https://www.statology.org/partial-least-squares-in-r/

Analytica Chimica Acta
Volume 185, 1986, Pages 1-17
Analytica Chimica Acta

Partial least-squares regression: a tutorial

PaulGeladi Bruce R.Kowalski

https://doi.org/10.1016/0003-2670(86)80028-9

Abstract

A tutorial on the partial least-squares (PLS) regression method is provided. Weak points in some other regression methods are outlined and PLS is developed as a remedy for those weaknesses. An algorithm for a predictive PLS and some practical hints for its use are given.

Harald Wold (PLS statistics) + Svante Arrhenius (Chemistry) -> Svante Wold (Chemometrics)

Partial Least Squares regression (PLS) is a regression method based on covariance. It is recommended in cases of regression where the number of explanatory variables is high, and where it is likely that there is multicollinearity among the variables, i.e. that the explanatory variables are correlated.
You can choose several response variables in one analysis, and different algorithms are available (NIPALS and SVD).

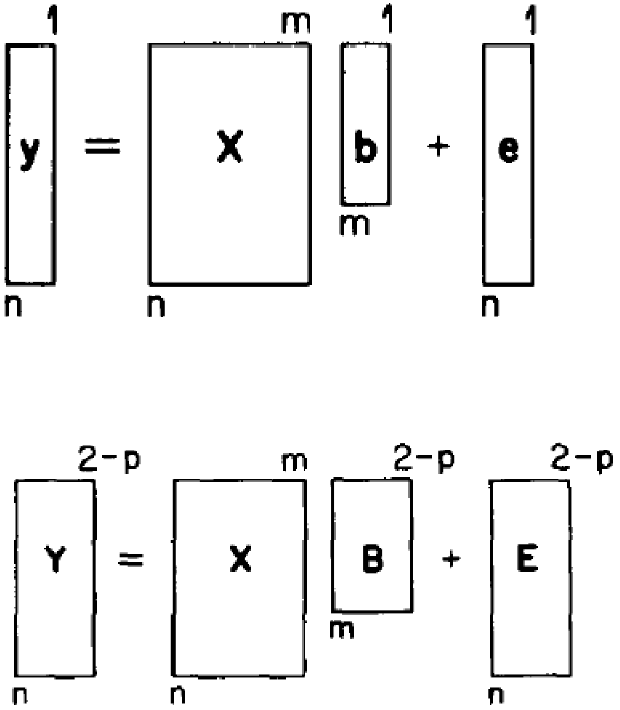
The **Partial Least Squares regression (PLS)** is a method which reduces the variables, used to predict, to a smaller set of predictors. These predictors are then used to perfom a regression.
The idea behind the **PLS regression** is to create, starting from a table with n observations described by p variables, a set of h components with the PLS 1 and PLS 2 algorithms
Some programs differentiate PLS 1 from PLS 2. **PLS 1** corresponds to the case where there is only **one dependent variable**. **PLS 2** corresponds to the case where there are **several dependent variables**.

Partial Least Squares regression model equations

In the case of the **Ordinary Least Squares** and **Principal Component Regression** methods, if models need to be computed for several dependent variables, the computation of the models is simply a loop on the columns of the dependent variables table Y.

In the case of PLS regression, the covariance structure of Y also influences the computations.

Ordinary Least Squares / Multiple Linear Regression



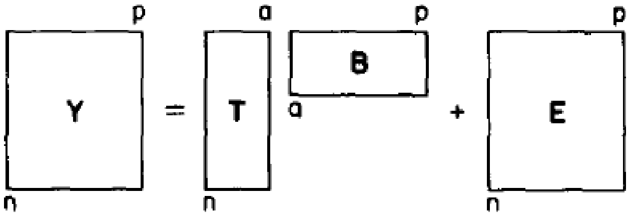
Summary: MLR

- For ***m* > *n*,** there is no unique solution unless one deletes independent variables.
- For ***m* = *n*,** there is one unique solution.
- For ***m* < *n*,** a least-squares solution is possible. For ***m* = *n* and *m* < *n*,** the matrix inversion can cause problems.
- MLR is possible with more than one dependent variable.

PaulGeladi Bruce R.Kowalski Analytica Chimica Acta Volume 185, 1986, Pages 1-17
Partial least-squares regression: a tutorial https://doi.org/10.1016/0003-2670(86)80028-9

PRINCIPAL COMPONENT REGRESSION (PCR)

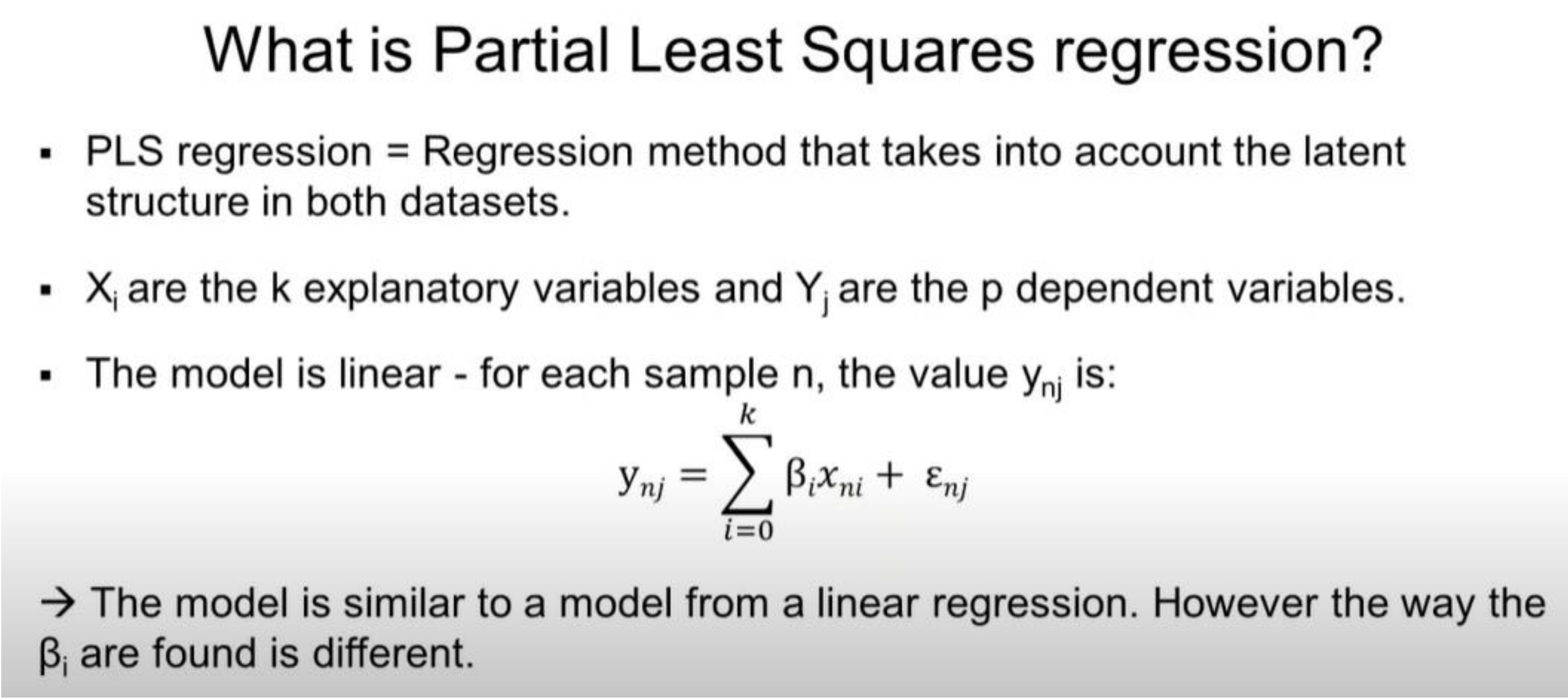
$Y = TB + E$



The variables of X are replaced by new ones that have better properties (orthogonality) and also span the multidimensional space of X.

Summary: PCR

- A data matrix can be represented by its score matrix.
- A regression of the score matrix against one or several dependent variables is possible, provided that scores corresponding to small eigenvalues are omitted.
- This regression gives no matrix inversion problems; it is well conditioned.



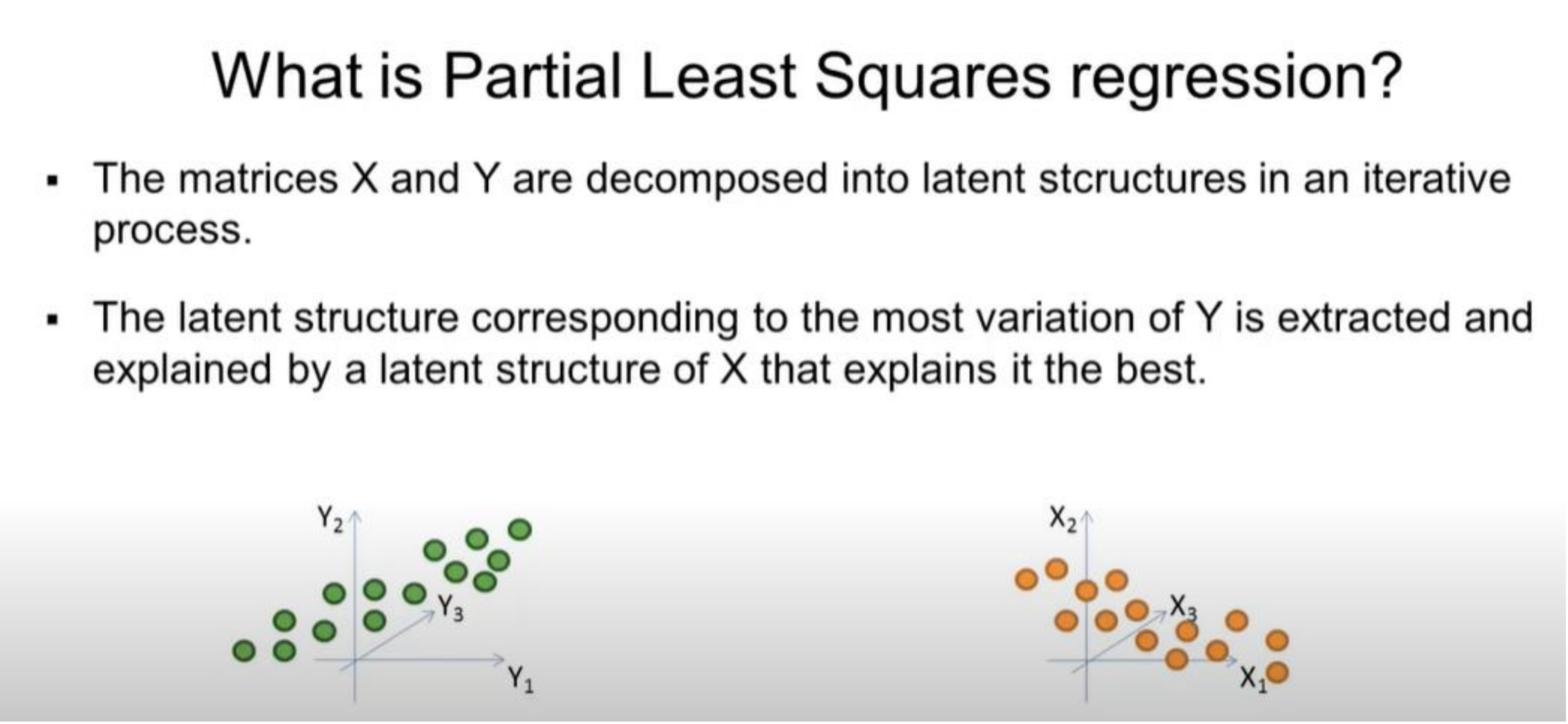
What is Partial Least Squares regression?

- PLS regression = Regression method that takes into account the latent structure in both datasets.
- X_i are the k explanatory variables and Y_j are the p dependent variables.
- The model is linear - for each sample n, the value y_{nj} is:

$$y_{nj} = \sum_{i=0}^{k} \beta_i x_{ni} + \varepsilon_{nj}$$

→ The model is similar to a model from a linear regression. However the way the β_i are found is different.

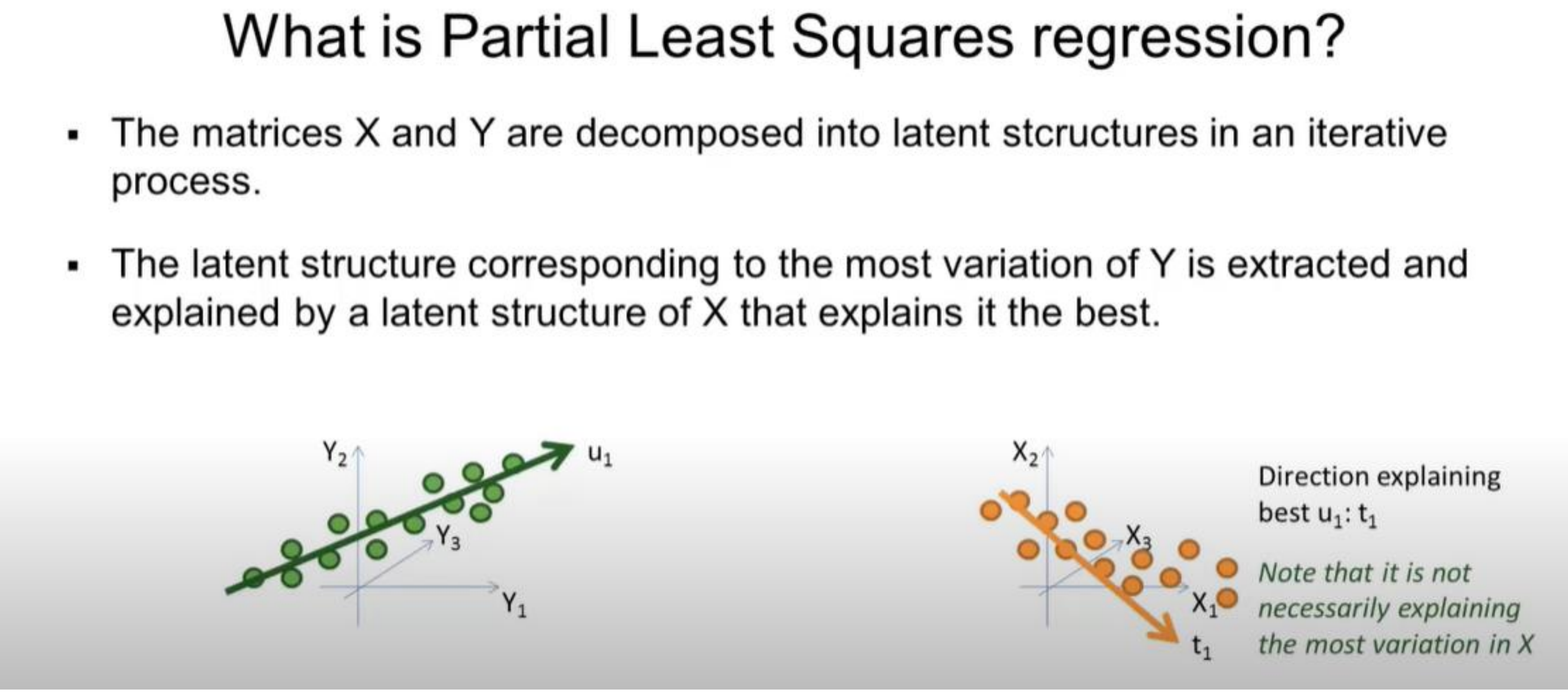
https://www.youtube.com/watch?v=WKEGhyFx0Dg



What is Partial Least Squares regression?

- The matrices X and Y are decomposed into latent stcructures in an iterative process.
- The latent structure corresponding to the most variation of Y is extracted and explained by a latent structure of X that explains it the best.

https://www.youtube.com/watch?v=WKEGhyFx0Dg



What is Partial Least Squares regression?

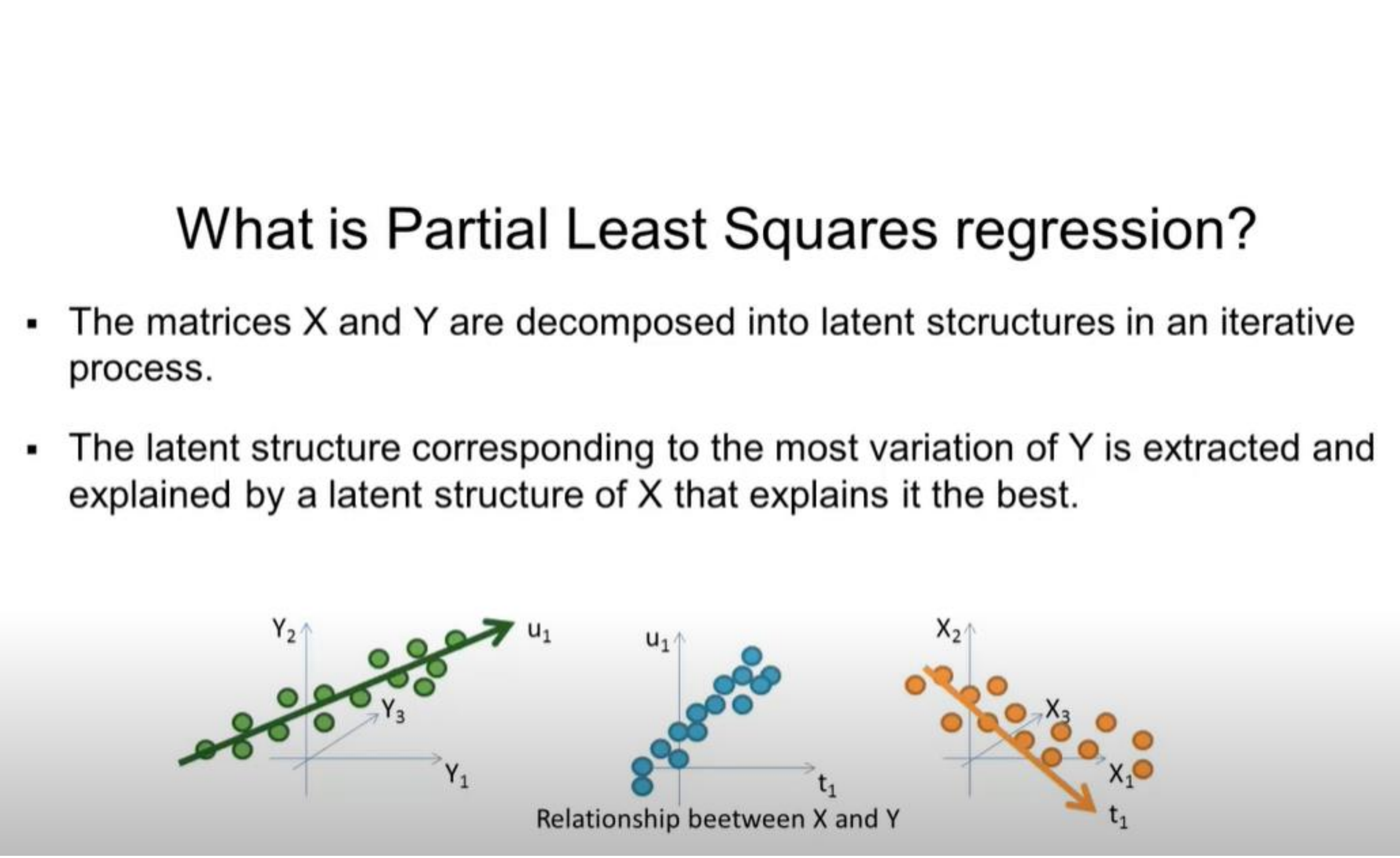
- The matrices X and Y are decomposed into latent stcructures in an iterative process.

- The latent structure corresponding to the most variation of Y is extracted and explained by a latent structure of X that explains it the best.

Direction explaining best u_1: t_1

Note that it is not necessarily explaining the most variation in X

https://www.youtube.com/watch?v=WKEGhyFx0Dg



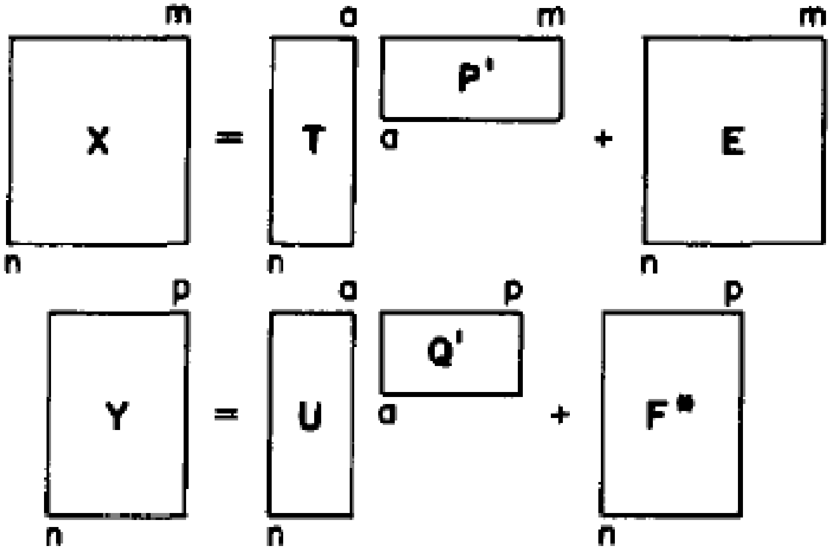
What is Partial Least Squares regression?

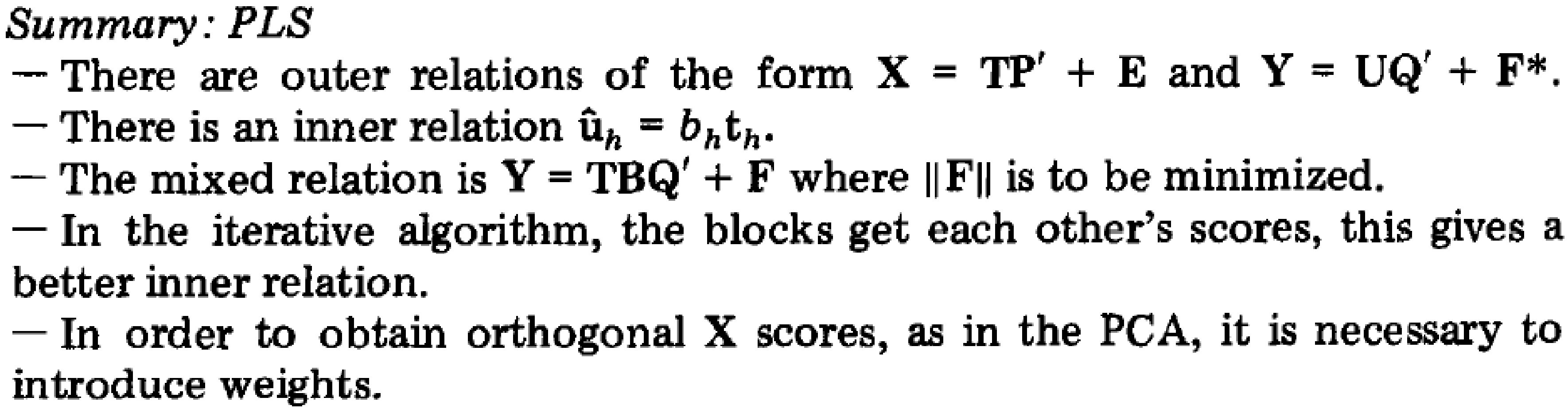
- The matrices X and Y are decomposed into latent stcructures in an iterative process.
- The latent structure corresponding to the most variation of Y is extracted and explained by a latent structure of X that explains it the best.

Relationship beetween X and Y

https://www.youtube.com/watch?v=WKEGhyFx0Dg

PLS





Summary: PLS

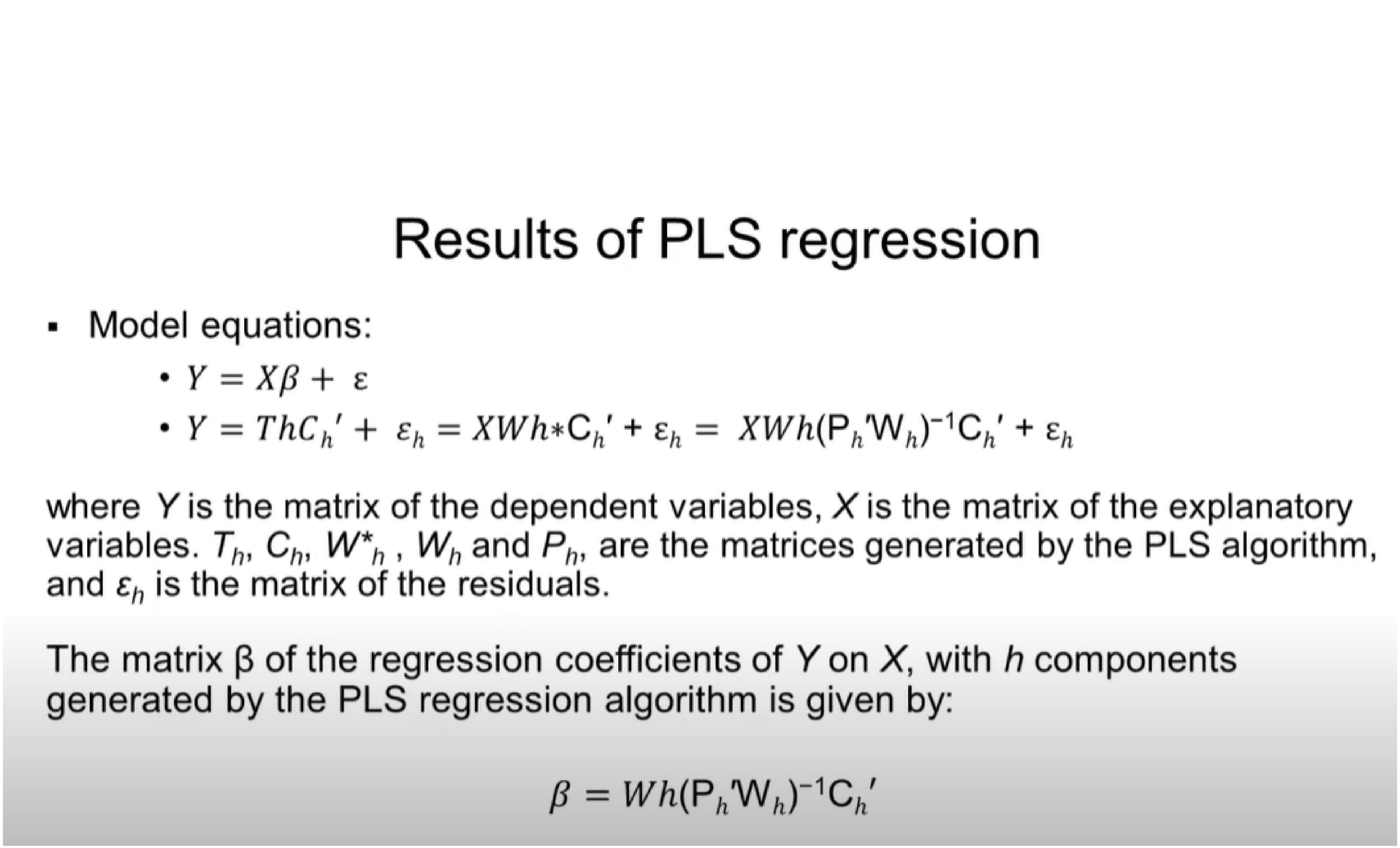
— There are outer relations of the form $\mathbf{X} = \mathbf{TP}' + \mathbf{E}$ and $\mathbf{Y} = \mathbf{UQ}' + \mathbf{F}^*$.

— There is an inner relation $\hat{\mathbf{u}}_h = b_h \mathbf{t}_h$.

— The mixed relation is $\mathbf{Y} = \mathbf{TBQ}' + \mathbf{F}$ where $\|\mathbf{F}\|$ is to be minimized.

— In the iterative algorithm, the blocks get each other's scores, this gives a better inner relation.

— In order to obtain orthogonal **X** scores, as in the PCA, it is necessary to introduce weights.



Results of PLS regression

- Model equations:
  - $Y = X\beta + \varepsilon$
  - $Y = ThC_h' + \varepsilon_h = XWh{*}C_h' + \varepsilon_h = XWh(P_h'W_h)^{-1}C_h' + \varepsilon_h$

where *Y* is the matrix of the dependent variables, *X* is the matrix of the explanatory variables. T_h, C_h, W^*_h , W_h and P_h, are the matrices generated by the PLS algorithm, and ε_h is the matrix of the residuals.

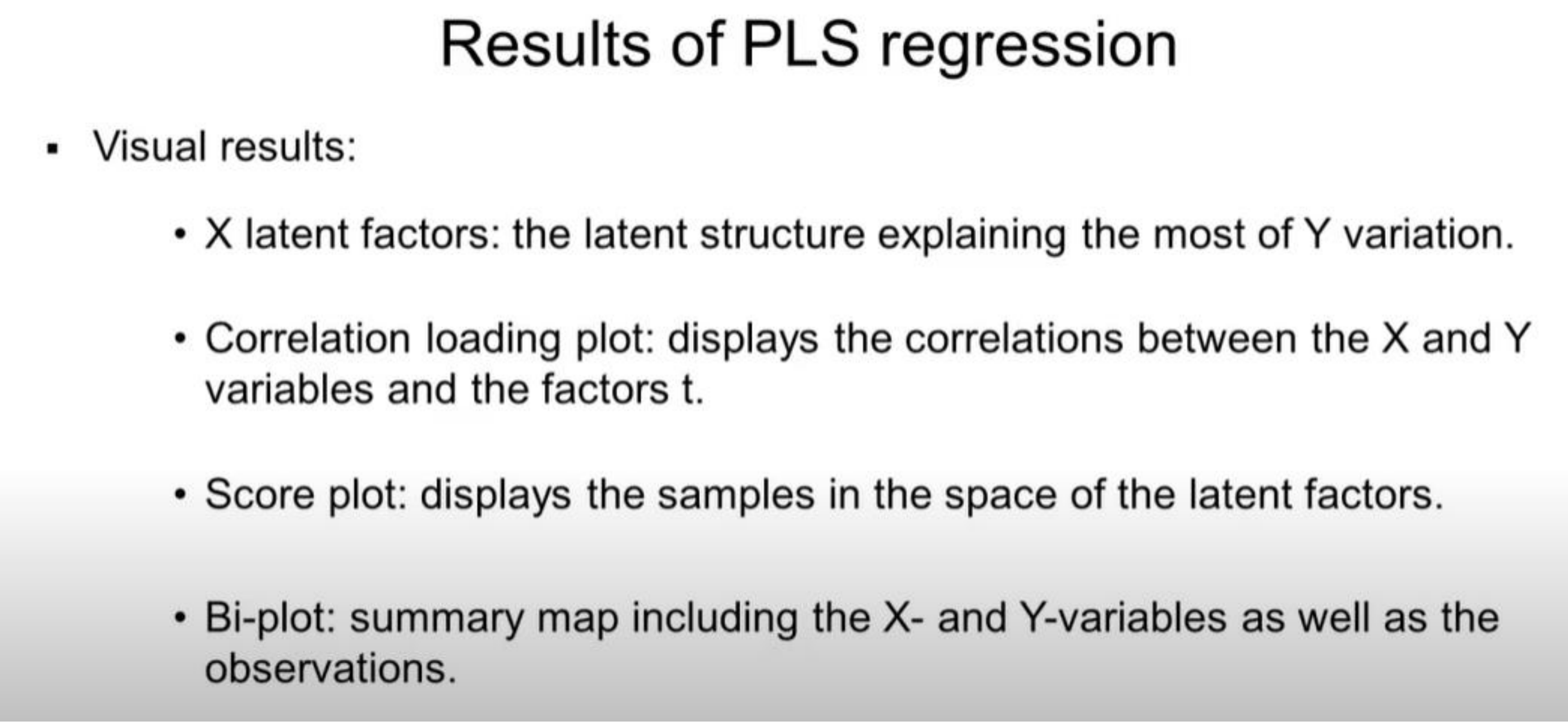
The matrix β of the regression coefficients of *Y* on *X*, with *h* components generated by the PLS regression algorithm is given by:

$$\beta = Wh(P_h'W_h)^{-1}C_h'$$

Algoritmo NIPALS (Nonlinear Iterative Partial Least Squares)
https://cran.r-project.org/web/packages/nipals/vignettes/nipals_algorithm.html

Algoritmo SVD https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wics.51

https://www.youtube.com/watch?v=WKEGhyFx0Dg



Results of PLS regression

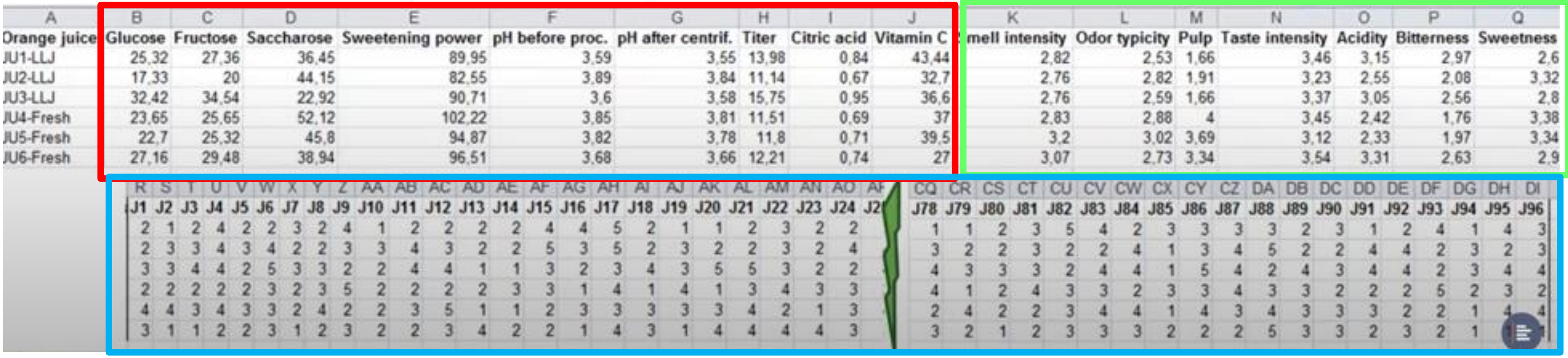
- Visual results:
  - X latent factors: the latent structure explaining the most of Y variation.
  - Correlation loading plot: displays the correlations between the X and Y variables and the factors t.
  - Score plot: displays the samples in the space of the latent factors.
  - Bi-plot: summary map including the X- and Y-variables as well as the observations.

https://www.youtube.com/watch?v=WKEGhyFx0Dg

Example of Partial Least Squares regression

- 6 orange juices
- 16 chemical and sensory characteristics
- Graded by 96 consumers

Model: Consumer preference = f(chemical, sensory)

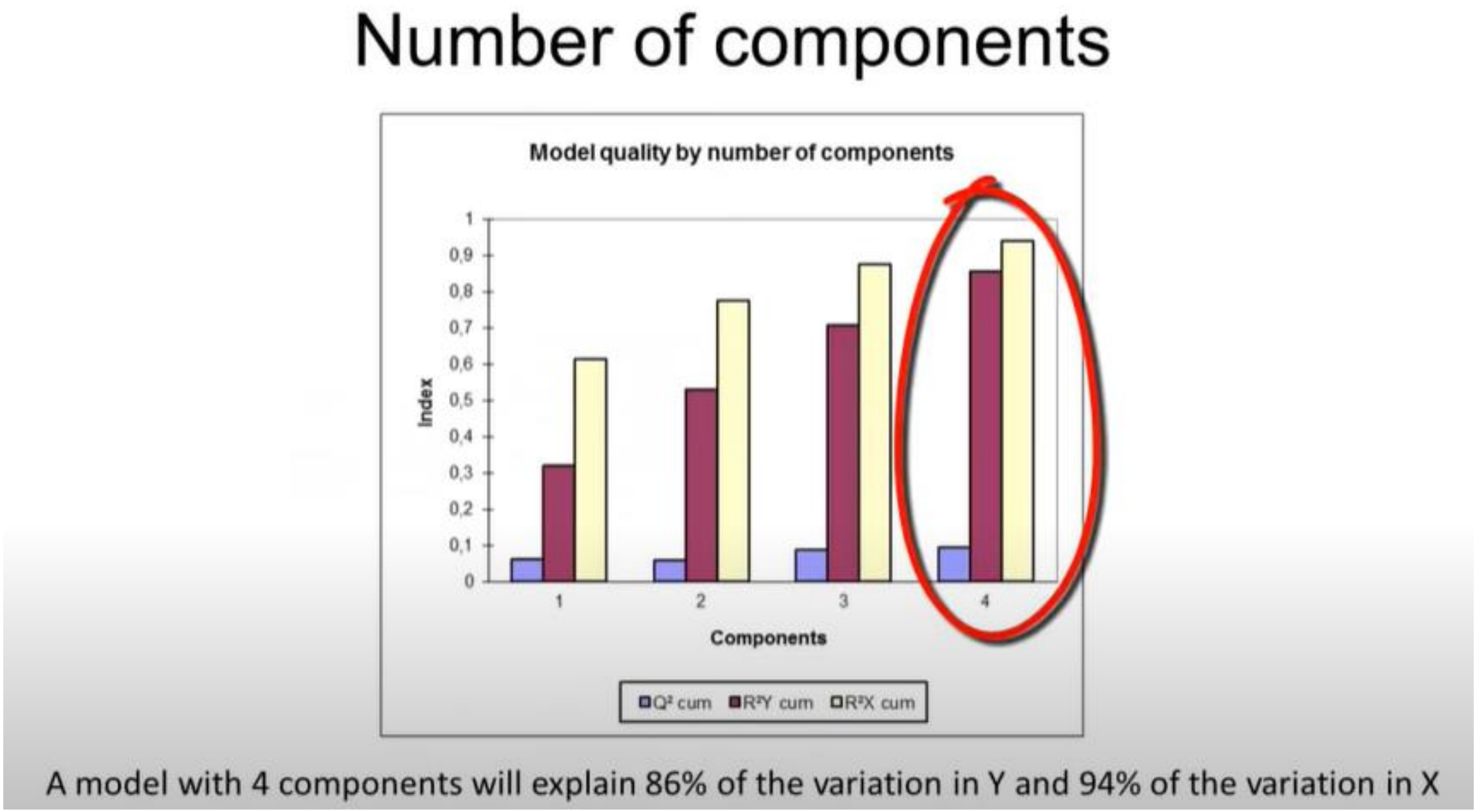


A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Orange juice	Glucose	Fructose	Saccharose	Sweetening power	pH before proc.	pH after centrif.	Titer	Citric acid	Vitamin C	Smell intensity	Odor typicity	Pulp	Taste intensity	Acidity	Bitterness	Sweetness
JU1-LLJ	25,32	27,36	36,45	89,95	3,59	3,55	13,98	0,84	43,44	2,82	2,53	1,66	3,46	3,15	2,97	2,6
JU2-LLJ	17,33	20	44,15	82,55	3,89	3,84	11,14	0,67	32,7	2,76	2,82	1,91	3,23	2,55	2,08	3,32
JU3-LLJ	32,42	34,54	22,92	90,71	3,6	3,58	15,75	0,95	36,6	2,76	2,59	1,66	3,37	3,05	2,56	2,8
JU4-Fresh	23,65	25,65	52,12	102,22	3,85	3,81	11,51	0,69	37	2,83	2,88	4	3,45	2,42	1,76	3,38
JU5-Fresh	22,7	25,32	45,8	94,87	3,82	3,78	11,8	0,71	39,5	3,2	3,02	3,69	3,12	2,33	1,97	3,34
JU6-Fresh	27,16	29,48	38,94	96,51	3,68	3,66	12,21	0,74	27	3,07	2,73	3,34	3,54	3,31	2,63	2,9

R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AF
J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12	J13	J14	J15	J16	J17	J18	J19	J20	J21	J22	J23	J24	J2
2	1	2	4	2	2	3	2	4	1	2	2	2	2	4	4	5	2	1	1	2	3	2	2	
2	3	3	4	3	4	2	2	3	3	4	3	2	2	5	3	5	2	3	2	2	3	2	4	
3	3	4	4	2	5	3	3	2	2	4	4	1	1	3	2	3	4	3	5	5	3	2	2	
2	2	2	2	2	3	2	3	5	2	2	2	2	3	3	1	4	1	4	1	3	4	3	3	
4	4	3	4	3	3	2	4	2	2	3	5	1	1	2	3	3	3	3	3	4	2	1	3	
3	1	1	2	2	3	1	2	3	2	2	3	4	2	2	1	4	3	1	4	4	4	4	3	

CQ	CR	CS	CT	CU	CV	CW	CX	CY	CZ	DA	DB	DC	DD	DE	DF	DG	DH	DI
J78	J79	J80	J81	J82	J83	J84	J85	J86	J87	J88	J89	J90	J91	J92	J93	J94	J95	J96
1	1	2	3	5	4	2	3	3	3	3	2	3	1	2	4	1	4	3
3	2	2	3	2	2	4	1	3	4	5	2	2	4	4	2	3	2	3
4	3	3	3	2	4	4	1	5	4	2	4	3	4	4	2	3	4	4
4	1	2	4	3	3	2	3	3	4	3	3	2	2	2	5	2	3	2
2	4	2	2	3	4	4	1	4	3	4	3	3	3	2	2	1	4	4
3	2	1	2	3	3	3	2	2	2	5	3	3	2	3	2	1	[illegible]	1

https://www.youtube.com/watch?v=WKEGhyFx0Dg



Number of components

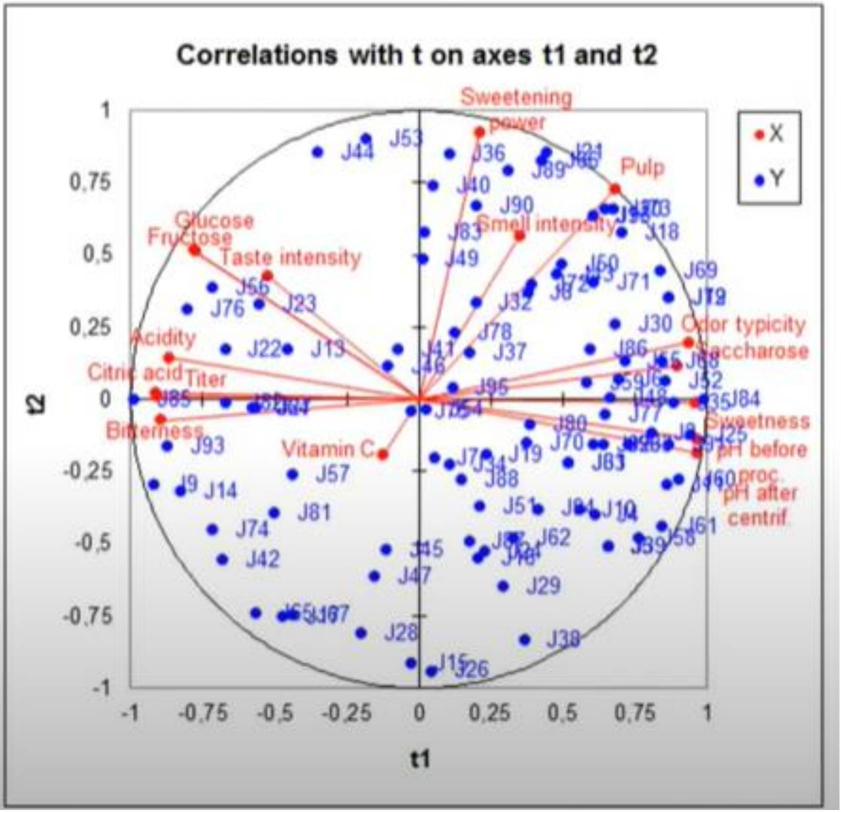
A model with 4 components will explain 86% of the variation in Y and 94% of the variation in X

https://www.youtube.com/watch?v=WKEGhyFx0Dg

Q2cum R2Ycum R2Xcum spiegare

Correlations

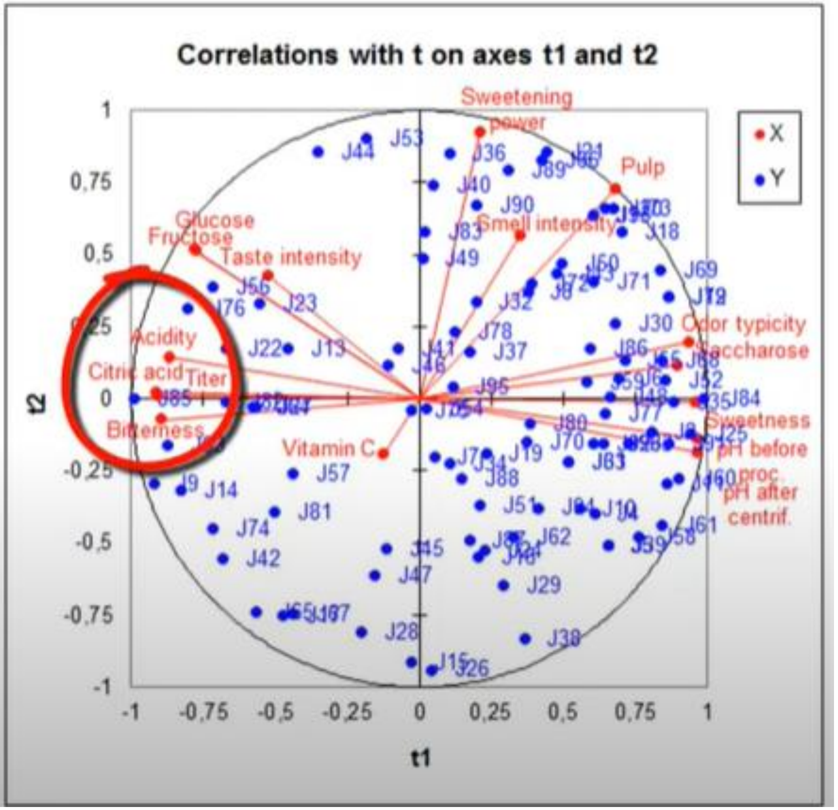
- Acidity and citric acid correlated
- Acidity and pH anticorrelated
- Bitterness and acidity correlated
- Glucose and fructose correlated
- …
- Correlation between consumer preferences and chemical and sensory characteristics



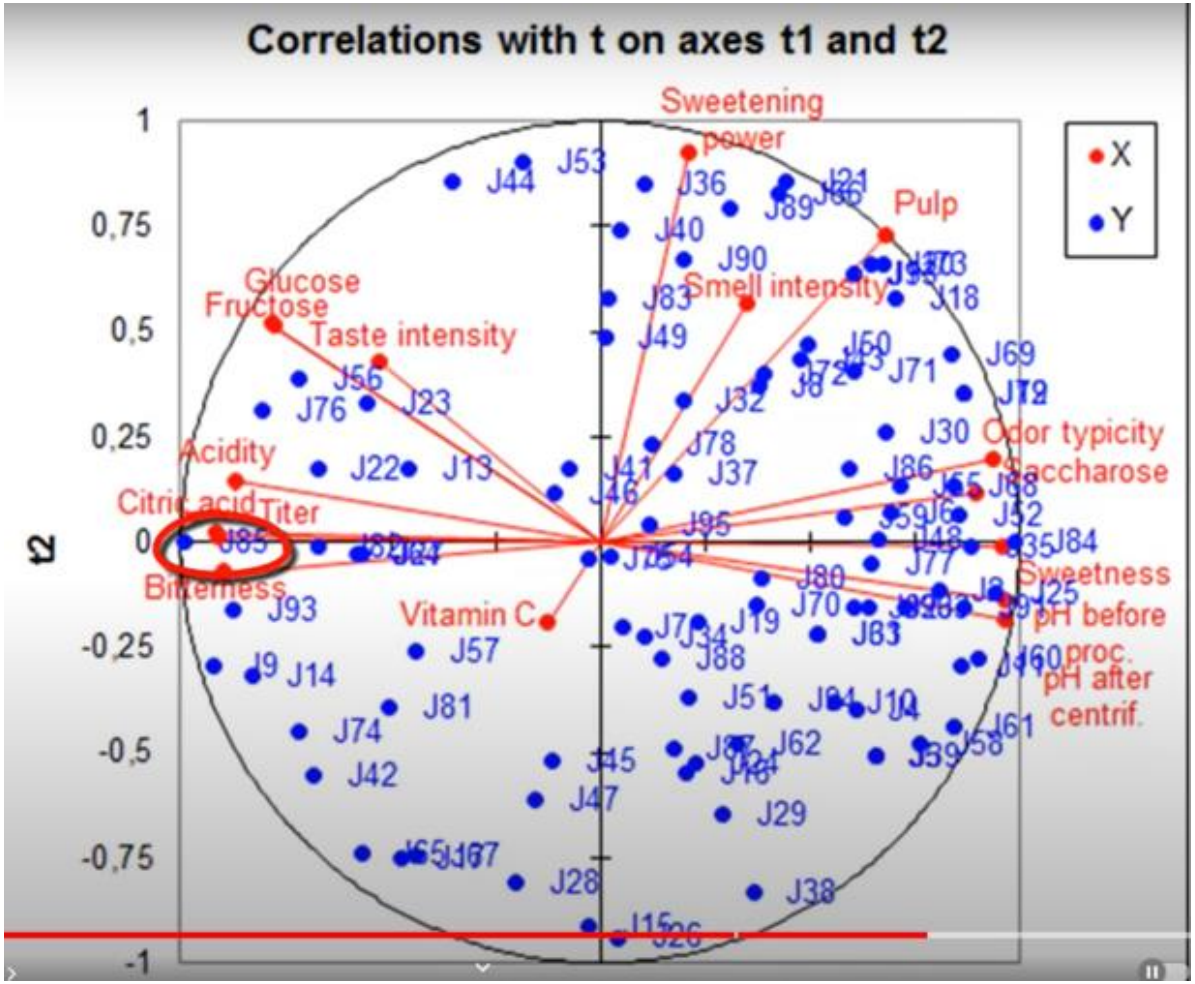
https://www.youtube.com/watch?v=WKEGhyFx0Dg

Correlations

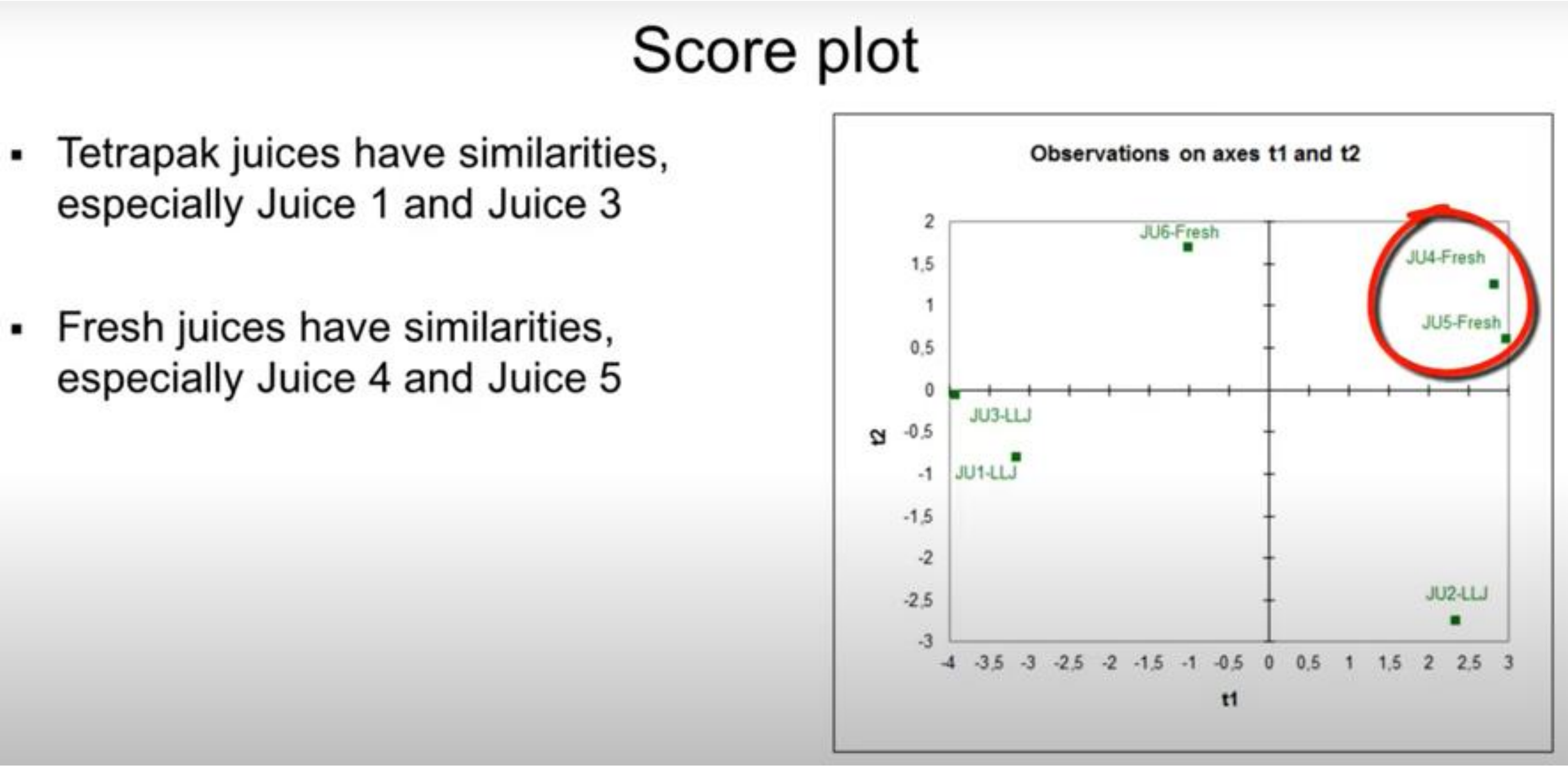
- Acidity and citric acid correlated
- Acidity and pH anticorrelated
- Bitterness and acidity correlated
- Glucose and fructose correlated
- …
- Correlation between consumer preferences and chemical and sensory characteristics



https://www.youtube.com/watch?v=WKEGhyFx0Dg



https://www.youtube.com/watch?v=WKEGhyFx0Dg



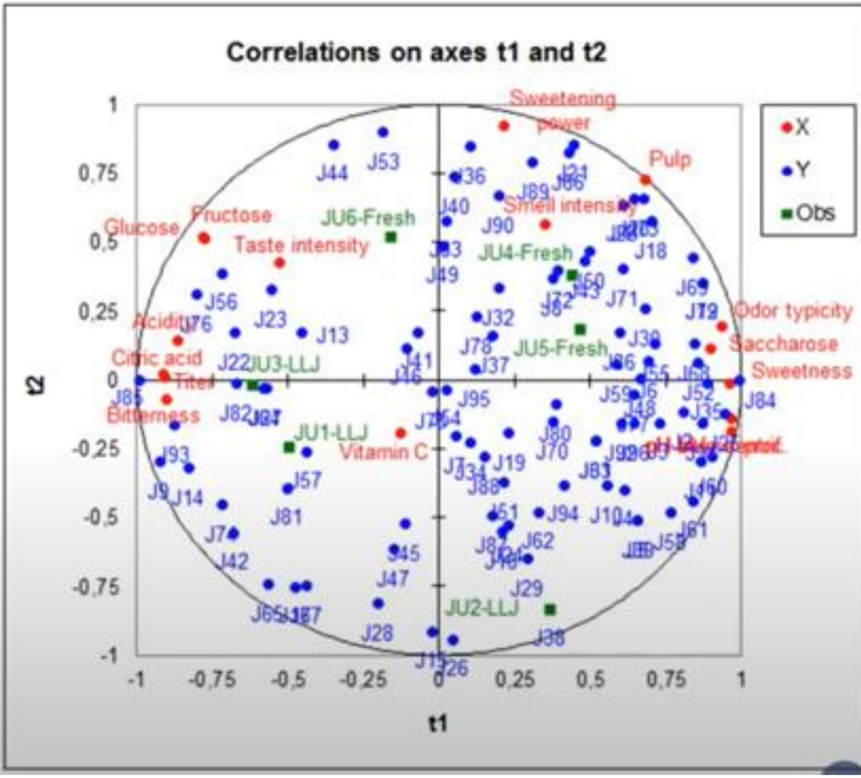
Score plot

- Tetrapak juices have similarities, especially Juice 1 and Juice 3
- Fresh juices have similarities, especially Juice 4 and Juice 5

https://www.youtube.com/watch?v=WKEGhyFx0Dg

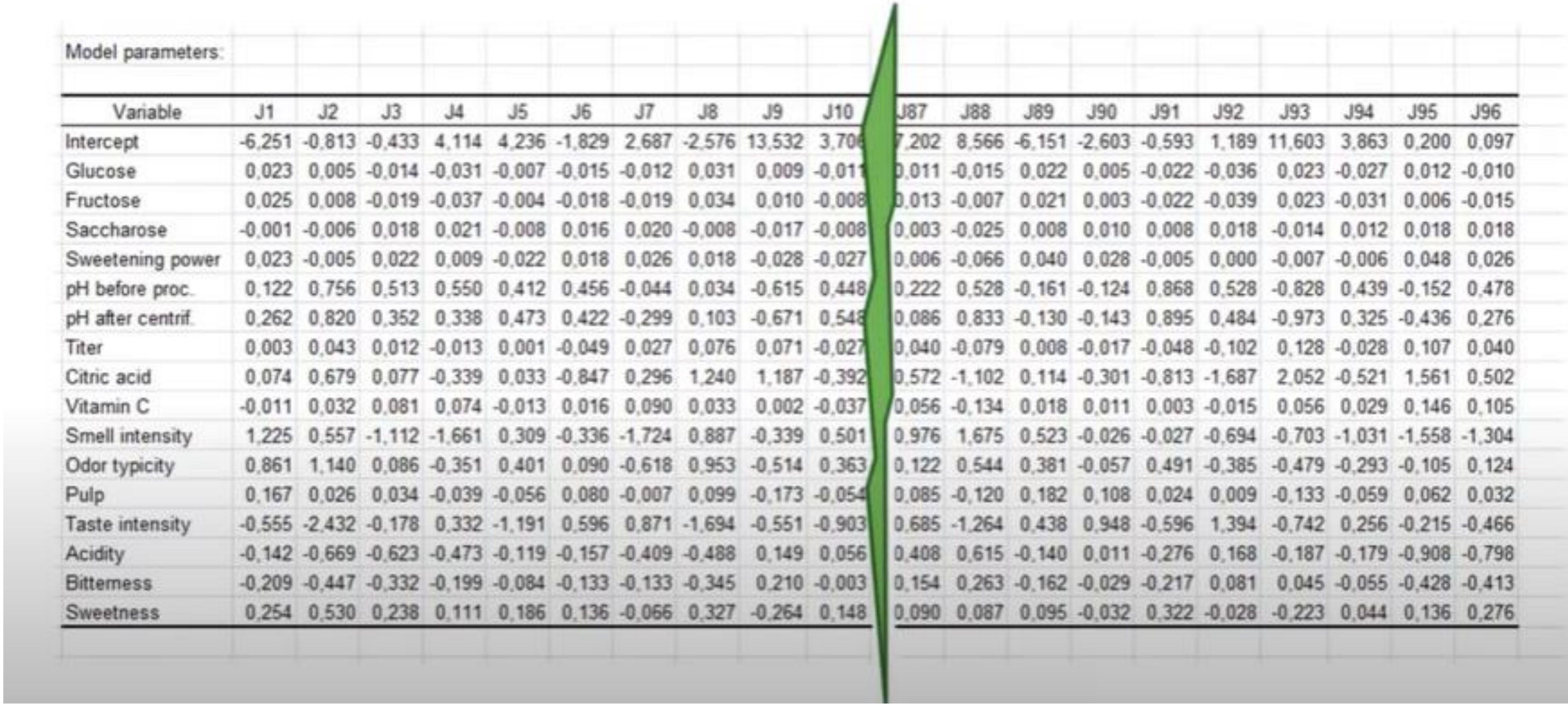
Biplot

- Fresh juices Juice 4 and Juice 5 have Pulp and have a high smell intensity, they are neither bitter, nor acidic
- JU3-LLJ is acidic and bitter, it does not have a typical smell
- Judge 85 likes bitter and acidic juices
- Judge 84 likes sweet juices
- …



https://www.youtube.com/watch?v=WKEGhyFx0Dg

Model equations



Model parameters:

Variable	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J87	J88	J89	J90	J91	J92	J93	J94	J95	J96
Intercept	-6,251	-0,813	-0,433	4,114	4,236	-1,829	2,687	-2,576	13,532	3,706	7,202	8,566	-6,151	-2,603	-0,593	1,189	11,603	3,863	0,200	0,097
Glucose	0,023	0,005	-0,014	-0,031	-0,007	-0,015	-0,012	0,031	0,009	-0,011	0,011	-0,015	0,022	0,005	-0,022	-0,036	0,023	-0,027	0,012	-0,010
Fructose	0,025	0,008	-0,019	-0,037	-0,004	-0,018	-0,019	0,034	0,010	-0,008	0,013	-0,007	0,021	0,003	-0,022	-0,039	0,023	-0,031	0,006	-0,015
Saccharose	-0,001	-0,006	0,018	0,021	-0,008	0,016	0,020	-0,008	-0,017	-0,008	0,003	-0,025	0,008	0,010	0,008	0,018	-0,014	0,012	0,018	0,018
Sweetening power	0,023	-0,005	0,022	0,009	-0,022	0,018	0,026	0,018	-0,028	-0,027	0,006	-0,066	0,040	0,028	-0,005	0,000	-0,007	-0,006	0,048	0,026
pH before proc.	0,122	0,756	0,513	0,550	0,412	0,456	-0,044	0,034	-0,615	0,448	0,222	0,528	-0,161	-0,124	0,868	0,528	-0,828	0,439	-0,152	0,478
pH after centrif.	0,262	0,820	0,352	0,338	0,473	0,422	-0,299	0,103	-0,671	0,548	0,086	0,833	-0,130	-0,143	0,895	0,484	-0,973	0,325	-0,436	0,276
Titer	0,003	0,043	0,012	-0,013	0,001	-0,049	0,027	0,076	0,071	-0,027	0,040	-0,079	0,008	-0,017	-0,048	-0,102	0,128	-0,028	0,107	0,040
Citric acid	0,074	0,679	0,077	-0,339	0,033	-0,847	0,296	1,240	1,187	-0,392	0,572	-1,102	0,114	-0,301	-0,813	-1,687	2,052	-0,521	1,561	0,502
Vitamin C	-0,011	0,032	0,081	0,074	-0,013	0,016	0,090	0,033	0,002	-0,037	0,056	-0,134	0,018	0,011	0,003	-0,015	0,056	0,029	0,146	0,105
Smell intensity	1,225	0,557	-1,112	-1,661	0,309	-0,336	-1,724	0,887	-0,339	0,501	0,976	1,675	0,523	-0,026	-0,027	-0,694	-0,703	-1,031	-1,558	-1,304
Odor typicity	0,861	1,140	0,086	-0,351	0,401	0,090	-0,618	0,953	-0,514	0,363	0,122	0,544	0,381	-0,057	0,491	-0,385	-0,479	-0,293	-0,105	0,124
Pulp	0,167	0,026	0,034	-0,039	-0,056	0,080	-0,007	0,099	-0,173	-0,054	0,085	-0,120	0,182	0,108	0,024	0,009	-0,133	-0,059	0,062	0,032
Taste intensity	-0,555	-2,432	-0,178	0,332	-1,191	0,596	0,871	-1,694	-0,551	-0,903	0,685	-1,264	0,438	0,948	-0,596	1,394	-0,742	0,256	-0,215	-0,466
Acidity	-0,142	-0,669	-0,623	-0,473	-0,119	-0,157	-0,409	-0,488	0,149	0,056	0,408	0,615	-0,140	0,011	-0,276	0,168	-0,187	-0,179	-0,908	-0,798
Bitterness	-0,209	-0,447	-0,332	-0,199	-0,084	-0,133	-0,133	-0,345	0,210	-0,003	0,154	0,263	-0,162	-0,029	-0,217	0,081	0,045	-0,055	-0,428	-0,413
Sweetness	0,254	0,530	0,238	0,111	0,186	0,136	-0,066	0,327	-0,264	0,148	0,090	0,087	0,095	-0,032	0,322	-0,028	-0,223	0,044	0,136	0,276

https://www.youtube.com/watch?v=WKEGhyFx0Dg

Advantages of PLS

- PLS regression:
  - Deals with multicolinearity
  - Allows taking into account the data structure
  - Provides visual results that help the interpretation
  - Can model several response variables at the same time taking into account their structure

General remarks about PLS regression

The three methods – Partial Least Squares regression (PLS), Principal Component regression (PCR), which is based on Principal Component analysis (PCA), and Ordinary Least Squares regression (OLS), which is the regular linear regression, - give the same results if the number of components obtained from the Principal Component analysis (PCA) in the PCR, or from the PLS regression is equal to the number of explanatory variables.

What is the difference between PCR and PLS regression?

The components obtained from the PLS regression, which is based on covariance, are built so that they explain as well as possible Y, while the components of the PCR are built to describe X as well as possible. This explains why the PLS regression outperforms PCR when the target is strongly correlated with a direction in the data that have a low variance.

https://www.xlstat.com/en/solutions/features/
partial-least-squares-regression

Ricapitolazione Programma 1

1. Approcci statistici classici: indicatori e statistiche univariate. Descrizione di descrittori numerici semplici delle raccolte di dati: statistiche descrittive parametriche e non parametriche. Visualizzazione di statistiche uni- e bi-variate. Esempi pratici in ambiente R software.
2. Visualizzazione di set di dati reali con diverse tecniche di rappresentazione grafica, analisi dei risultati tramite esplorazione visiva, tecniche di individuazione di dati anomali. Illustrazione del concetto di carta di controllo dei dati. Esempi pratici in ambiente R software.
3. Metodi di raggruppamento e classificazione: aspetti teorici ed applicativi dell'analisi di raggruppamento gerarchico e non gerarchico e dei metodi di classificazione supervisionata. Esempi pratici in ambiente R software.
4. Analisi delle componenti principali e metodi fattoriali: aspetti teorici ed applicativi dell'analisi delle componenti principali e dei metodi fattoriali per la compressione ed interpretazione dell'informazione contenuta nei dati. Esempi pratici in ambiente R software.
5. Metodi di regressione: aspetti teorici ed applicativi del modellamento di dati e predizione con metodi di regressione multilineare e, calibrazioni strumentali e diagnostiche di regressione. Esempi pratici in ambiente R software.

Ricapitolazione 2

…

6. Metodi di regressione delle componenti principali: aspetti teorici ed applicativi del modellamento di dati e predizione con algoritmi principal component regression (PCR) e partial-least square regression (PLS). Esempi pratici in ambiente R software.

7. La progettazione degli esperimenti: aspetti teorici ed applicativi di design fattoriale e screening, controllo di fattori e variabili dipendenti, metodi avanzati. Esempi pratici in ambiente R software.

8. Modelli di relazione quantitativa struttura-attività (QSAR-Quantitative Structure Activity Realtionship): descrittori molecolari, aspetti teorici ed applicativi di classificazione, regressione e predizione. Esempi pratici in free software dedicati.

9. Panoramica sui metodi di analisi di dati esplorativa e supervisionata con reti neurali artificiali. Software dedicati ed esempi di applicazione.