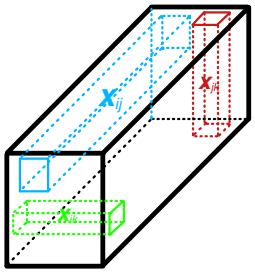
Multiple imputation and Three-mode analysis.

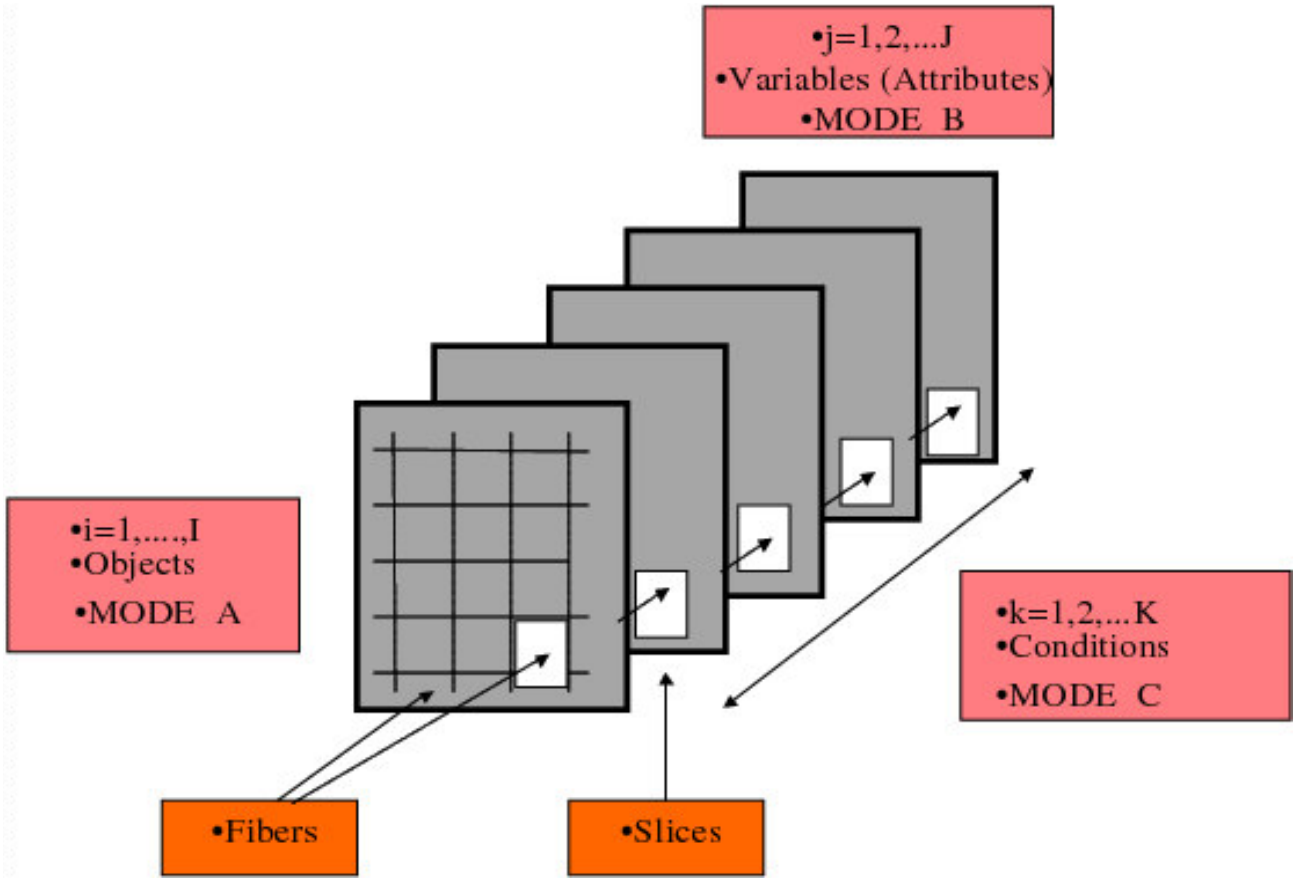
A research programme

Pieter M. Kroonenberg

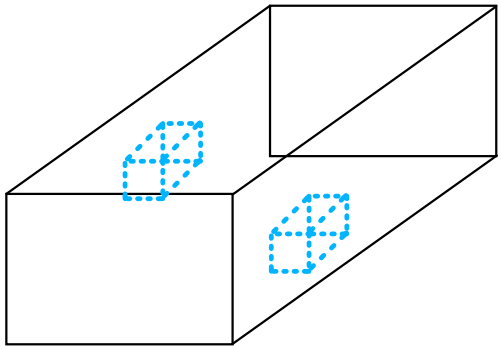
Leiden University



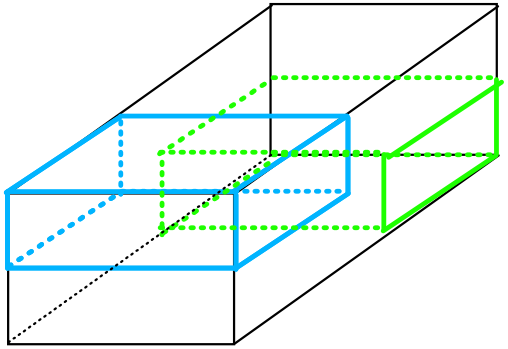
Three-mode data



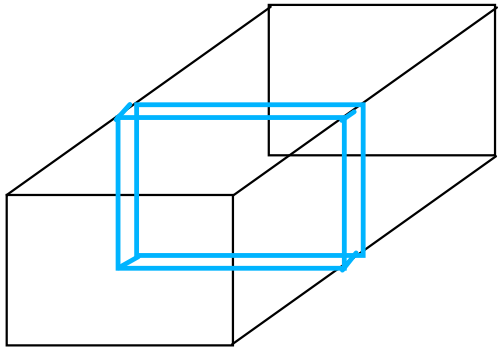
Missing data



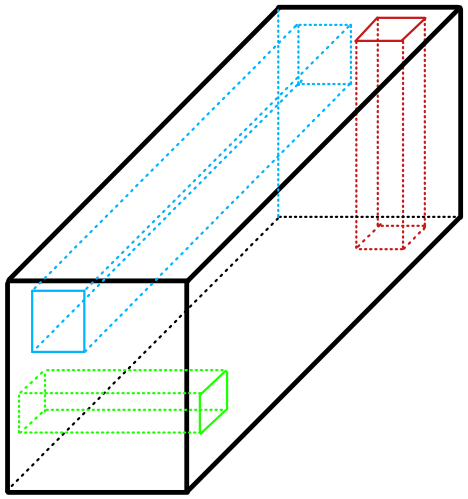
single observations



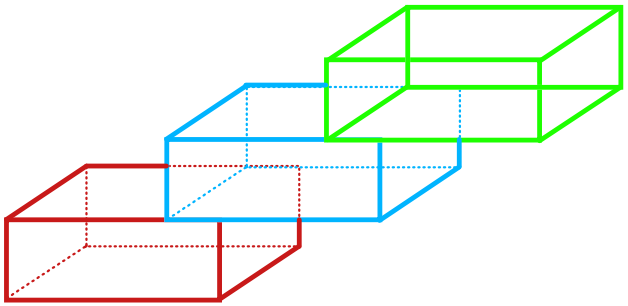
sub-blocks



slices

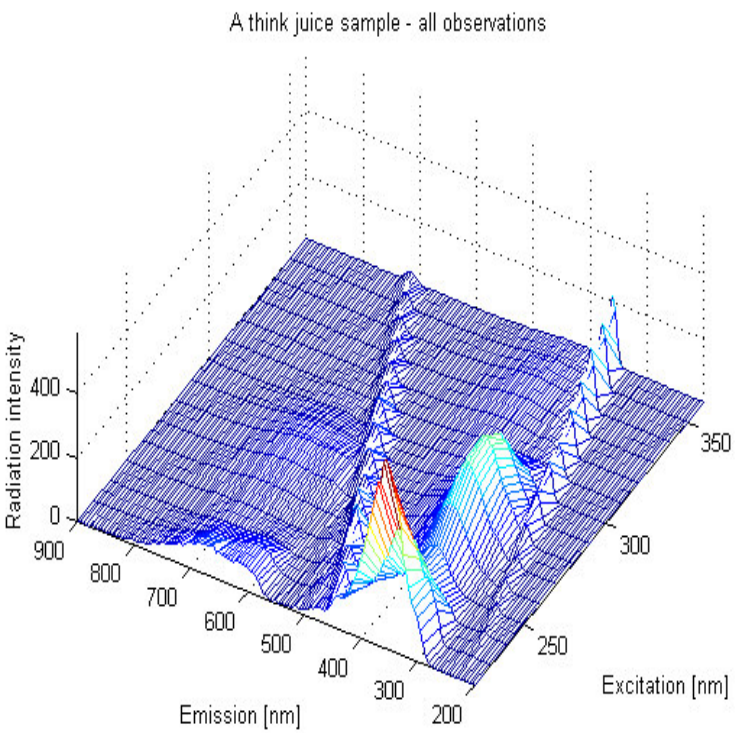


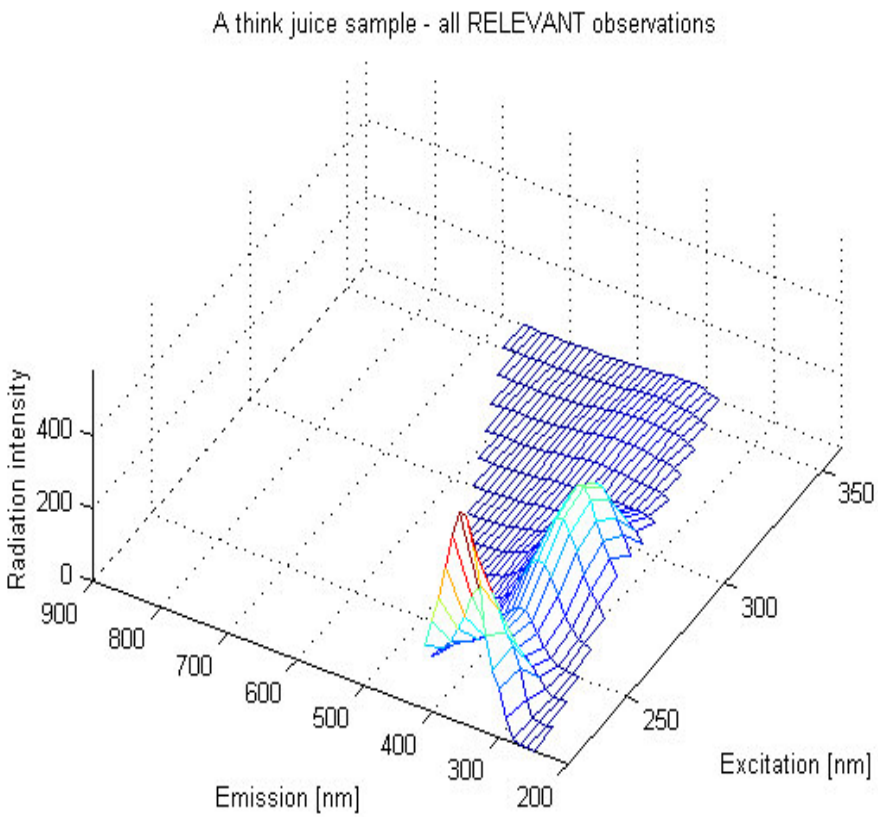
Columns, rows, tubes



linked modes (Harshman)

Creation of missing data





Missing values

- Second-order signals
- Light scattering
- Detector out of range

Source data:KVL, Bro & Ander

Origin of missing data

- **Missing completely at random**
  - Data are missing because of a random generating process
  - Cause of missingness is unrelated to the variable with the missing data
  - Deleting cases with missing data has no influence on representativeness, but diminishes power
- **Missing at random**
  - Cause of missing is systematic and correlated with the variable containing the missing data.
  - Cause is accessible and can be included in the analysis to correct for bias
- **Missing not at random**
  - Cause of missing is systematic and correlated with the variable containing the missing data. Often the variable is the cause itself and thus the cause not accessible
  - Cause is not accessible and cannot be included in the analysis to correct for bias

Little & Rubin (1987). *Statistical analysis with missing data*. Wiley;
Schafer(1997). *Analysis of incomplete multivariate data*. Chapman & Hall

Procedures

- **Expectation-Maximisation (EM)** via three-mode model:
  Estimate the missing data during iterations to determine the estimates of the model parameters

- **Multiple imputation** via data augmentation:
  Create several data sets with different values for the missing data and analyse each of them with a three-mode model, then combine the results

E(xpectation)-M(aximization)

Tucker3 Model: $x_{ijk} = \sum_p \sum_q \sum_r a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk}$

1. *Tuckals*: Express **G** in **A**, **B**, **C** and **X**; *Gepcam*: Skip this step
2. (Preprocess: centre and normalise)
3. Find reasonable *starting values* for **A**, **B**, **C(, G)** and for missing data.
4. Estimate *model parameters* of three-mode model
5. Estimate *missing values* using model parameters
6. (Recentre and renormalise)
7. Iterate till convergence

Eigenvalue-eigenvector based (Kroonenberg & De Leeuw)
Regression based (Weesie & Van Houwelingen - Gepcam).
Missing data estimates are continously updated.

E(xpectation)-M(aximization)

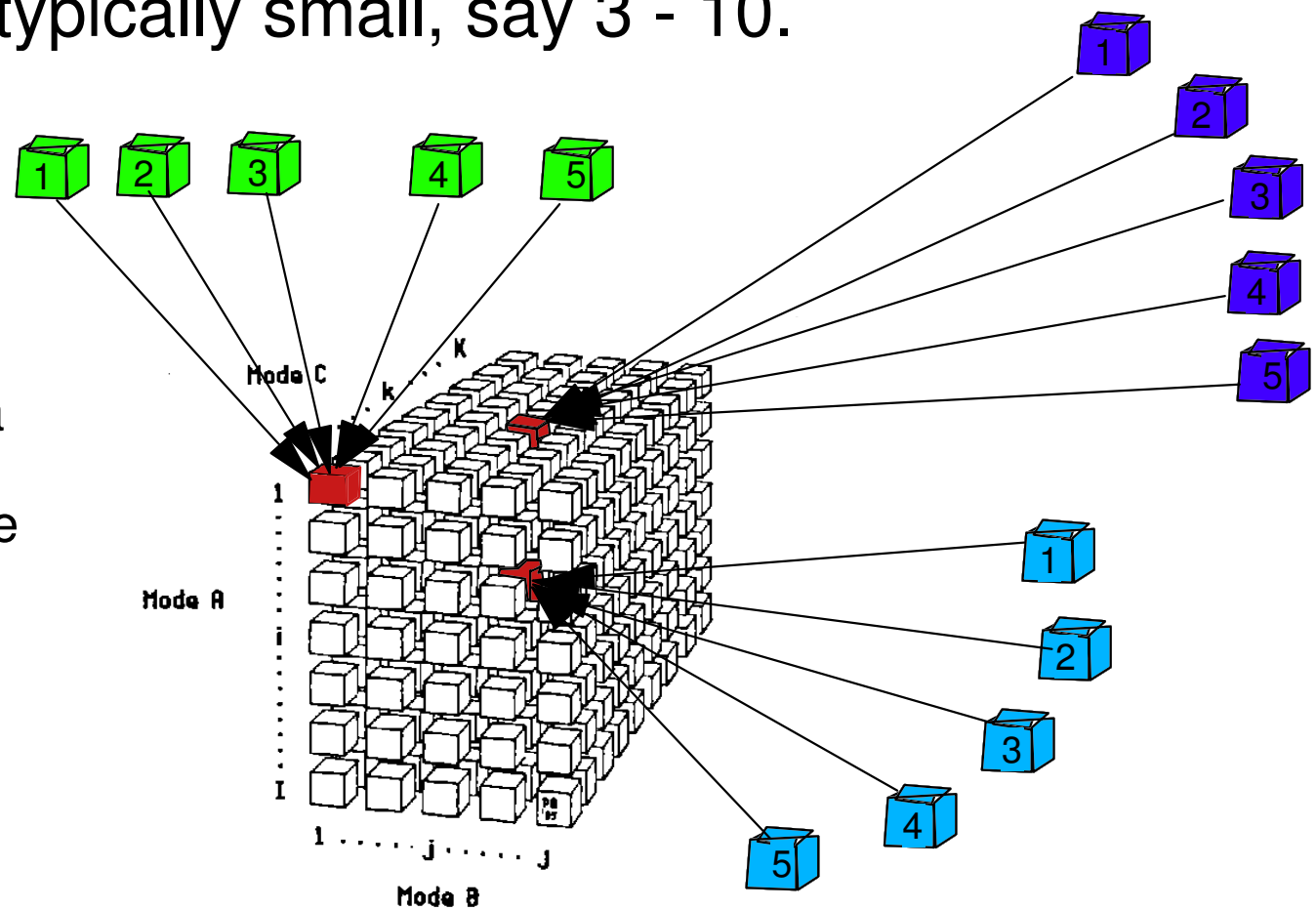
Limitations

- Single imputation
- Missing data estimates are tailored to the model.
- Model fits the (augmented) data too well
- Underestimation of sampling variability
- No estimate of uncertainty due to missing data
- Missing data estimates have no sampling errors

Multiple imputation: Basics

Multiple imputation is a Monte Carlo technique in which missing data are replaced by $m > 1$ simulated versions, where m is typically small, say 3 - 10.

Creation of 5 data sets with different imputations for the missing data



missing data point

Multiple imputation: Basics

- Validity imputations depends on the method of generation of the imputations
- Often normality of the original scores assumed
- Rubin:
  - Specify a parametric model for the complete data
  - Apply a prior distribution to unknown model parameters
  - Simulate *m* independent draws from conditional distribution of missing values given the observerd ones by Bayes' theorem

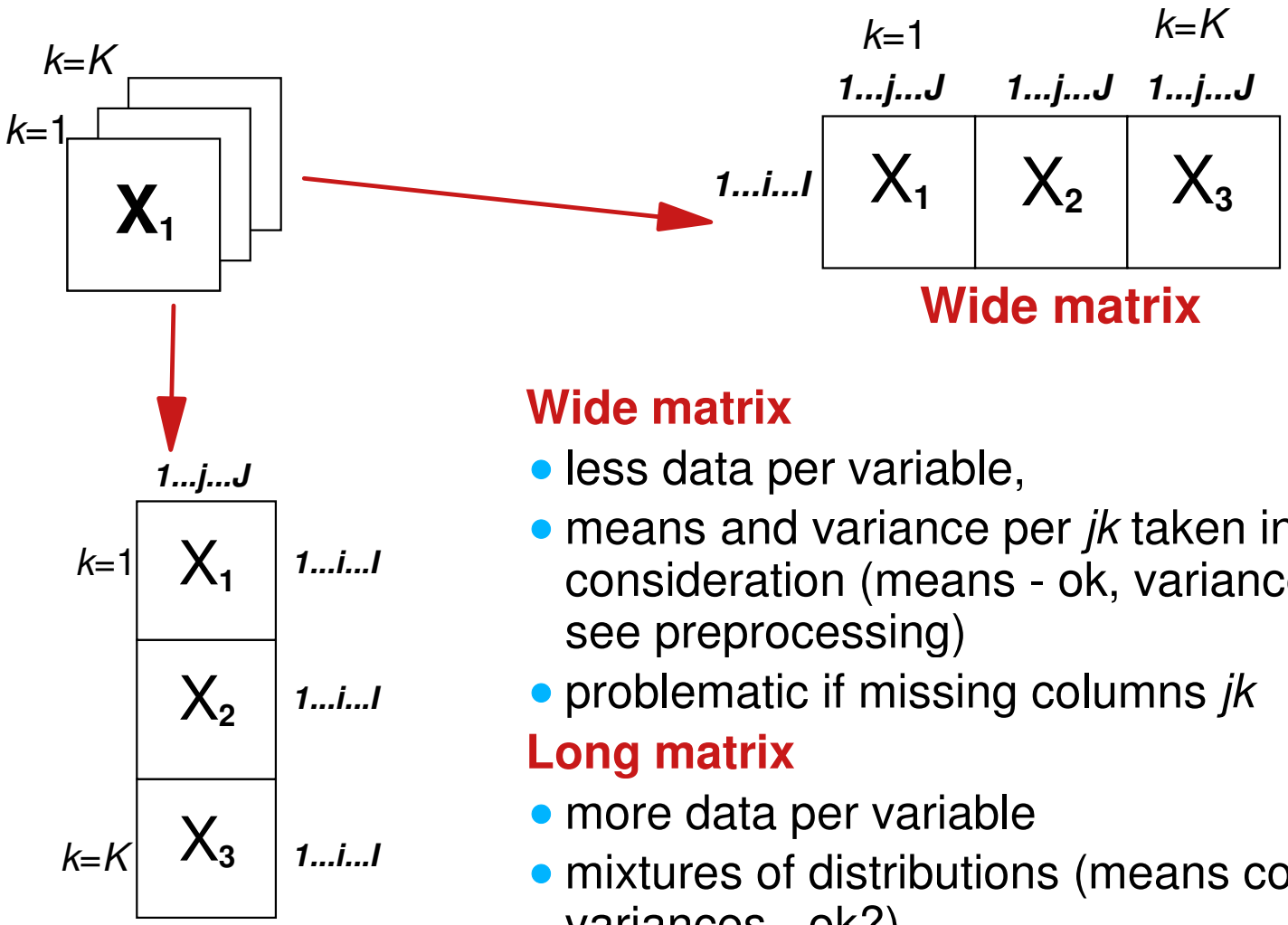
Multiple imputation: Generation

Schafer (using Tanner-Wong's data augmentation procedure)(1997)*

- **iterative two-step process:**
  - alternatingly sample missing values from their conditional predictiv distribution
  - then sample unknown parameters from a simulated complete-data posterior distribution.
- given initial values of the parameters this defines a Markov chair which converges to a stationary distribution of the missing values and the parameters, given the observed data
- iteration produces a draw of the parameters from its observed da posterior distribution and a draw of the missing values from the distribution of the missing values given the observed ones

*Description taken from Schafer, J. L. (1999) in *Stat. Meth. Med. Res., 8,* 3-15

Multiple imputation: Three-way



Wide matrix

- less data per variable,
- means and variance per *jk* taken into consideration (means - ok, variances - not?; see preprocessing)
- problematic if missing columns *jk*

Long matrix

- more data per variable
- mixtures of distributions (means confounded, variances - ok?)
- missing column = missing slice =>delete it

Special procedures necessary?

Multiple imputation: Stochastics

- **With sampling framework**
  - cases x variables x conditions
- **Without sampling framework**
  (single observation (or mean) per cell)
    - varieties x attributes x locations
    - wavelengths x wavelengths x concentrations
    - solutes x eluents x adsorbents
  - Distributional assumptions for multiple imputation valid?
  - Estimate missing values some way and add normal error distributions per cell of three-way array with external standard errors for parameters to create multiple data sets? (add measurement error)

Multiple data sets, multiple solutions

- 10 imputed data sets
- 10 Tucker3 (Parafac) solutions
  - 10 Solutes component spaces
  - 10 Adsorbents component spaces
  - 10 Eluents component spaces
  - 10 Fit measures

How to combine it all?

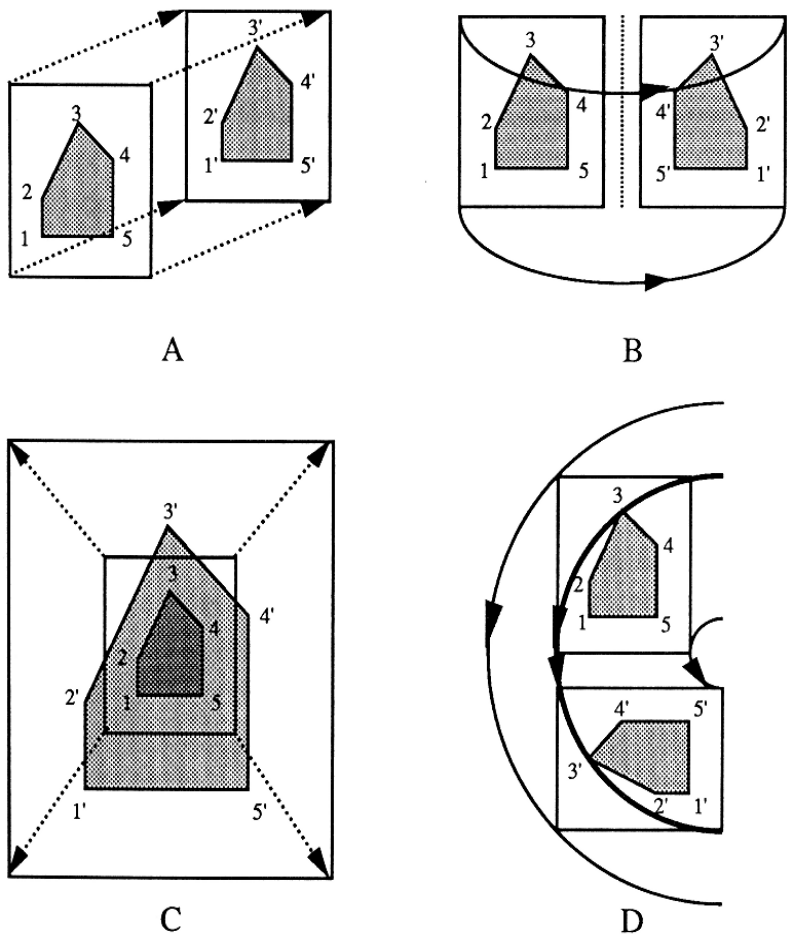
Standard MI - per parameter standard errors

Here: Invariant subspaces with rotatable axes

Options

- Generalised `Prokrousthj` analysis on all imputed spaces
  - including the E-M solution
  - only imputed data, fit E-M solution into the centroid space for comparison
- First E-M solution and use that solution as target from imputed data: Target rotations

Matching spaces via Generalised `Prokrousthj` analysis



~~A. Translation~~
B. Reflection
~~C. Isotropic Scaling~~
D. Rotation

First find iteratively a centroid
then determine the optimal
transformation to the centroid

Examples

Chromatography

- Data from De Ligny et al.
- Liquid chromatography

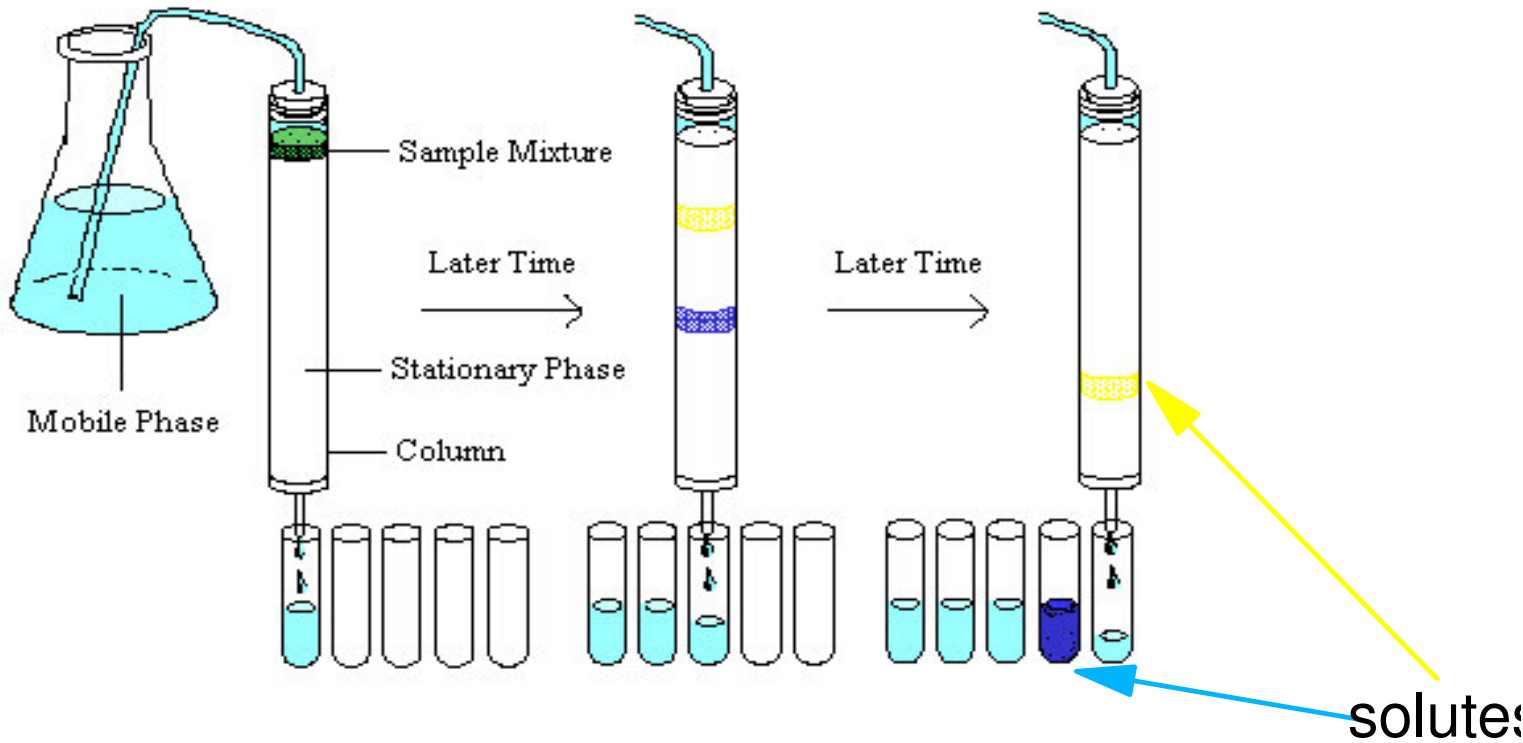
Child development

- Data from the child care study of the NICHD
- Development in family background variables

Chromatography

De Ligny, Spanjer, et al.

Liquid chromatography



1st mode: Solutes - monosubstituted phenols, anilines, pyridines
2nd mode: Stationary phase = adsorbents
3rd mode: Mobile phase = eluents
Measurement: Retention rate = log (net retention volume)/weight of absorben

Source picture: http://falcon.sbuniv.edu/~ggray/CHE3345/chp24.html

Data De Ligny et al.

Data

- *Dependent variable*: **Retention rate** in High Performance Liquid Chromatography (HPLC)
- 39 solutes (bisubstituted benzenes) x 3 adsorbents x 2 eluents
- 21 missing data (= 9%); *cause? retention too long?*
- 5 rows for the 1st eluent have 1 valid and 2 missing observations. No missing for 2nd eluent.
- No preprocessing (=> 1st components primarily means)

Purpose of the orginal analysis (De Ligny et al.)

Get estimates for the missing data, but not today

Structure is also interesting; present focus.

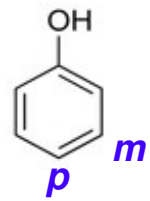
Question

How does the presence of the missing data influence the relationships between solutes?

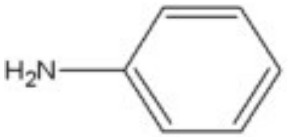
Source data: De Ligny, C.L., et al. (1984). *Journal of Chromatography*, *301*,311-323

Data description

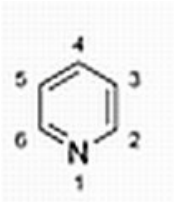
	Solutes	
Phenols	**Anilines**	**Pyridines**
*m*F	*m*F	
*p*F	*p*F	
*m*Cl	*m*Cl	3Cl
*p*Cl	*p*Cl	
*m*Br	*m*Br	3Br
*p*Br	*p*Br	
*m*CH3	*m*CH3	
*p*CH3	*p*CH3	4CH3
*m*OCH3	*m*OCH3	
*p*OCH3	*p*OCH3	
*m*NO2	*m*NO2	
*p*NO2	*p*NO2	
*m*CN	*m*CN	3CN
*p*CN	*p*CN	4CN
*m*COOCH		
*p*COOCH		
*m*COCH3	*m*COCH3	
*p*COCH3	*p*COCH3	



phenol



analine



pyridine

p=para *m*=meta

Adsorbents = stationary phase

Octadecyl-silica
N-cyanoethyl-N-methylamino-silicia
Aminobutyl-silica

Eluents = mobile phase

---

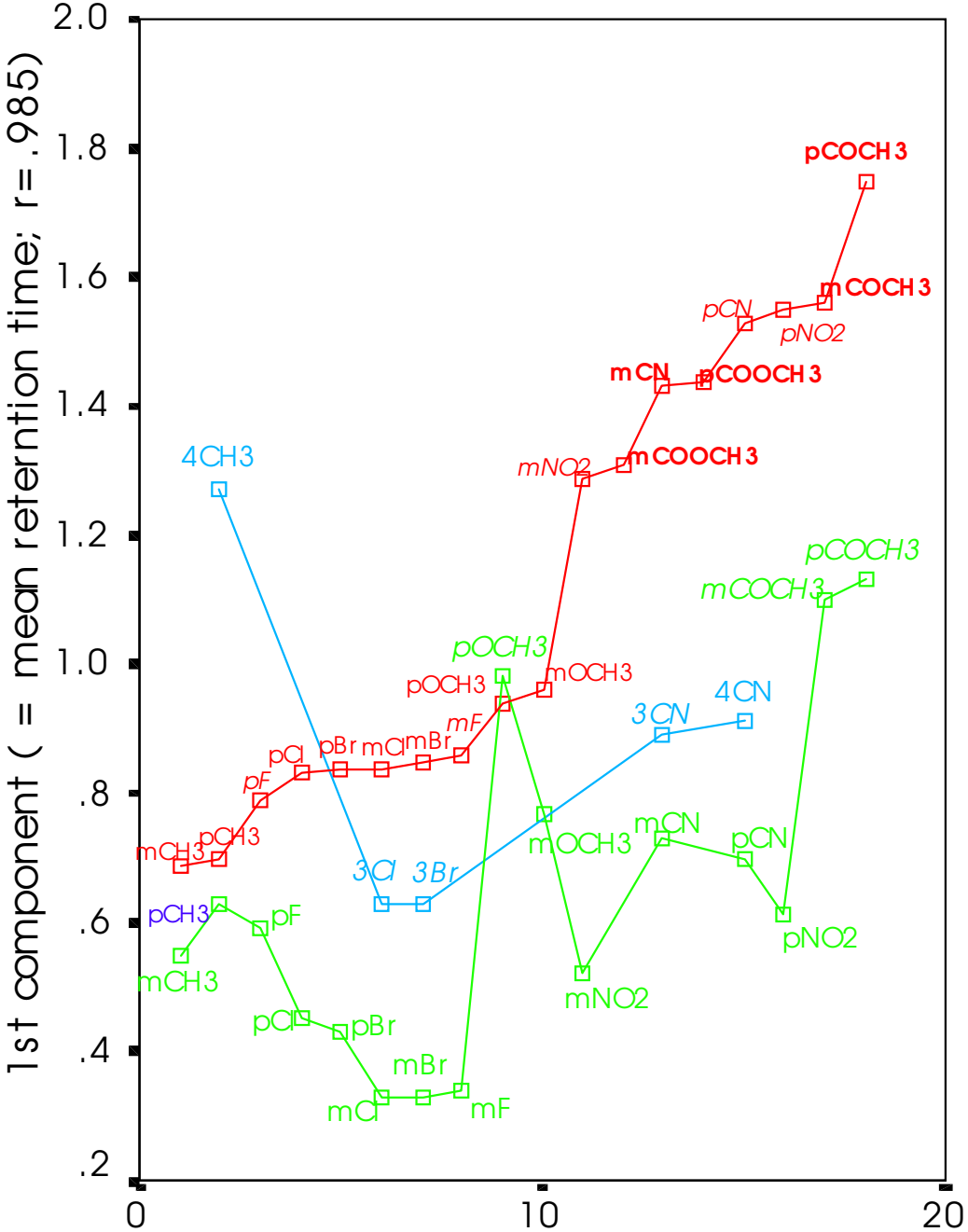
35 v/v% methylene chloride in n-hexar
pure methylene chloride

E-M solution

- Parameter estimation via
  - Tuckals algorithm -- eigendecomposition-based
  - Gepcam algorithm -- regression-based
  - Gepcam slightly more stable with very high fit

- Solution:
  - no preprocessing - all means included
  - 3 solutes components
  - 2 adsorbants components
  - 2 eluent components ($K = R$)
  - Proportion fitted sum of squares =
    .9978 -- based on valid data
    .9984 -- SS(Total) includes estimates missing data

Solutes 1st component



Means eluents: approx. equal

Means adsorbents:
35% vs. 100% methylene chloride =1

italics = 1 missing data point
bold = 2 missing data points
out of three measurements

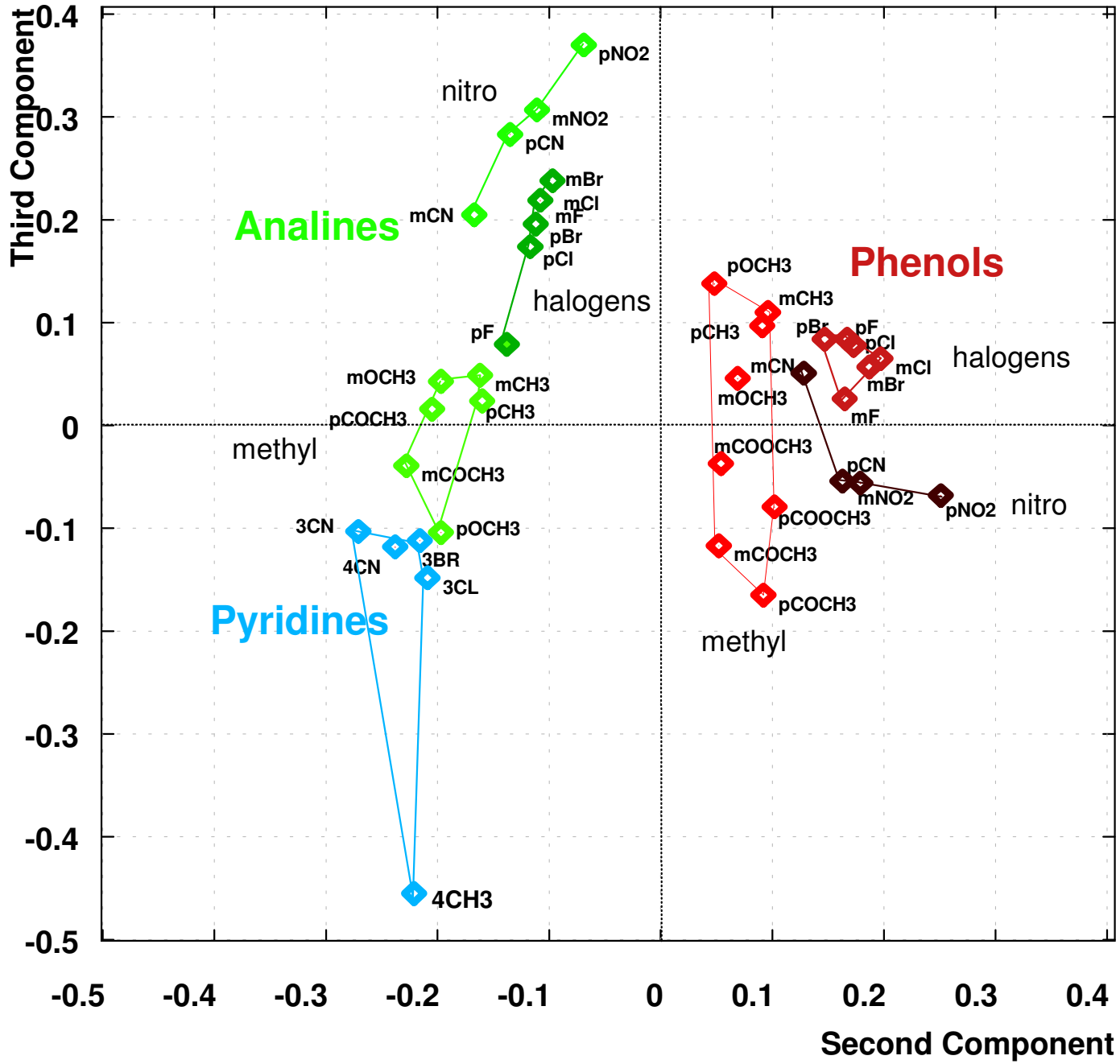
Solutes with substituents containing CH_3 or Nitrogen retention for para-isomers seems longer than for meta-isomers

Phenols

Pyridines

Anilines

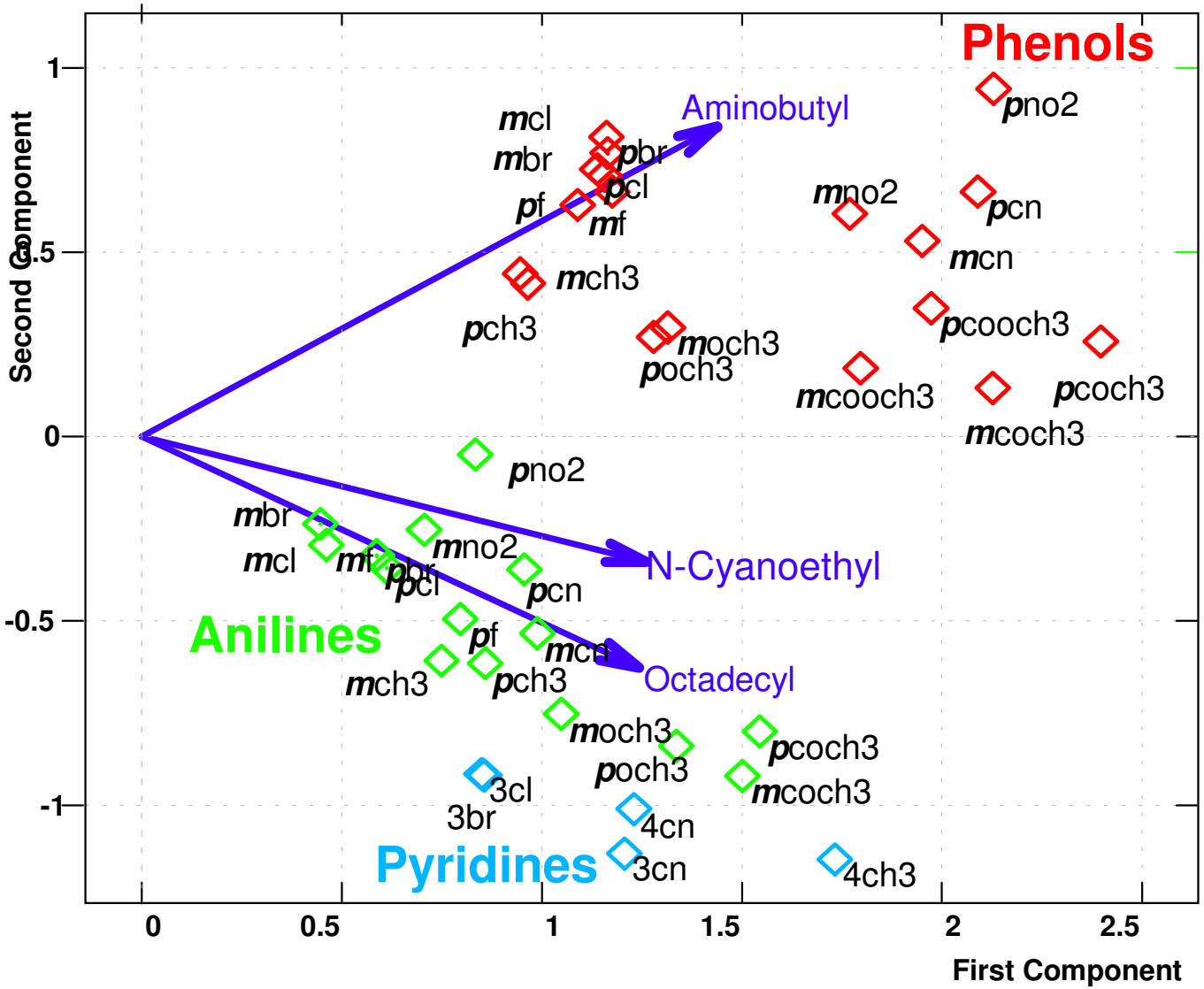
Solutes - 2nd and 3rd components



06/08/05 13:14:36

Joint plot for solutes and adsorbents

(First (consensus) component eluents - 97.4%)



Anilines & Pyridines tend to have relatively longer retention rates for both eluents with N-Cyanoethyl and Octadecyl silicates, while the Phenols tend to have relatively longer retention rates with Aminobutyl silicate, especially those with substituents containing nitrogen.

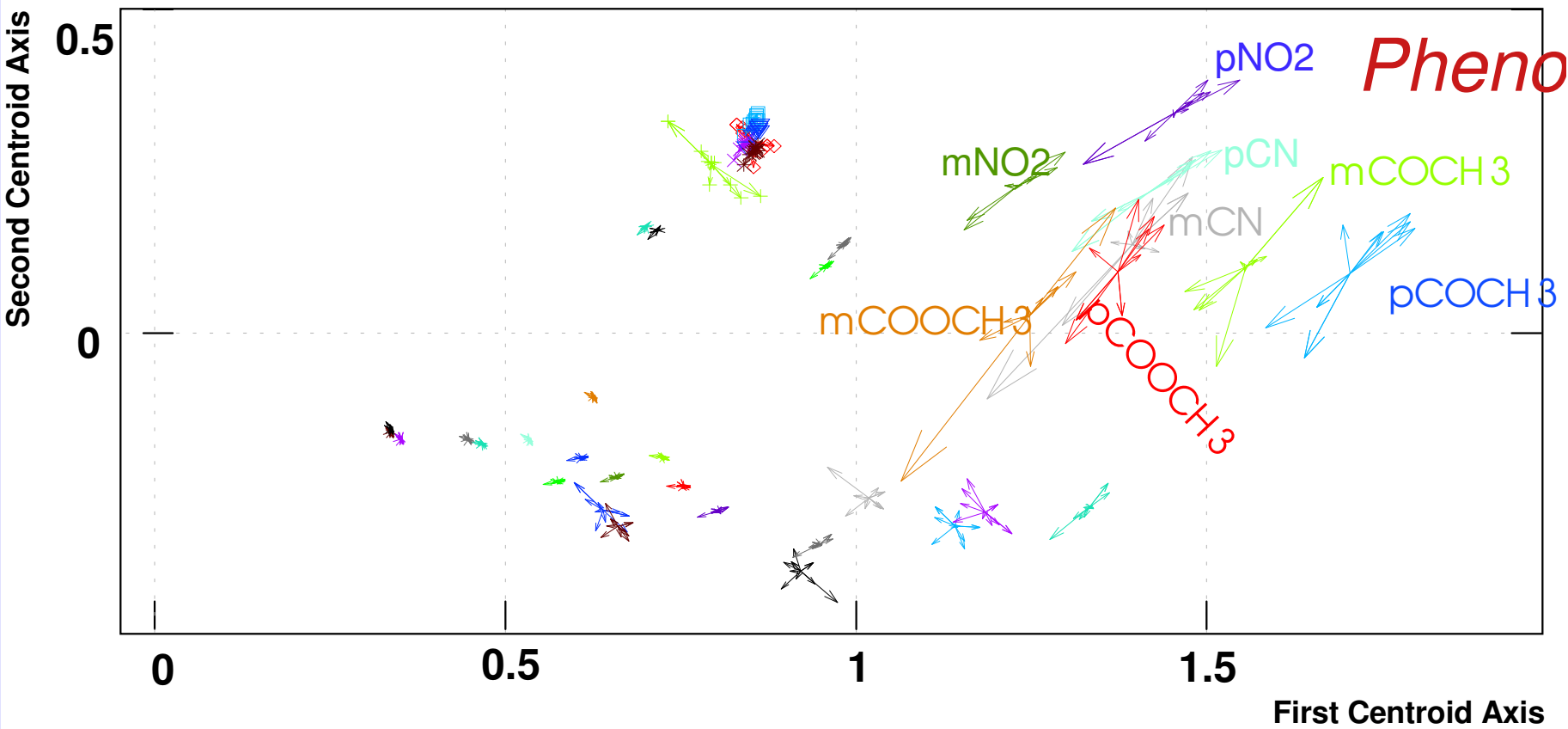
Proportional fit - MI & E-M
(Components)

		Minimum	Maximum	Mean	Std. Deviation	E-M
Total		.993	.997	.995	.0013	**.998**
Solutes						
	1	.914	.928	.922	.0041	**.928**
	2	.052	.055	.054	.0013	**.057**
	3	.016	.025	.020	.0034	**.014**
Adsorbents						
	1	.937	.946	.942	.0023	**.942**
	2	.051	.056	.053	.0013	**.057**
Eluents						
	1	.964	.972	.968	.0029	**.974**
	2	.025	.030	.027	.0020	**.024**

E-M is generally higher because no error for missing data

Optimally matched configurations

(Variability due to imputed values)



Points coloured by solutes; arrows start in mean configuration; bold = 2 missing & 1 valid value

Measure for stability of solutes (cases):

Sum (Average) of the squared distances to centroid (or E-M solution)

Unfinished business

- ***Compare estimates missing data and their standard errors for***:
  - E-M solution 3x2x3-solution (De Ligny et al.)
  - E-M solution 3x2x2-solution
  - Multiple imputation estimates
  - Estimated data values from the analyses of the 10 imputated data sets
  - Evaluate the location of E-M solution with respect to the solutions of imputed data sets

Child Development

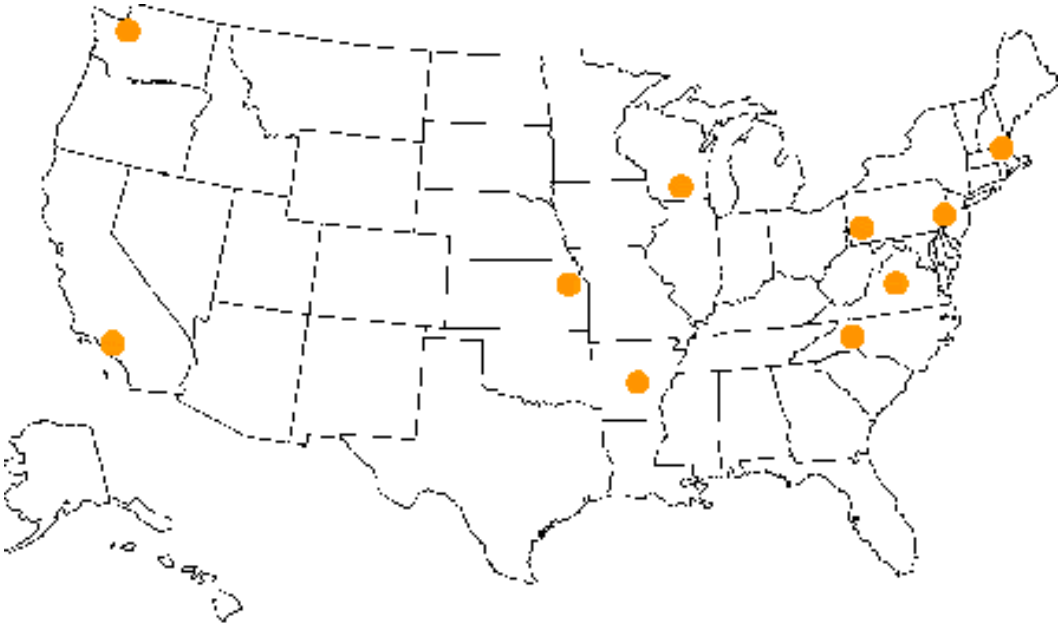
NICHD

The National Institute of Child and Human Development
Study of Early Child Care and Youth Development

Study of Early Child Care and Youth Development

The SECC is a large longitudinal study started in 1989 to answer all kinds of questions with respect to the effects of child care.

Origin of the samples



http://secc.rti.org

Data description

Present subset of the roughly 1300 families:

150 Afro-American families

11 Variables (see next slide)

4 Points in time: 6, 15, 24, 36 months after birth

Purpose of the analysis

Determining the structure of the family situation and its changes in the first three years after birth of the baby.

Questions

How does the presence of the missing data influence the relationships between variables?

Does the structure of the variables change over time?

Missing data

Missing Patterns

Number of Cases	HealthBaby 36	HealthMother 36	SatisfiedWork 36	HoursWork/Week 36	SocialSupport 36	FinanacialResources 36	Maternal Deprression 36	Parenting 36	LogTotalIncome 36	HealthMother 24	SatisfiedWork 24	HealthBaby 24	HoursWork/week 24	LogTotalIncome 24	MaternalDepression 24	FinancialResources 24	SocialSupprt 24	Parenting 24	LogIncome/NeedRatio 36	LogIncome/NeedRatio 24	LogIncome/NeedRatio 15	LogIncome/NeedRatio 6
5																				X		
3																				X		X
5																						X
5																					X	X
15																			X	X	X	X
10															X	X	X	X				
7	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		

Patterns with less than 2% cases (2 or fewer) are not displayed.

Variables & their means over time

Abbreviation	Description	months 06	15	24	36
HrWrkM-xx	Hours/week mother works-all jobs	17.1	21.3	21.0	20.9
Satisf-xx	Mom satisfied with own work schedule	3.6	3.8	3.6	3.5
Depres-xx	Maternal depression	11.9	11.2	13.1	11.9
Suppor-xx	Social Support	5.0	4.8	4.6	4.7
PStres-xx	Parenting stress*	51.0	34.3	35.7	34.7
HealtM-xx	Health of mother	3.2	3.1	3.0	2.9
HealtB-xx	Health of baby	3.3	3.1	3.2	3.2
HrCare-xx	Hours/week in care	23.6	26.1	24.4	26.8
Financ-xx	Financial resources	9.3	9.3	9.2	9.4
Income-xx	Log total income	9.7	9.7	9.8	9.9
Need -xx	Log income to need ratio	.3	.2	.4	.4

xx = 06, 15, 24 or 36; indicating observed in the xx month after birth.

*different instrument at 6 months

E-M results

Fundamental results

- Number of components for a Tucker3 model:
  3 (subjects) x 3 (variables) x 1 (time)

- The coefficients are virtually equal for the four time points: Structure variables hardly changed over time.

- We might as well average over time points:.
  Tucker3 analysis is then equivalent to an SVD (PCA) on the subject-x-variable matrix averaged over time.

- Multiple imputation over wide matrix, thereafter standard preprocessing

$$\tilde{x}_{ijk} = (x_{ijk} - \bar{x}_{jk}) / s_j$$

Fit results

5 Imputed data sets

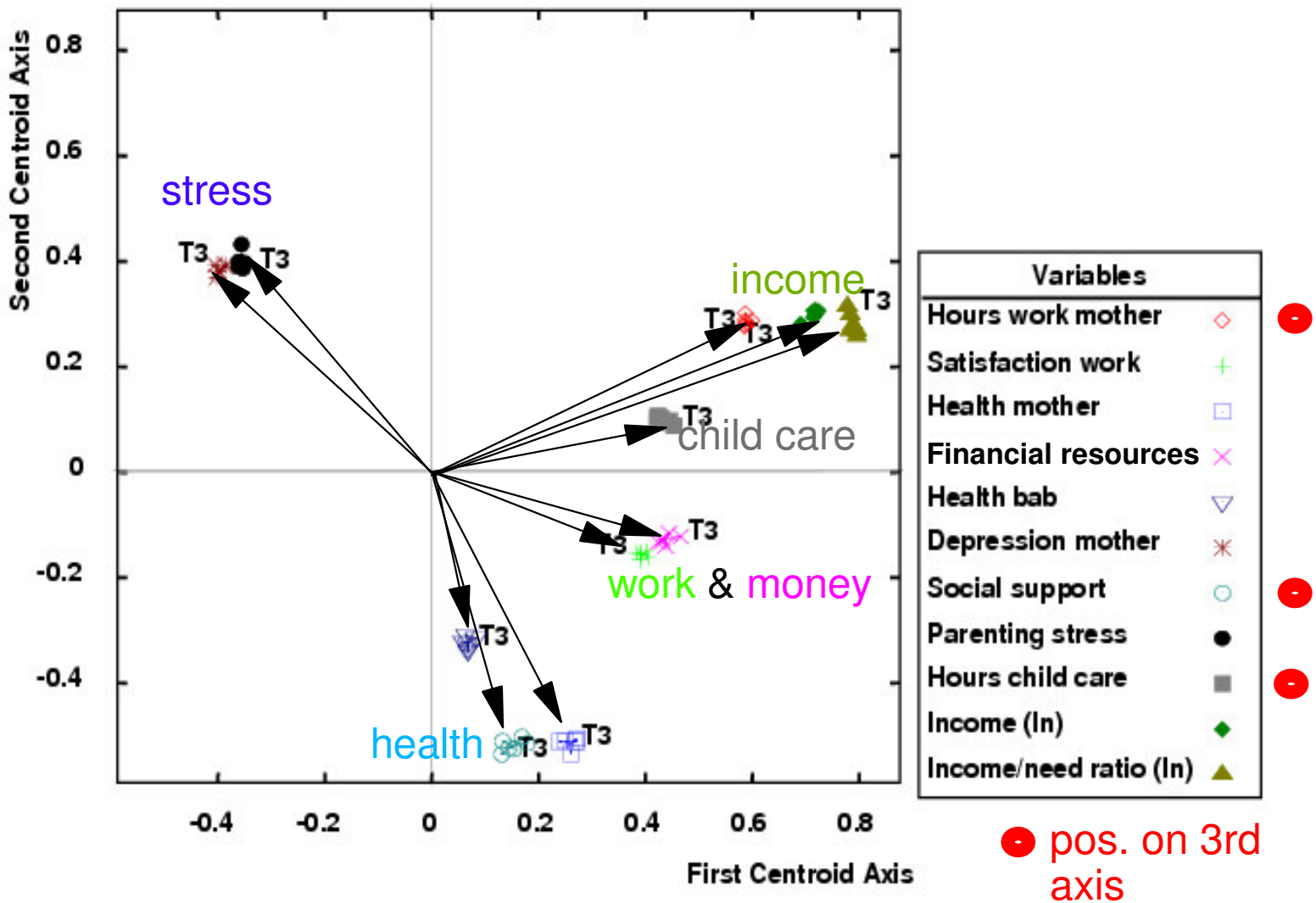
Solution	SS(Fit)	Proportional fit per component		
		1	2	3
Base solution	.415 (.382)	.229	.123	.063
Equal weights	.414 (.381)	.228	.122	.063
Time component				
Imputation 1	.388	.220	.109	.059
Imputation 2	.383	.216	.110	.058
Imputation 3	.380	.211	.113	.056
Imputation 4	.382	.212	.111	.059
Imputation 5	.390	.218	.113	.058

excluding missing values from SS(Total)

including missing values from SS(Total)

First components of E-M explain relatively more; probably due to the tailoring of missing data to the model (to be seriously investigated)

Variables spaces



- pos. on 3rd axis

Income/need ratio had 183 missing compared to Stress,Support, Depression with around 50.

Research programme

- **Large scale questions**
  - Multiple imputation via wide or long matrix?
  - Multiple imputation and means, standard deviations, and recommended preprocessing, i.e. three-way multiple imputation?
  - Multiple imputation and lack of stochastics in three-way data? Use external information, e.g. standard deviations from earlier studies, in multiple imputations?
  - Rotation to a target (=E-M solution) rather than to centroid?

Some (random?) comments

"I must add that even doing multiple imputation relatively crudely, using simple methods, is very likely to be inferentially far superior to any other equally easy method to implement (e.g., complete-cases, available cases, single imputation, Last Value Carried Forward) because the multiple copies of the data set allow the uncertainty about the values of the missing data to be incorporated into the final inferences;"

Rubin on www.statsol.ie/solas/rubin1.htm

The results suggest a reliable and efficacious basis for *imputation method for repeated measures data* is to substitute a missing datum with a value from another individual who has the closest scores on the same variable measured at other timepoints, or the average value of four individuals who have the closest scores on the same variable at other timepoints.

Elliott P, Hawthorne G. Aust N Z J Psychiatry. 2005 Jul;39(7):575-82.

A final comment

"Analysing data that you do not have is so obviously impossible that it offers endless scope for expert advice on how to do it."

Ranald R. MacDonald, University of Stirling,UK.
www.psychology.stir.ac.uk/staff/rmacdonald/Missing.htm;
seen 30/8/2005