

**Linear Mixing Models: an overview
with application to air pollution**

Werner Stahel

Seminar für Statistik, ETH Zürich

April 19. 2006

1 Introduction

Large, automatically recorded data sets on air pollution.

Chemical “compounds”:

– Main pollutants: NO_x , CO_2 , SO_2 , O_3 ,

– Volatile Organic Compounds (VOC):

ethane, ..., benzene, toluene

Here: **Data** from a monitoring station

20 VOC, 1 obs./hour, for 1 year (≈ 8000 observations)

● Pollution comes from “sources”!

– Exhaust from gasoline cars

– Exhaust from diesel vehicles (trucks)

– Evaporation, mainly gasoline (?)

– Solvents in paint etc.

Who contributes how much?

● Assumption: Sources emit compounds in

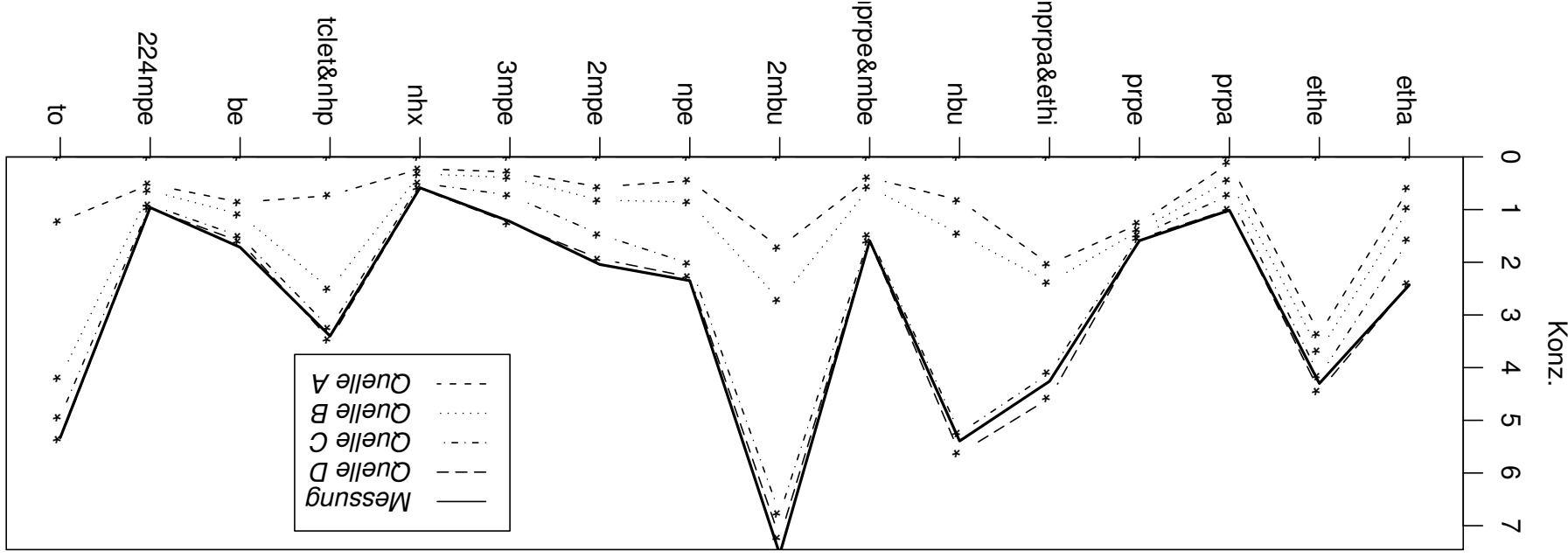
(more or less) **constant proportions**

... that remain unchanged during transport

← “Fingerprints”, “Profiles”.

Visualization of the basic idea

Observation "profile" $[X_{(1)}^i, \dots, X_{(m)}^i]$ is the sum of contributions of the sources, and these are multiples of the "source profiles" $[C_{(1)}^k, \dots, C_{(m)}^k]$



Goal: Identify and quantify sources and their contribution to pollution!

Applications

- Geology: rocks are mixtures of basic rock types
- Hydrochemistry: Concentrations of minerals in spring water reflect sediments that the water passed and length of stay there.
- Chemistry: Spectra of mixtures of chemical compounds concentrations can be monitored online!
- (Probability: Discrete distributions as mixtures of basic types.)
- Sociology: Household budgets as mixtures of basic types.

Lit.: Chemical Mass Balance Models, Linear Unmixing
Factor Analysis in Chemistry, Linear Mixing Model:

Renner (1993), Weltje (1997) (Geology)

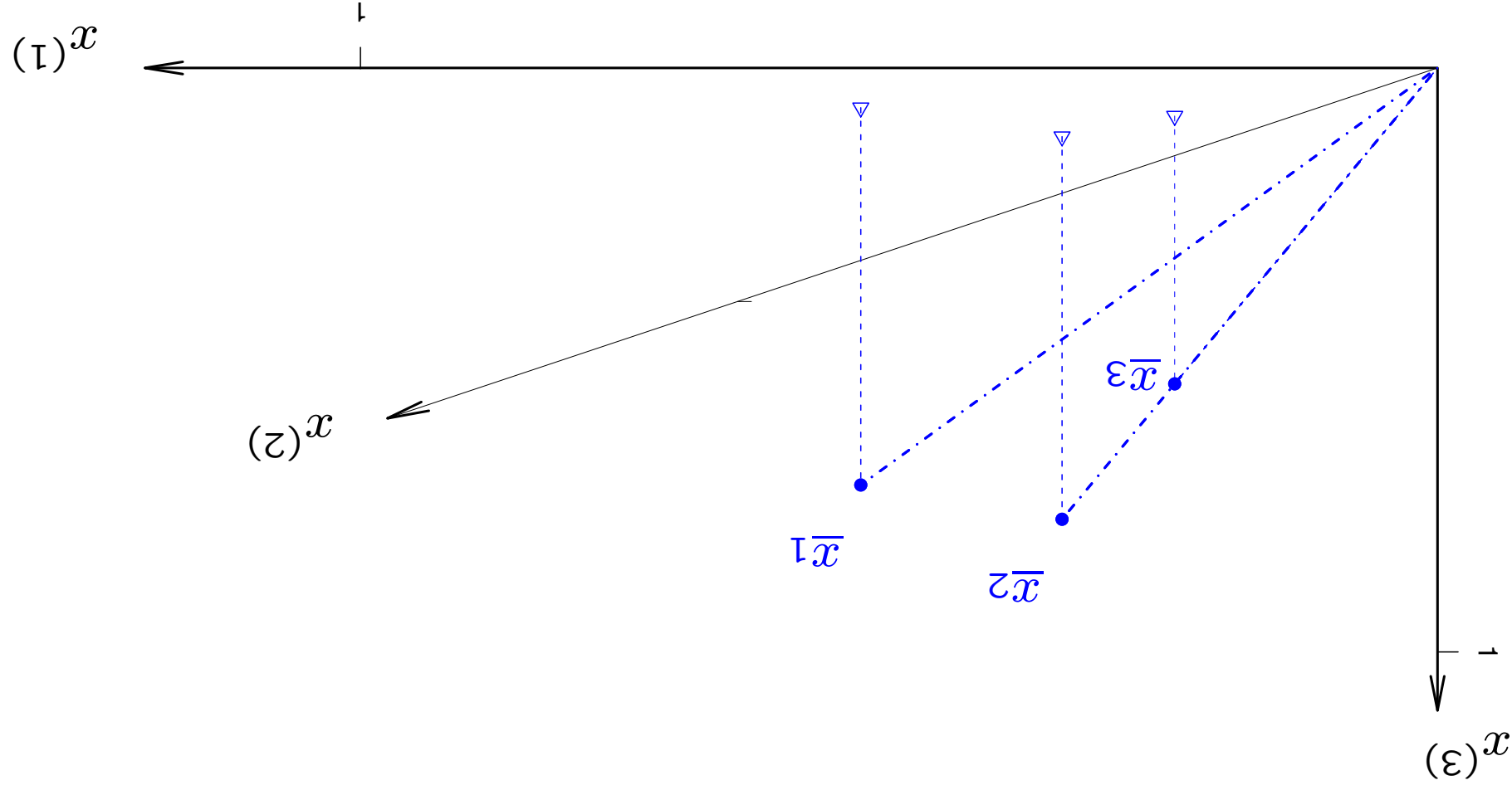
C.Henry, Lewis, Hopke and Williamson (1984), Hopke (1991),
Malinowski (1991).

Overview

2. Models
3. Methods
4. Results
5. Extensions

2 Models

Visualization: 3 compounds. Observations $\bar{x}_i = [x_{(1)}^i, x_{(2)}^i, x_{(3)}^i]$

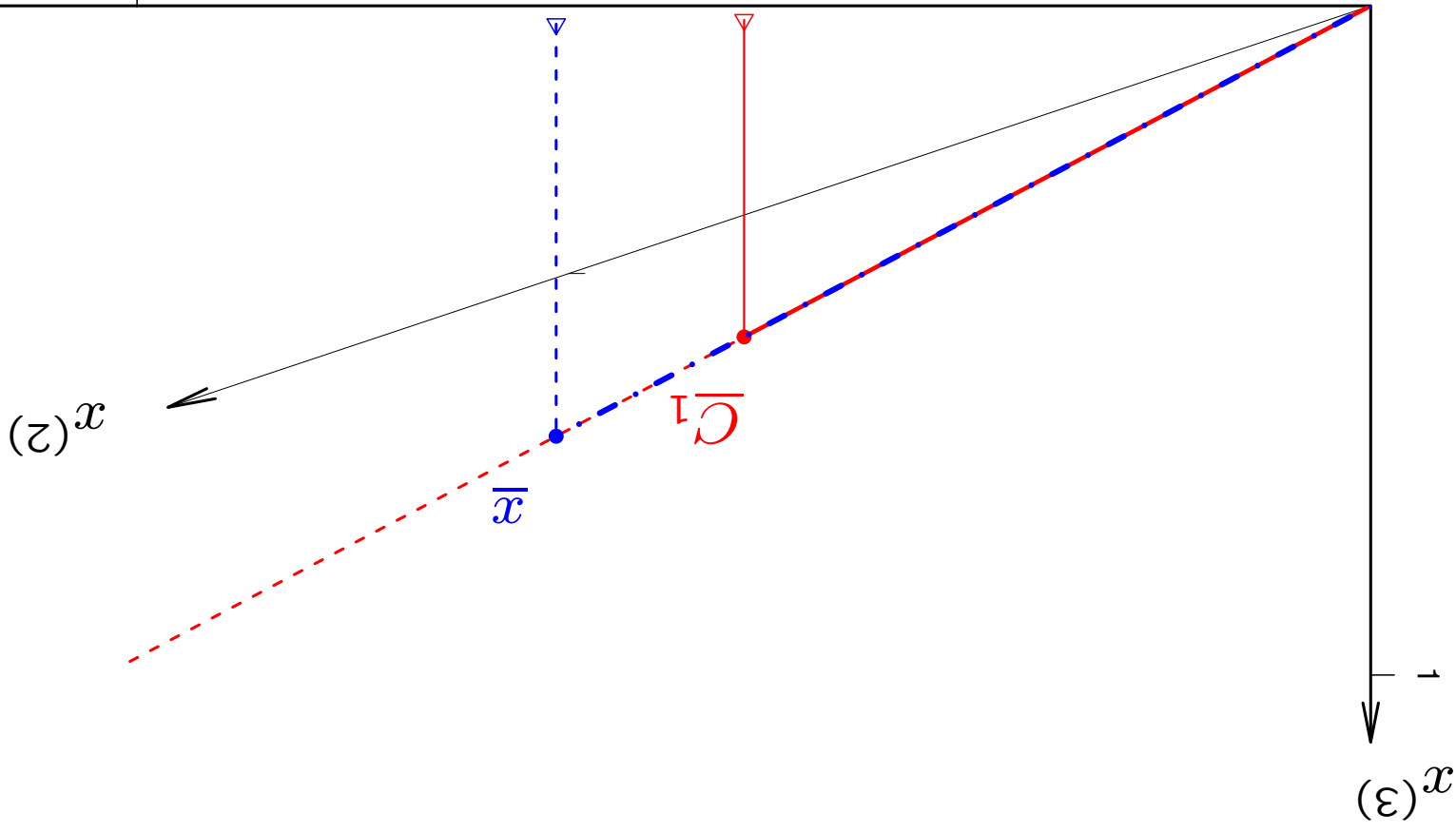


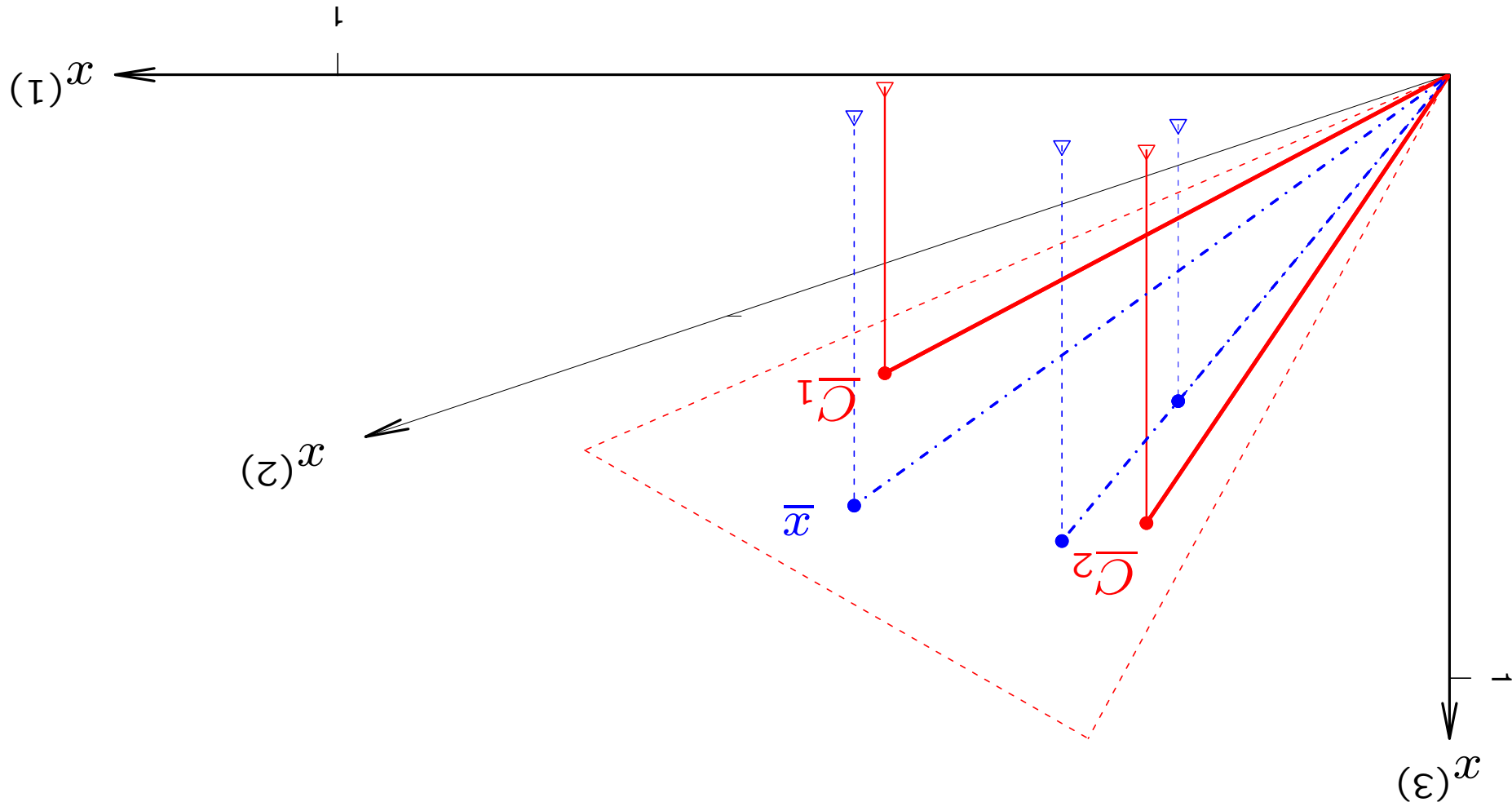
1 source \rightarrow Data are multiples of \bar{C} (up to meas. error) $X^{(j)} \approx S_j C^{(j)}$

concentrations of compounds j per unit emission of source k .

Source profile $\bar{C}^k = [C_{(1)}^k, C_{(2)}^k, C_{(3)}^k]$:

Source profile $\bar{C}^k = [C_{(1)}^k, C_{(2)}^k, C_{(3)}^k] :$
 concentrations of compounds j per unit emission of source k .
 1 source \rightarrow Data are multiples of \bar{C} (up to meas. error) $X_{(j)}^i \approx S_i \cdot C_{(j)}^i$





2 sources → Data on plane "between" \overline{C}_1 and \overline{C}_2

$X_{(j)}^i$: conc. of compound j for **observation** (time) i
 $C_{(j)}^k$: conc. ... j in unit emission from **source** k
 $S_{(k)}^i$: **activity** of source k ("score") for obs. i
 $E_{(j)}^i$: measurement error.

$$X_{(j)}^i = S_{(1)}^i C_{(j)}^1 + S_{(2)}^i C_{(j)}^2 + E_{(j)}^i, \quad j = 1, 2, 3$$

$$X_{(j)}^i = S_{(1)}^i C_{(j)}^1 + S_{(2)}^i C_{(j)}^2 + E_{(j)}^i, \quad j = 1, 2, 3$$

$X_{(j)}^i$: conc. of compound j for **observation** i
 $C_{(j)}^k$: conc. ... j in unit emission from **source** k
 $S_{(k)}^i$: **activity** of source k ("score") for obs. i
 $E_{(j)}^i$: measurement error.

General Model: $X = SC + E$.

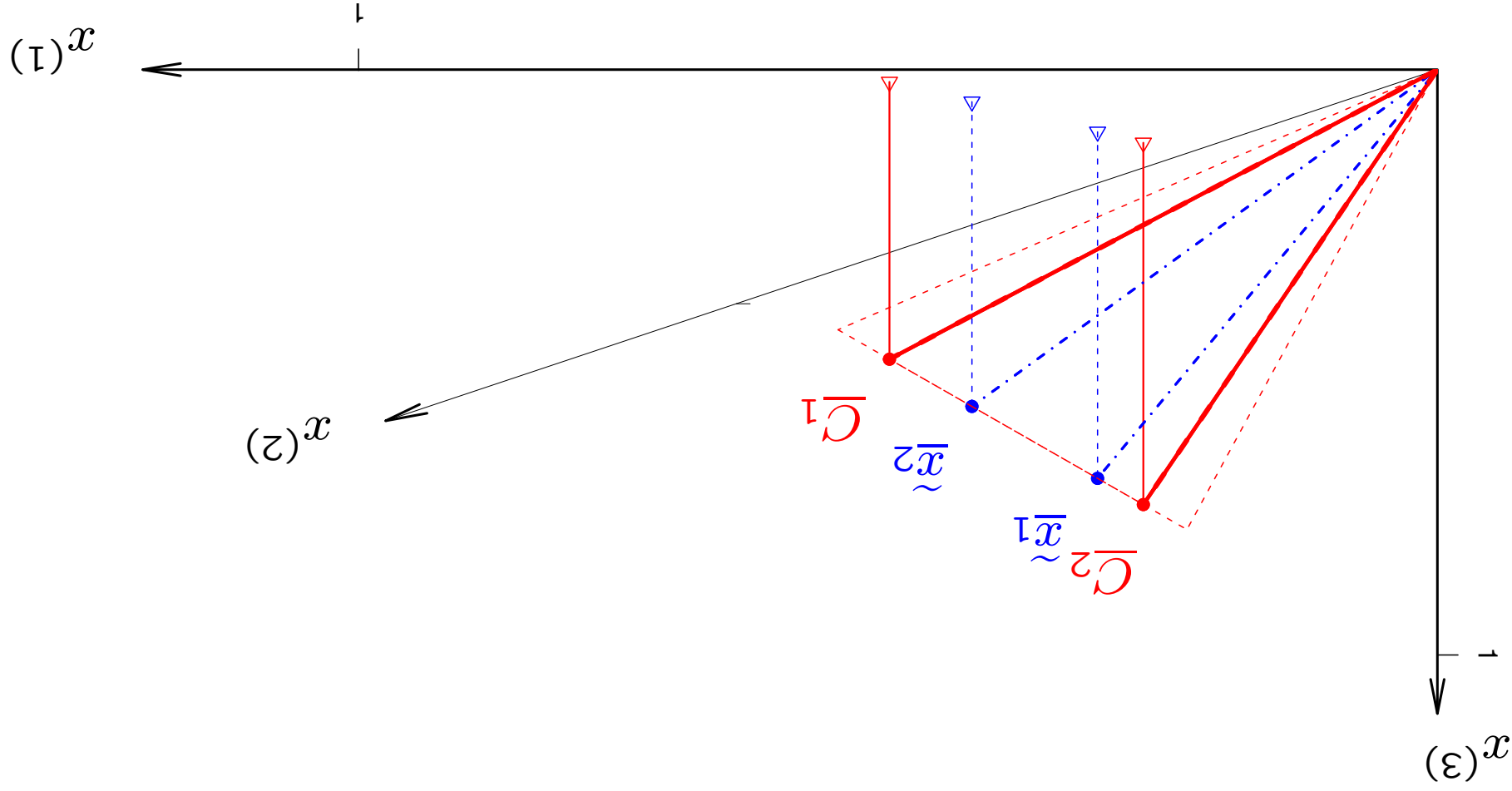
Data are in **p -dimensional subspace** – up to measurement error.
 Source profiles = edges of a "cone" in the subspace
 that contains all observations

Constraints: $X_{(j)}^i \geq 0, C_{(j)}^k \geq 0, S_{(k)}^i \geq 0$.

Standardized data.

Divide profiles $X_{(j)}^i$ by their length $\sum_j X_{(j)}^i$ \leftarrow $\tilde{X}_{(j)}^i = X_{(j)}^i / \sum_j X_{(j)}^i$ (In composition of rocks, data are already in this form.)

\leftarrow Data on "simplex" \leftarrow approx. on subspace of dim. $d - 1$.



Model contains parameters C , S ,

and parameters for the distribution of the errors E

(e.g., $E \sim \mathcal{N}(\bar{0}, \text{diag}(\sigma_1^2, \dots, \sigma_m^2))$)

S : **incidental parameters** (p more for each add. obs.)

← **“functional”** model.

Alternatively, the $S_{(j)}^?$ are random.

Wolbers (and Stahel, 2005): $[\log(S_{(j)}^?) \mid \Phi] \sim \mathcal{N}(\bar{\mu}, \Phi)$

← **“structural”** model.

(Notions used in “errors-in-variables” regression models.)

Structural Model.

$$[\log(S_i^{(j)})] \sim \mathcal{N}(\bar{\mu}, \Psi)$$

Errors are multiplicative, lognormal:

$$\bar{X}_i = \mathbf{C}^T \bar{S}_i \circ \bar{E}_i \quad (\circ: \text{elementwise multiplication})$$
$$\log(E_i^{(j)}) \sim \mathcal{N}(0, \sigma_j)$$

Lognormal distributions are much more plausible than normal distributions for many applications.

← Log transformation!

But: **Mixing occurs for untransformed X 's!**

No physically interpretable linear model for $\log(X_i^{(j)})$!

3 Methods

Model: $X = SC + E$.

If C is known \rightarrow "Chemical Mass Balance".

(Not useful in our case, since "known" profiles proved inappropriate!)

If C is unknown \rightarrow "(Bi-)linear Unmixing", Factor Analysis Model

Approaches.

- 2 steps: (A) Find subspace, (B) find suitable source profiles.
- \rightarrow estimate parameters of structural model (by max. likelihood).

The Graphical Approach

... suitable for the Functional Model

Step 0: Standardization to $\sum_j x_i^{(j)} = 1$ \longleftarrow Compositional data.

1 dimension disappears (for centered data).

Step 1: Find subspace which contains the data (approx.)

= reduction of dimension – by

- Principal Components Analysis (PCA)

- traditional Factor Analysis (FA)

$$X = S_o C_o + E \quad E_i^{(j)} \sim N(0, \sigma_j^2), \text{ indep.}$$

FA estimates σ_j 's from the data.

- PCA (implicitly) assumes equal σ_j 's.

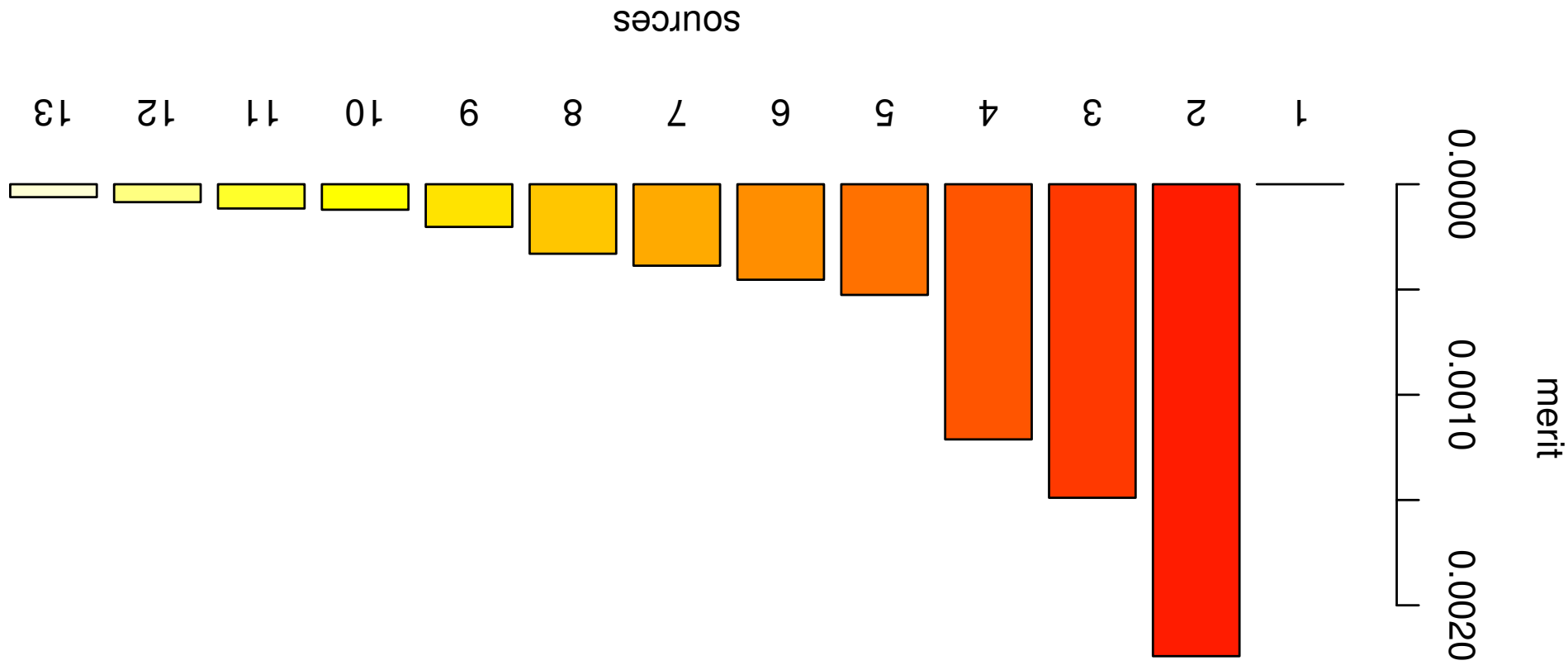
- Weighted PCA assumes known (proportions of) σ_j 's.

- Optimal subspace assuming $E_i^{(j)} \sim N(0, \sigma_j^2 \cdot \widehat{X}_i^{(j)})$.

Choice of dimension d ?
– “elbow” in “scree” plot

Choice of dimension d ?
– “elbow” in “screen” plot

PCA: Variance explained by additional dim.

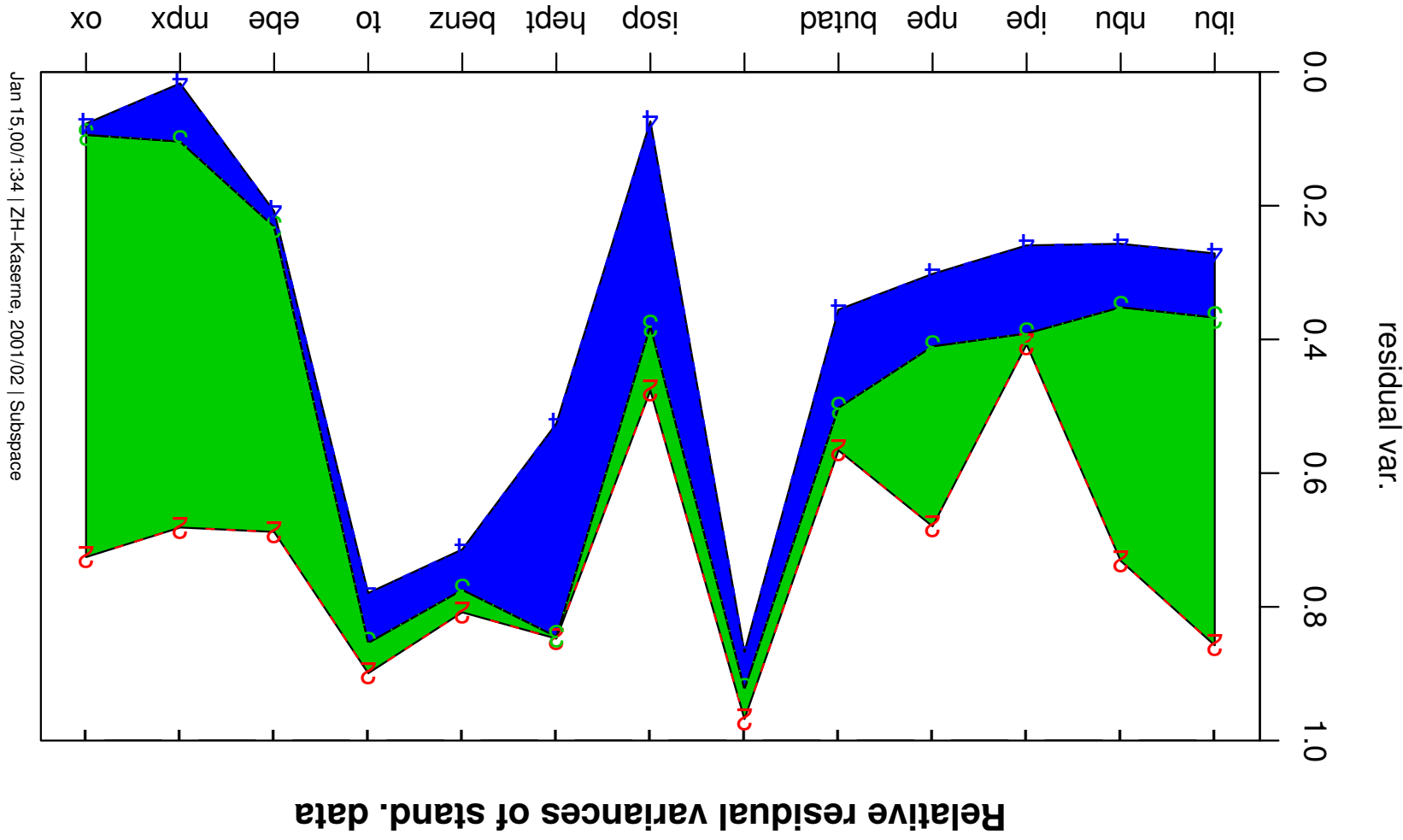


Choice of dimension d ?

- “elbow” in “scree” plot
- explained variance of the variables,

Choice of dimension p ?

- “elbow” in “scree” plot
- explained variance of the variables,

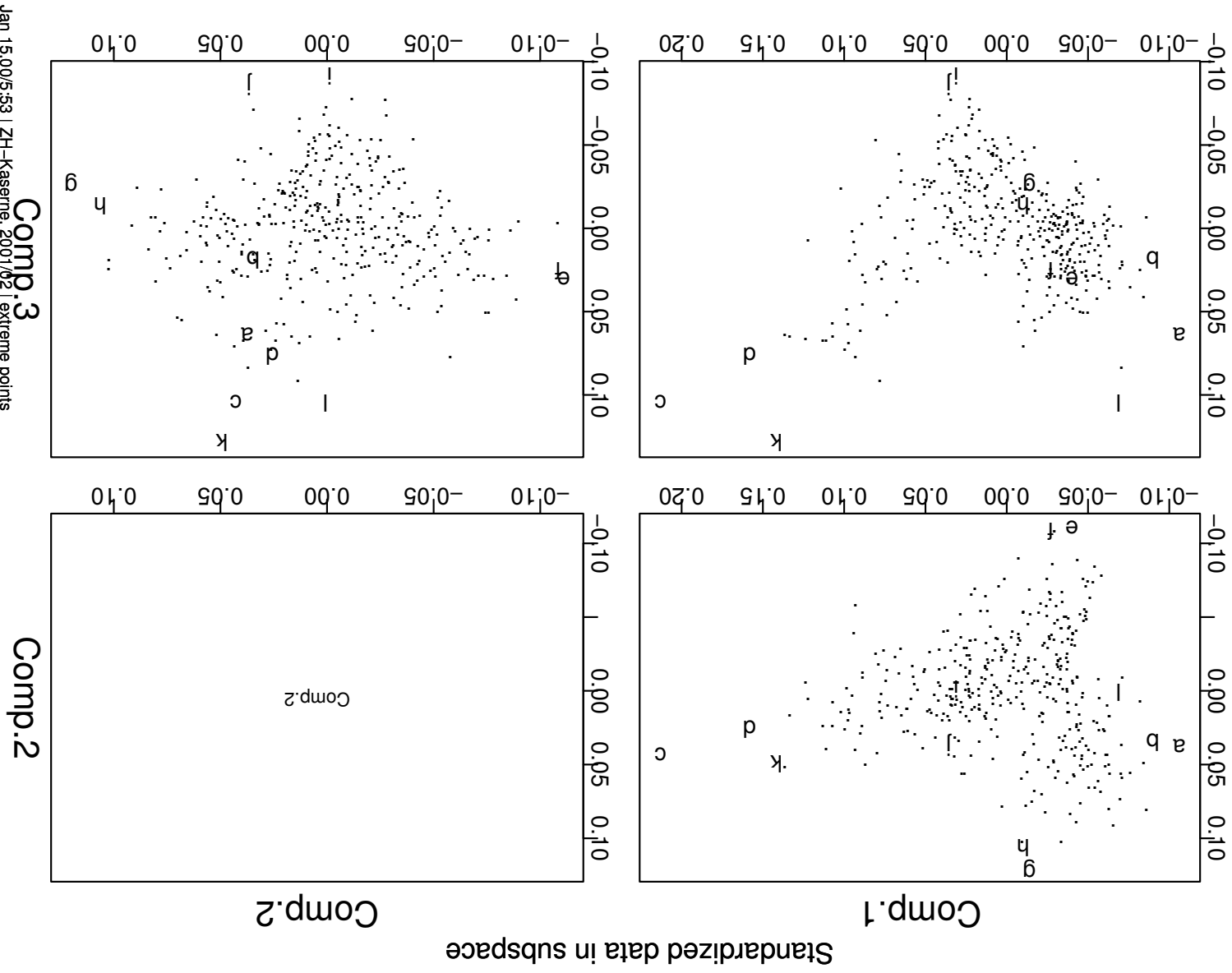


Choice of dimension d ?

- “elbow” in “scree” plot
- explained variance of the variables,
- time patterns in scores,
- lack of patterns in residuals.

Choose $d = 4$

Coordinates in subspace



Step 2:

“Abstract” factors C_o can NOT be source profiles.
→ Convert C_o into “real factors” C .

Change basis of subspace determined in step 1

(In FA, this is called “(oblique) rotation methods”).

Parameters S and C are not identifiable !

→ Need additional criteria for uniqueness (below)

or ad-hoc selection, e.g.

stable point of a specific iterative algorithm.

In our application: Use extreme points based on graphical inspection
(find polyhedron which contains the data).

→ 4 “corner points” = source profiles.

Further **ideas for Step 2** (“rotation methods”)

- Exploit non-negativity constraints! (Solution remains indeterminate)

Lit.: Lawton and Sylvestre (1971), Henry and Kim (1990),

Schostack and Malinowski (1991).

Alternating Least Squares: Modify initial solution

as little as possible to fulfill nonneg. constraints.

- Tracer compound: Koutrakis and Spengler (1987).

- “target testing”, Malinowski (1991).

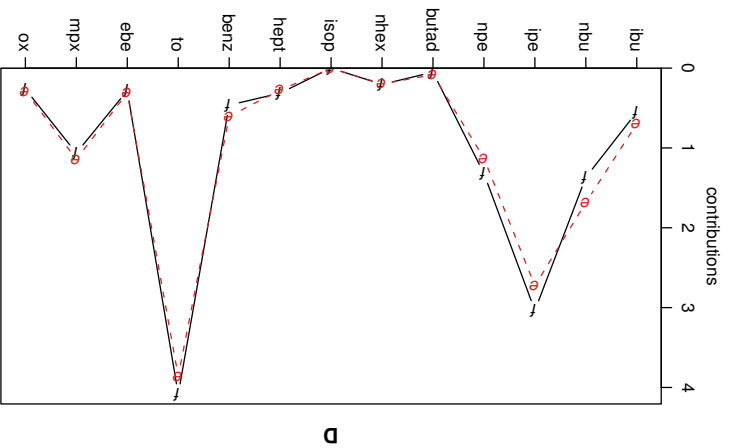
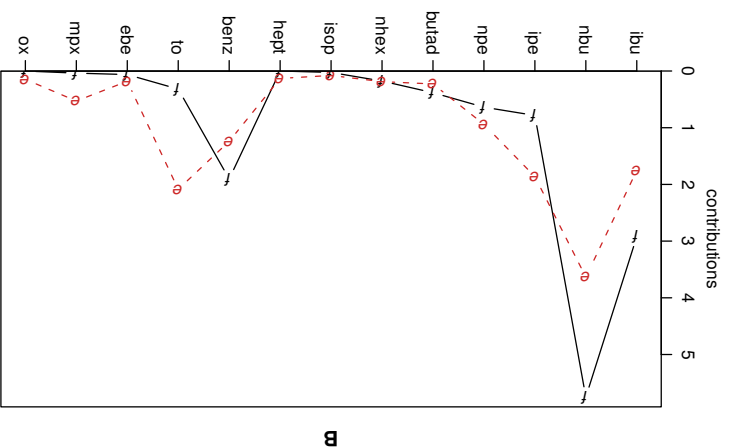
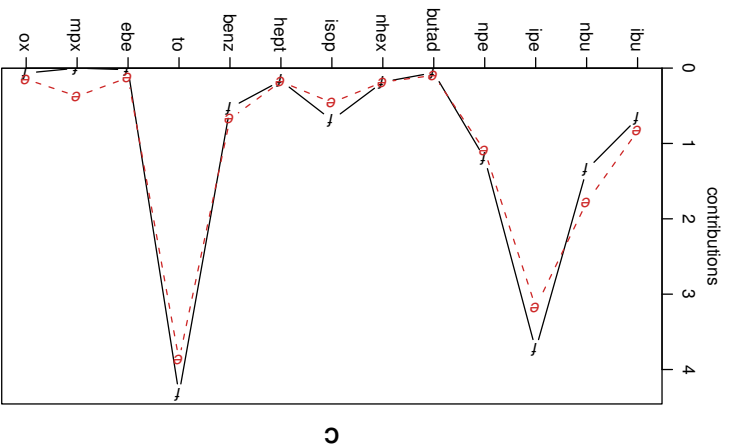
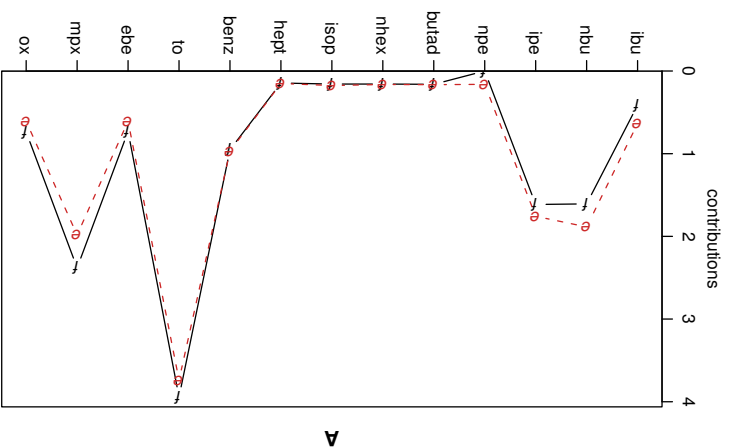
Direct factorization: Combines steps 1 and 2. → Positive Matrix Factorization

Anttila, Paatero, Tapper and Järvinen (1995), Paatero (1996).

4 Results

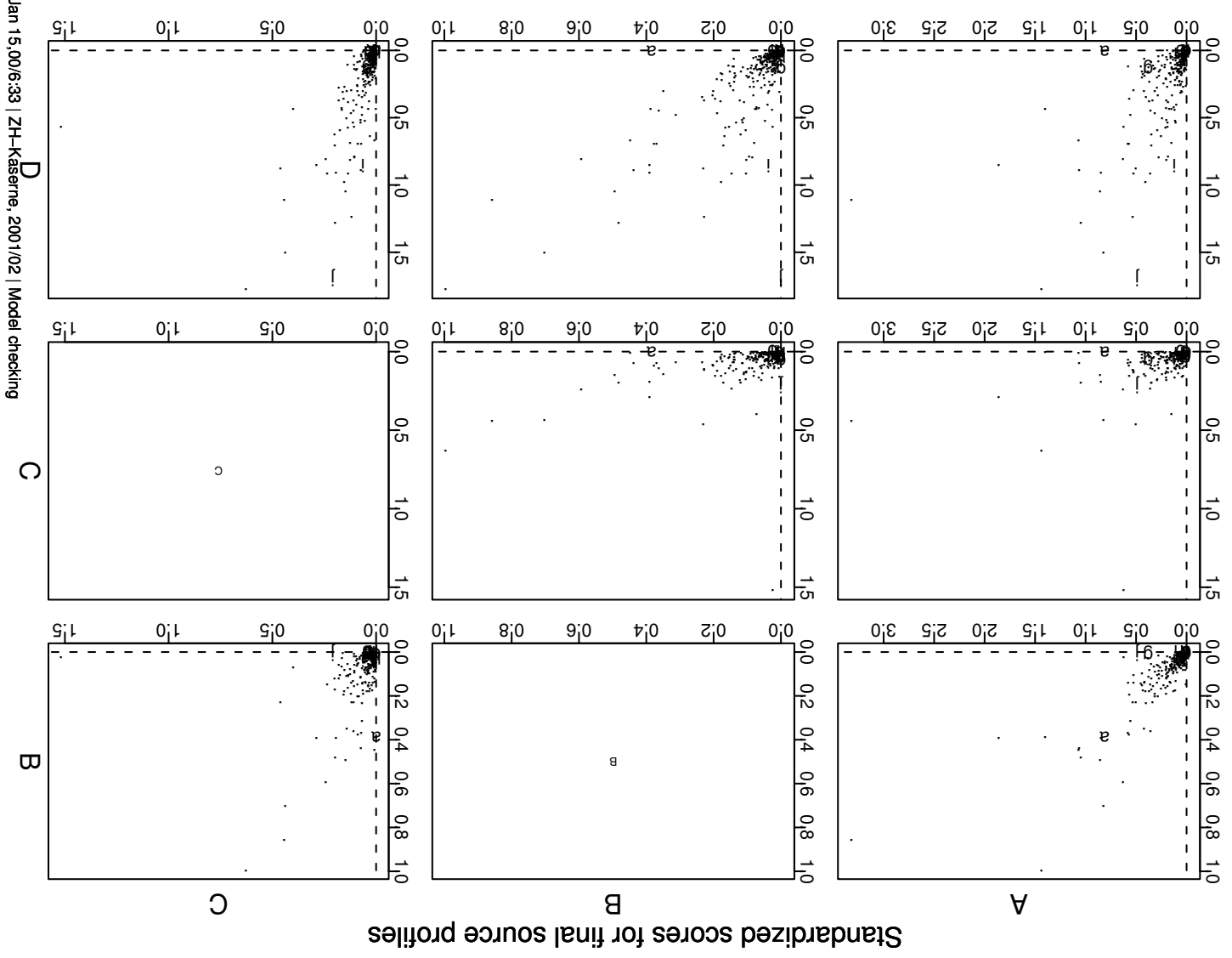
Source profiles.

Trial and final profiles

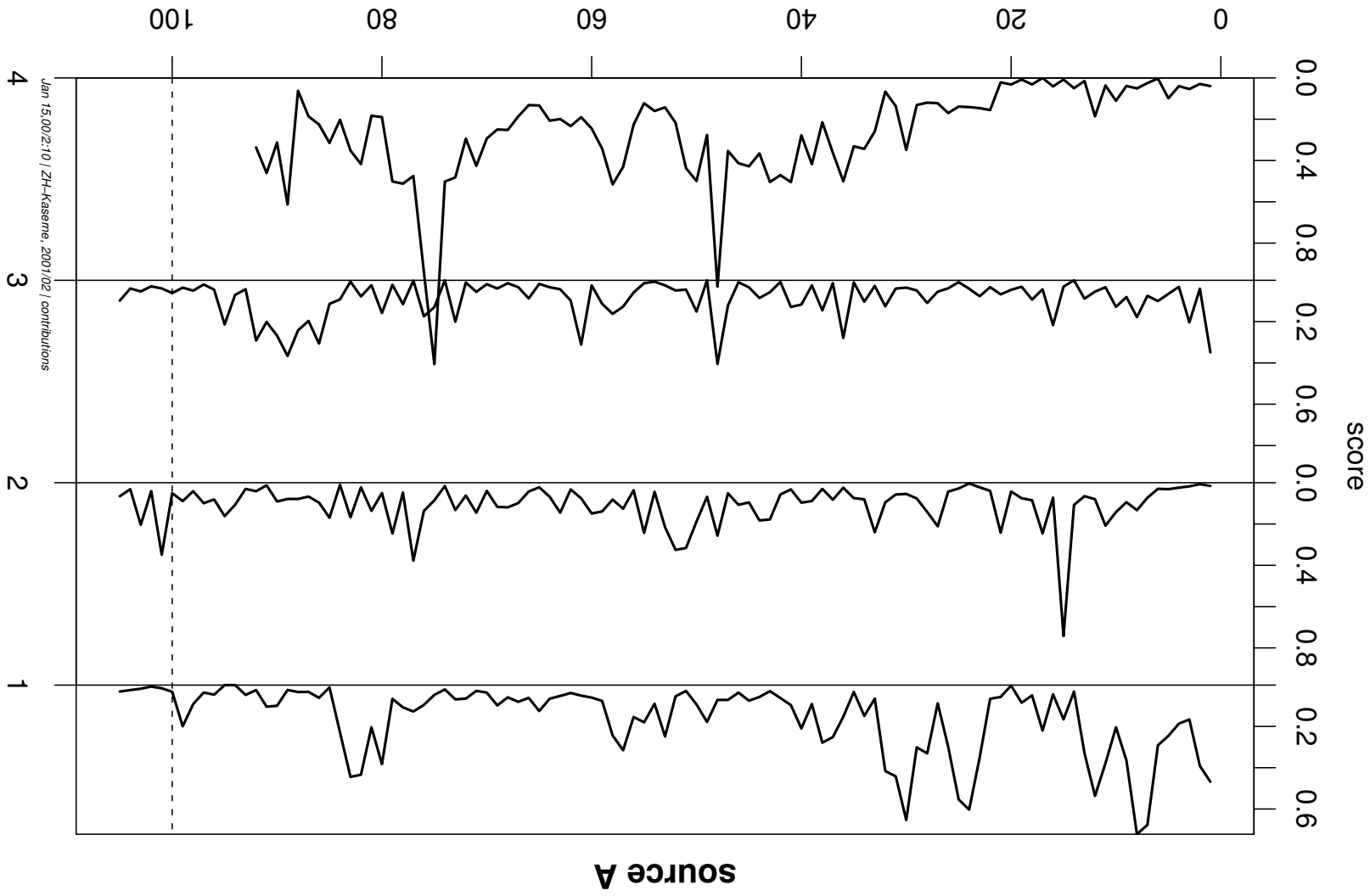


Activity scores

= coordinates with respect to new coordinate system given by corner points



Activities in time order



Interpretation of estimated source profiles:

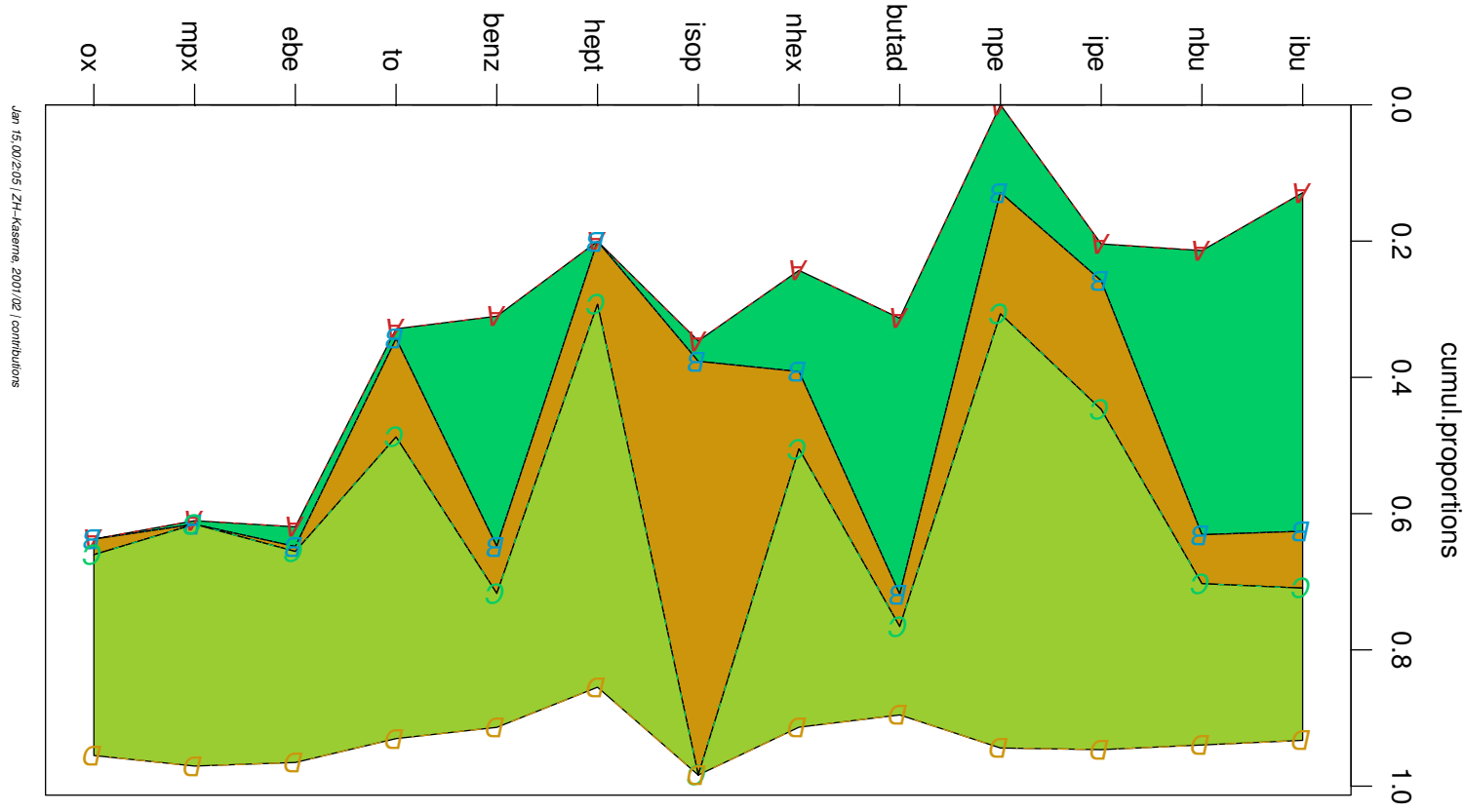
- **Prior knowledge** about composition of source profiles.
 - Calculate scores from unstandardized data. Interpret weekly pattern.
 - Profile b agrees with combustion profile of **passenger cars** from tunnel study, and is high at noon
 - Profile d: evaporation (and some combustion?)
 - Profile a: found also in urban places. Unclear...?
 - Profiles c and e ?
 - **Regression** of scores time series on explanatory time series like meteorology, activities, weekly patterns.

Contributions.

Mean Contributions of source k to variable j : $\gamma_k^{(j)} = \frac{\sum_{i=1}^n S_i^{(k)} C_i^{(j)} / n}{\sum_{i=1}^n X_i^{(j)} / n}$

If sums over variables make sense, calculate overall $\gamma_k = \sum_j \gamma_k^{(j)}$.

Mean contributions of the sources



5 Extensions

- $X = SC + T + \tilde{E}$ with T unknown sources, all $T_{(j)}^i$ are positive, \tilde{E} measurement error, $\tilde{E}_{(j)}^i \sim N\left(0, \sigma_{(j)}^2\right)$
 - together: allow $E_{(j)}^i = T_{(j)}^i + \tilde{E}_{(j)}^i$ to be > 0 more liberally than to be > 0
 - Robust “regression” quantiles.

- C corresponding to partly,

approximately known source profiles C_* ,

S random, obeying a regression model

with given explanatory variables (s.above).

→ Penalized likelihood, penalty for

- deviations $|C^{(j)} - C^{(j)*}|$

- deviations of scores from expected patterns.

- Treatment of values below detection limit possible.

Code in R available (at own risk).

Messages

- Data on many pollution compounds allow for **identification** of source profiles and contributions.
- There are applications of Bilinear Unmixing in several scientific areas.
- PCA or “factor analysis” (first step) are usually descriptive methods. Here, there is a stochastic model with **clear theoretical basis** ... and there is a parametric model.
- **Graphical methods** help a lot to **understand** and solve the problem.
- Extensions to more realistic assumptions and **use of additional information** are easy to design.

Literature

- Anttila, P., Paatero, P., Tapper, U. and Järvinen, O. (1995). Source identification of bulk wet deposition in finland by positive matrix factorization, [Atmospheric Environment](#) **29**(14): 1705–1718.
- C.Henry, R., Lewis, C. W., Hopke, P. K. and Williamson, H. J. (1984). Review of receptor model fundamentals, [Atmospheric Environment](#) **18**(8): 1507 –1515.
- Henry, R. C. and Kim, B. M. (1990). Extension of self-modeling curve resolution to mixtures of more than three components, [Chemometrics and Intelligent Laboratory Systems](#) **8**: 205–216.
- Hopke, P. K. (1991). [Receptor modeling for air quality management](#), Elsevier, Amsterdam.
- Koutrakis, P. and Spengler, J. D. (1987). Source apportionment of ambient particles in Steubenville, OH using specific rotation factor analysis, [Atmospheric Environment](#) **21**(7): 1511–1519.

Lawton, W. H. and Sylvestre, E. A. (1971). Self modeling curve resolution, *Technometrics* **13**: 617–633.

Malinowski, E. R. (1991). *Factor Analysis in Chemistry*, John Wiley & Sons.

Paatero, P. (1996). Least squares formulation of robust non-negative factor analysis, *Chemometrics and Intelligent Laboratory Systems* **762**: 1–13.

Renner, R. M. (1993). The resolution of a compositional data set into mixtures of fixed source compositions, *Appl. Statist.* **42**: 615–631.

Schostack, K. J. and Malinowski, E. R. (1991). Theory of evolutionary factor analysis for resolution of ternary mixtures, *Chemometrics and Intelligent Laboratory Systems* **10**: 303–324.

Weltje, G. J. (1997). End-member modeling of compositional data: Numerical-statistical algorithms for solving the explicit mixing problem, *Math. Geology* **29**: 503–549.

Wolbers, M. and Stahel, W. A. (2005). Linear unmixing of multivariate observations: A structural model, *J. of the American Statistical Association* **100**: 1328–1342.