Lecture 17

Factor Analysis

Syllabus

Lecture 01	Describing Inverse Problems
Lecture 02	Probability and Measurement Error, Part 1
Lecture 03	Probability and Measurement Error, Part 2
Lecture 04	The L_2 Norm and Simple Least Squares
Lecture 05	A Priori Information and Weighted Least Squared
Lecture 06	Resolution and Generalized Inverses
Lecture 07	Backus-Gilbert Inverse and the Trade Off of Resolution and Variance
Lecture 08	The Principle of Maximum Likelihood
Lecture 09	Inexact Theories
Lecture 10	Nonuniqueness and Localized Averages
Lecture 11	Vector Spaces and Singular Value Decomposition
Lecture 12	Equality and Inequality Constraints
Lecture 13	L_1 , L_∞ Norm Problems and Linear Programming
Lecture 14	Nonlinear Problems: Grid and Monte Carlo Searches
Lecture 15	Nonlinear Problems: Newton's Method
Lecture 16	Nonlinear Problems: Simulated Annealing and Bootstrap Confidence Intervals
Lecture 17	**Factor Analysis**
Lecture 18	Varimax Factors, Empircal Orthogonal Functions
Lecture 19	Backus-Gilbert Theory for Continuous Problems; Radon's Problem
Lecture 20	Linear Operators and Their Adjoints
Lecture 21	Fréchet Derivatives
Lecture 22	Exemplary Inverse Problems, incl. Filter Design
Lecture 23	Exemplary Inverse Problems, incl. Earthquake Location
Lecture 24	Exemplary Inverse Problems, incl. Vibrational Problems

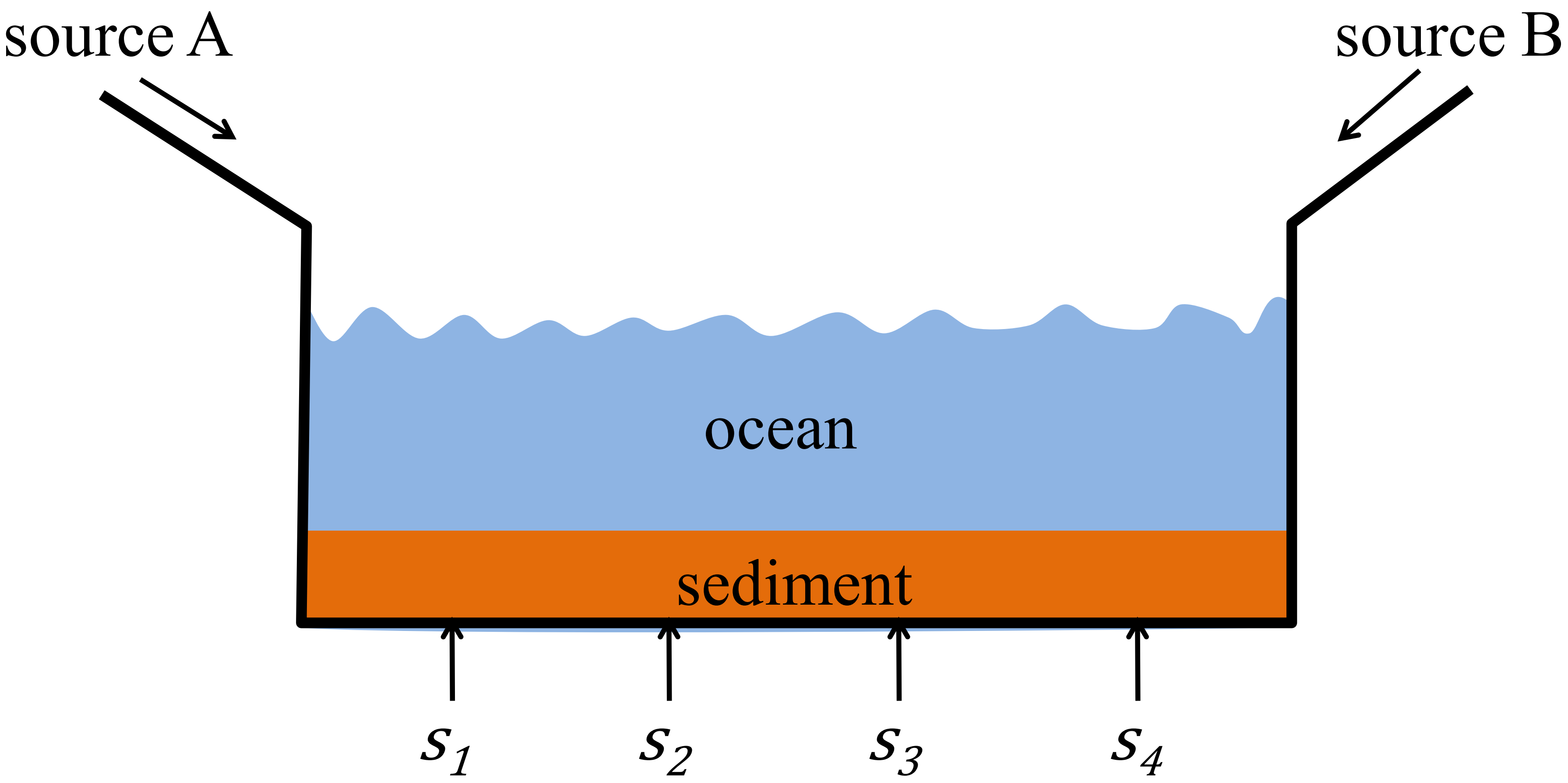
Purpose of the Lecture

Introduce Factor Analysis

Work through an example

Part 1

Factor Analysis


source A
source B
ocean
sediment
s_1
s_2
s_3
s_4

sample matrix **S**

S_{ij} is the fraction of element j in sample i:

$$\mathbf{S} = \begin{bmatrix} \text{element 1 in sample 1} & \cdots & \text{element } M \text{ in sample 1} \\ \text{element 1 in sample 2} & \cdots & \text{element } M \text{ in sample 2} \\ \vdots & \ddots & \vdots \\ \text{element 1 in sample } M & \cdots & \text{element } M \text{ in sample } N \end{bmatrix}$$

S arranged row-wise
but we'll use a column vector $\mathbf{s}^{(i)}$ for individual samples)

theory

samples are a linear mixture of sources

$$\mathbf{S} = \mathbf{C}\,\mathbf{F}$$

theory

samples are a linear mixture of sources

$$\mathbf{S} = \mathbf{C}\,\mathbf{F}$$

samples contain "elements"

theory

samples are a linear mixture of sources

$$\mathbf{S} = \mathbf{C}\,\mathbf{F}$$

sources called “factors”
factors contain “elements”

factor matrix **F**

F_{ij} is the fraction of element j in factor i:

$$\mathbf{F} = \begin{bmatrix} \text{element 1 in factor 1} & \cdots & \text{element } M \text{ in factor 1} \\ \text{element 1 in factor 2} & \cdots & \text{element } M \text{ in factor 2} \\ \vdots & \ddots & \vdots \\ \text{element 1 in factor } p & \cdots & \text{element } M \text{ in factor } p \end{bmatrix}$$

F arranged row-wise
but we'll use a column vector $\mathbf{f}^{(i)}$ for individual factors

theory

samples are a linear mixture of sources

$$\mathbf{S} = \mathbf{C}\,\mathbf{F}$$

coefficients
called "loadings"

loading matrix **C**

C_{ij} is the fraction of factor i in sample j:

$$\mathbf{C} = \begin{bmatrix} \text{factor 1 in sample 1} & \cdots & \text{factor } p \text{ in sample 1} \\ \text{factor 1 in sample 2} & \cdots & \text{factor } p \text{ in sample 2} \\ \vdots & \ddots & \vdots \\ \text{factor 1 in sample } N & \cdots & \text{factor } p \text{ in sample } N \end{bmatrix}$$

inverse problem

given $\mathbf{S}$

find $\mathbf{C}$ and $\mathbf{F}$

so that $\mathbf{S}=\mathbf{C}\mathbf{F}$

very non-unique

given $\mathbf{T}$ with inverse $\mathbf{T}^{-1}$

if $\mathbf{S}=\mathbf{CF}$

then $\mathbf{S}=[\mathbf{C}\,\mathbf{T}^{-1}][\mathbf{TF}]=\mathbf{C'F'}$

very non-unique

so a priori information needed to select a solution

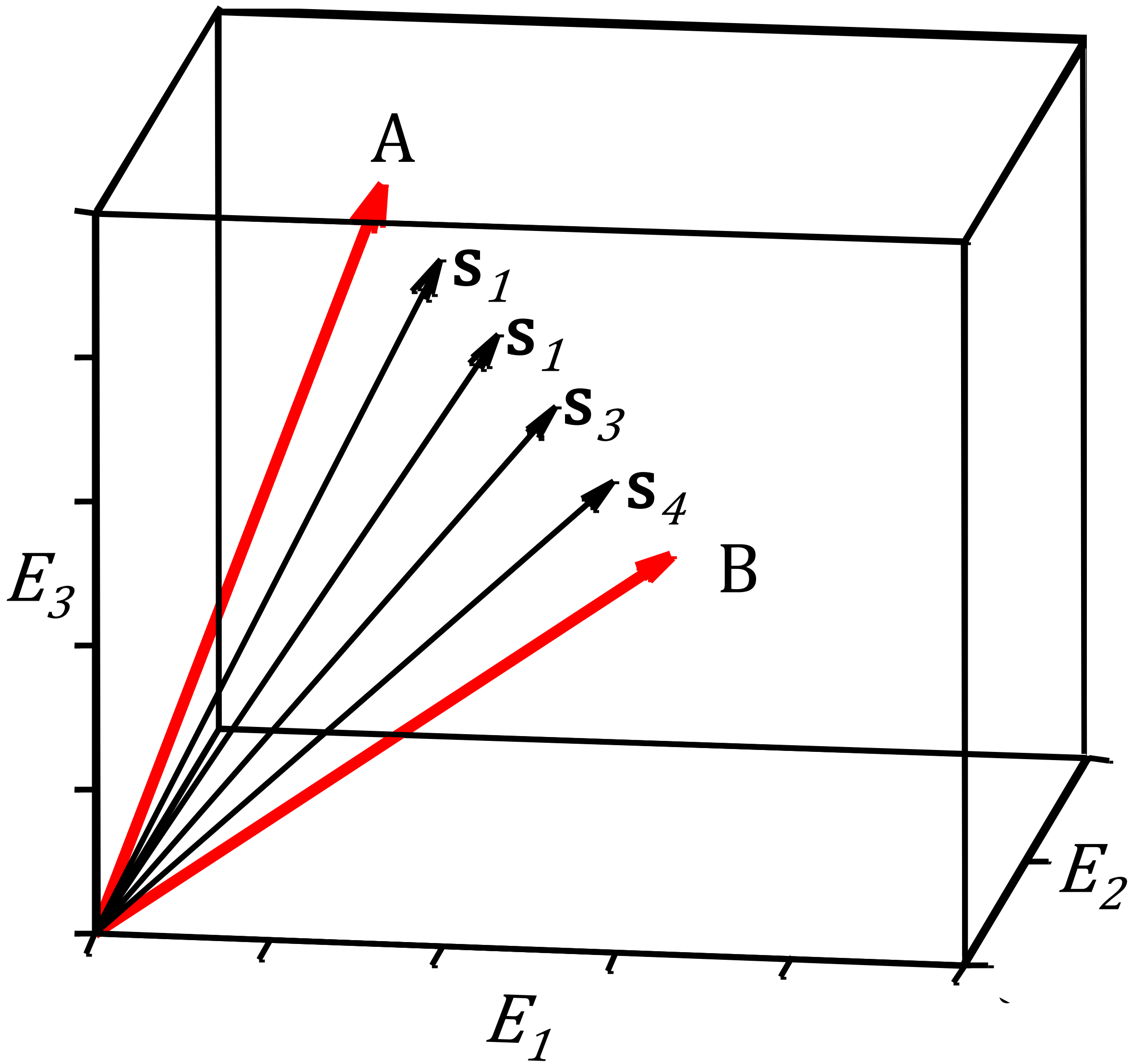
simplicity

what is the minimum number of factors needed

call that number p

does $\mathbf{S}$ span the full space of M elements?

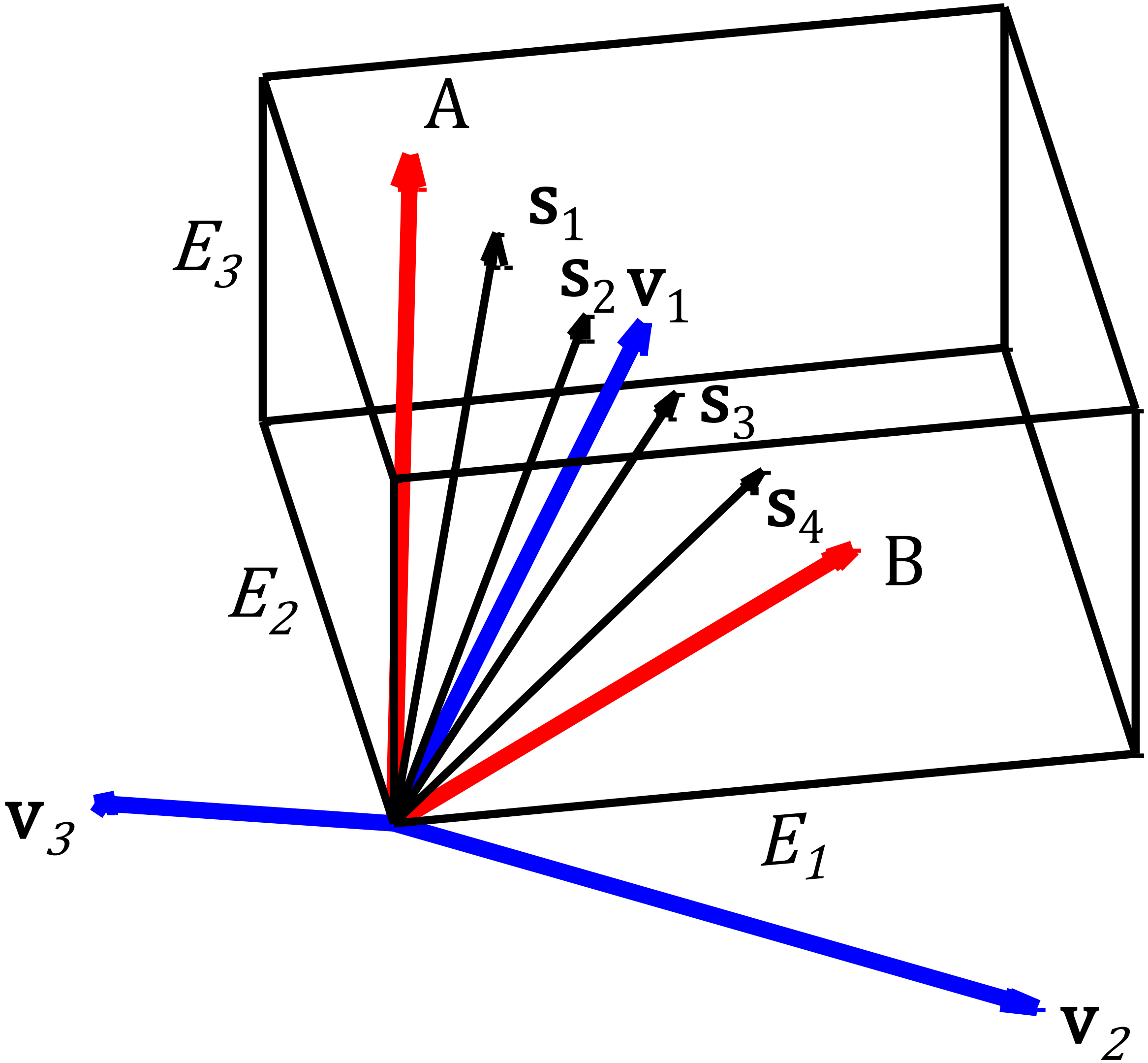
or just a p–dimensional subspace?


A
$\mathbf{s}_1$
$\mathbf{s}_1$
$\mathbf{s}_3$
$\mathbf{s}_4$
B
E_3
E_2
E_1

we know how to answer this question

$$\mathbf{S} = \mathbf{U}_\mathrm{p}\mathbf{\Lambda}_\mathrm{p}\mathbf{V}_\mathrm{p}^\mathrm{T} = \left(\mathbf{U}_\mathrm{p}\mathbf{\Lambda}_\mathrm{p}\right)\left(\mathbf{V}_\mathrm{p}^\mathrm{T}\right) = \mathbf{CF}$$

p is the number of non-zero singular values


A
$\mathbf{s}_1$
E_3
$\mathbf{s}_2$
$\mathbf{v}_1$
$\mathbf{s}_3$
$\mathbf{s}_4$
B
E_2
$\mathbf{v}_3$
E_1
$\mathbf{v}_2$

SVD identifies a subspace

but the SVD factors

$$\mathbf{f}^{(i)} = \mathbf{v}^{(i)} \quad i=1, p$$

not unique

usually not the "best"

factor $\mathbf{f}^{(1)}$

$\mathbf{v}$ with the largest singular value

usually near the mean sample

sample mean <**s**>
minimize

$$\sum_i \left[\mathbf{s}^{(i)} \cdot \langle \mathbf{s} \rangle\right]$$

eigenvector <**v**>
minimize

$$\sum_i \left[{\mathbf{s}^{(i)}}_i \cdot \mathbf{v}\right]^2$$

factor $\mathbf{f}^{(1)}$
v with the largest singular value

usually near the mean sample

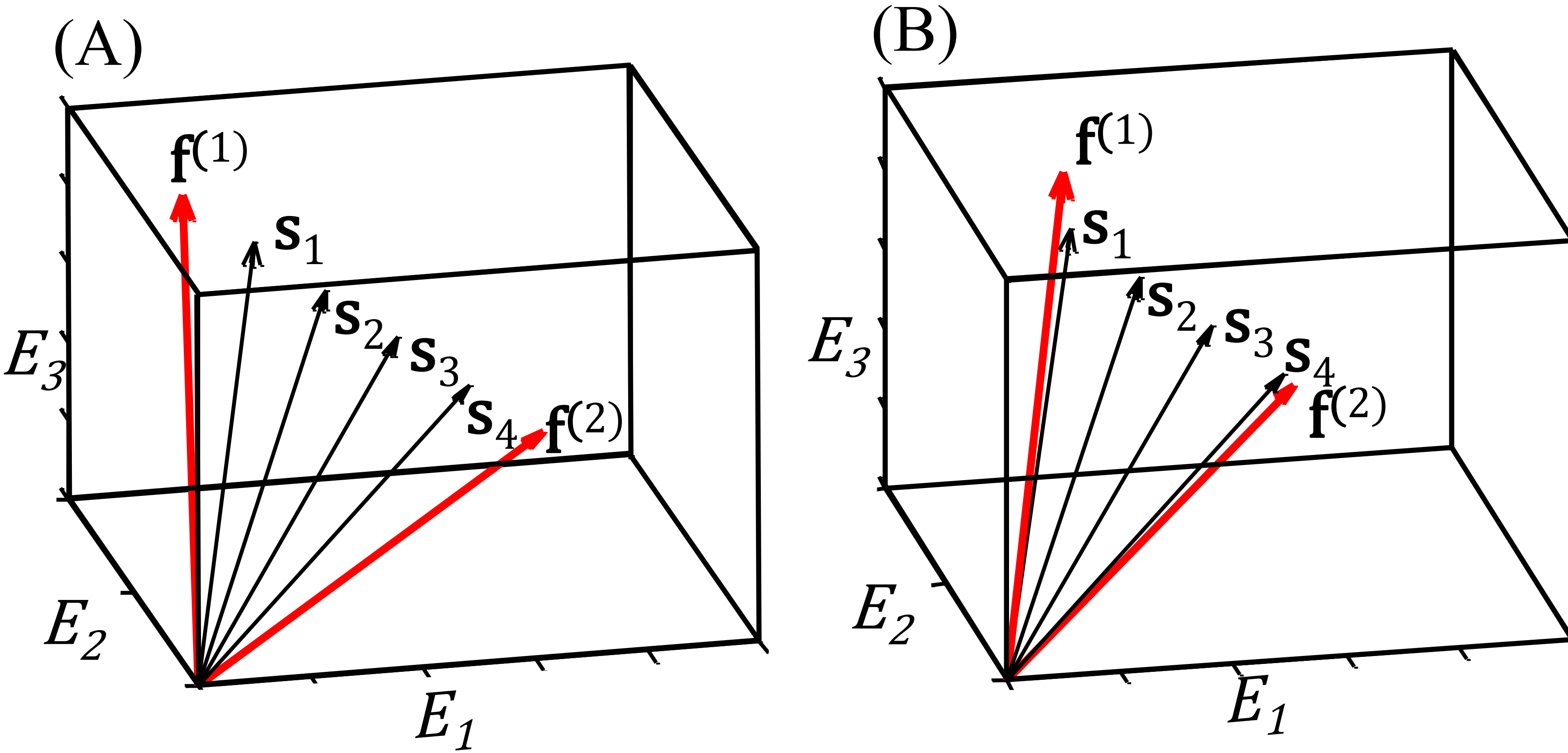
sample mean <**s**>
minimize

$$\sum_i \left[\mathbf{s}^{(i)} \cdot \langle \mathbf{s} \rangle\right]$$

eigenvector <**v**>
minimize

$$\sum_i \left[\mathbf{s}^{(i)}{}_i \cdot \mathbf{v}\right]^2$$

about the same if samples are clustered


(A)
f(1)
s1
s2
s3
s4
f(2)
E3
E2
E1
(B)
f(1)
s1
s2
s3
s4
f(2)
E3
E2
E1

in *MatLab*

```
[U, LAMBDA, V] = svd(S,0);
lambda = diag(LAMBDA);
F = V';
C = U*LAMBDA;
```

in *MatLab*

"economy" calculation
LAMBDA is $M \times M$

```
[U, LAMBDA, V] = svd(S,0);
lambda = diag(LAMBDA);
F = V';
C = U*LAMBDA;
```

since samples have measurement noise

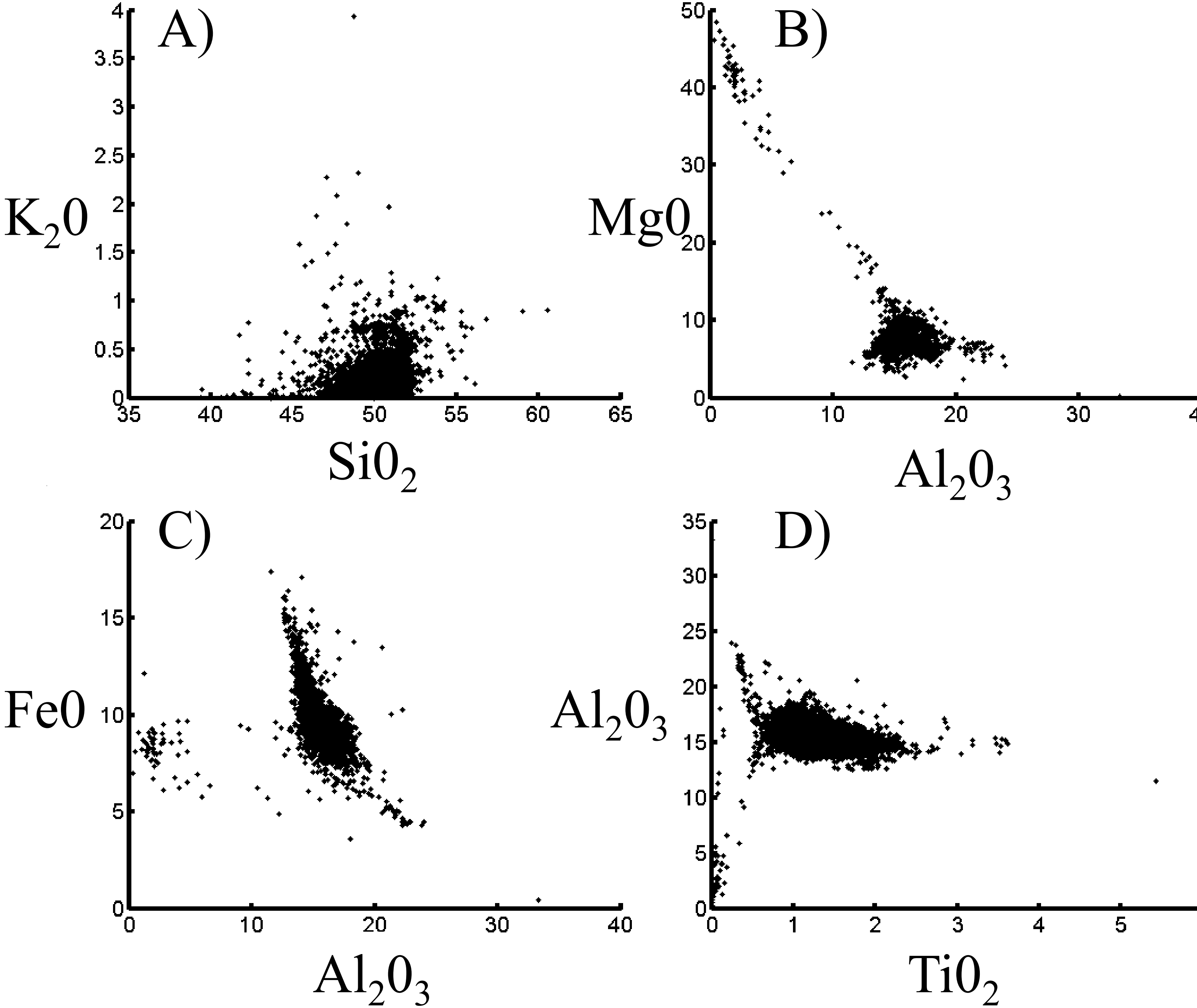
probably no exactly singular values
just very small ones

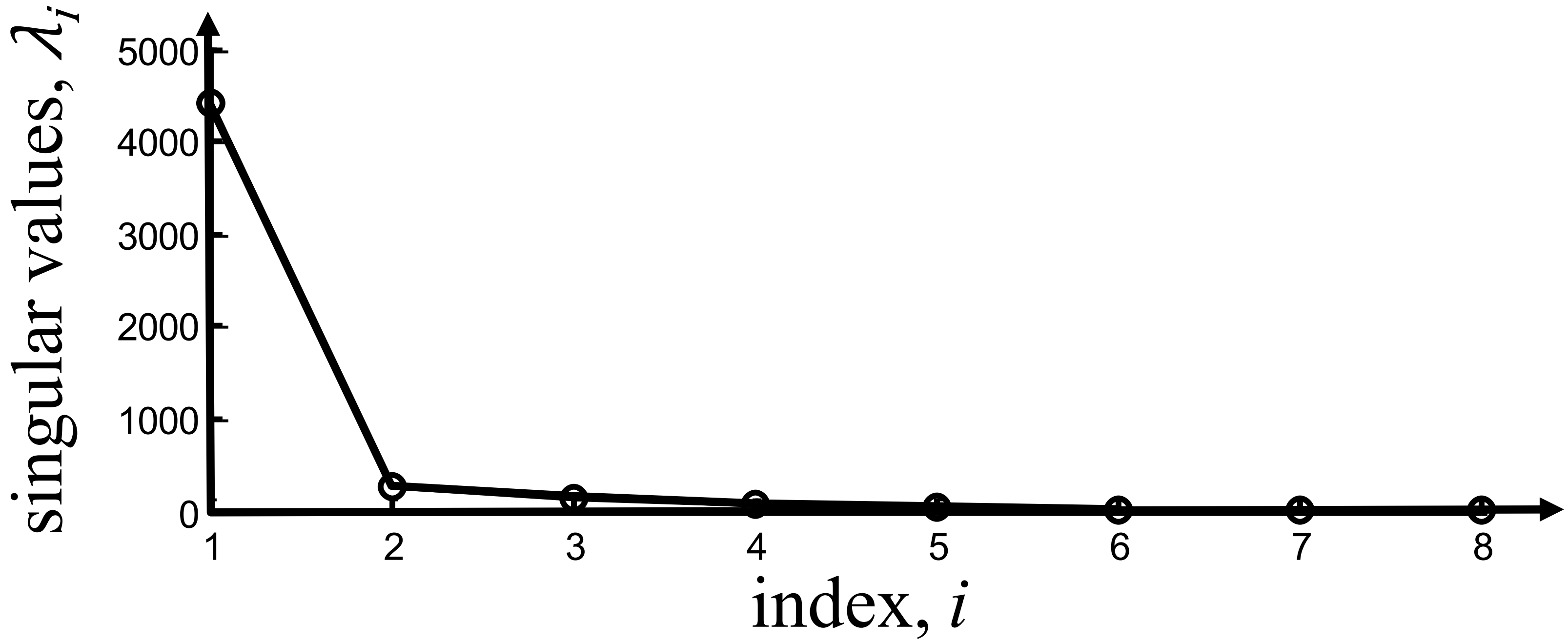
so pick p
for which
$\mathbf{S} \approx \mathbf{CF}$
is an adequate approximation

Atlantic Rock Dataset

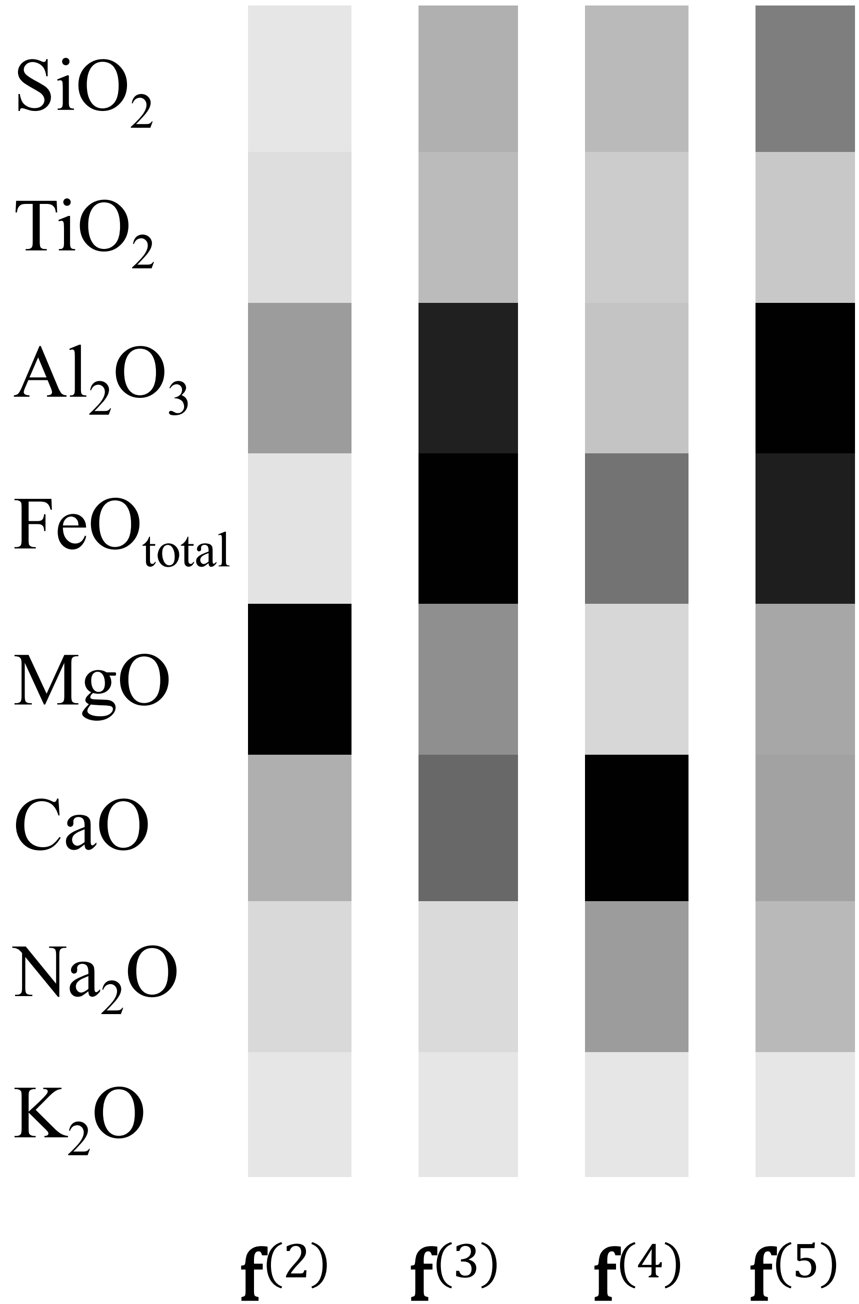
SiO_2	TiO_2	Al_2O_3	FeO_t	MgO	CaO	Na_2O	K_2O
51.97	1.25	14.28	11.57	7.02	11.67	2.12	0.07
50.21	1.46	16.41	10.39	7.46	11.27	2.94	0.07
50.08	1.93	15.6	11.62	7.66	10.69	2.92	0.34
51.04	1.35	16.4	9.69	7.29	10.82	2.65	0.13
52.29	0.74	15.06	8.97	8.14	13.19	1.81	0.04
49.18	1.69	13.95	12.11	7.26	12.33	2	0.15
50.82	1.59	14.21	12.85	6.61	11.25	2.16	0.16
49.85	1.54	14.07	12.24	6.95	11.31	2.17	0.15
50.87	1.52	14.38	12.38	6.69	11.28	2.11	0.17

(several thousand more rows)

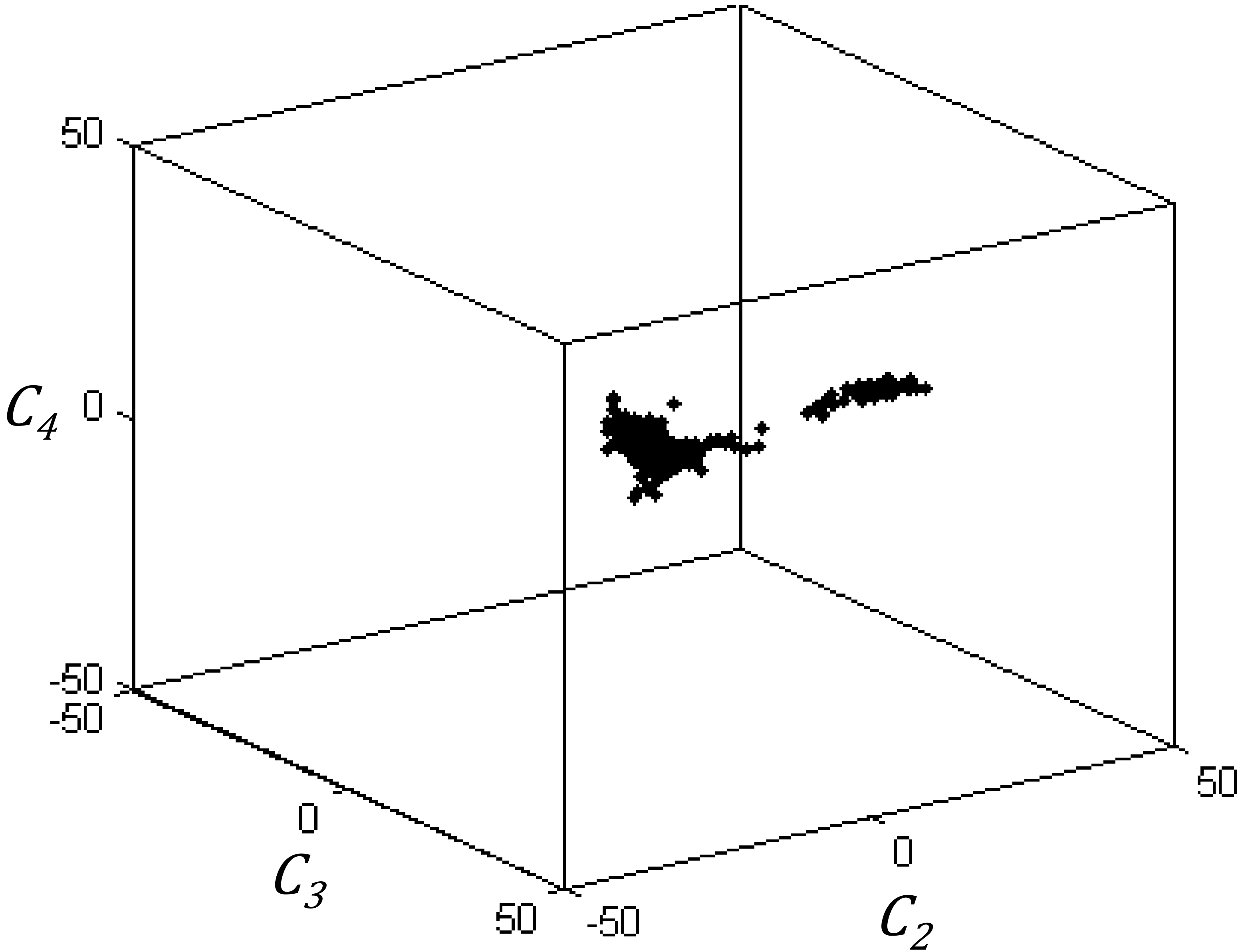

A)
K20
Si02
B)
Mg0
Al203
C)
Fe0
Al203
D)
Al203
Ti02



singular values, λ_i
5000
4000
3000
2000
1000
0
1
2
3
4
5
6
7
8
index, i



SiO2
TiO2
Al2O3
FeOtotal
MgO
CaO
Na2O
K2O
f(2)
f(3)
f(4)
f(5)


50
C4 0
-50
-50
0
C3
50
-50
0
C2
50