I_N \otimes I_n) \qquad \Sigma = \mathrm{diag}(\sigma_1^2, \dots, \sigma_N^2)$$

## Random inverse gamma variances

$$\mathbb{E}(\sigma_i^2) = 1 \qquad \text{and} \qquad \mathrm{Var}(\sigma_i^2) \in \{0, 1, 10\}$$

Homoscedastic as well as mild and strong heteroscedastic.

# Factor scenarios

| | Scenario | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| # Undetectable | 1 | 1 | 1 | 1 | 1 | 1 |
| # Harmful | 1 | 1 | 1 | 3 | 3 | 6 |
| # Useful | 6 | 4 | 3 | 1 | 3 | 1 |
| # Strong | 0 | 2 | 3 | 3 | 1 | 0 |

True rank is 8

Always one undetectable.

Vary the number of harmful, $1$ to $6$.

Vary the mix useful/strong among the non-harmful.

Some algorithms have trouble with strong factors.

# $10$ sample sizes, $5$ aspect ratios

| N | n | $\gamma$ |
|---|---|---|
| 20 | 1000 | 1/50 |
| 20 | 100 | 1/5 |
| 50 | 50 | 1 |
| 100 | 20 | 5 |
| 1000 | 20 | 50 |
| 100 | 5000 | 1/50 |
| 200 | 1000 | 1/5 |
| 500 | 500 | 1 |
| 1000 | 200 | 5 |
| 5000 | 100 | 50 |

# Random signal

$$Y = \sqrt{n}UDV^{\mathsf{T}} + \Sigma^{1/2}E, \quad \text{versus}$$
$$Y = \Sigma^{1/2}\bigl(\sqrt{n}UDV^{\mathsf{T}} + E\bigr)$$

Second version supports use of RMT to judge size (undetectable, harmful, helpful, strong).

## The signal $X$

Diagonals of $D$ spread through range given by factor strength.

$U^*$ and $V$ uniform on Stiefel manifolds.

Solve $\Sigma^{-1/2}U^*DV^{\mathsf{T}} = U\widetilde{D}\widetilde{V}^{\mathsf{T}}$ for $U$.

Deliver $X = \sqrt{n}\Sigma^{1/2}UDV^{\mathsf{T}}$

The point is to avoid having large $\sigma_i^2$ detectable from large variane of row $i$ of $Y$, making the problem too easy.

# Relative estimation error (REE)

$$\text{REE}(\hat{k}) = \frac{\|\widehat{X}(\hat{k}) - X\|_F^2}{\|\widehat{X}(k^*) - X\|_F^2} - 1$$

Here $k^*$ is the rank that attains the best reconstruction of $X$.

REE measures relative error compared to an oracle. The oracle uses early stopping but can try every rank and see which is closest to truth.

## Cases

6 factor scenarios $\times$ 3 noise scenarios $\times$ 10 sample sizes $=$ 180 cases, each done $100$ times

# Worst case REE

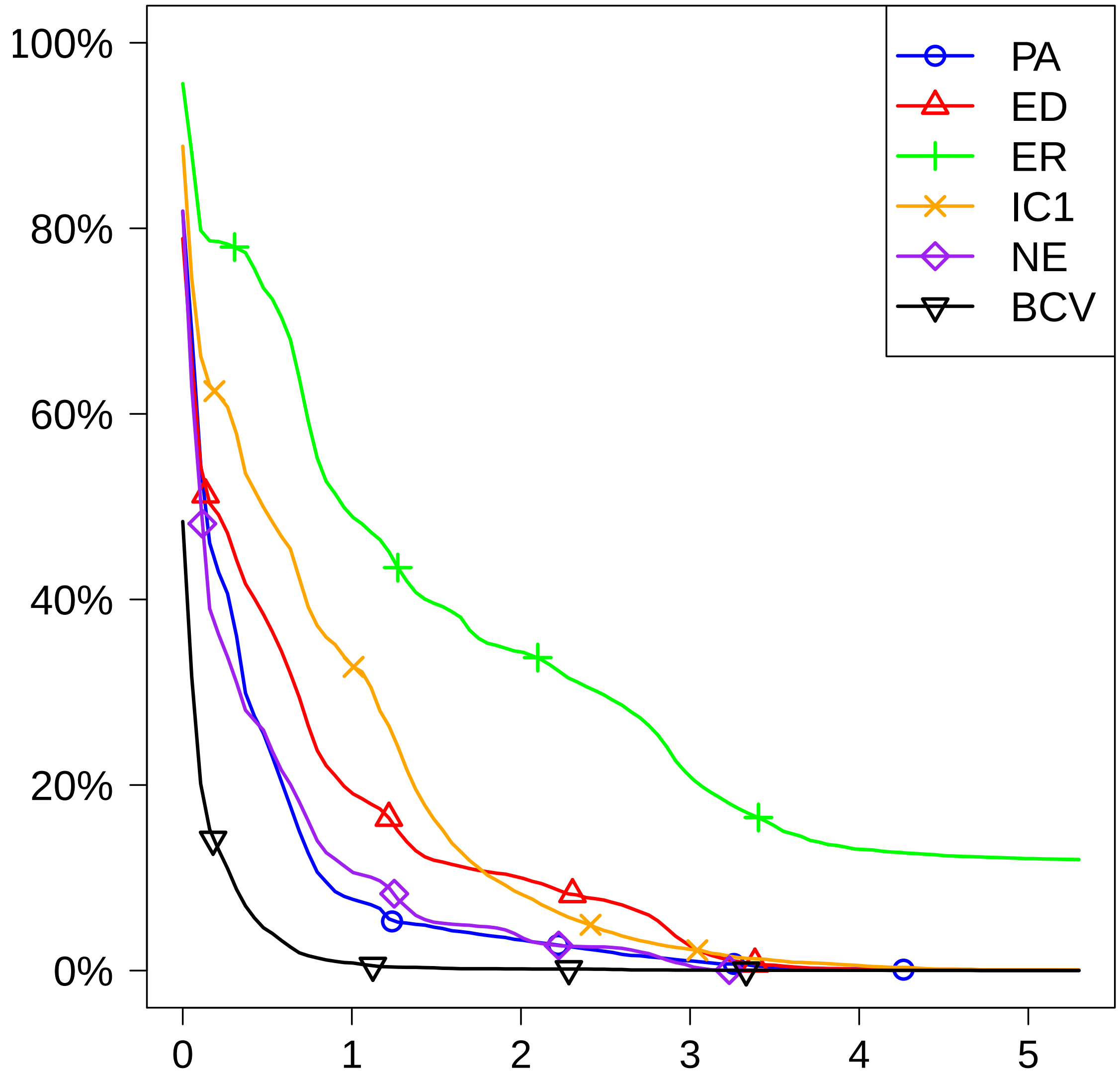
| $\mathrm{var}(\sigma_i^2)$ | PA | ED | ER | IC1 | NE | BCV |
|---|---|---|---|---|---|---|
| 0 | 1.99 | 1.41 | 49.61 | 1.13 | 0.12 | 0.29 |
| 1 | 2.89 | 2.42 | 25.02 | 3.11 | 2.45 | 0.37 |
| 10 | 3.66 | 2.28 | 15.62 | 4.46 | 2.10 | 0.62 |

For each of $6$ methods and $3$ noise levels we find the worst $\overline{\mathrm{REE}}$. The average is over $100$ trials. The maximum is over the $10$ sample sizes and $6$ factor strength scenarios.

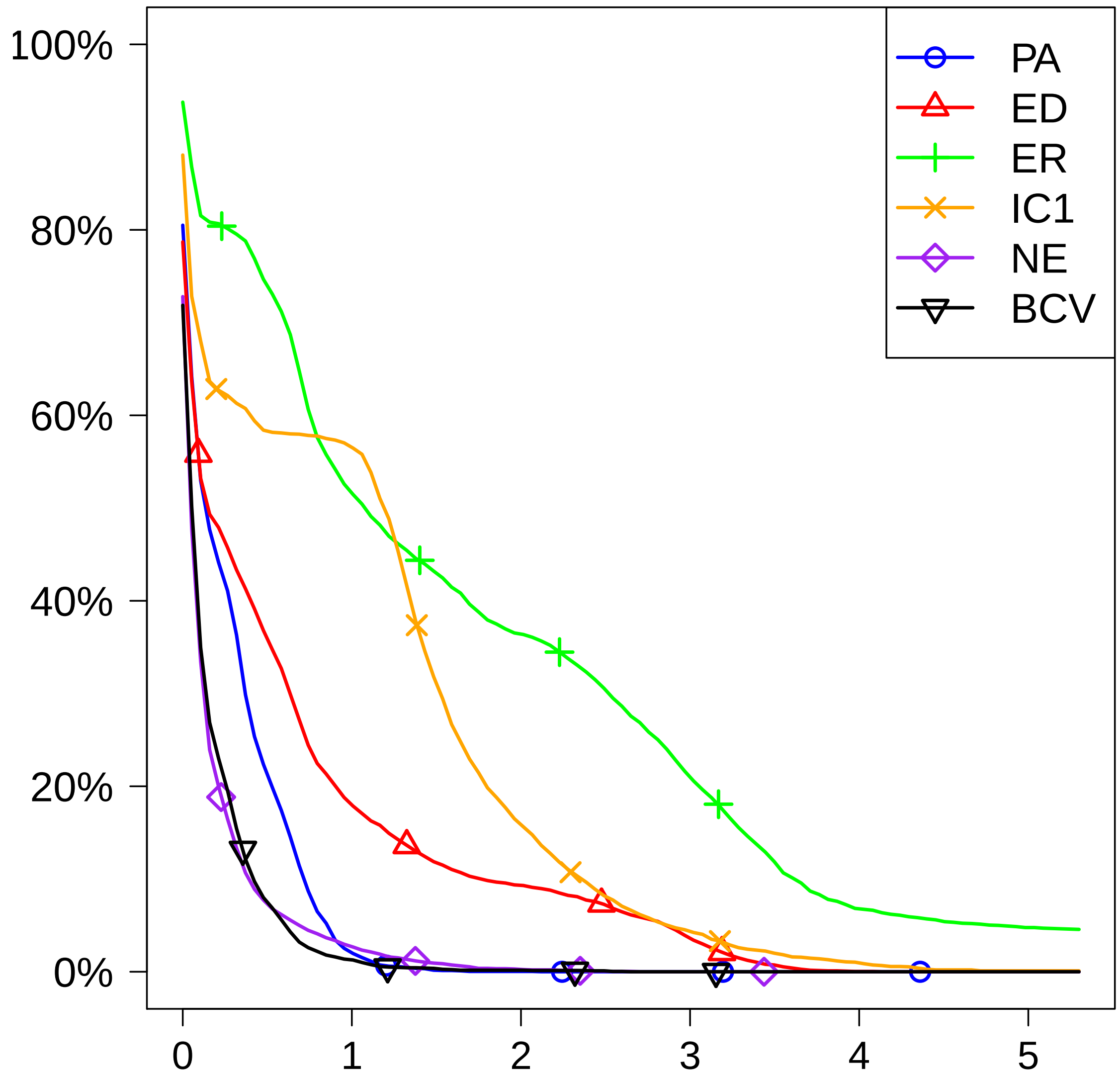
NE is best on problems with constant error variance.

BCV is best overall.

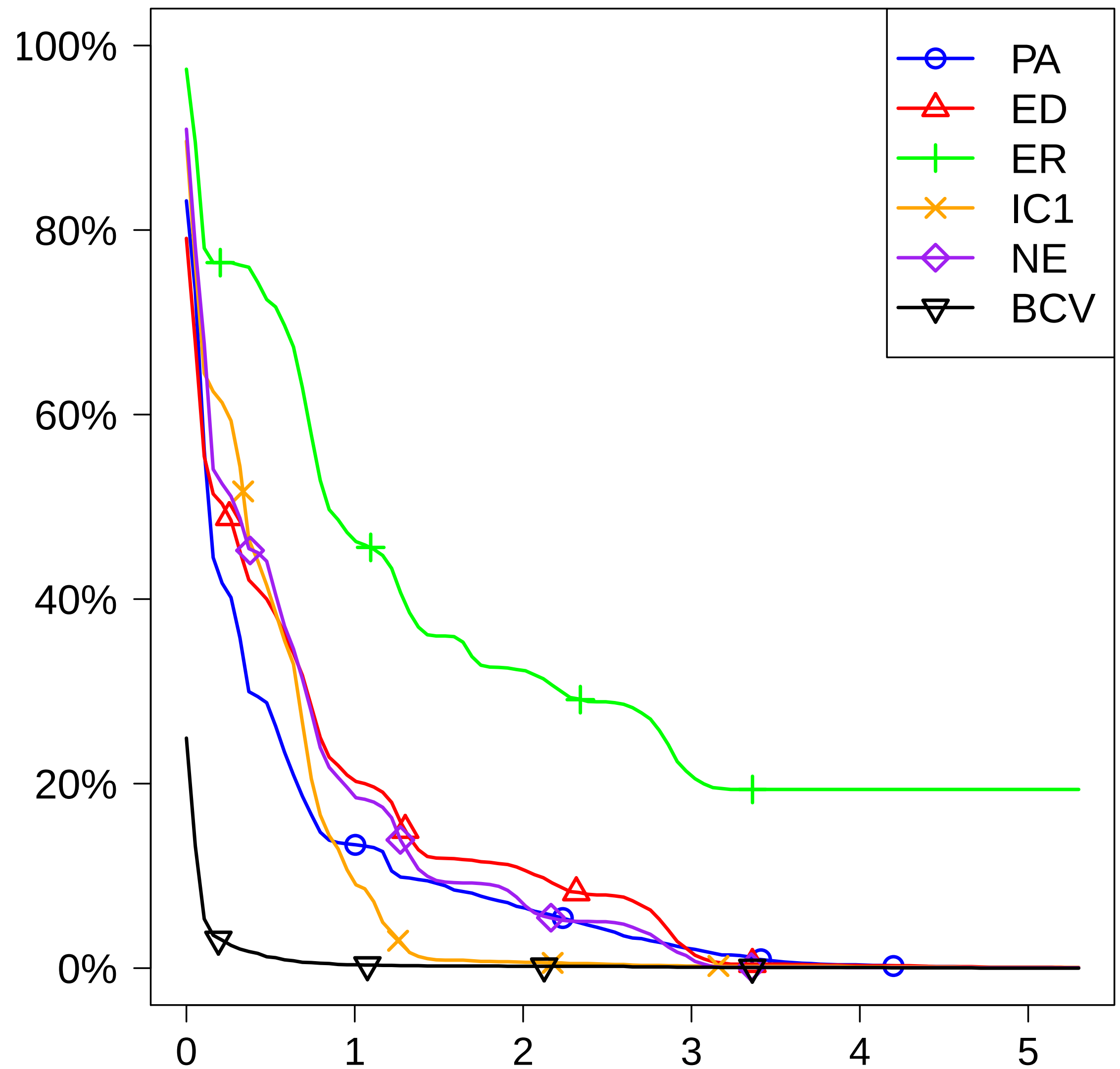
# REE survival plots



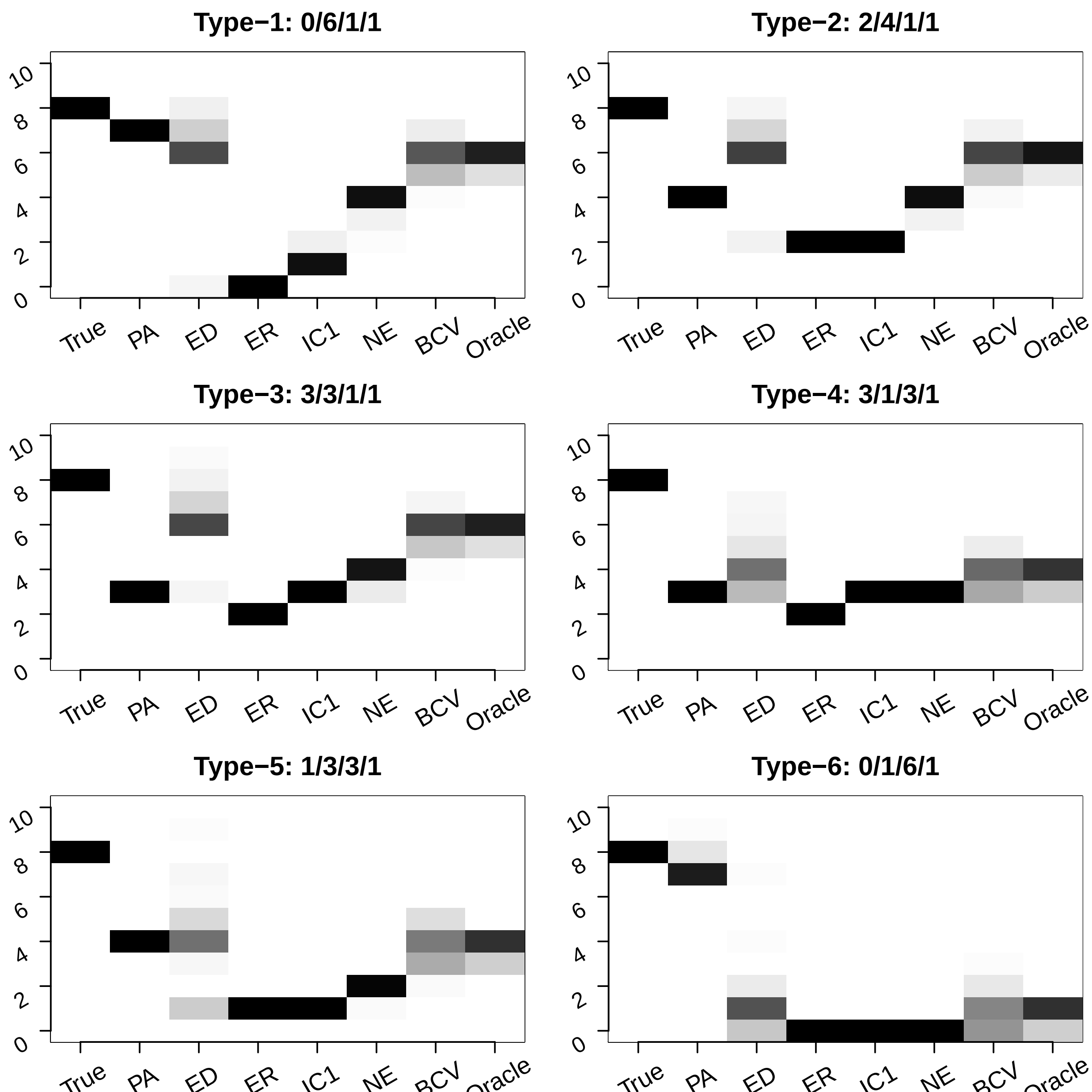
# Small data sets only



# Large data sets only



# Rank distribution example




$N = 5000 \; n = 100 \; \mathrm{Var}(\sigma_i^2) = 1$

# Meteorite data

$64 \times 128 = 8192$ pixel map of a meteorite

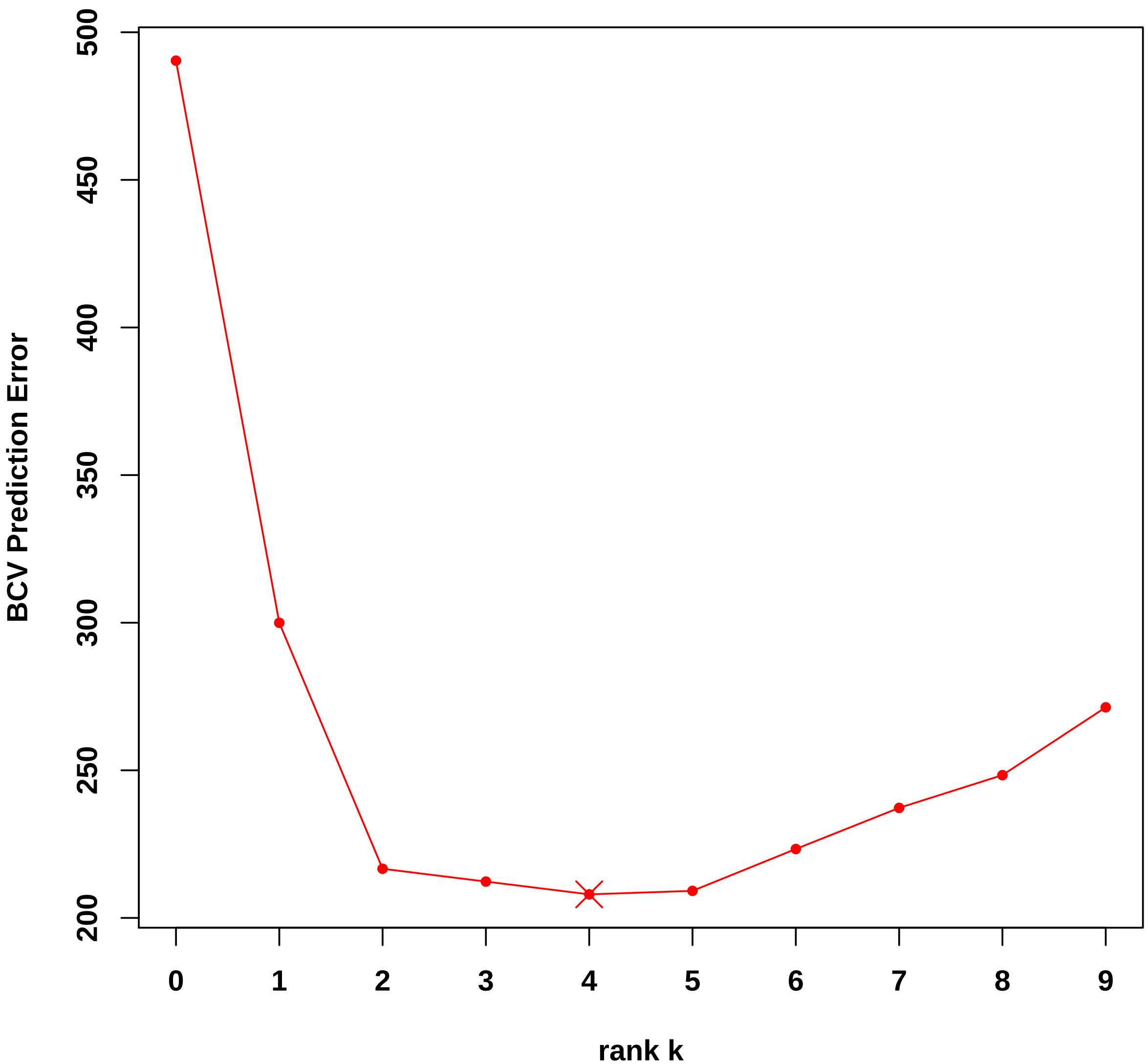
Concentraion of $15$ elements measured at each pixel.

Factors $\approx$ chemical compounds.

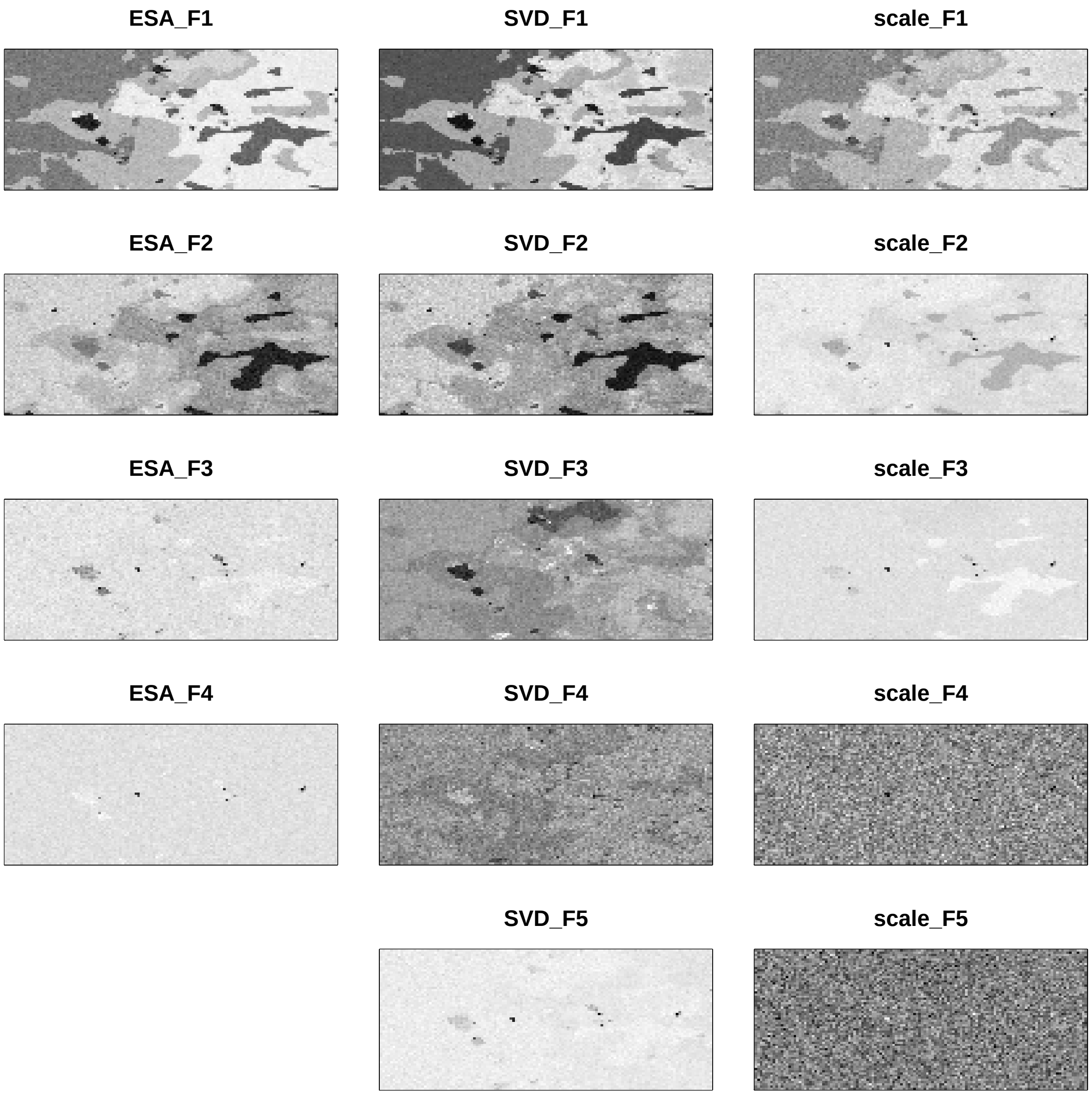
We center each of $15$ variables.

BCV chooses $k = 4$ while PA chooses $k = 3$.

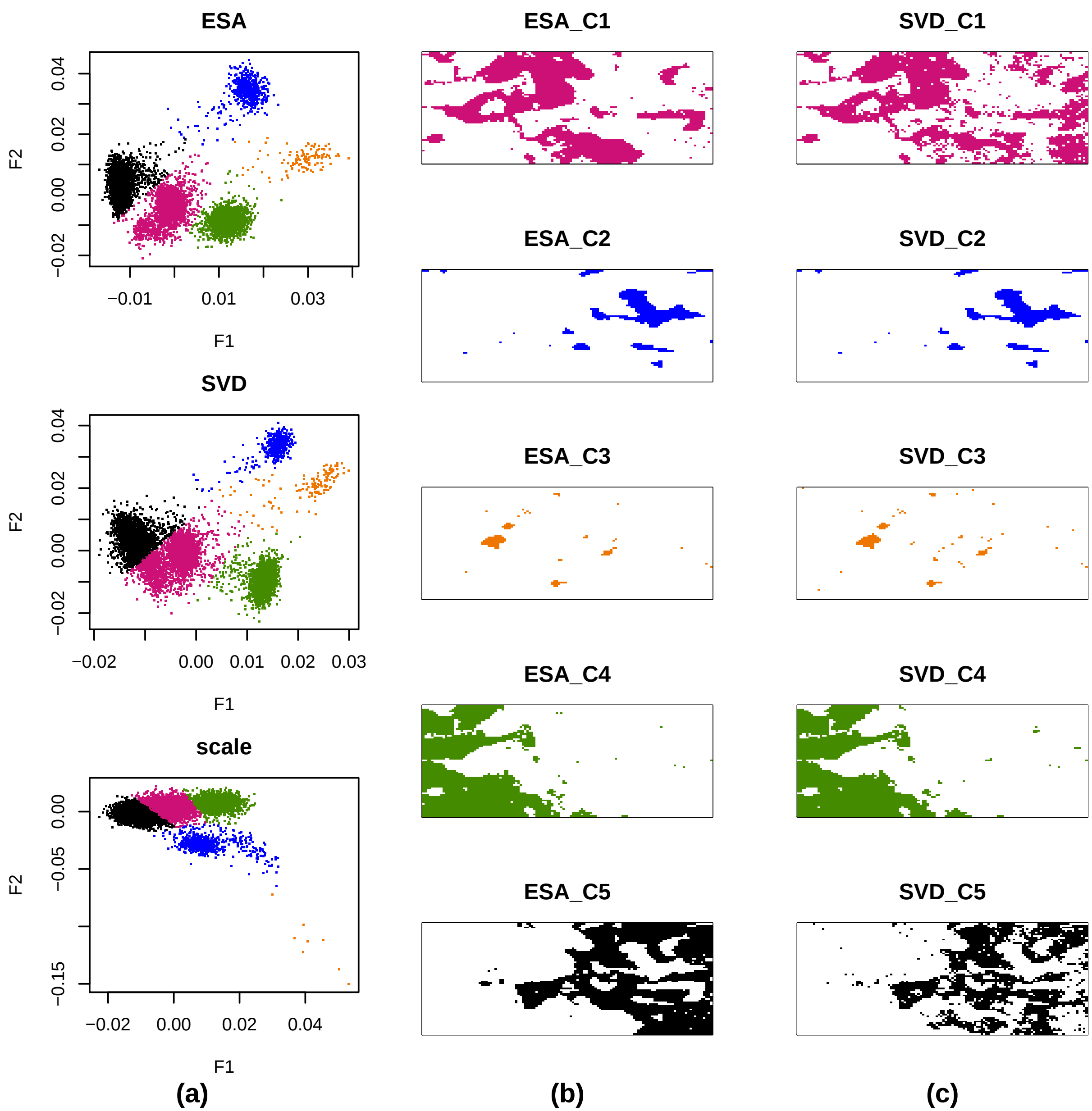
# BCV vs $k$


BCV Prediction Error
500
450
400
350
300
250
200
0
1
2
3
4
5
6
7
8
9
rank k

# Early stopping vs SVD



# ESA vs SVD ($k$-means)



# Early stopping: why $3$?

| Measurements | White Noise | Heteroscedastic Noise | |
|---|---|---|---|
| | $\mathrm{Var}(\sigma_i^2) = 0$ | $\mathrm{Var}(\sigma_i^2) = 1$ | $\mathrm{Var}(\sigma_i^2) = 10$ |
| $\frac{\text{Err}_X(m=3)}{\text{Err}_X(m=m_{\text{Opt}})}$ | $1.01 \pm 0.01$ | $1.00 \pm 0.01$ | $1.00 \pm 0.01$ |
| $\frac{\text{Err}_X(m=3)}{\text{Err}_X(m=1)}$ | $0.93 \pm 0.09$ | $0.90 \pm 0.11$ | $0.89 \pm 0.12$ |
| $\frac{\text{Err}_X(m=3)}{\text{Err}_X(m=50)}$ | $0.87 \pm 0.21$ | $0.87 \pm 0.21$ | $0.87 \pm 0.21$ |
| $\frac{\text{Err}_X(m=3)}{\text{Err}_X(\text{SVD})}$ | $1.03 \pm 0.06$ | $0.81 \pm 0.20$ | $0.75 \pm 0.22$ |
| $\frac{\text{Err}_X(m=3)}{\text{Err}_X(\text{QMLE})}$ | $1.02 \pm 0.05$ | $0.95 \pm 0.15$ | $0.91 \pm 0.19$ |
| $\frac{\text{Err}_X(m=3)}{\text{Err}_X(\text{oSVD})}$ | $1.03 \pm 0.06$ | $1.03 \pm 0.07$ | $1.03 \pm 0.08$ |

Table 1: Average and standard deviation over $6000$ simulations. PCA has $m = 1$. QMLE $\doteq$ EM. oSVD$=$SVD for oracle given $\Sigma$. Oracle's $k$ in numer. and denom.

Five is right out. M. Python (1975)

# Thanks

- Co-author Jingshu Wang
- Data: Ray Browning
- NSF for funds