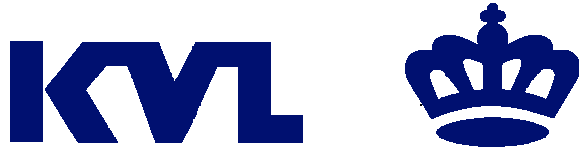


Theory session I: Multivariate data analysis using Principal Component Analysis



Frans van den Berg

The Royal Veterinary and Agricultural University (KVL), Denmark
Dept. of Dairy and Food Science, Food Technology
fb@kvl.dk

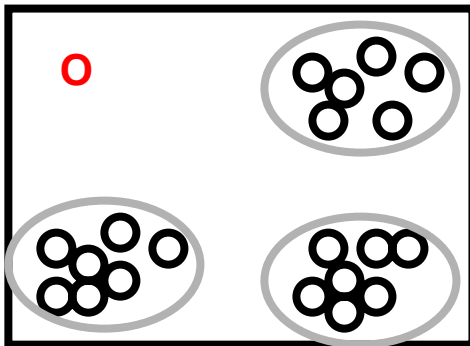
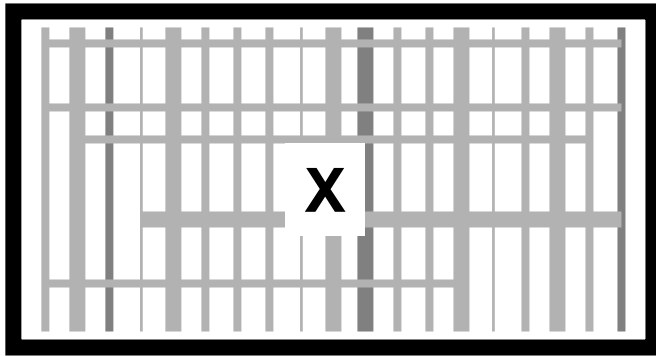


CENTRE FOR ADVANCED
FOOD STUDIES

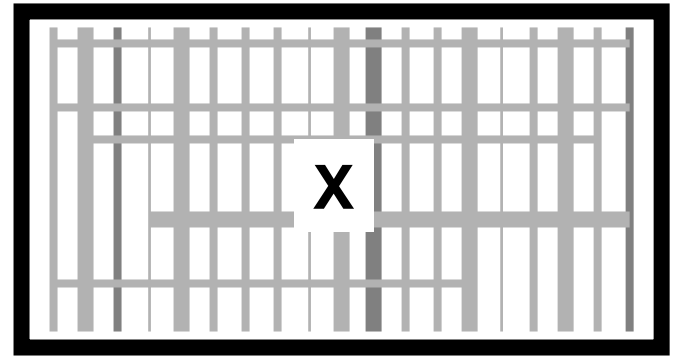
$$X = \begin{matrix} \text{---} \\ | \\ \text{---} \end{matrix} + \begin{matrix} \text{---} \\ | \\ \text{---} \end{matrix} + E$$
A diagram illustrating the PCA equation. On the left is a grey square with a thick black border containing the letter 'X'. To its right is an equals sign. This is followed by a vertical bar with horizontal lines above and below it, representing a matrix. To the right of this bar is a plus sign. This is followed by another vertical bar with horizontal lines above and below it, representing a second matrix. To the right of this second bar is another plus sign. Finally, on the far right, is a grey square with a thick black border containing the letter 'E', representing the error term.

'Visualization and Data Mining'

Visualization: certainly!



Data Mining: maybe?
(depends on the definition)



Contents

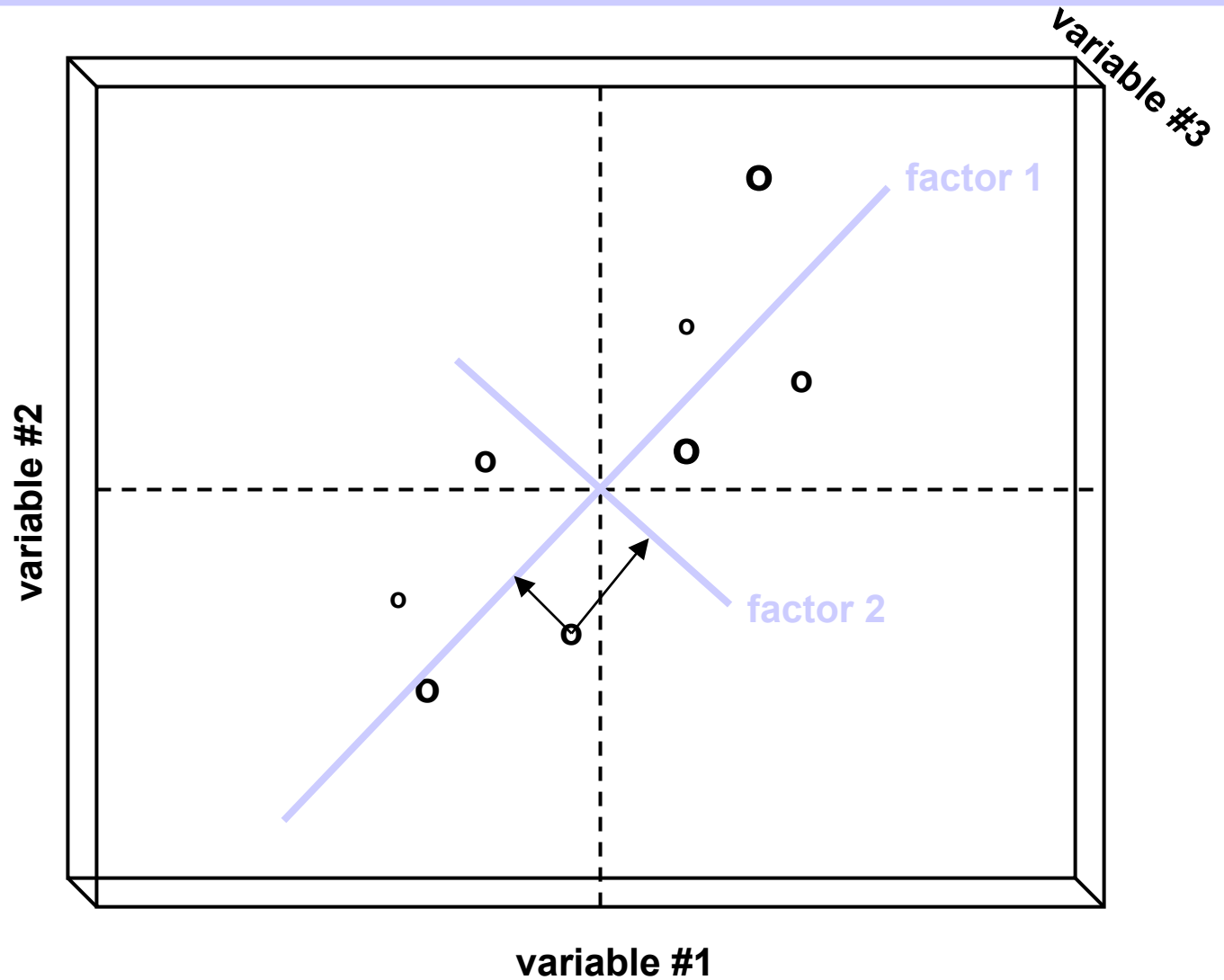
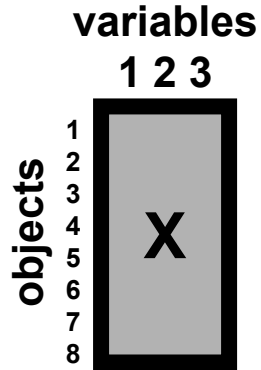
- **Principal Component Analysis**
 - **Example: demographical data**
 - **PCA - some details**
 - **Example: styrene reactor**
 - **(Multi-block and 'higher order' PCA)**
-

Principal Component Analysis

Data table/matrix (X)

		variables		
		#1	#2	#3
objects (samples)	1	x	x	x
	2	x	x	x
	3	x	x	x
	4	x	x	x
	5	x	x	x
	6	x	x	x
	7	x	x	x
	8	x	x	x

Principal Component Analysis



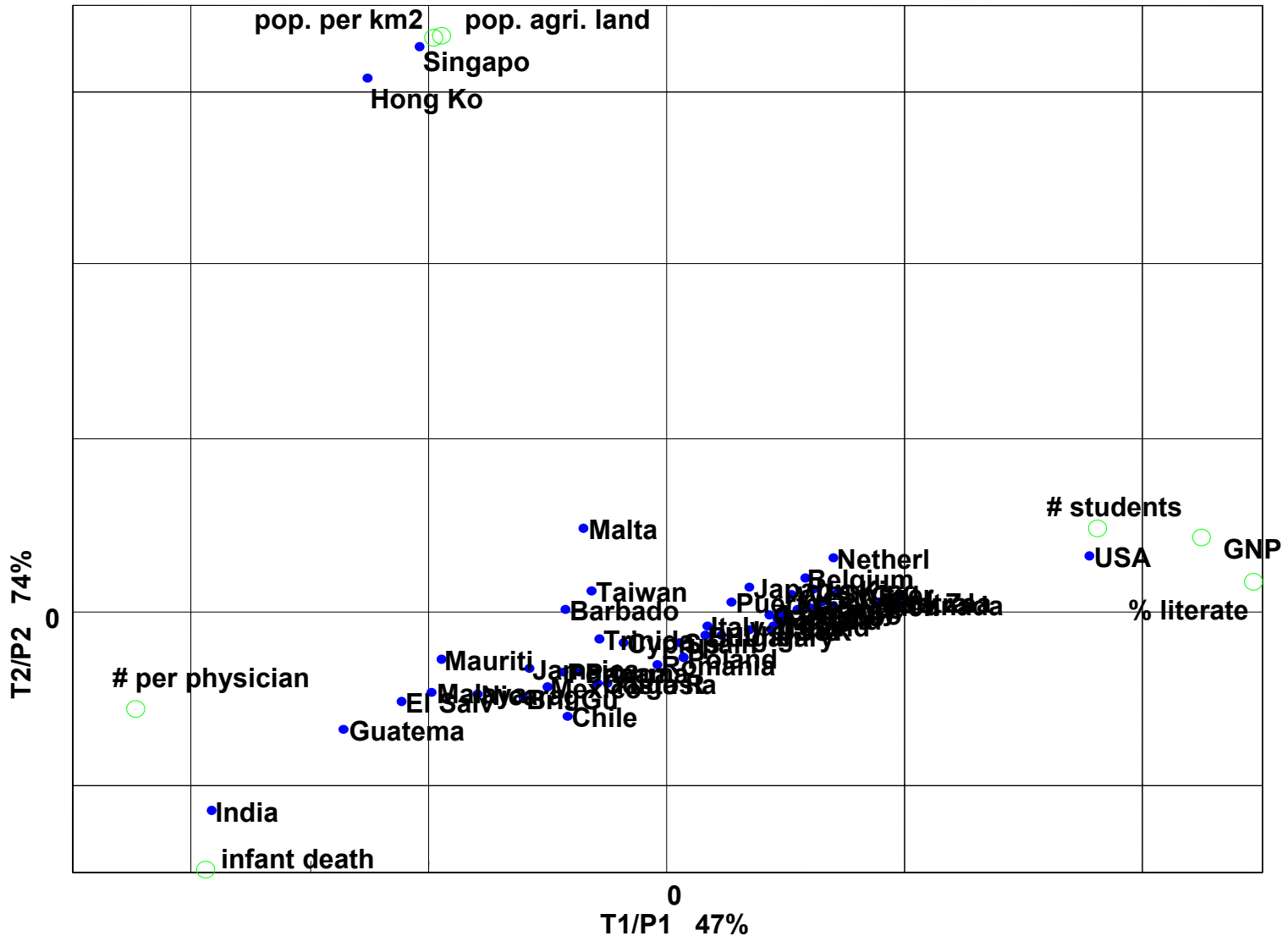
Demographical data

<i>countries</i>	<i>Infant death per 1000 live births</i>	<i># of inhabitants per physician</i>	<i>population per square kilometer</i>	<i>population per hectares agri. land</i>	<i>percentage literate above age 14</i>	<i># students higher education per 100k</i>	<i>GNP per capita</i>
	7 variables						
Austral	0.0195	0.8600	0.0010	0.0210	0.0985	0.8560	1.3160
Austria	0.0375	0.6950	0.0840	1.7200	0.0985	0.5460	0.6700
Barbado	0.0604	3.0000	0.5480	7.1210	0.0911	0.0240	0.2000
Belgium	0.0354	0.8190	0.3010	5.2570	0.0967	0.5360	1.1960
Brit Gu	0.0671	3.9000	0.0030	0.1920	0.0740	0.0270	0.2350
Bulgari	0.0451	0.7400	0.0720	1.3800	0.0850	0.4560	0.3650
Canada	0.0273	0.9000	0.0020	0.2570	0.0975	0.6450	1.9470
Chile	0.1279	1.7000	0.0110	1.1640	0.0801	0.2570	0.3790
Costa R	0.0789	2.6000	0.0240	0.9480	0.0794	0.3260	0.3570
Cyprus	0.0299	1.4000	0.0620	1.0420	0.0605	0.0780	0.4670
Czechos	0.0310	0.6200	0.1080	1.8210	0.0975	0.3980	0.6800
Denmark	0.0237	0.8300	0.1070	1.4340	0.0985	0.5700	1.0570
El Salv	0.0763	5.4000	0.1270	1.4970	0.0394	0.0890	0.2190
Finland	0.0210	1.6000	0.0130	1.5120	0.0985	0.5290	0.7940
France	0.0274	1.0140	0.0830	1.2880	0.0964	0.6670	0.9430
...
West Ge	0.0338	0.7980	0.2170	3.6310	0.0985	0.5280	0.9270
Yugosla	0.1000	1.6370	0.0730	1.2150	0.0770	0.5240	0.2650

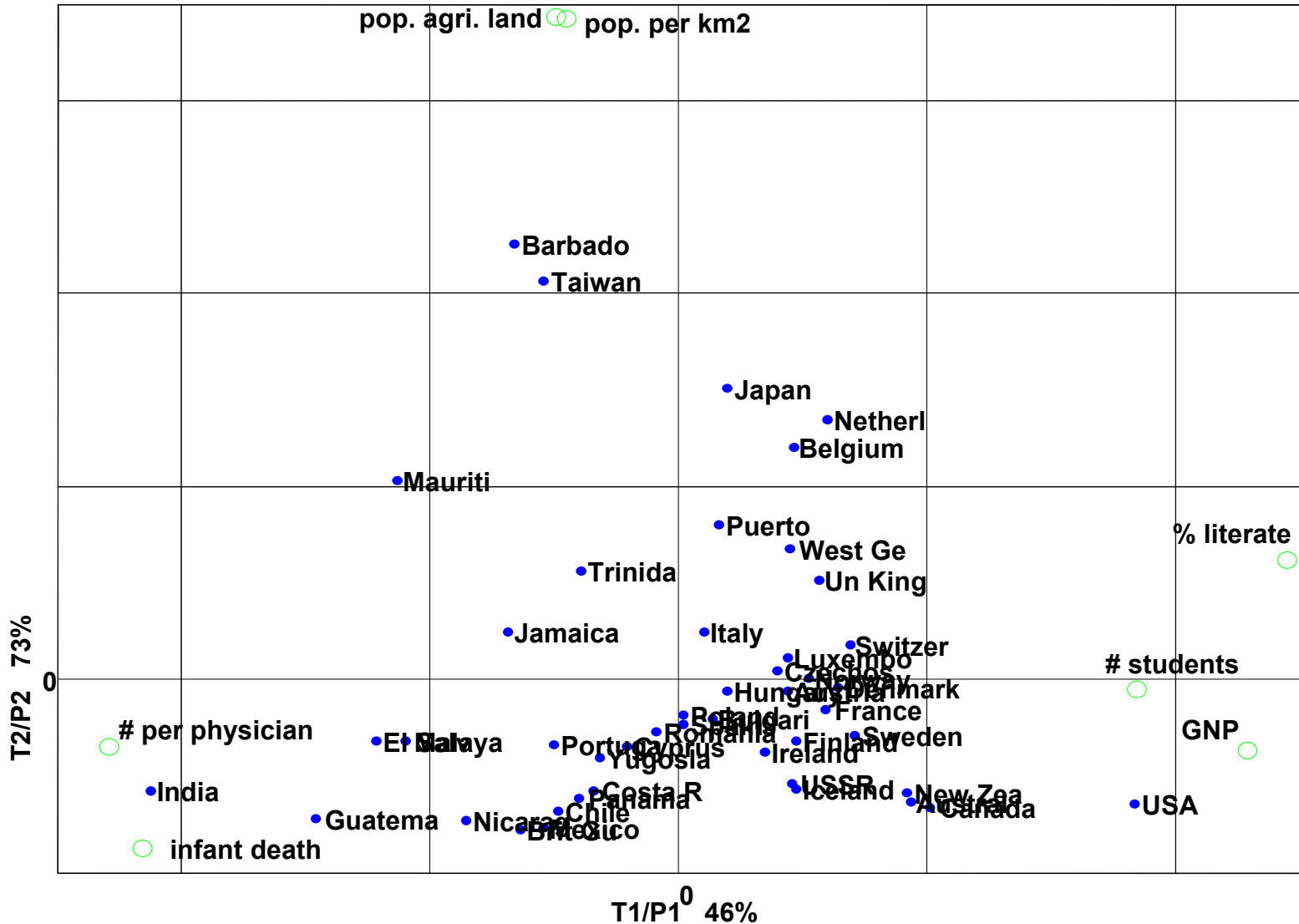
49 objects

x1000

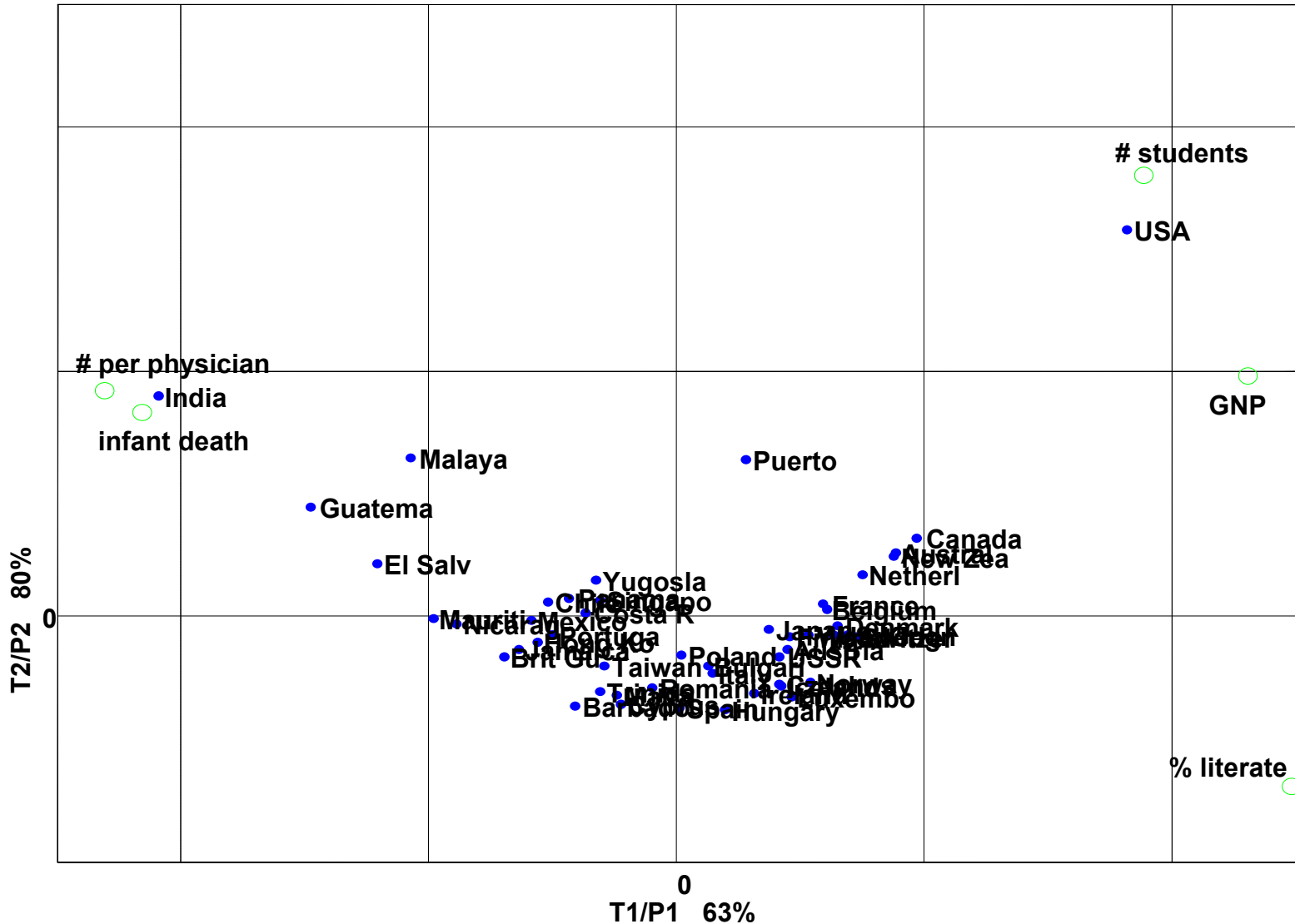
Demographical data



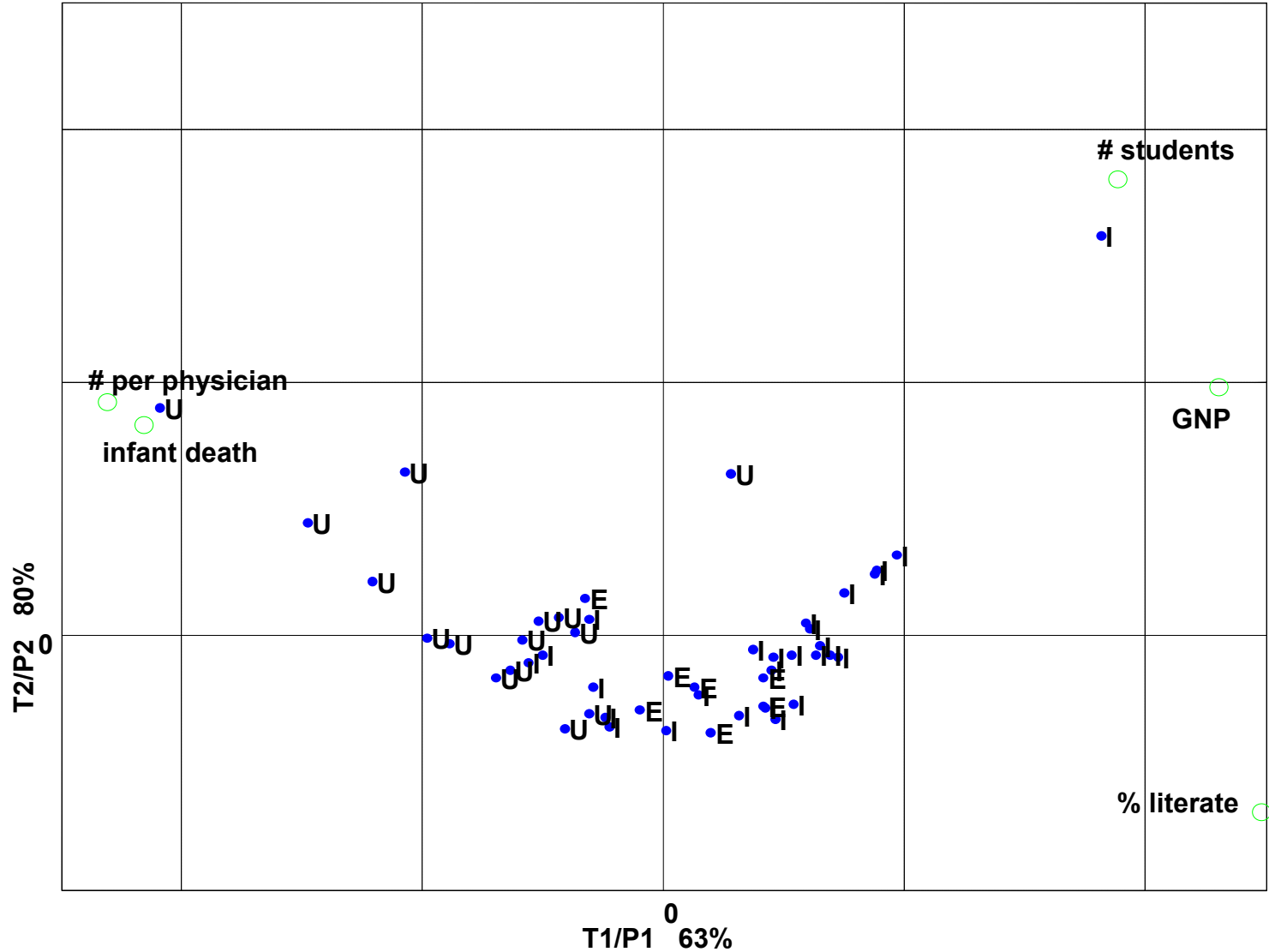
Demographical data



Demographical data



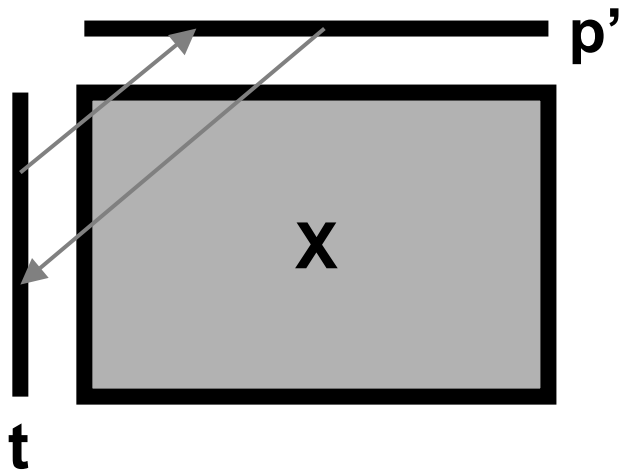
Demographical data



PCA - some details: Math's

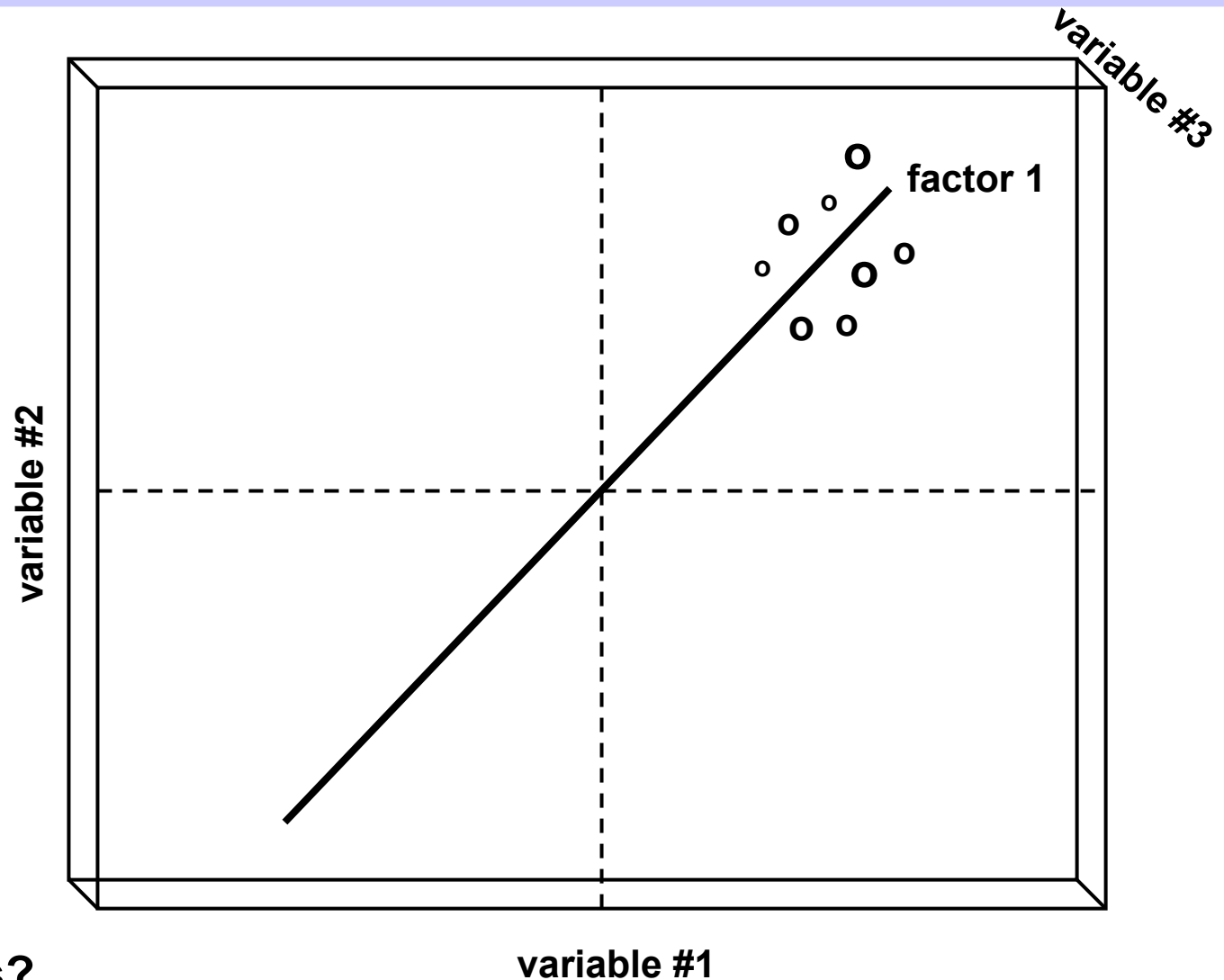
$$\min || X - T.P' ||^2 \quad t_i \cdot t_j' = 0 / p_i \cdot p_j' = 0 \quad || p_i || = 1$$

$$X = t.p' + E \quad \rightarrow \quad E = X - t.p'$$



- **NIPALS-algorithm**
(power method)
- **Eigenvalue decomposition**
 $(X.X') \cdot \lambda = t \cdot \lambda \quad p = t' \cdot X$
- **Singular Value Decomposition**
 $X = U \cdot D \cdot P' = T \cdot P'$
- **Alternating Least Squares**
 $X \approx A \cdot B' \quad \text{rotate} \rightarrow X \approx T \cdot P'$
- **Principal Factor Analysis**
Orthogonal/Non-orthogonal rotation

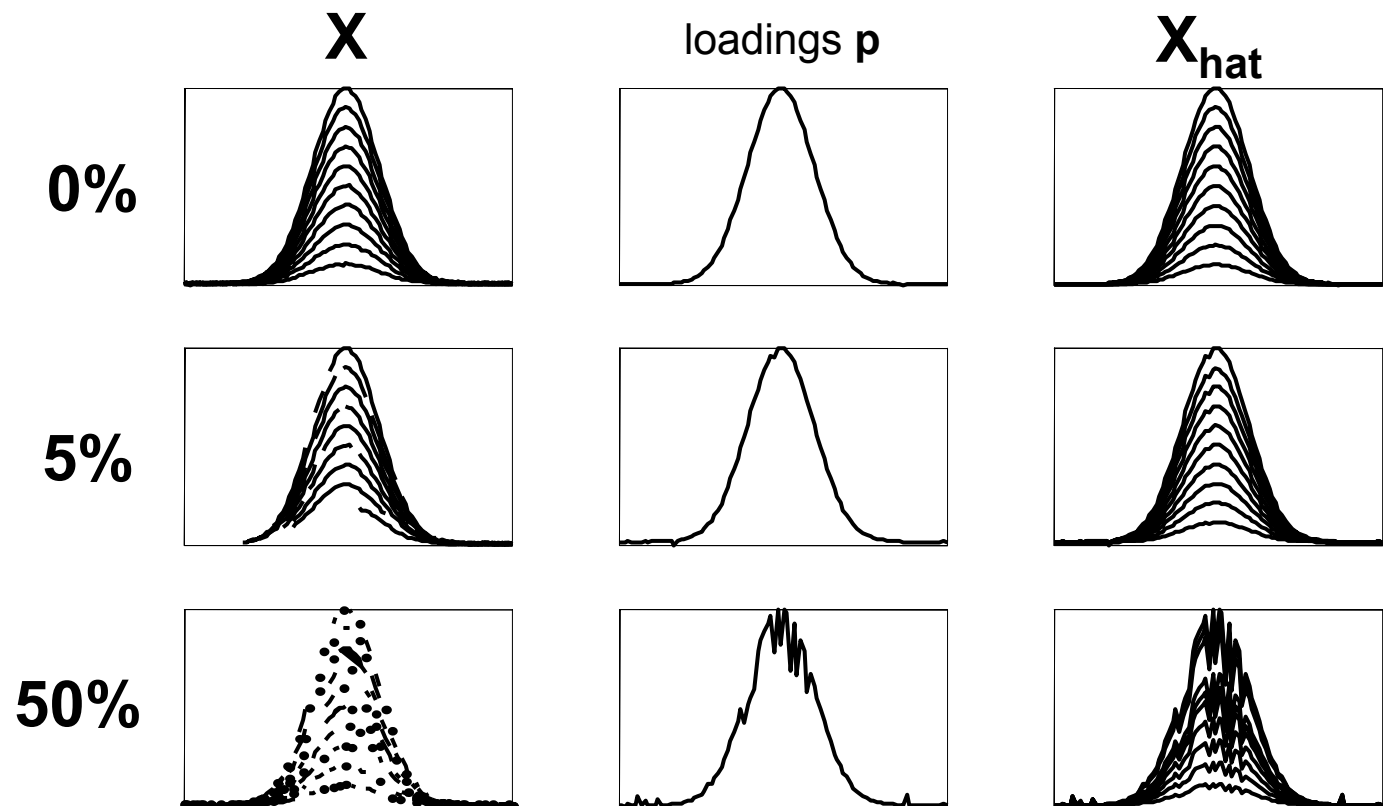
PCA - some details: Preprocessing



- Mean center
- Auto scale
- Transformations?

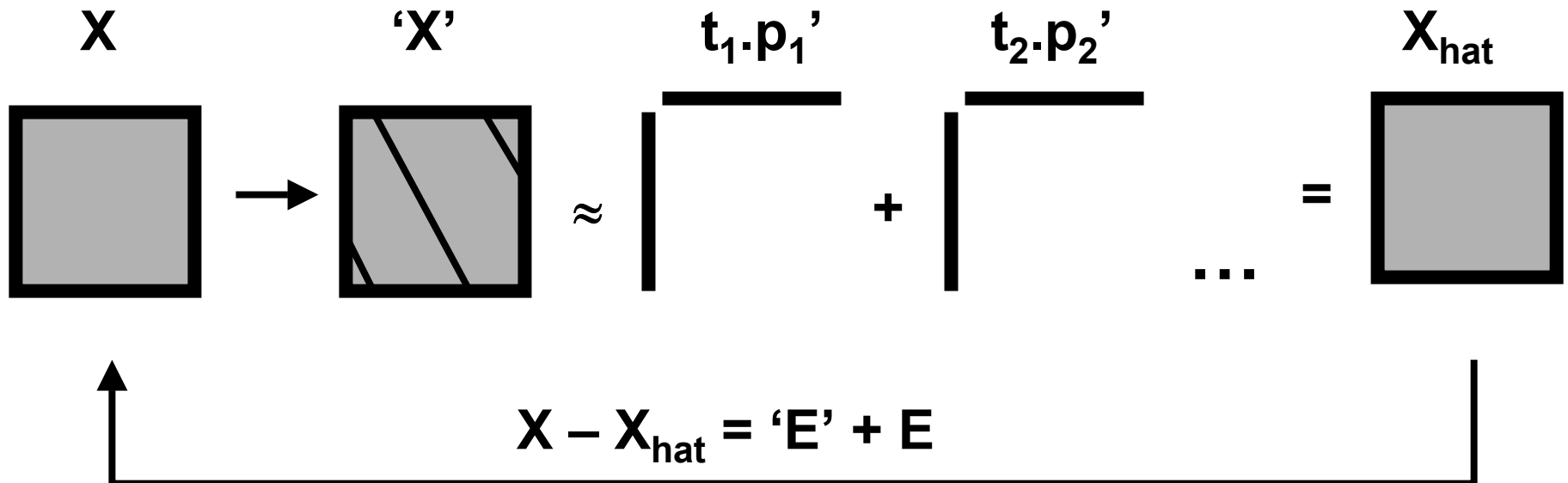
PCA - some details: Missing values

- PCA models can handle a modest amount of missing data
- Imputation, expectation maximization, ...
- Depends on the structure of the 'non-missing' data.



PCA - some details: Number of factors

- Mathematical rank v. 'true rank' (e.g. chemical rank)
- Percentage explained variance per factor
- Cross validation

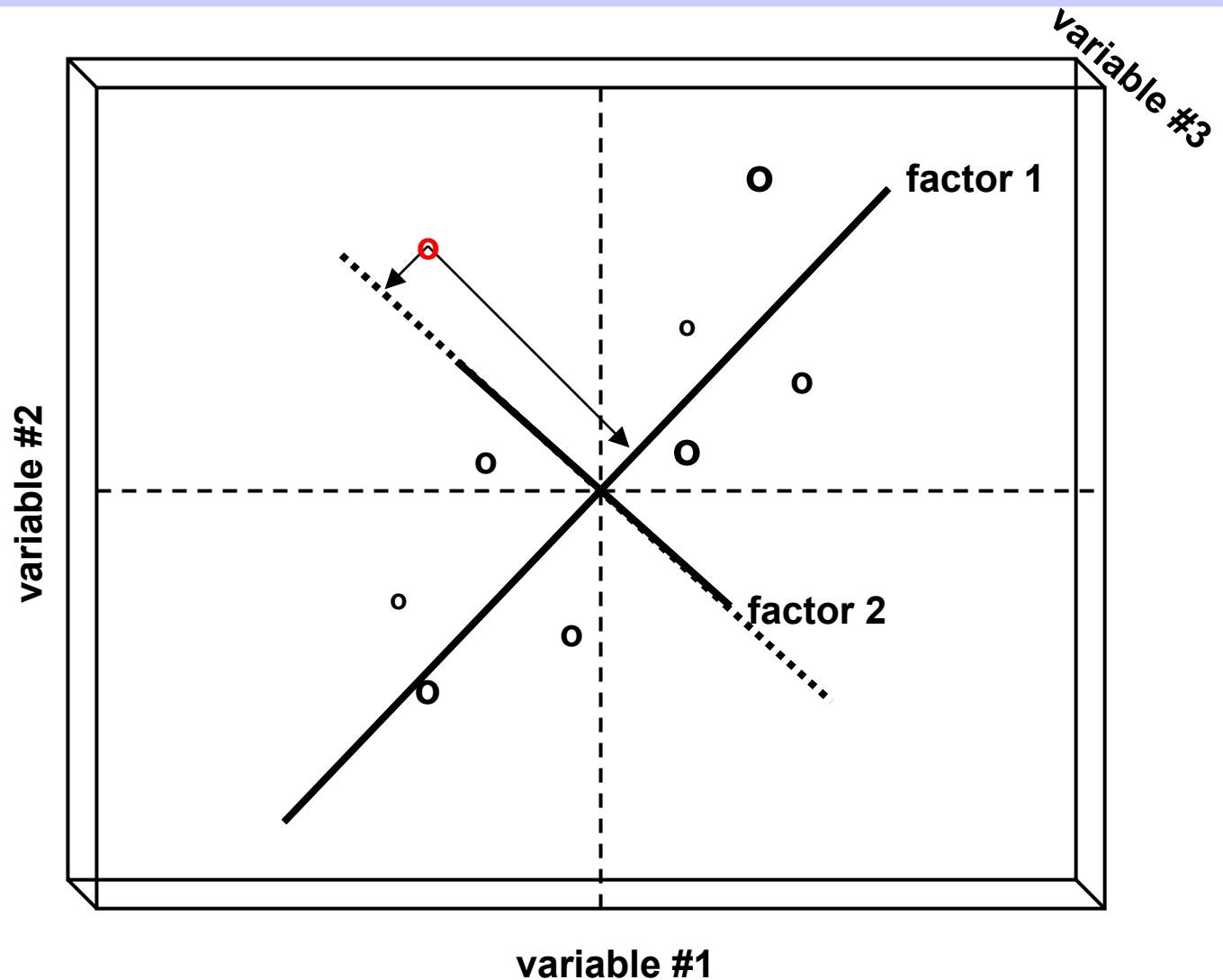


PCA - some details: New data

- New object x' :

$$t = P' \cdot x$$

$$t = \begin{bmatrix} \\ \\ \end{bmatrix}$$



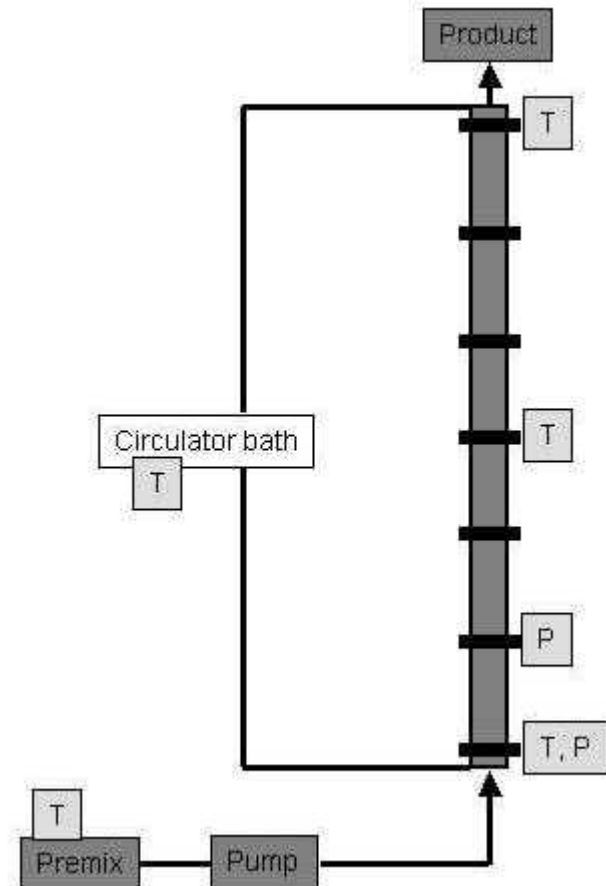
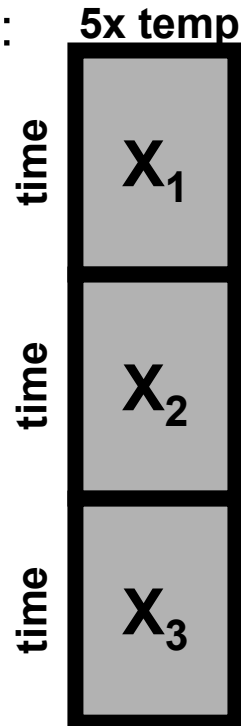
- MSPM/MSPC

PCA - some details: Diagnostics

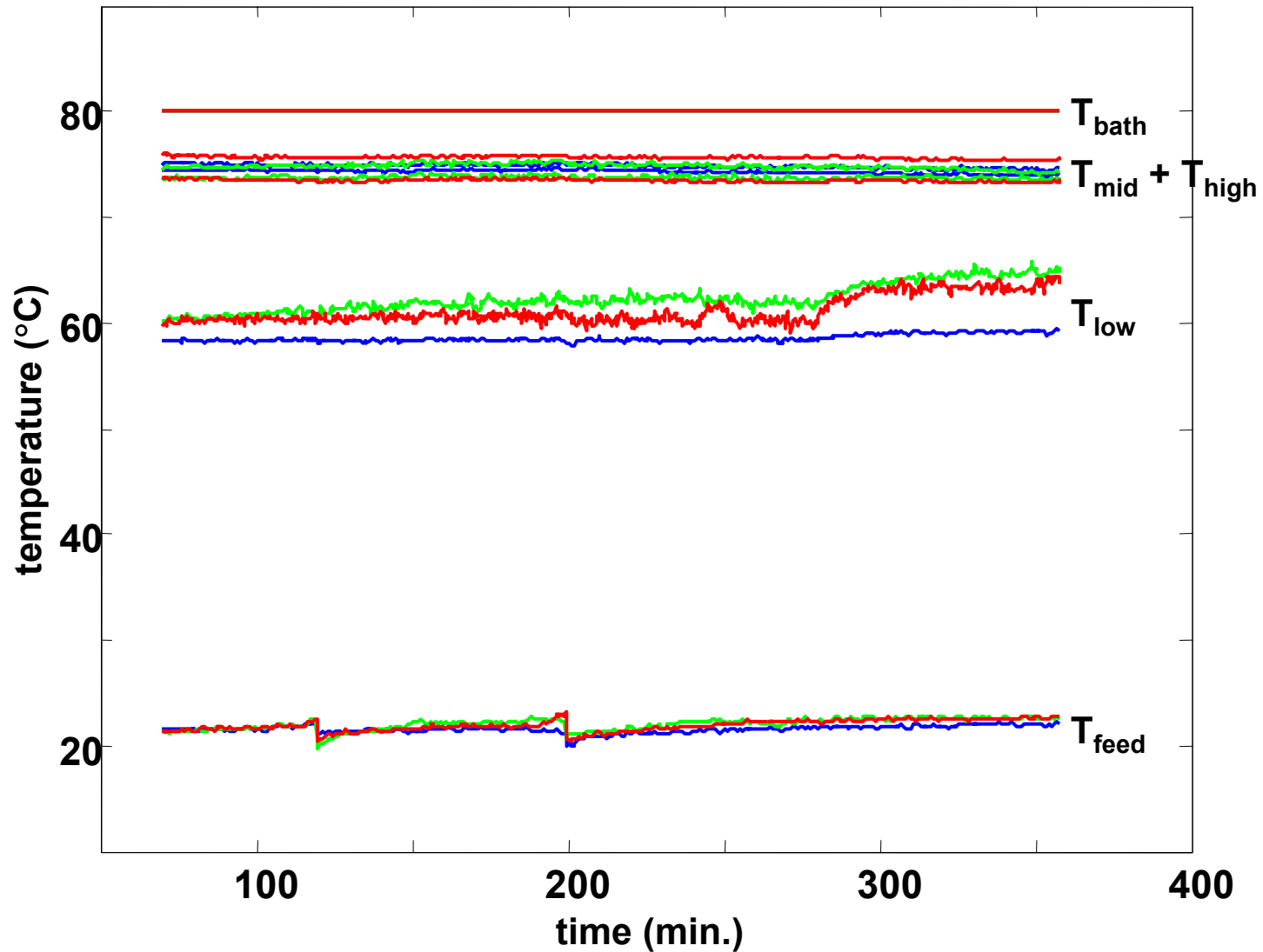
- **Score values t_a**
Position of objects in the new factor-space
- **Loading values p_a**
Role of original variables in determining the new factor-space
- **Percentage explained variance**
$$\sum \sum x_{ij}^2 = \sum \sum (\mathbf{t} \cdot \mathbf{p}')^2 + \sum \sum e_{ij}^2$$
- **Leverage h_a**
How important is an object compared the rest of the data set
$$h_{i,a} \approx \sum t_{i,a} / (\mathbf{t}_a' \cdot \mathbf{t}_a)$$
- **Residuals e_a**
How much structure/information remains after n-factors
objects: $\sum e^2$ over rows variables: $\sum e^2$ over columns

Styrene reactor

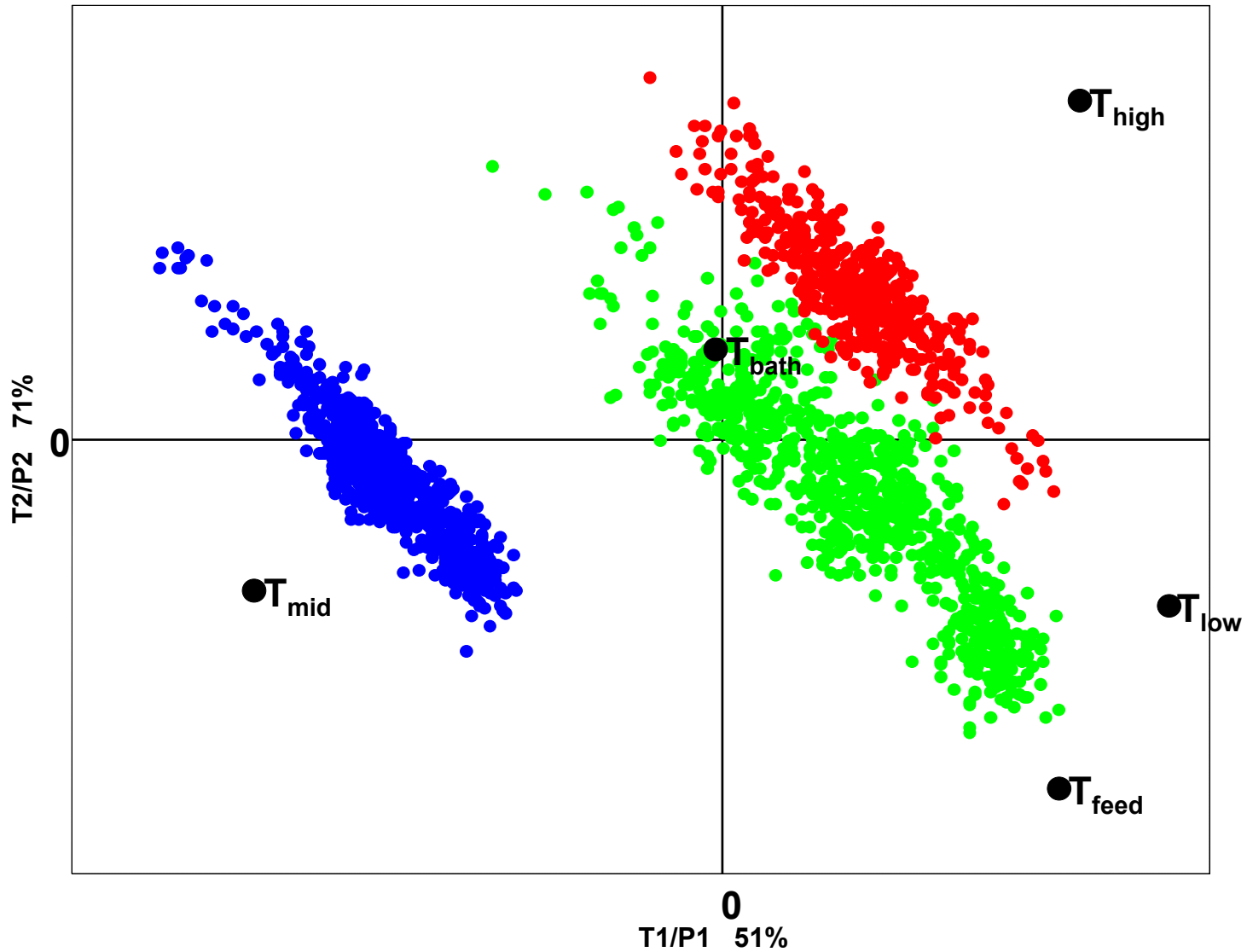
- AIBN initiator-driven polymerisation of styrene in a tubular reactor
- 5 temperatures
- 3 experiments, each with a duration of 850 time points (5 hours)
- Data matrices: **5x temp.**



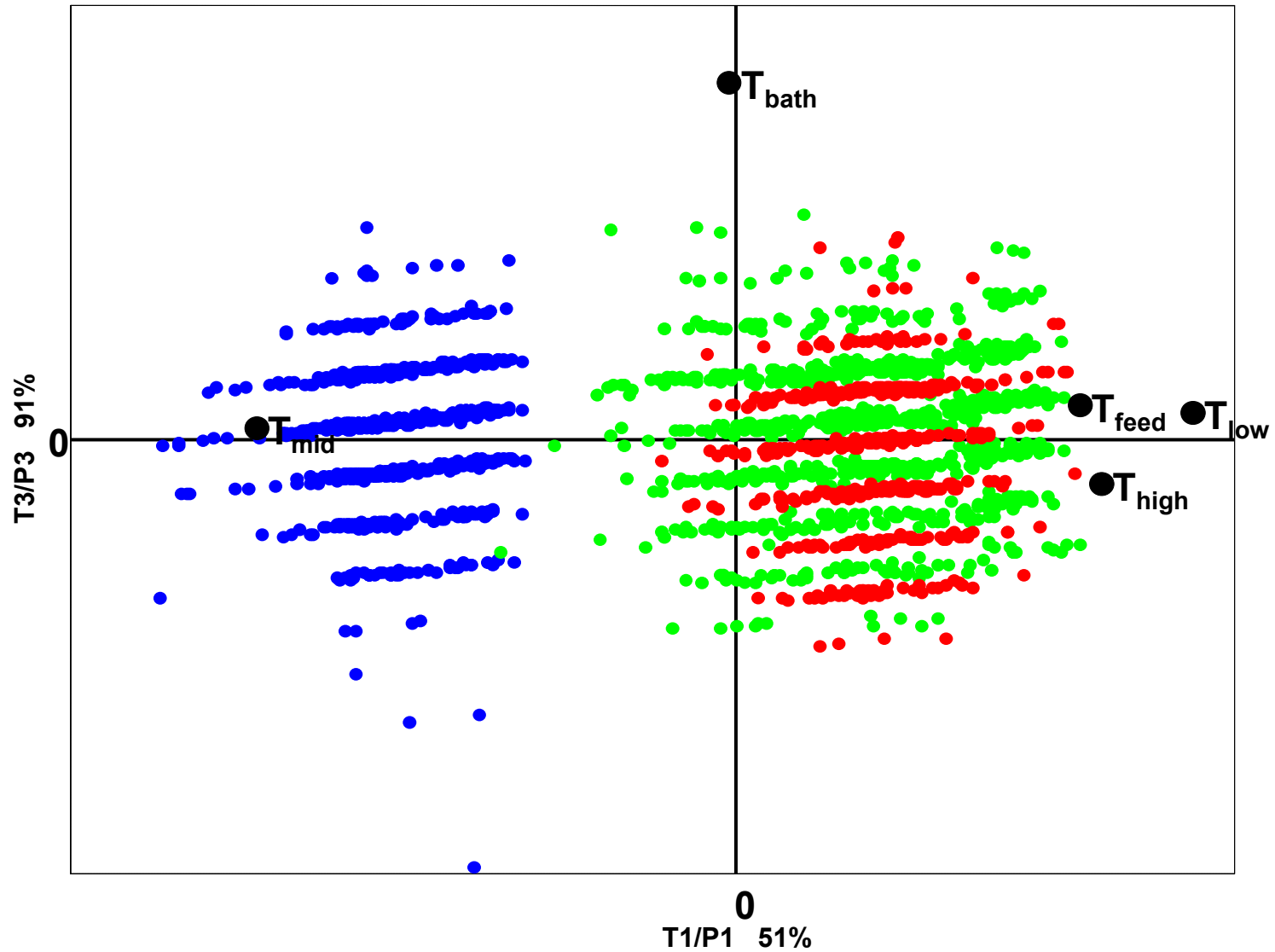
Styrene reactor



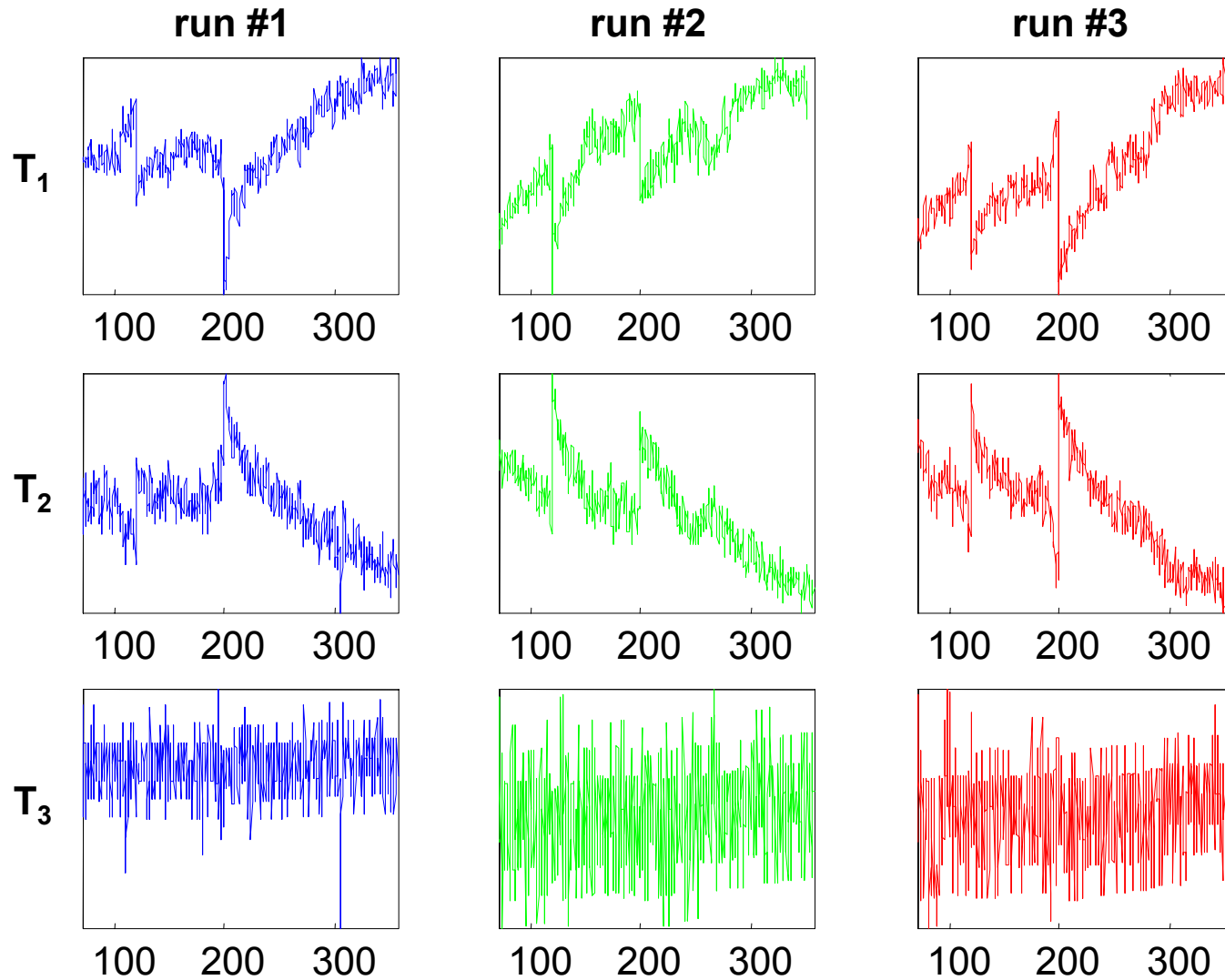
Styrene reactor



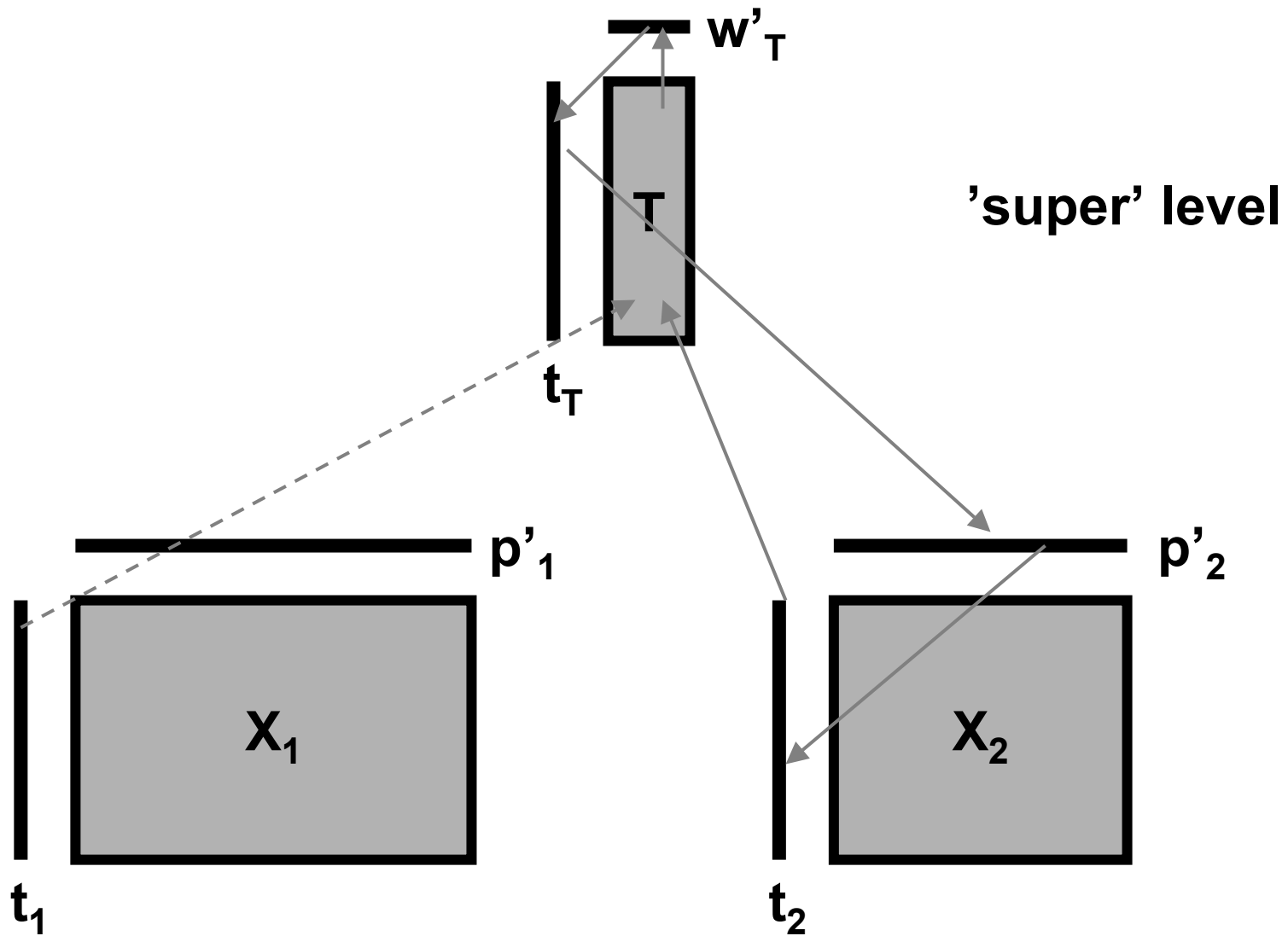
Styrene reactor



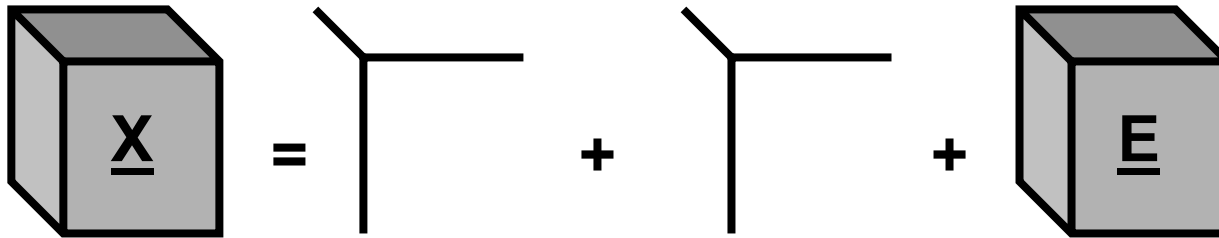
Styrene reactor



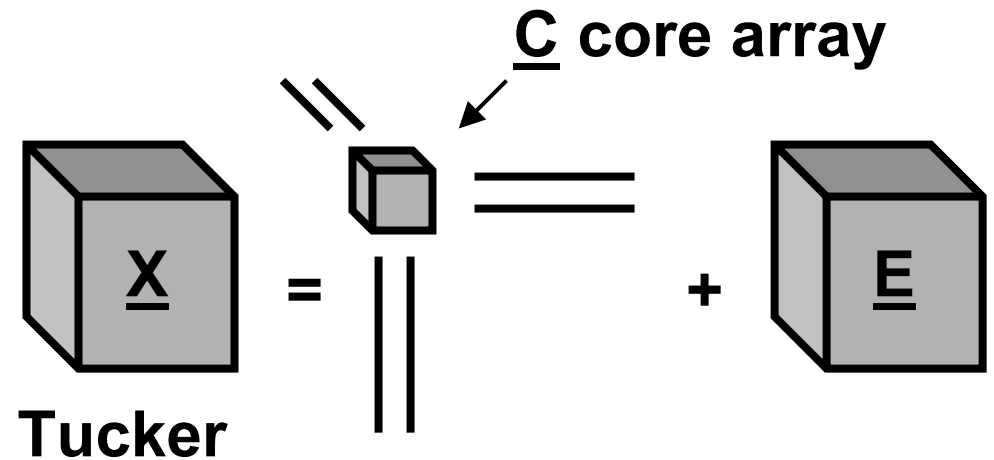
Multi-block PCA



'Higher order' PCA

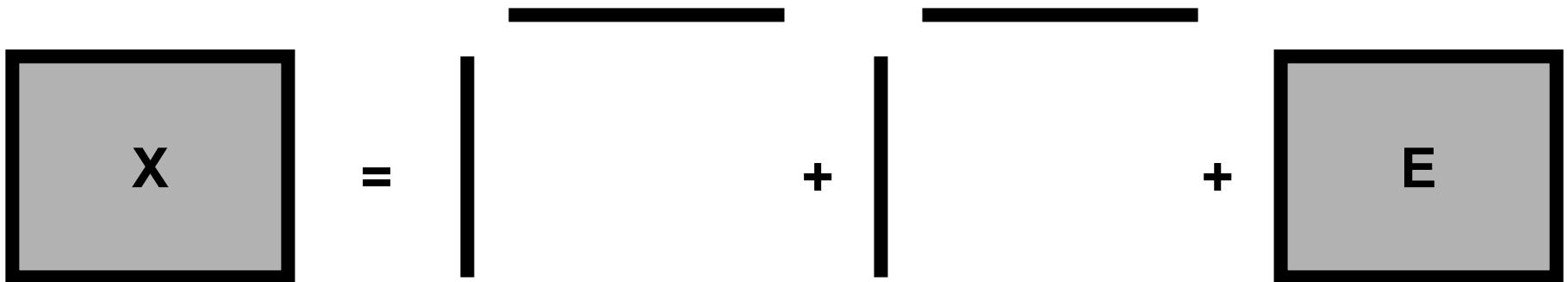


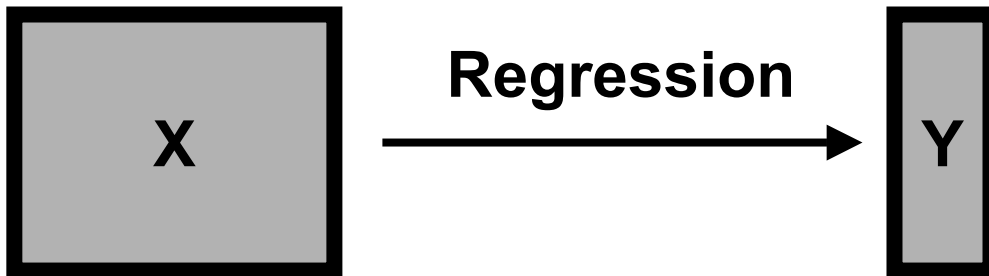
PARAFAC

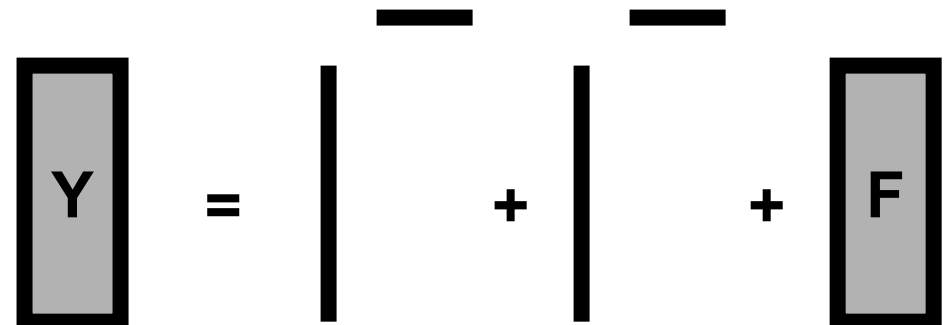


Tucker

Theory session II: Multivariate Regression using Partial Least Squares

$$X = \begin{matrix} \text{---} \\ | \\ \text{---} \end{matrix} + \begin{matrix} \text{---} \\ | \\ \text{---} \end{matrix} + E$$
A diagram illustrating the decomposition of a matrix X. On the left is a gray square box labeled 'X'. To its right is an equals sign. This is followed by a vertical bar with a horizontal line above it, a plus sign, another vertical bar with a horizontal line above it, another plus sign, and finally a gray square box labeled 'E'.



$$Y = \begin{matrix} \text{---} \\ | \\ \text{---} \end{matrix} + \begin{matrix} \text{---} \\ | \\ \text{---} \end{matrix} + F$$
A diagram illustrating the decomposition of a vector Y. On the left is a gray vertical rectangle labeled 'Y'. To its right is an equals sign. This is followed by a vertical bar with a horizontal line above it, a plus sign, another vertical bar with a horizontal line above it, another plus sign, and finally a gray vertical rectangle labeled 'F'.

Contents

- **Partial Least Squares Regression**
 - **PLS - some details**
 - **Example: demographical data**
 - **(Multi-block and 'higher order' PLS)**
 - **Computer exercises**
-

PLS - some details: Math's

$$y = X.b + f$$

$$X'.y = X'.X.b$$

$$(X'.X)^{-1}.X'.y = \underbrace{(X'.X)^{-1}.(X'.X)}_{“(X'.X) / (X'.X) = 1”}.b$$

$$(X'.X)^{-1}.X'.y = b$$

↑
(OLS/MLR solution)

$$y \text{ | } = \boxed{X} \text{ | } b$$

$$\left[\begin{array}{c} \boxed{X} \\ \boxed{X'} \end{array} \right]^{-1}$$

PLS - some details: Math's

$$X = T.P' + E \quad (\text{pca})$$

$$y = T.q + f$$

$$T'.y = T'.T.q$$

$$(T'.T)^{-1}.T'.y = (T'.T)^{-1}.(T'.T).q$$

$$(T'.T)^{-1}.T'.y = q$$

↑
(PCR solution, $y = X.b = X.P.q$)

$$y = T.q$$

$$\begin{bmatrix} T' & T \end{bmatrix}^{-1}$$

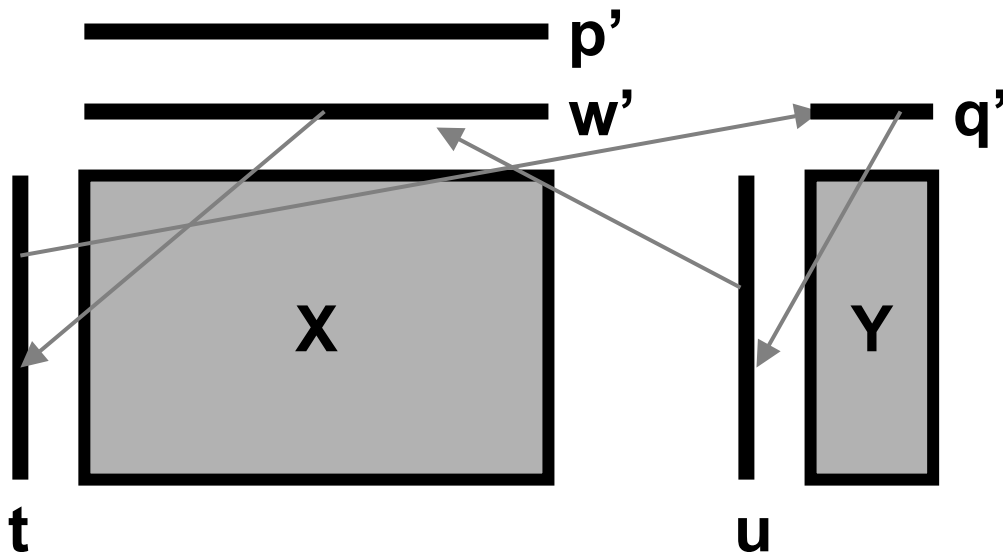
PLS - some details: Math's

BUT...

$$\max (\text{cov}(t,u) \mid Xw = t, \parallel w \parallel = 1)$$

$$X = tp' + E \quad \rightarrow \quad E = X - tp'$$

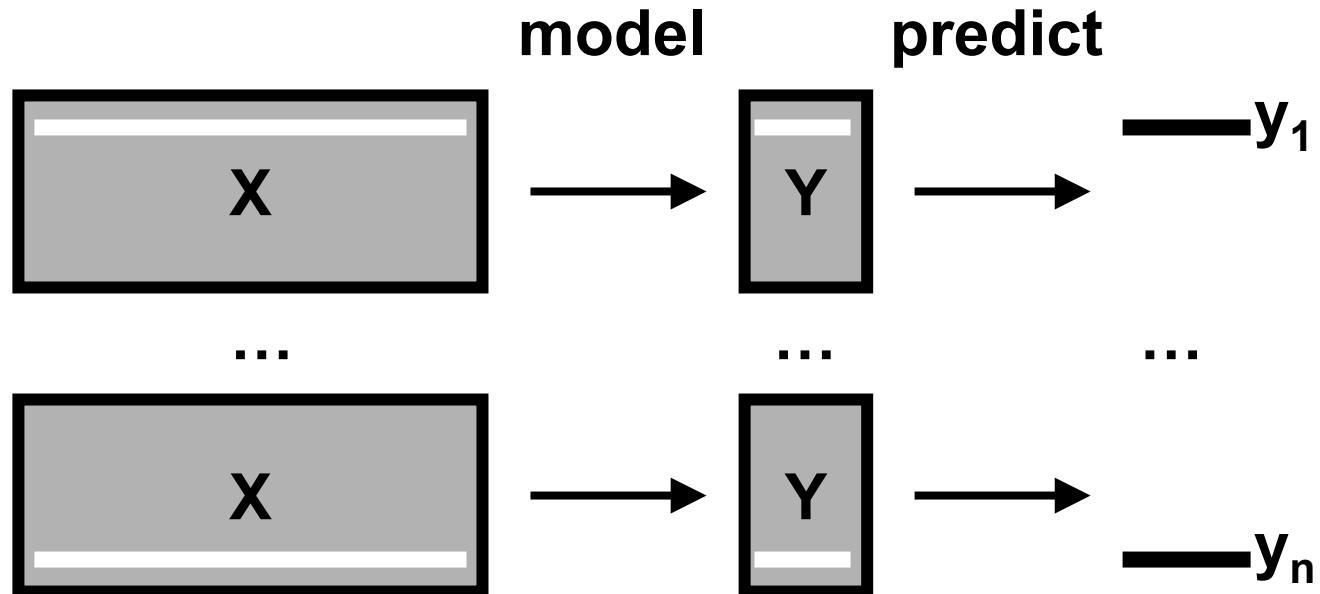
$$Y = tq' + F \quad \rightarrow \quad F = Y - tq'$$



$$(\text{PLS solution, } y = X.b = X.W.(P'.W)^{-1}.q)$$

PLS - some details: Number of factors

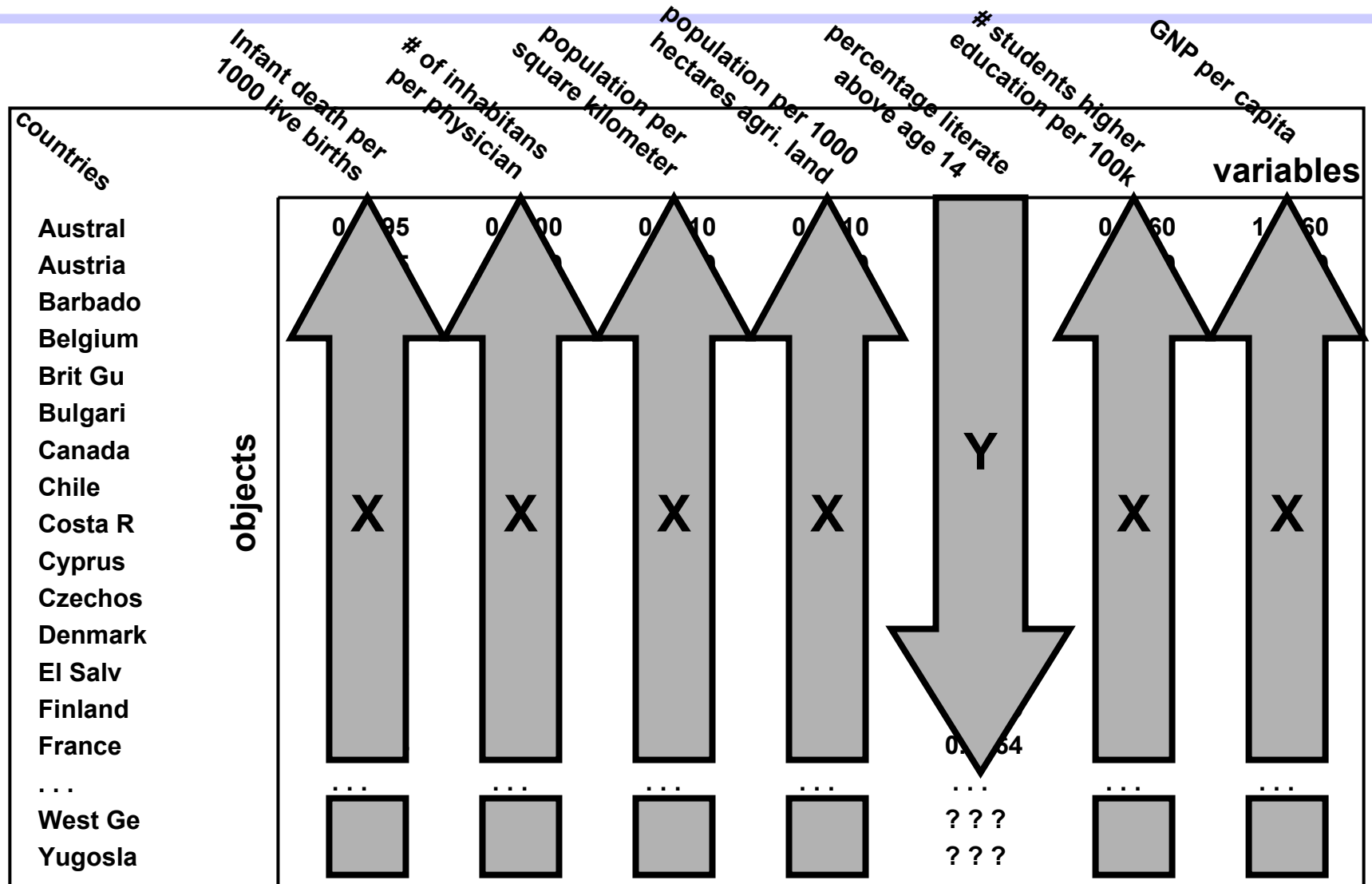
- Leave-one-out cross validation



- $\text{RMSP}_{cv} = \sqrt{(\sum(y - y_{\text{hat}})^2/n)}$

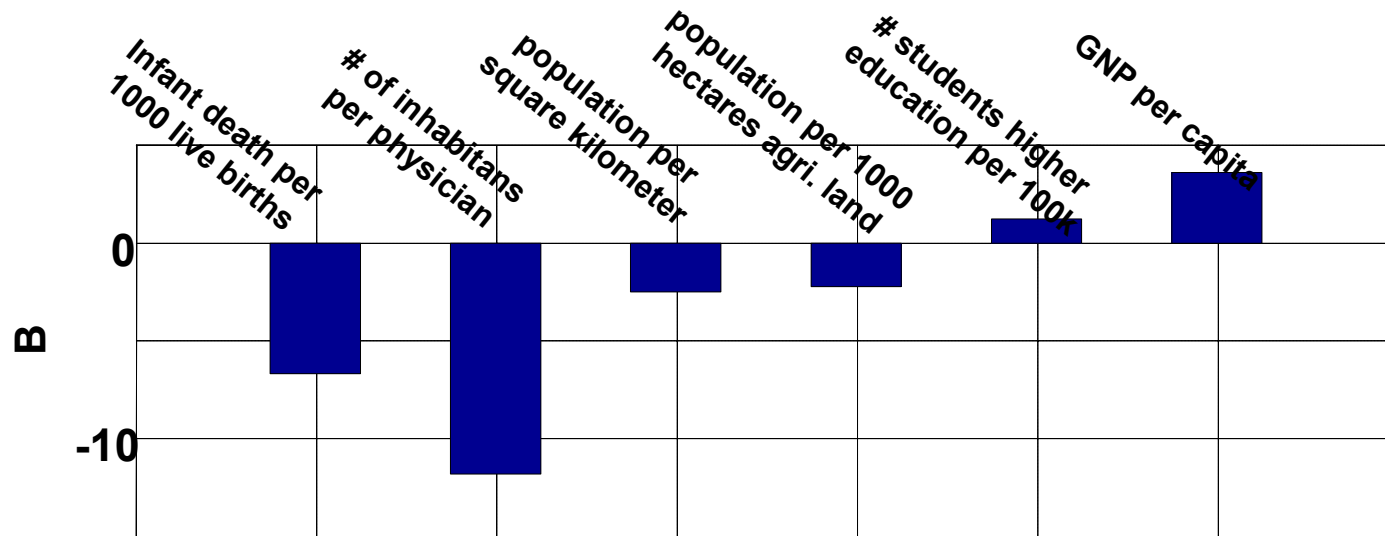
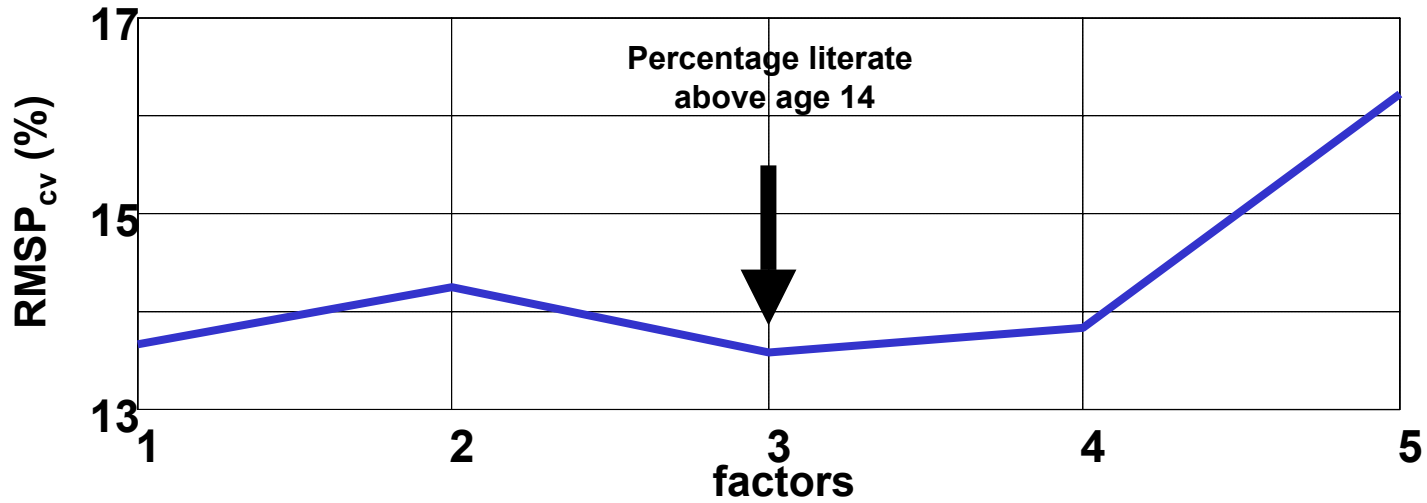
root mean squared error of prediction

Demographical data

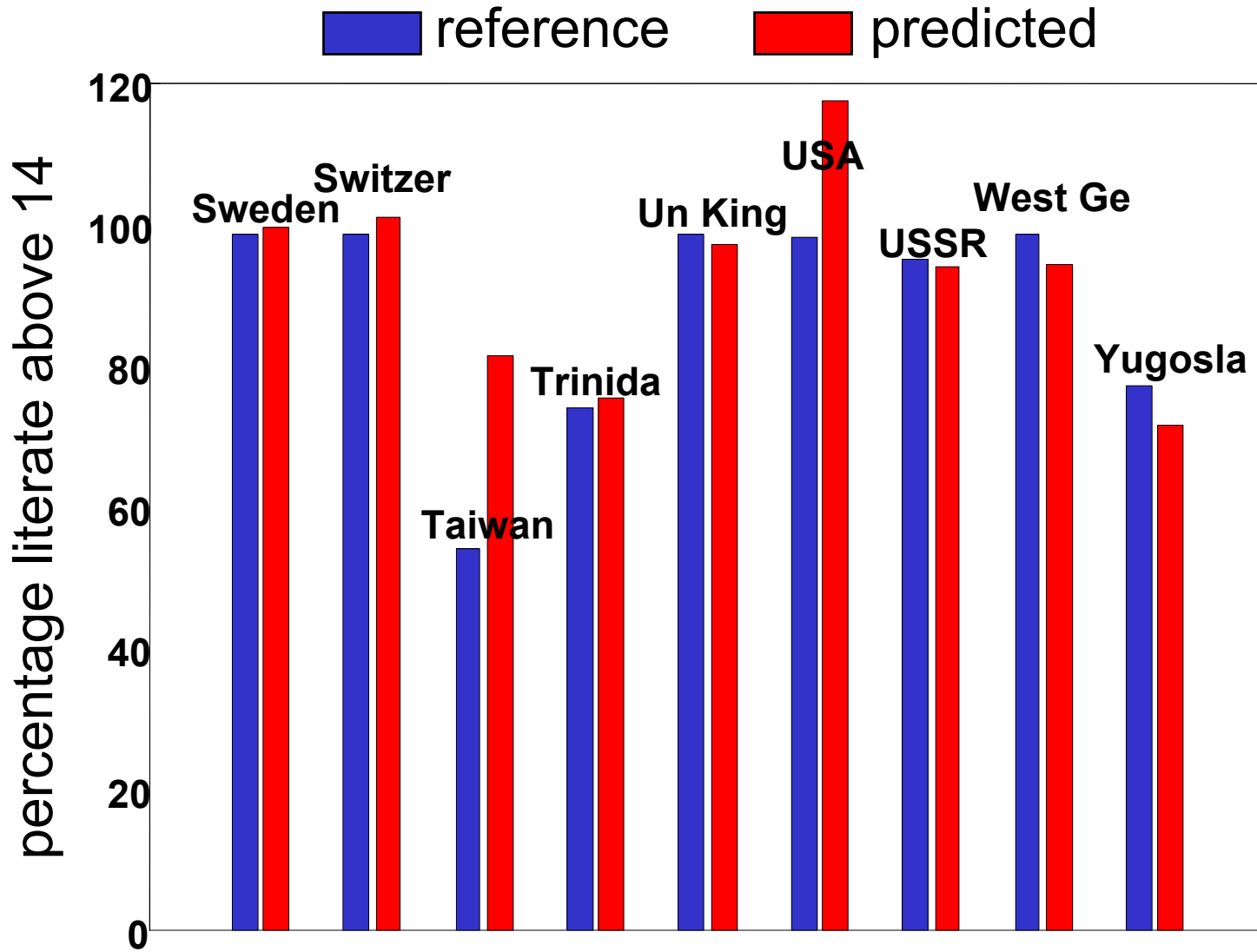


x1000

Demographical data

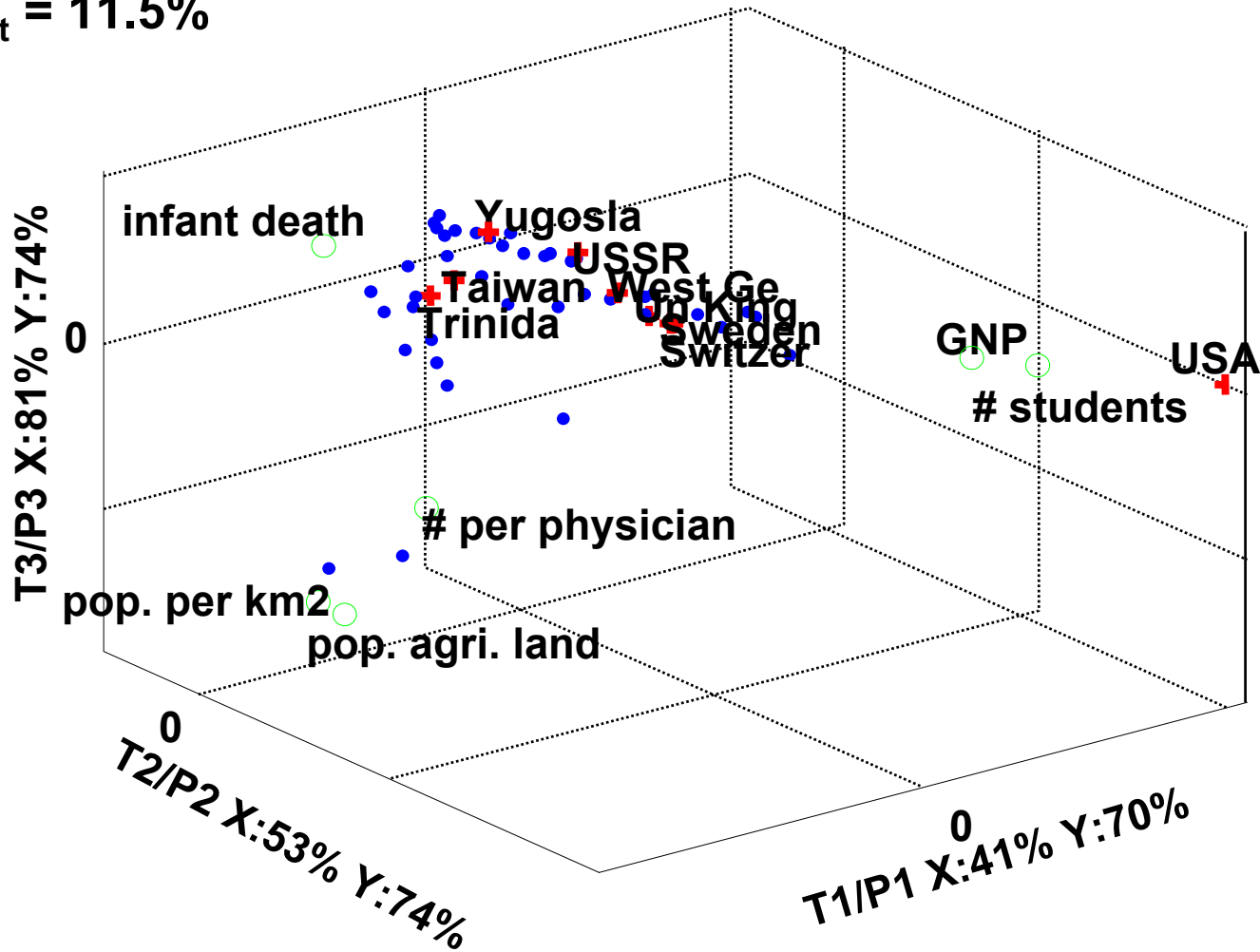


Demographical data

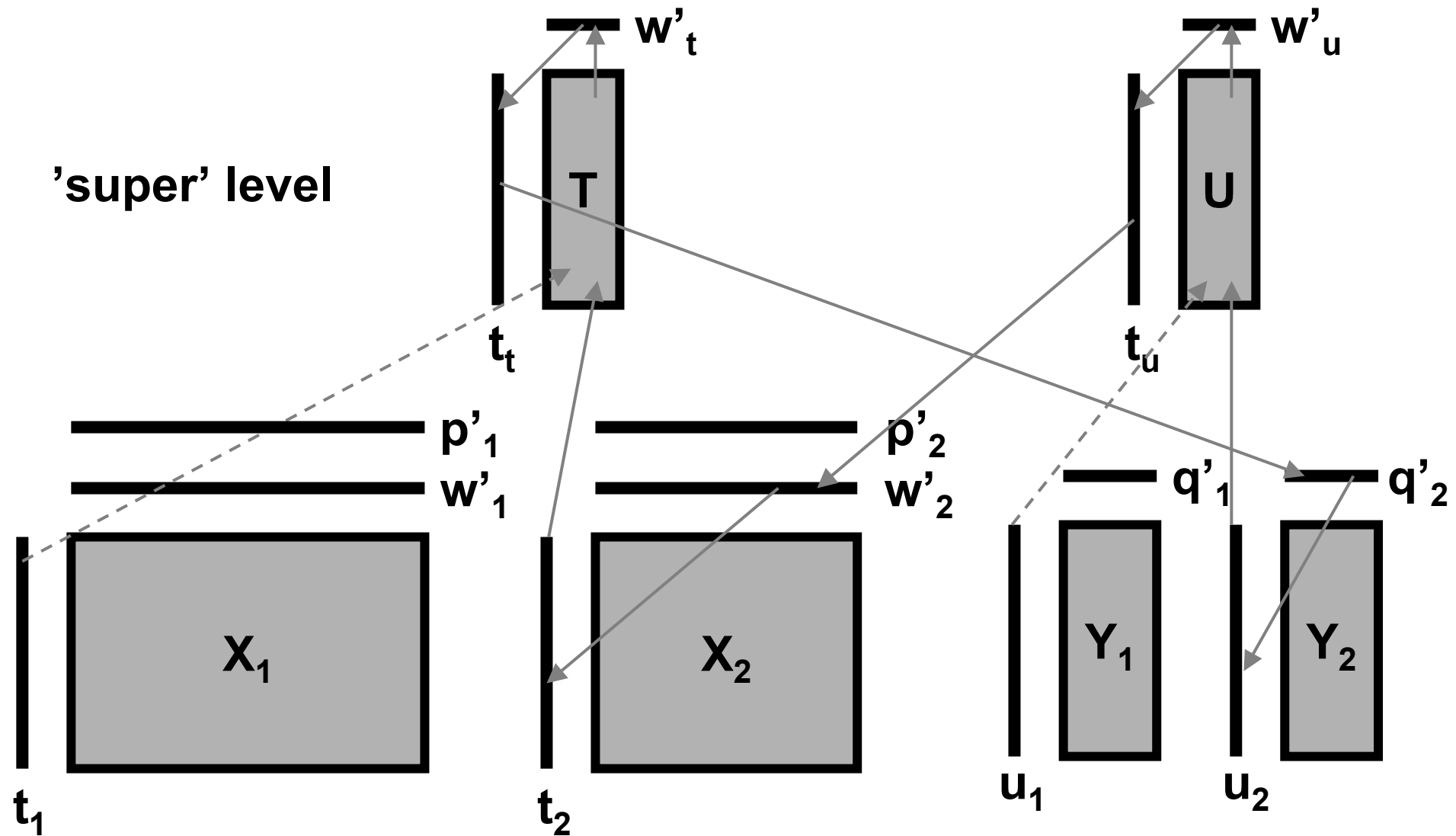


Demographical data

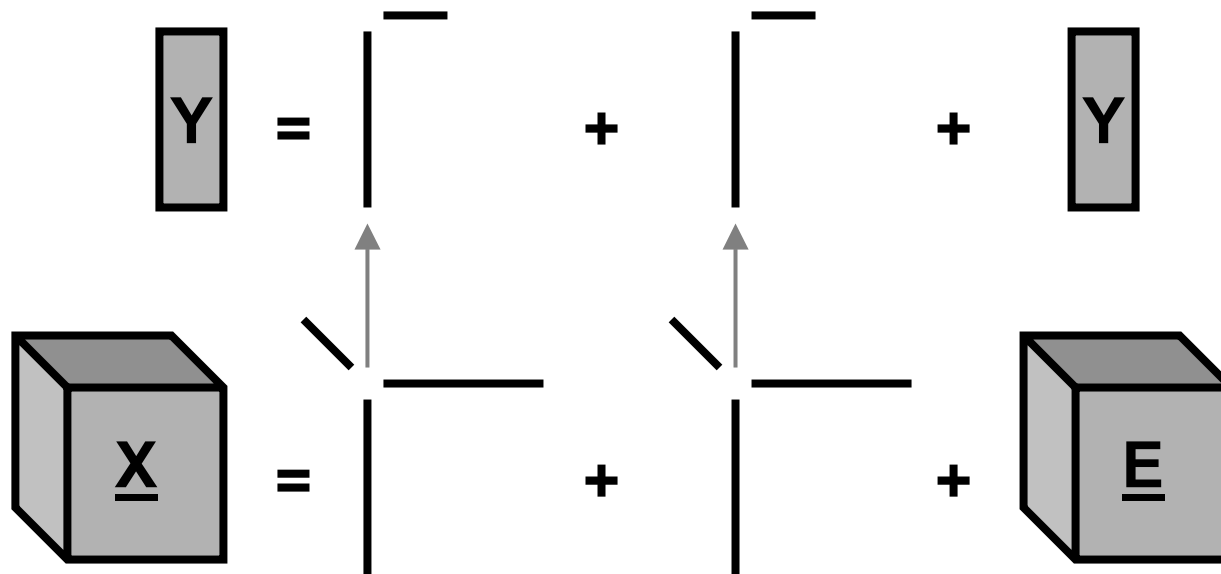
RMSP_{test} = 11.5%



Multi-block PLS



'Higher order' PLS



Multi linear PLS

Computer exercises 'Foods'

- 16 countries x 20 food/consumption related variables

The Unscrambler - [Foods]												
File Edit View Plot Modify Task Results Window Help												
[Toolbar icons]												
		Gr_Coffe	Inst_Coffe	Tea	Sweetner	Biscuits	Pa_Soup	Ti_Soup	In_Potat	Fro_Fish	Fro_Veg	App
		1	2	3	4	5	6	7	8	9	10	11
Germany	1	90.0000	49.0000	88.0000	19.0000	57.0000	51.0000	19.0000	21.0000	27.0000	21.0000	81.0
Italy	2	82.0000	10.0000	60.0000	2.0000	55.0000	41.0000	3.0000	2.0000	4.0000	2.0000	67.0
France	3	88.0000	42.0000	63.0000	4.0000	76.0000	53.0000	11.0000	23.0000	11.0000	5.0000	87.0
Holland	4	96.0000	62.0000	98.0000	32.0000	62.0000	67.0000	43.0000	7.0000	14.0000	14.0000	83.0
Belgium	5	94.0000	38.0000	48.0000	11.0000	74.0000	37.0000	23.0000	9.0000	13.0000	12.0000	76.0
Luxembou	6	97.0000	61.0000	86.0000	28.0000	79.0000	73.0000	12.0000	7.0000	26.0000	23.0000	85.0
England	7	27.0000	86.0000	99.0000	22.0000	91.0000	55.0000	76.0000	17.0000	20.0000	24.0000	76.0
Portugal	8	72.0000	26.0000	77.0000	2.0000	22.0000	34.0000	1.0000	5.0000	20.0000	3.0000	22.0
Austria	9	55.0000	31.0000	61.0000	15.0000	29.0000	33.0000	1.0000	5.0000	15.0000	11.0000	49.0
Switzerl	10	73.0000	72.0000	85.0000	25.0000	31.0000	69.0000	10.0000	17.0000	19.0000	15.0000	79.0
Sweden	11	97.0000	13.0000	93.0000	31.0000	m	43.0000	43.0000	39.0000	54.0000	45.0000	56.0
Denmark	12	96.0000	17.0000	92.0000	35.0000	66.0000	32.0000	17.0000	11.0000	51.0000	42.0000	81.0
Norway	13	92.0000	17.0000	83.0000	13.0000	62.0000	51.0000	4.0000	17.0000	30.0000	15.0000	61.0
Finland	14	98.0000	12.0000	84.0000	20.0000	64.0000	27.0000	10.0000	8.0000	18.0000	12.0000	50.0
Spain	15	70.0000	40.0000	40.0000	m	62.0000	43.0000	2.0000	14.0000	23.0000	7.0000	59.0
Ireland	16	30.0000	52.0000	99.0000	11.0000	80.0000	75.0000	18.0000	2.0000	5.0000	3.0000	57.0

Computer exercises 'Foods'

2.0000	10.0000	60.0000	2.0000	55.0000	41.0000	3.0000	2.0000	4.0000	2.0000	67.0000	71.0000
3.0000	42.0000	63.0000	4.0000	76.0000	53.0000	11.0000	23.0000	11.0000	5.0000	87.0000	84.0000
6.0000	62.0000	98.0000	32.00							83.0000	89.0000
4.0000	38.0000	48.0000	11.00							76.0000	76.0000
7.0000	61.0000	86.0000	28.00							85.0000	94.0000
7.0000	86.0000	99.0000	22.00							76.0000	68.0000
2.0000	26.0000	77.0000	2.00							22.0000	51.0000
6.0000	31.0000	61.0000	15.00							49.0000	42.0000
3.0000	72.0000	85.0000	25.00							79.0000	70.0000
7.0000	13.0000	93.0000	31.00							56.0000	78.0000
6.0000	17.0000	92.0000	35.00							81.0000	72.0000
2.0000	17.0000	83.0000	13.00							61.0000	72.0000
3.0000	12.0000	84.0000	20.00							50.0000	57.0000
0.0000	40.0000	40.0000								59.0000	77.0000
0.0000	52.0000	99.0000	11.00							57.0000	52.0000

Principal Component Analysis

Samples Variables

Variable Set:
 [20]

Keep Out of Calculation:

Weights:

Validation Method

Leverage Correction

Cross Validation

Uncertainty test: --- PCs

Test Set

Model Size: Num PCs:

Center Data

Add Start Noise

Issue Warnings

Computer exercises 'Foods'

- **Is Finland part of Scandinavia, and why not?**
 - **Can the data divide Europe in 'food-compartments'?**
 - **And which products make the split?**
 - **Can you predict coffee consumption?**
 - **...**
-

Computer exercises 'Octane' (Back-up)

- 39 gasoline samples x NIR spectra (= 226 variables)
- Octane number
- Additives

