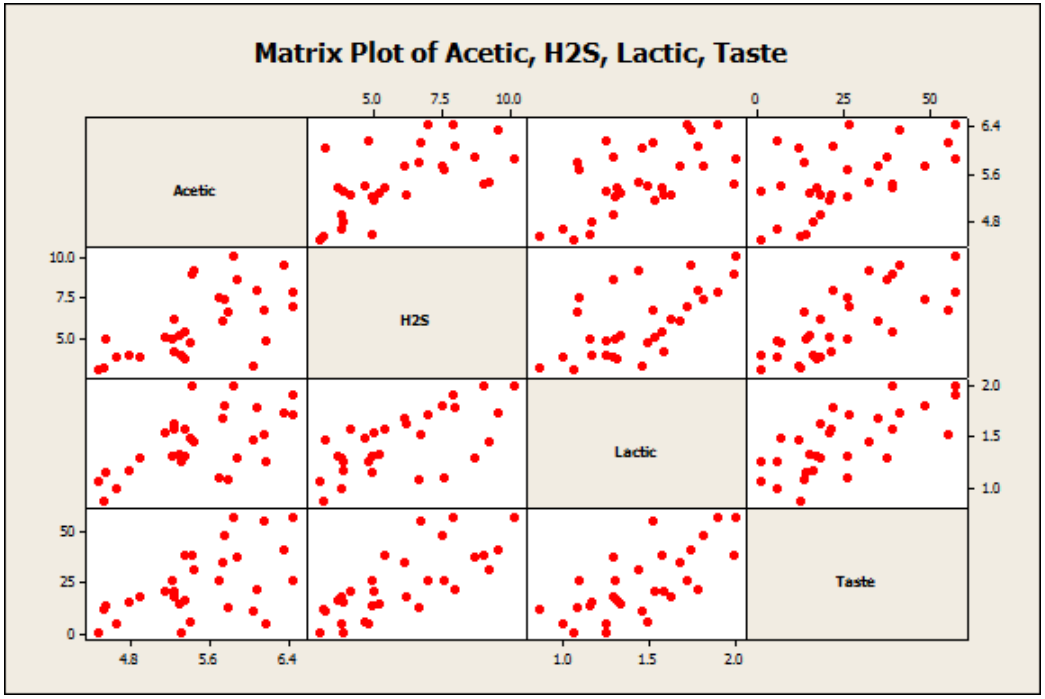
Example: Taste of cheese

- Description
  - Three x's: Concentrations of acetic acid, H_2S, and lactic acid in 30 samples of mature cheddar cheese.
  - One y: a subjective taste value is also provided for each sample.



Example: Taste of cheese

- Build MLR model

The regression equation is
Taste = - 28.9 + 0.31 Acetic + 3.92 H2S + 19.7 Lactic

Predictor	Coef	SE Coef	T	P
Constant	-28.85	19.74	-1.46	0.156
Acetic	0.315	4.464	0.07	0.944
H2S	3.920	1.248	3.14	0.004
Lactic	19.674	8.647	2.28	0.031

S = 10.1341 R-Sq = 65.2% R-Sq(adj) = 61.1%

PRESS = 3406.38 R-Sq(pred) = 55.55%

Example : Taste of cheese

- Correlation coefficients

	Acetic	H2S
H2S	**0.618**	
Lactic	**0.604**	**0.644**

Example : Taste of cheese

- Build PCR model with A=1

The regression equation is
Taste = 24.5 - 8.41 PC1

Predictor	Coef	SE Coef	T	P
Constant	24.534	1.909	12.85	0.000
PC1	-8.410	1.296	-6.49	0.000

S = 10.4546 R-Sq = 60.1% R-Sq(adj) = 58.6%

PRESS = 3495.98 R-Sq(pred) = 54.38%

Example : Taste of cheese

- Build PCR model with A=2

The regression equation is
Taste = 24.5 - 8.41 PC1 + 5.25 PC2

Predictor	Coef	SE Coef	T	P
Constant	24.534	1.839	13.34	0.000
PC1	-8.410	1.249	-6.73	0.000
PC2	5.248	2.954	1.78	0.087

S = 10.0740 R-Sq = 64.2% R-Sq(adj) = 61.6%

PRESS = 3392.43 R-Sq(pred) = 55.73%

Example : Taste of cheese

- Build PLS

Model Selection and Validation for Taste

Components	X Variance	Error SS	R-Sq
1	0.746493	2935.08	0.617007
2	0.879418	2670.75	0.651498

Regression Coefficients

	Taste	Taste standardized
Constant	-28.9465	0.000000
Acetic	0.2476	0.008694
H2S	3.8541	0.504261
Lactic	20.2673	0.377578

Example : Taste of cheese

- Comparison of MLR, PCR, PLS
  - Confidence interval for predicted y
  - Obs. 20, taste = 38.90

MLR		PCR (A=1)	
predicted	**C.I (95%)**	**predicted**	**C.I (95%)**
47.54	**(36.89, 58.20)**	**40.04**	**(33.77, 46.30)**

PCR (A=2)		PLS (A=2)	
predicted	**C.I (95%)**	**predicted**	**C.I (95%)**
47.54	**(36.97, 58.11)**	**47.66**	**(37.77, 57.55)**

Example : Taste of cheese

- FYI
  - PLS

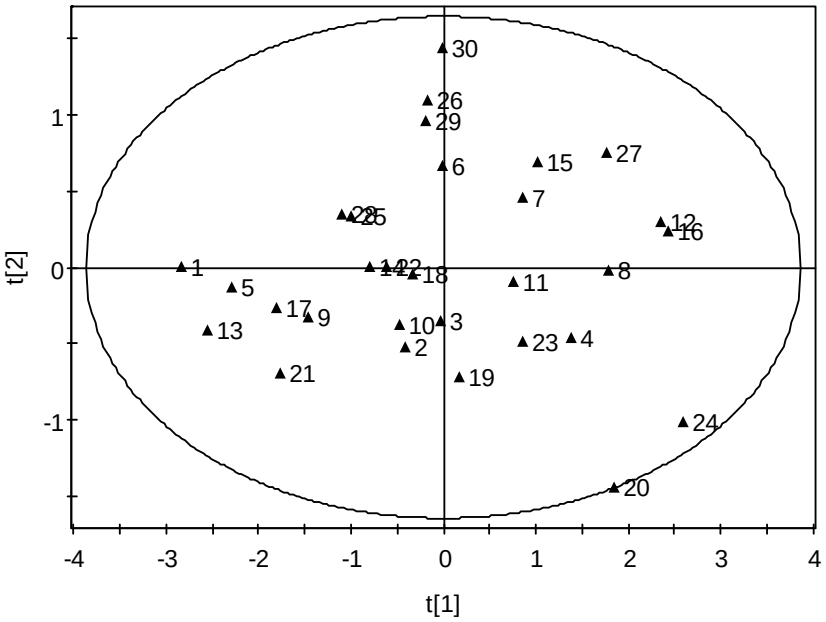
A	R2X	R2X (cum)	eigenvalue	R2Y	R2Y (cum)	Q2	Q2 (cum)
1	0.746	0.746	2.239	0.617	0.617	0.592	0.592
2	0.133	0.879	0.399	0.034	0.651	0.025	0.602

  - PCA

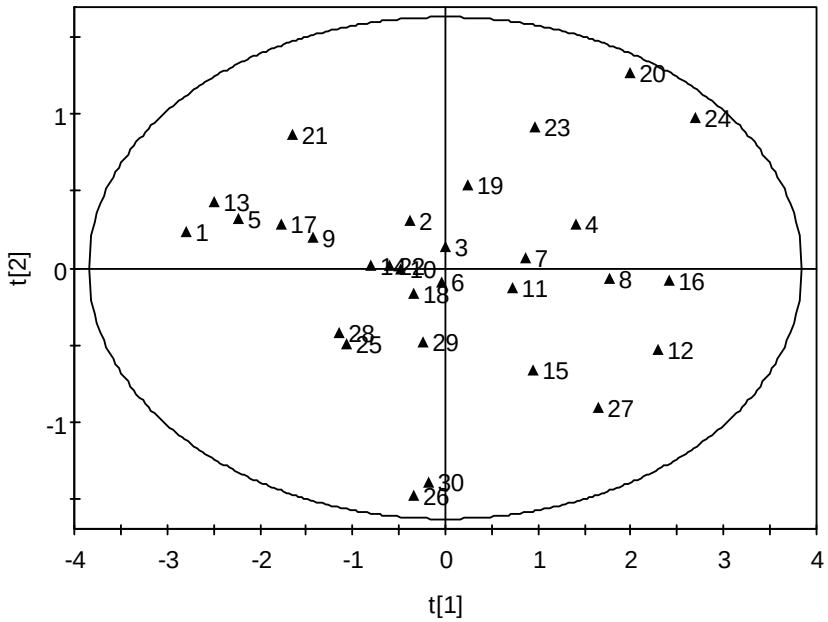
A	R2X	R2X (cum)	eigenvalue	Q2	Q2 (cum)
1	0.748	0.748	2.244	0.417	0.417
2	0.134	0.882	0.401	-0.253	0.417

Example : Taste of cheese

PCA score plot (t_1 va. t_2)

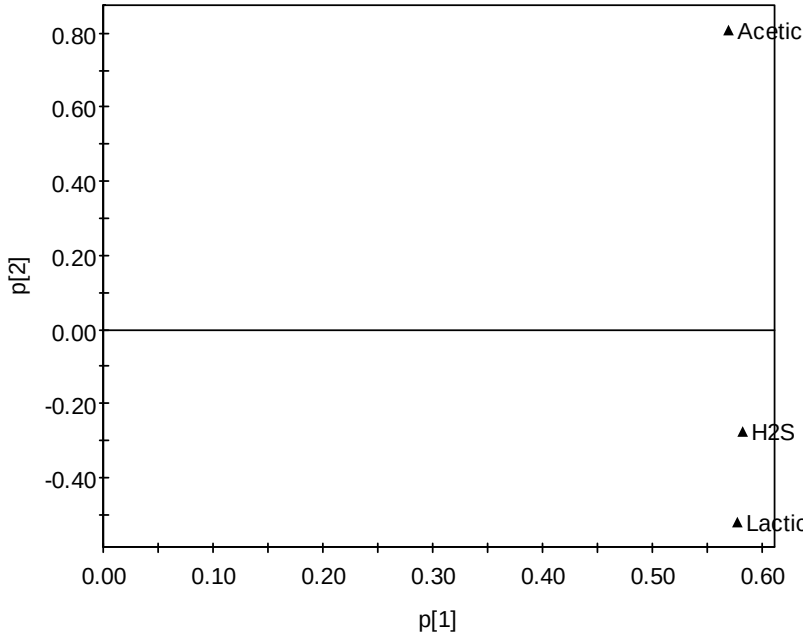


PLS score plot (t_1 vs. t_2)

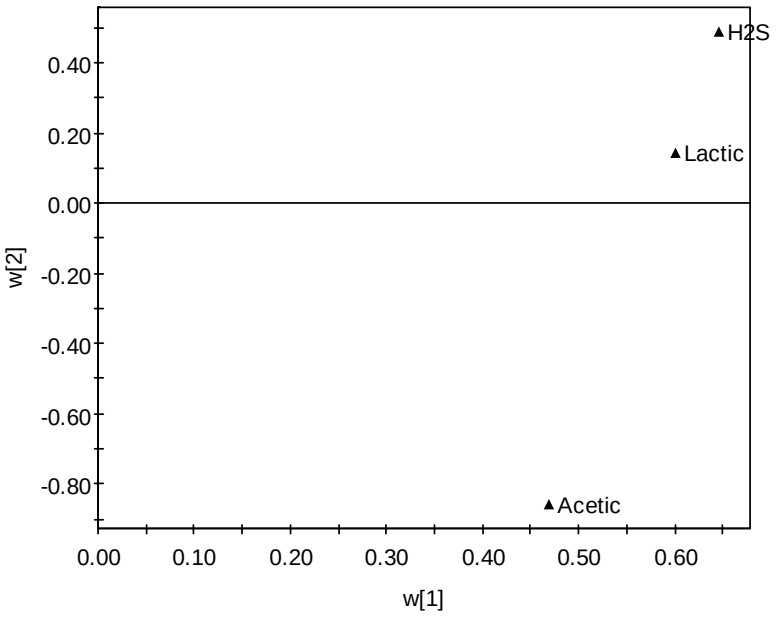


Example : Taste of cheese

PCA weights plot (p_1 vs. p_2)



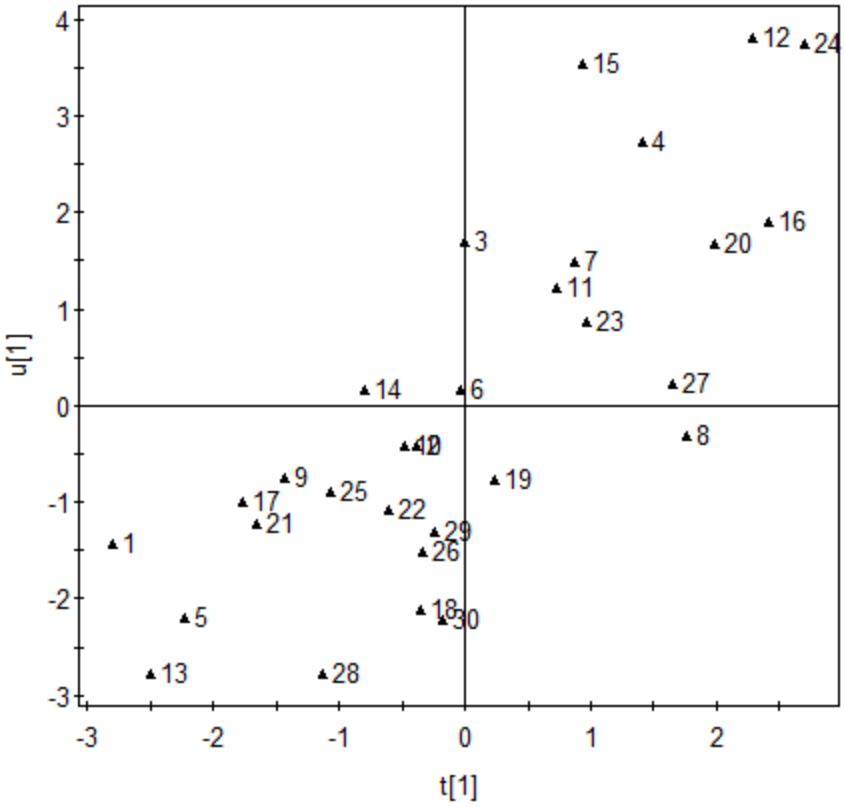
PLS weights plot (w_1 vs. w_2)

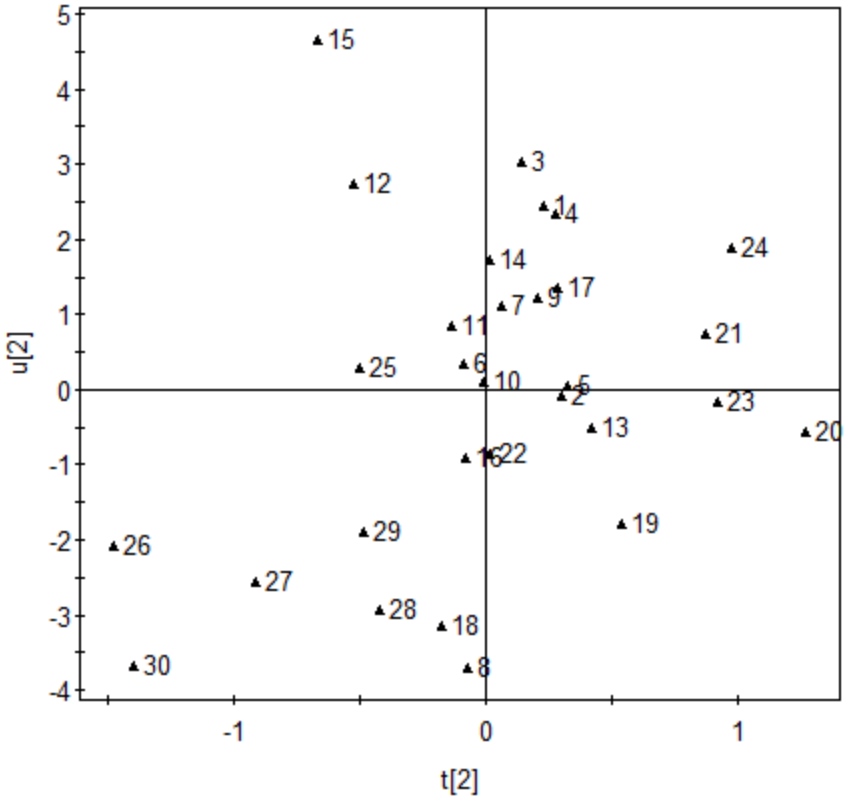


Different. Why?

Example : Taste of cheese

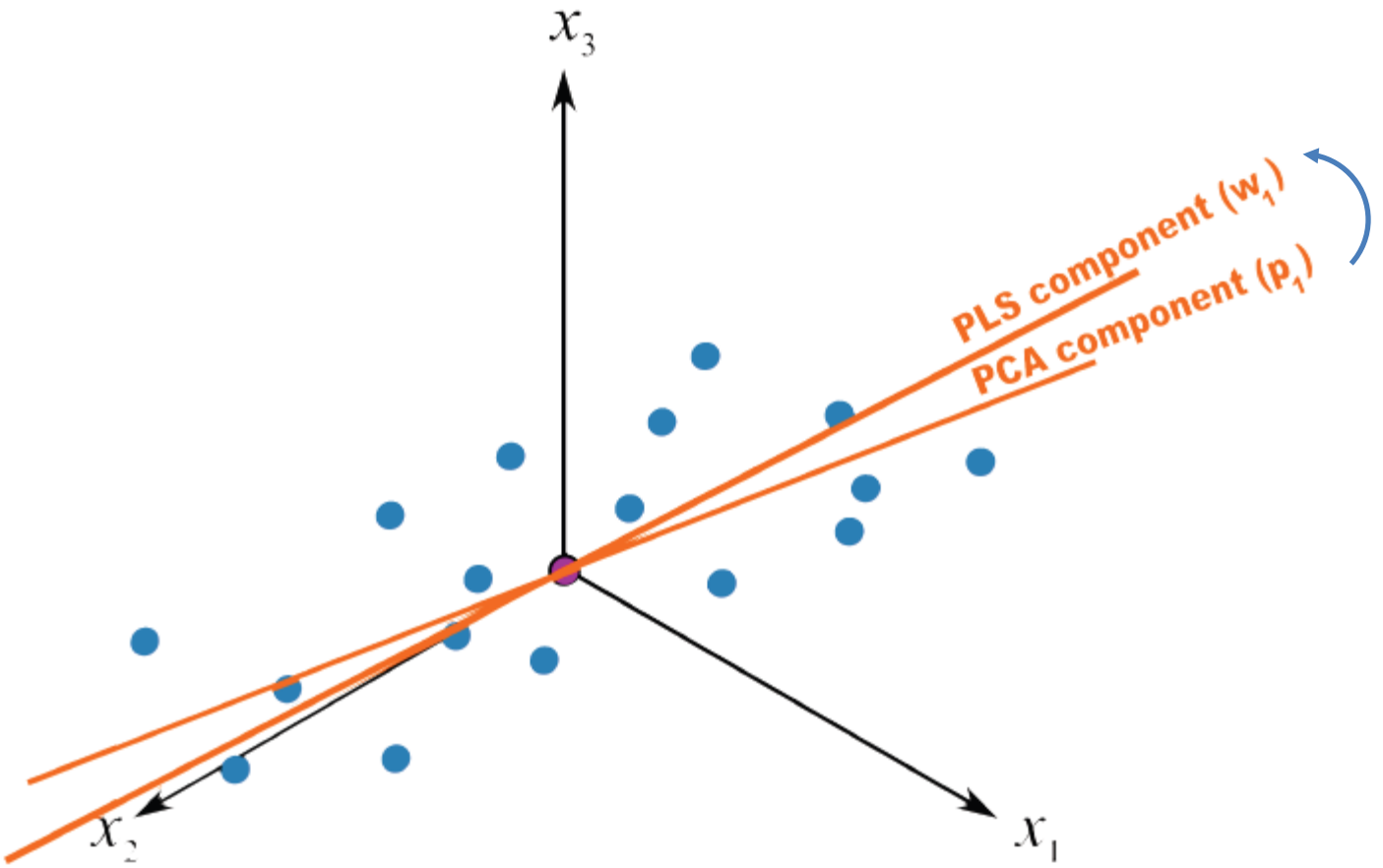
- PLS Score plots ($\mathbf{t}_i$ vs. $\mathbf{u}_i$)





Interpreting scores in PLS

- PLS scores interpreted exactly the same as PCA scores
- Verify correlation: plot $\mathbf{t}_a$ against $\mathbf{u}_a$: 45 degree line
- Usually just visualize the $\mathbf{t}_a$ scores



Interpreting the loadings in PLS

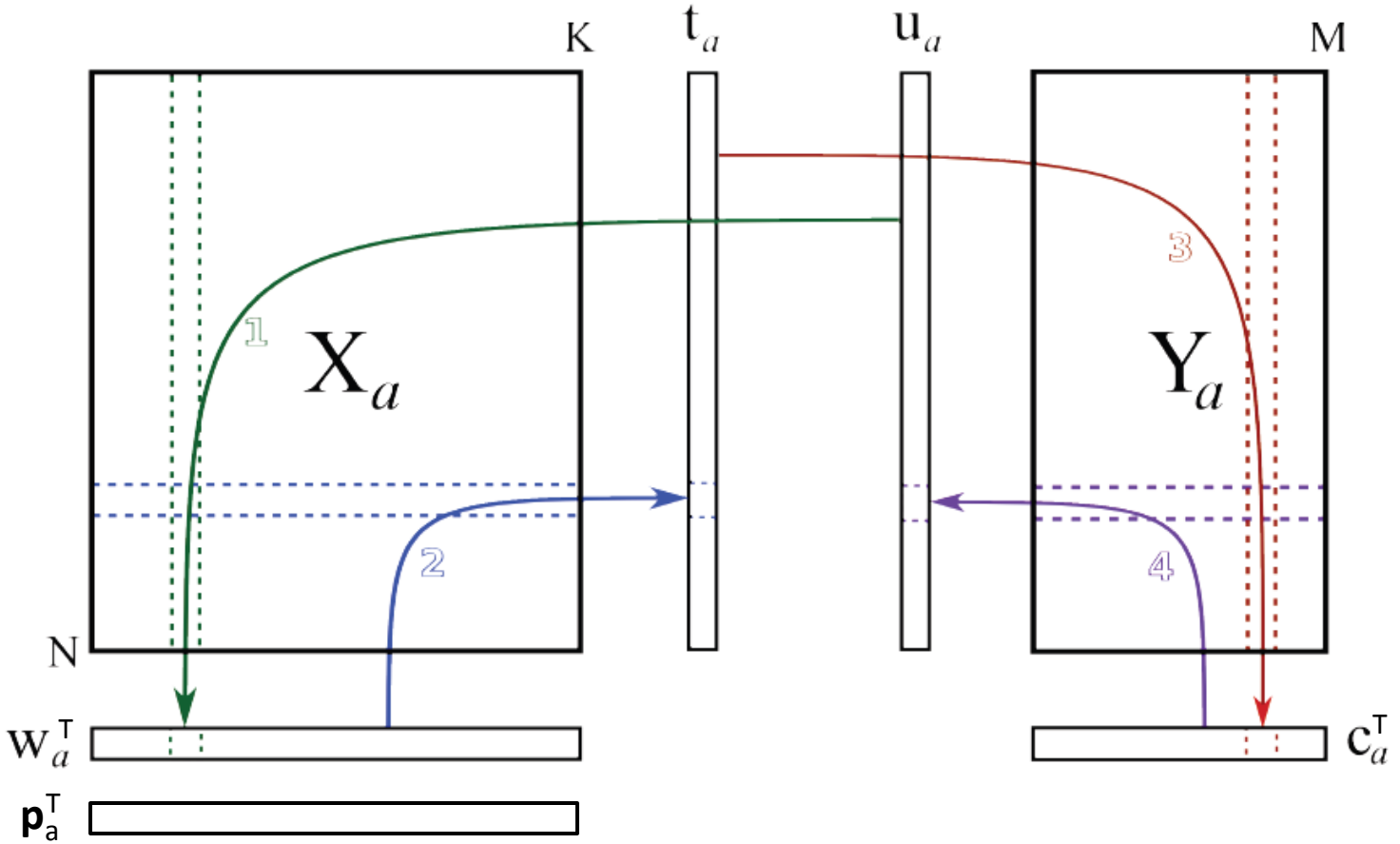
- The loadings: $\mathbf{w}_a$ (usually called weights in PLS)
- Interpreted in the same way as PCA loadings
- Weights for **X** and for **Y**: superimpose them
  - $\mathbf{w}*$ weights for **X**
  - **c** weights for **Y**
  - called a $\mathbf{w} * \mathbf{c}$ plot

Where did $\mathbf{w}*$ come from?

- $\mathbf{w}*_1 = \mathbf{w}_1$
- $\mathbf{w}*_a \neq \mathbf{w}_a$ for $a > 1$
- Explained next

How the PLS model is calculated

- NIPALS algorithm



How the PLS model is calculated

(0. start with **u**: a column of **Y**)

1. Regress columns of **X** on **u**: $\mathbf{w} = \mathbf{X}^T\mathbf{u} / \mathbf{u}^T\mathbf{u}$

1-1. Normalize **w**: $\|\mathbf{w}\| = 1.0$

2. Calculate scores **t**: $\mathbf{t} = \mathbf{X}\mathbf{w} / \mathbf{w}^T\mathbf{w}$

3. Regress columns of **Y** on **t**: $\mathbf{c} = \mathbf{Y}^T\mathbf{t} / \mathbf{t}^T\mathbf{t}$

4. Calculate new **u**: $\mathbf{u} = \mathbf{Y}\mathbf{c} / \mathbf{c}^T\mathbf{c}$

5. Repeat 1 ~ 4 until converge

6. Calculate X loadings after convergence: $\mathbf{p} = \mathbf{X}^T\mathbf{t} / \mathbf{t}^T\mathbf{t}$

How the PLS model is calculated

7. Deflate **X** and **Y** (take residuals):

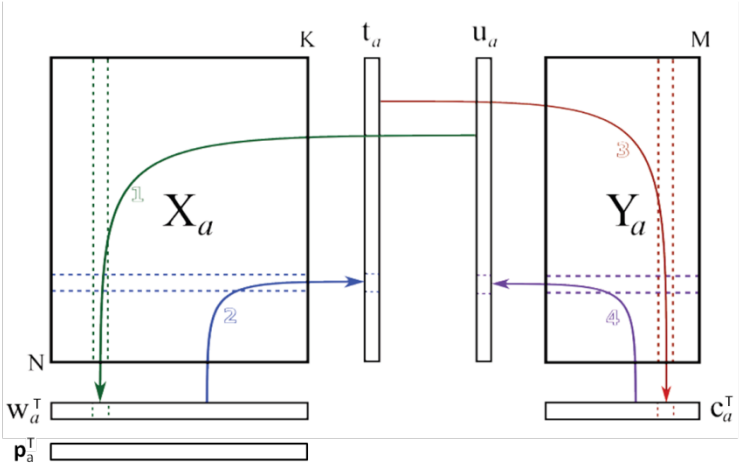
$$\mathbf{E}_1 = \mathbf{X} - \hat{\mathbf{X}} = \mathbf{X} - \mathbf{t}\mathbf{p}^T$$

$$\mathbf{F}_1' = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{t}\mathbf{c}^T$$

8. Set **X** = $\mathbf{E}_1$ & **Y** = $\mathbf{F'}_1$; go to step 1 and iterate for next component

How the PLS model is calculated

- Difference between **p** and **w**
  - **w**: regression coefficients of columns of **X** on **u**.
  - **p**: regression coefficients of columns of **X** on **t** and is Computed only at convergence.



- $\mathbf{tp}^T$ is best approximation of **X** at each stage.
- Therefore **p** is used to calculate residuals.

$$\mathbf{X}_a = \mathbf{X}_{a-1} - \mathbf{t}_a\mathbf{p}_a^T$$

How the PLS model is calculated

- **w***?
  - Calculated on deflated matrices
    - $\mathbf{t}_1 = \mathbf{X}_{a=1}\mathbf{w}_1 = \mathbf{X}\mathbf{w}_1$
    - $\mathbf{t}_2 = \mathbf{X}_{a=2}\mathbf{w}_2 = (\mathbf{X} - \mathbf{t}_1\mathbf{p}^T_1)\mathbf{w}_2$
  - $\mathbf{w}_2$: relates score $\mathbf{t}_2$ to $\mathbf{X}_{a=2}$
  - This is hard to interpret. We would like
    - $\mathbf{t}_1 = \mathbf{X}_{a=1}\mathbf{w}_1 = \mathbf{X}\mathbf{w}^*_1$
    - $\mathbf{t}_2 = \mathbf{X}_{a=2}\mathbf{w}_2 = \mathbf{X}\mathbf{w}^*_2$
    - …
  - Compare to PCA:
    - $\mathbf{t}_a = \mathbf{X}\mathbf{p}_a$

In the next lecture

- A bit more on PLS
- Tutorials
- Process monitoring (MSPC)
- Assignment #2