Introduction to Partial Least Squares Regression

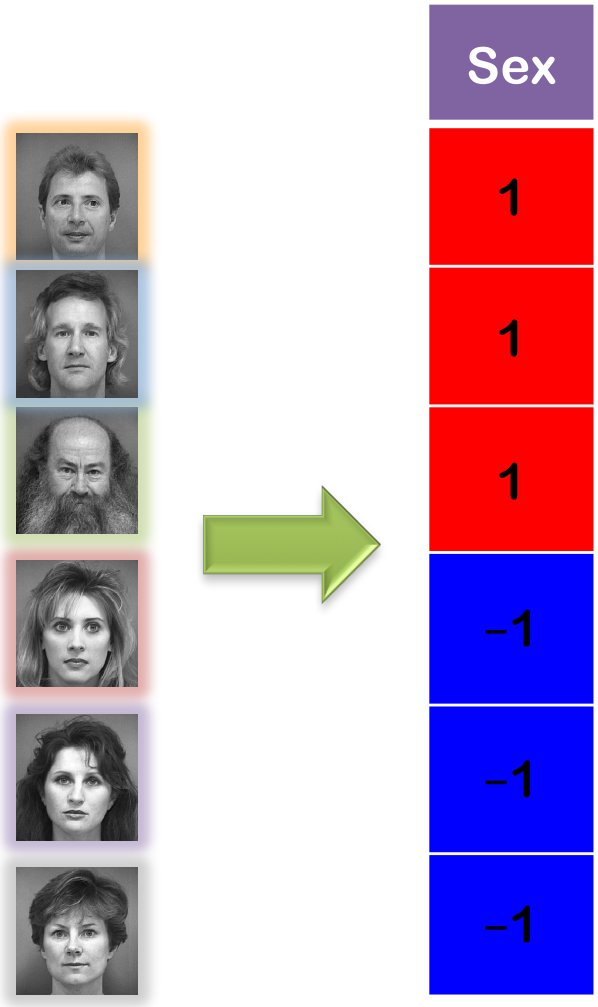
Hervé Abdi, Ph.D.
The University of Texas at Dallas

Partial Least Squares Methods

- **Partial Least Squares**
  - PLS Regression – Predict performance
  - PLS Correlation – Extract commonalities
- **Partial Least Squares Regression**
  - Econometrics; Chemometrics
    - H. Wold (1982), M. Martens, H. Martens, S. Wold (1983)

Partial Least Squares Regression

- Goal of PLSR
  - Predict one table from another
- Components
  - Latent variables
- 6 Faces
  - Boys +1
  - Girls –1
- Constraint
  - Best prediction of Y



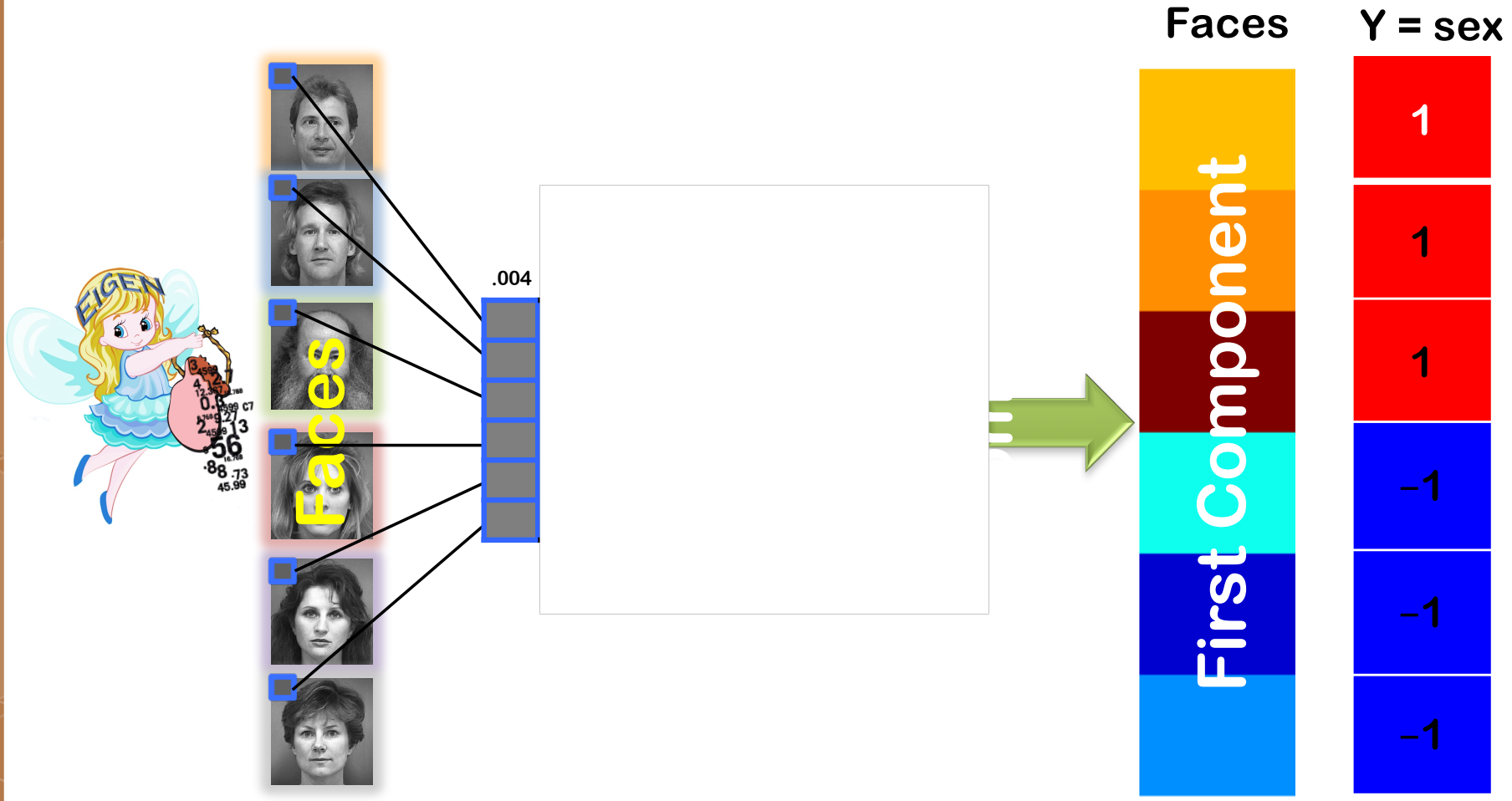
What in these faces separate boys from girls

- To find out:
  - Mix the faces up
    - with the constraint that this mixture gives the best prediction of y

- This gives the first *latent* face:
  - from X contributing to the prediction of y

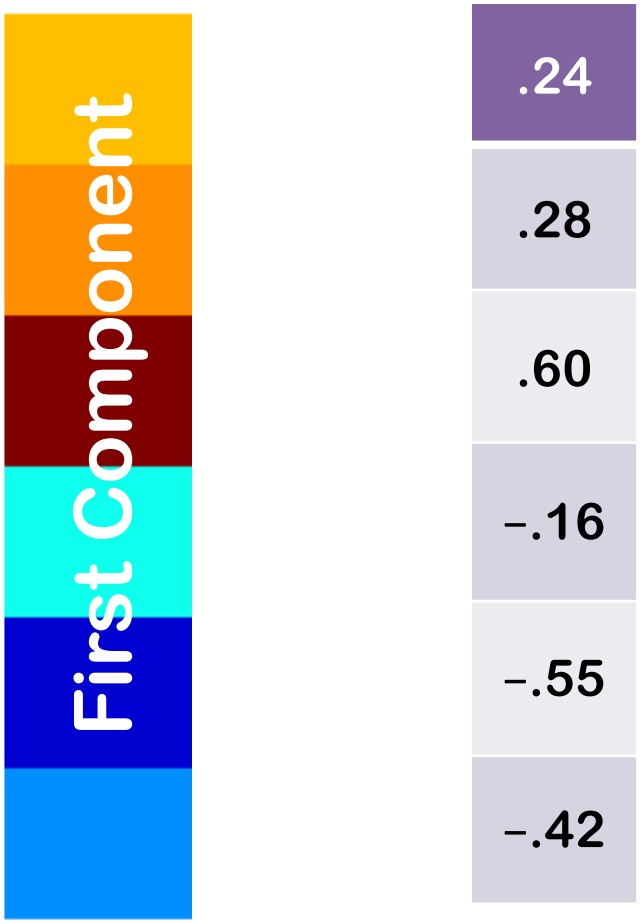
Remove What is Common

Mixing pixels to get sex …



With Numbers: Latent Pixels

Faces



How does it Look? The pixel Coefficients

.004 .005 003

55,200 of these ...

Looks like this ...

Call it a latent face ...

Double Duty for t: Faces to Latent-Face

.24 ×

.28 ×

.60 ×

−.16 ×

−.55 ×

−.42 ×

So the first latent face

Is a linear combination of the faces

with *positive* and *negative* weights

Back to the latent Pixel or X-latent variable

t_1
.24
.28
.60
−.16
−.55
−.42

WHAT DO WE WANT TO KNOW?

t_1
.24
.28
.60
−.16
−.55
−.42

Sex
1
1
1
−1
−1
−1

… PREDICT Y FROM T

Solution

It's a simple (linear) regression problem!

$$\hat{y}_1 = b_1$$

t_1
.24
.28
.60
−.16
−.55
−.42

Why no a?

PREDICT Y FROM X

$\hat{y}_1$
.55
.63
1.35
−.36
−1.23
−.94

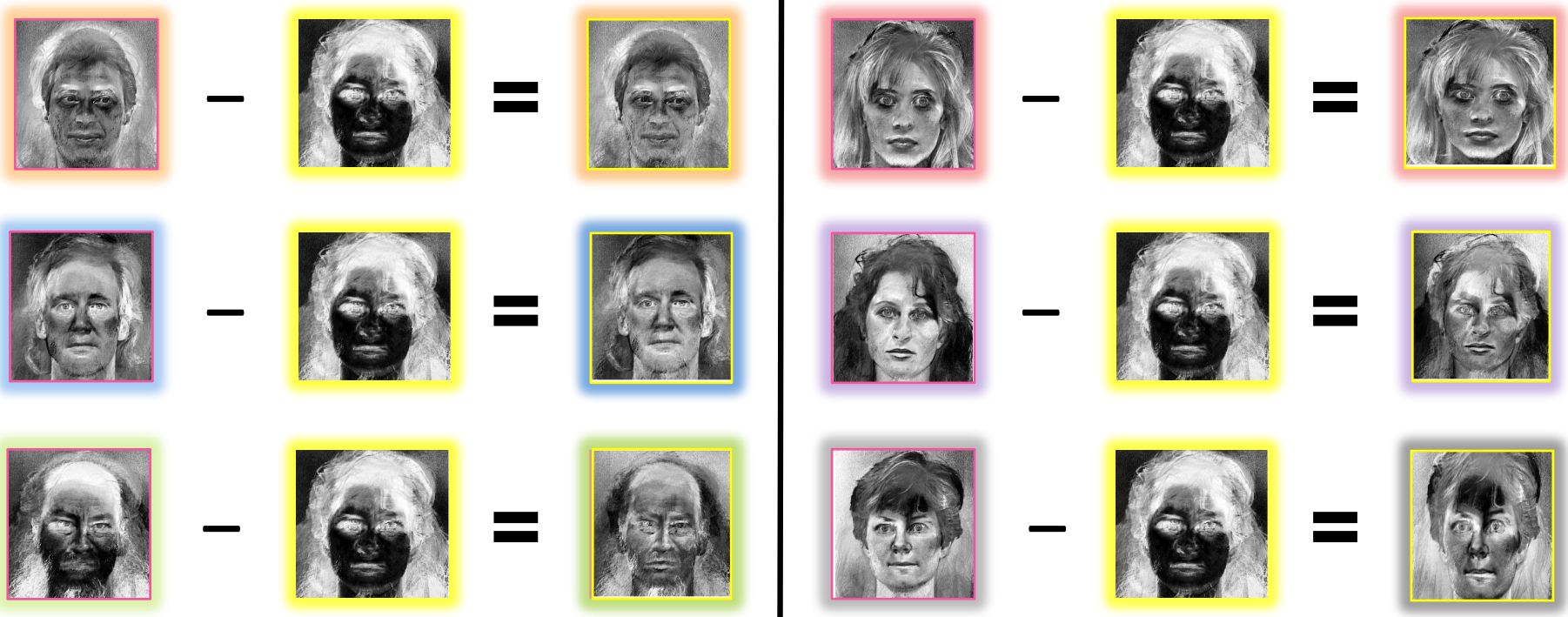
2.25

t_1
.24
.28
.60
−.16
−.55
−.42

What's Left Behind in Y?

y		$\hat{y}_1$		$y - \hat{y}_1$
1		.55		.45
1		.63		.37
1	−	1.35	=	−.35
−1		−.36		−.64
−1		−1.23		.23
−1		−.94		−.06

What is Left Behind in X?



Subtract = to partial out

$$\mathbf{Y}_2 = \mathbf{Y}_1 - \hat{\mathbf{Y}}_1$$

$$\mathbf{X}_2 = \mathbf{X}_1 - \hat{\mathbf{X}}_1$$

So we partialed out the effect of the latent variable from X and Y

What's in a name

And then we keep on doing the same stuff

Hence: *partial-least square* regression

Note: We need an SVD for every-iteration

Second Latent Face

What's The Next Best Prediction?

.35 ×

.44 ×

−.31 ×

−.67 ×

.34 ×

−.14 ×

Solution

$B_2 = .95$

$$\hat{y}_2 = b_2$$

t_2
.35
.44
−.31
−.67
.34
−.14

Predict Y_2

$\hat{y}_2$
.33
.42
−.30
−.64
.32
−.14

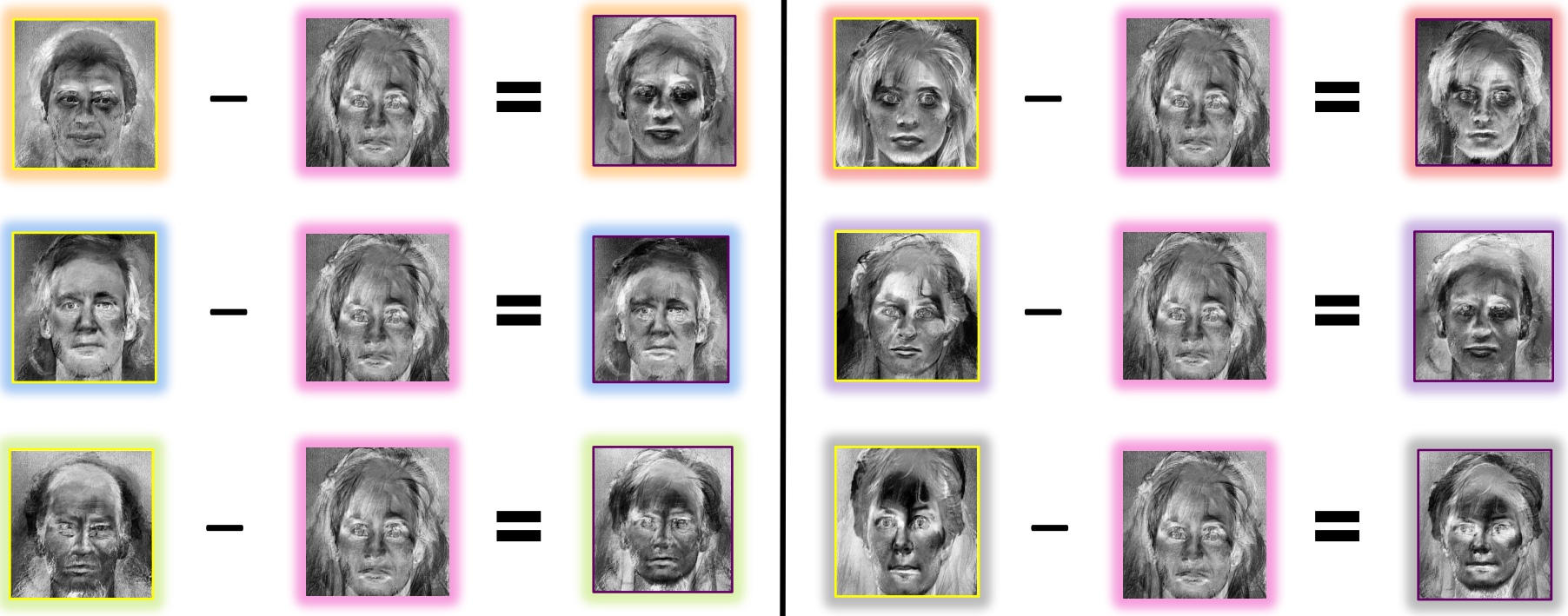
= .95 ×

t_2
.35
.44
−.31
−.67
.34
−.14

What is Left Behind in Y_1?

y_1		$\hat{y}_2$		$y_1 - \hat{y}_2$
.45		.33		.12
.37		.42		−.05
−.35		−.30		−.06
−.64		−.64		−.00
.23		.32		−.09
−.06		−.14		.08

What is Left Behind in X?



And so on

Keep on, keep on!

Subtract, predict, partial out …

… Till nothing is left …

When Do We Stop?

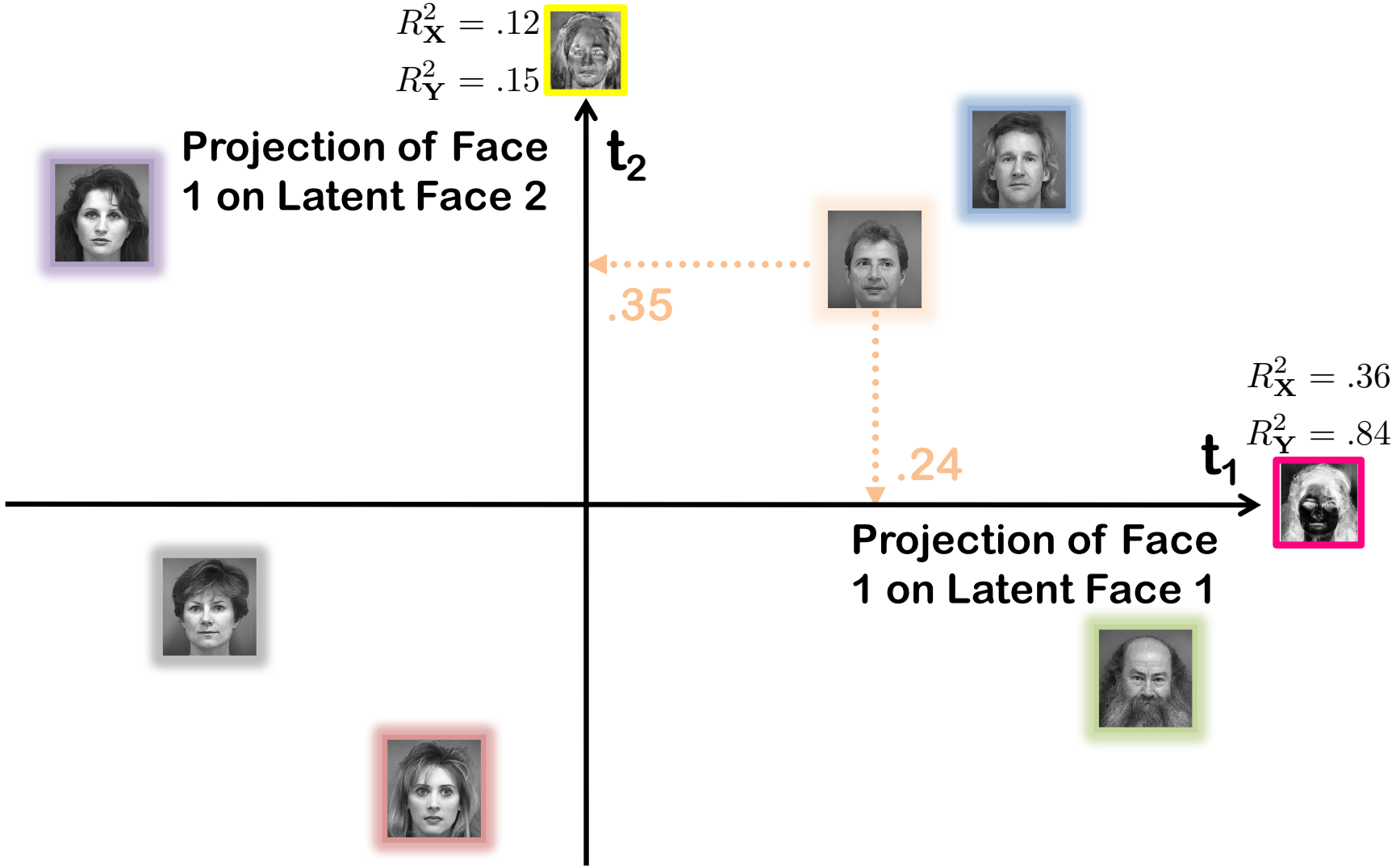
$Y = b \times T$ ←

							X ↑
	2.25	.95	.18	.04	.01	.00	
+1	.24	.35	.56	.49	.32	-.46	
+1	.28	.44	-.11	-.73	.11	-.44	
+1	.60	-.31	-.35	.26	-.43	-.30	
-1	-.16	-.67	-.01	-.15	.58	-.27	
-1	-.55	.34	-.56	.31	.03	-.52	
-1	-.42	-.14	.48	-.19	-.61	-.32	
	t_1	t_2	t_3	t_4	t_5	t_6	

HOW GOOD IS THE PREDICTION?

	2.25	.95	.18	.04	.01	.00
						
Y % Variance	84	15	1	0	0	0
Sum Y % Variance	84	99	100	100	100	100
X % Variance	36	12	21	16	15	0
Sum X % Variance	36	48	69	85	100	100
Sum X eigenvalues	94	96	97	98	99	100
	t_1	t_2	t_3	t_4	t_5	t_6

WHAT TO DO WITH LATENT VARIABLES?



How Do We Do It?

Get Formal!

PLS is a 2-step procedure

Kind of Double SVD (not a real one)

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathsf{T}} \text{ with } \mathbf{T}^{\mathsf{T}}\mathbf{T} = \mathbf{I}$$

$$\hat{\mathbf{Y}} = \mathbf{T}\mathbf{B}\mathbf{C}^{\mathsf{T}}$$

As a Regression Problem (pseudo-inverse)

$$\hat{\mathbf{Y}} = \mathbf{T}\mathbf{B}\mathbf{C}^{\mathsf{T}} = \mathbf{X}\mathbf{B}_{\mathbf{PLS}}$$

$$\mathbf{B}_{\mathbf{PLS}} = \mathbf{X}^{+}\mathbf{Y} = \mathbf{P}^{\mathsf{T}+}\mathbf{B}\mathbf{C}^{\mathsf{T}}$$

How to Predict Y?

- From the latent vectors, use:

$$\hat{\mathbf{Y}} = \mathbf{TBC}^{\mathsf{T}}$$

- From X, with *Z*-scores (for X and Y), use:

$$\hat{\mathbf{Z}}_{\mathbf{Y}} = \mathbf{Z}_{\mathbf{X}}\mathbf{B}_{\mathbf{PLS}}$$

- From X, use:

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B}^{*}_{\mathbf{PLS}}$$

NIPALS: The Original Algorithm

- Get a linear combination from X
- With maximal covariance with Y

Start with E = X, and F = Y. Initiate a random u

- Step 1. $w \propto E^T u$ (estimate X weights)
- Step 2. $t \propto Ew$ (estimate X factor scores)
- Step 3. $c \propto F^T t$ (estimate Y weights)
- Step 4. $u = Fc$ (estimate Y scores)

Iterate till t converges, then compute

$$\beta = t^T u, \text{ and } p = E^T t$$

Now partial out t from both E and F as:

$$E = E - tp^T, \text{ and } F = F - \beta tc^T.$$

How Good is the Prediction

- The problem:
  - The prediction is sample-dependent
- The question:
  - How would we do on a *new* sample?
- One answer:
  - Cross-validation, e.g. *Jackknife*
  - A.K.A Leave-One-Out (LOO)

The Jackknife

John W. Tukey
(1915-2000)

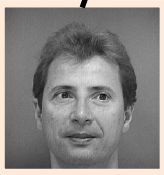
A Jackknife

Jackknife

The training set

The testing set

$$\hat{\mathbf{Y}}_1 = \mathbf{X}_1^{\mathsf{T}} \mathbf{B}^*_{\mathbf{PLS}}$$

−.09 =  × 

The latent faces:

1

2

Jackknife

The training set

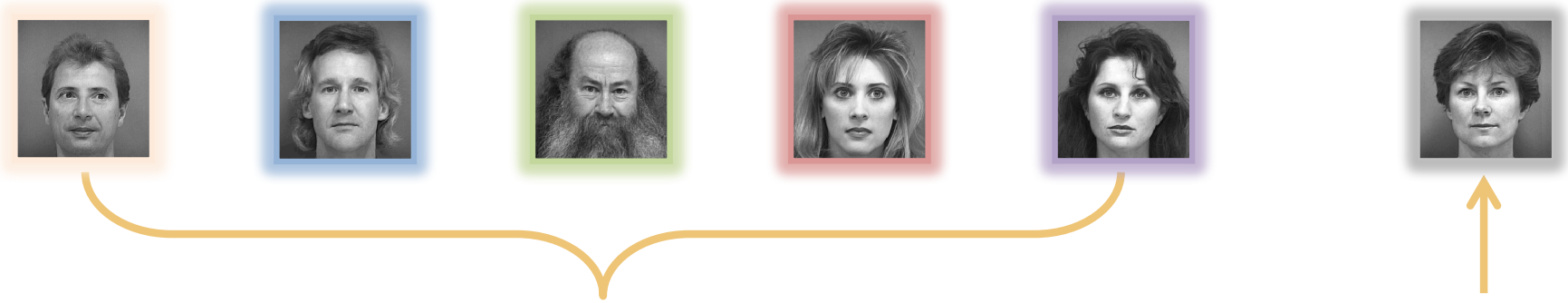
The testing set

$$\hat{\mathbf{Y}}_2 = \mathbf{X}_2^{\mathsf{T}} \mathbf{B}^*_{\mathbf{PLS}}$$

−.05 =  × 

The latent faces:  1  2

Jackknife



The training set

The testing set

$$\hat{\mathbf{Y}}_6 = \mathbf{X}_6^{\mathsf{T}} \mathbf{B}^*_{\mathbf{PLS}}$$

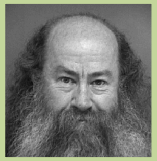
−.48 =  × 

The latent faces:


1

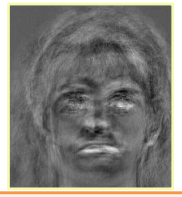

2

Jackknife Summary

The faces						
The predictors: B_{PLS}						
1st latent faces						
2nd latent faces						

More About Sex

134 Faces
83 Women
51 Men

	2	4	3	6	39
PCR					
$R^2_{\mathbf{Y}}$	.29	.13	.09	.05	.03
Sum	.29	.42	.51	.52	.59
$R^2_{\mathbf{X}} = \lambda$	.02	.01	.01	.00+	.00+
PLSR					
$R^2_{\mathbf{Y}}$	.50	.18	.06	.07	.05
Sum	.50	.69	.75	.82	.87
$R^2_{\mathbf{X}}$	.16	.09	.10	.06	.04

Jackknife for PCR and PLSR

Jackknife (2 latent vector solution)						
$\hat{Y}$	–.12	–.08	2.08	.40	–.50	–.60
Y	1	1	1	–1	–1	–1

$r_{Y,\hat{Y}} = .48$ $r^2_{Y,\hat{Y}} = .23$

Sample (2 latent vector solution)						
$\hat{Y}$	.79	.96	.97	–.91	–.83	–.98
Y	1	1	1	–1	–1	–1

$r_{Y,\hat{Y}} = .99$ $r^2_{Y,\hat{Y}} = .98$

QUALITY OF CLASSIFICATION

PCR			
Fixed Effect		Jackknife	
R^2	Misclassifications	R^2	Misclassifications
.77	17/134	.74	20/134

PLSR			
Fixed Effect		Jackknife	
R^2	Misclassifications	R^2	Misclassifications
.93	1/134	.79	14/134

Additional Problem:

The Classification is Optimal for 2 LVs

But we used the data to find this number

This is another case of "double dipping"

Three Samples ...

We need (at least) three independent samples

The Classification is Optimal for 2 LVs

But we used the data to find this number

This is another case of "double dipping"

A (Faked) Wine Example

What Do We Want To Do?

- Predict subjective evaluation of 5 wines
- The dependent variables for each wine:
  - Likeability,
  - How well it goes with meat
  - How well it goes with dessert
- The predictors are:
  - Price
  - Sugar
  - Alcohol
  - Acidity

The Predictors

Wine	Price	Sugar	Alcohol	Acidity
1	7	7	13	7
2	4	3	14	7
3	10	5	12	5
4	16	7	11	3
5	13	3	10	3

The Dependent Variables

Wine	Hedonic	Goes with Meat	Goes with Dessert
1	14	7	8
2	10	7	6
3	8	5	5
4	2	4	7
5	6	2	4

First Step: Get Z-scores

The Predictors Z-scores

Wine	Price	Sugar	Alcohol	Acidity
1	–0.63	1	0.63	1
2	–1.26	–1	1.26	1
3	0	0	0	0
4	1.26	1	–0.63	–1
5	.0.63	–1	–1.26	–1

The Dependent Variables Z-scored

Wine	Hedonic	Goes with Meat	Goes with Dessert
1	1.34	0.94	1.26
2	0.45	0.94	0
3	0	0	–0.63
4	–1.34	–0.47	0.63
5	–0.45	–1.41	–1.26

The "R" matrix

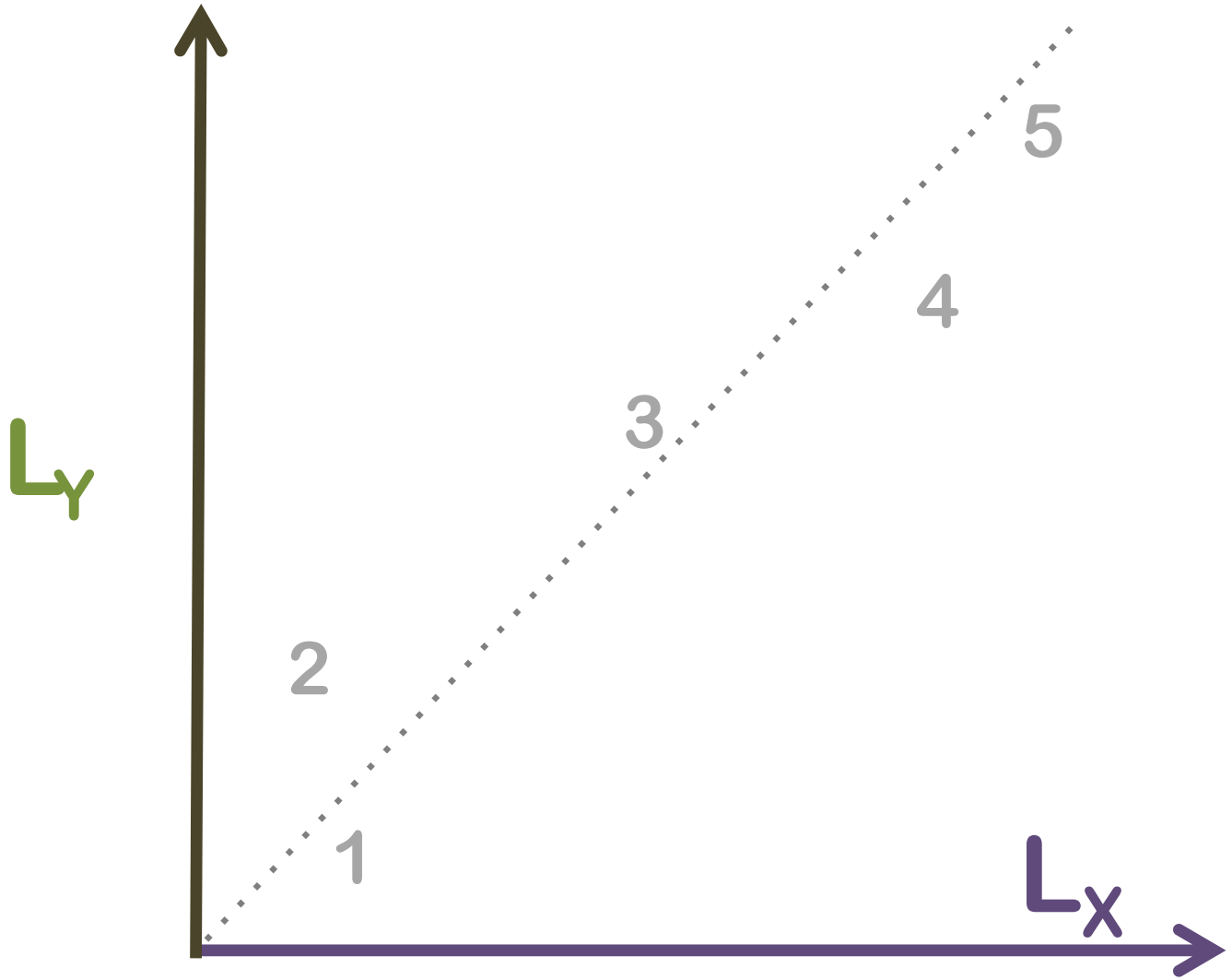
		Y		
		Hedonic	Meat	Dessert
X	Price	–3.39	–3.28	–0.80
	Sugar	0.00	0.94	3.16
	Alcohol	2.82	3.88	2.00
	Acidity	3.58	3.77	1.90

$$\mathbf{R} = \mathbf{Z}_{\mathbf{X}}^{\top}\mathbf{Z}_{\mathbf{Y}} = \mathbf{R}^{*} \times (I - 1)$$

(with $\mathbf{R}^{*}$ being correlation matrix)

Plot them

L_X vs L_Y



The New "R" matrix

		Y		
		Hedonic	Meat	Dessert
X	Price	–0.25	–0.34	1.10
	Sugar	–0.08	0.85	3.11
	Alcohol	–0.45	0.09	0.02
	Acidity	0.24	–0.08	–0.12

$$\mathbf{R}_1 = \mathbf{E}_1^{\top}\mathbf{F}_1$$

Keep on doing that till X is Gone

The Latent Variables

Wine	t_1	t_2	t_3
1	0.4538	−0.4662	0.5716
2	0.5399	0.4940	−0.4631
3	0	0	0
4	−0.4304	−0.5327	−0.5301
5	−0.5633	−0.5049	0.4217

Weights for X Variables

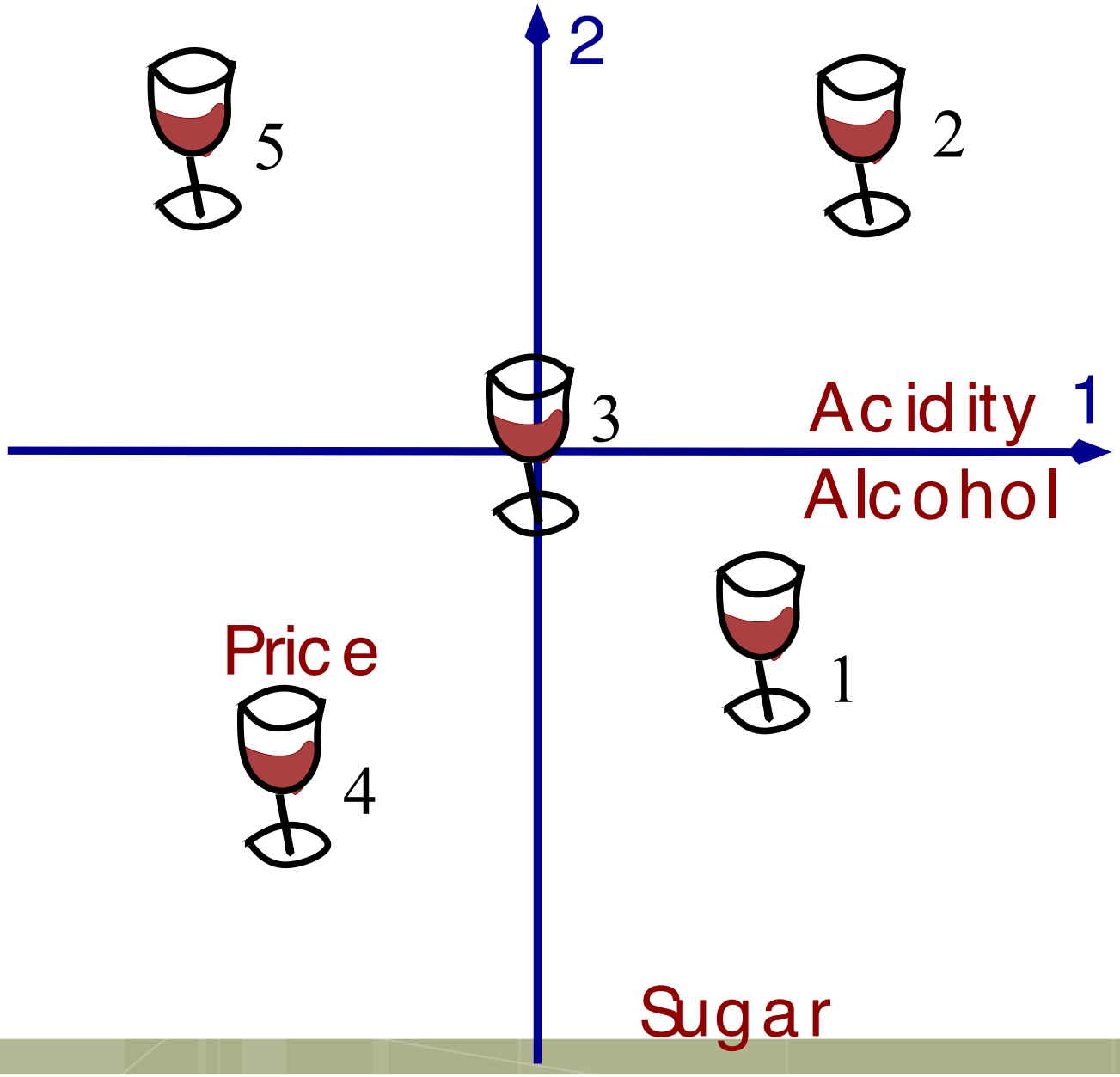
Wine	w_1	w_2	w_3
Price	–.51	–.34	–.35
Sugar	.20	–.94	.16
Alcohol	.57	–.02	–.82
Acidity	.61	.04	.42

Weights for Y Variables

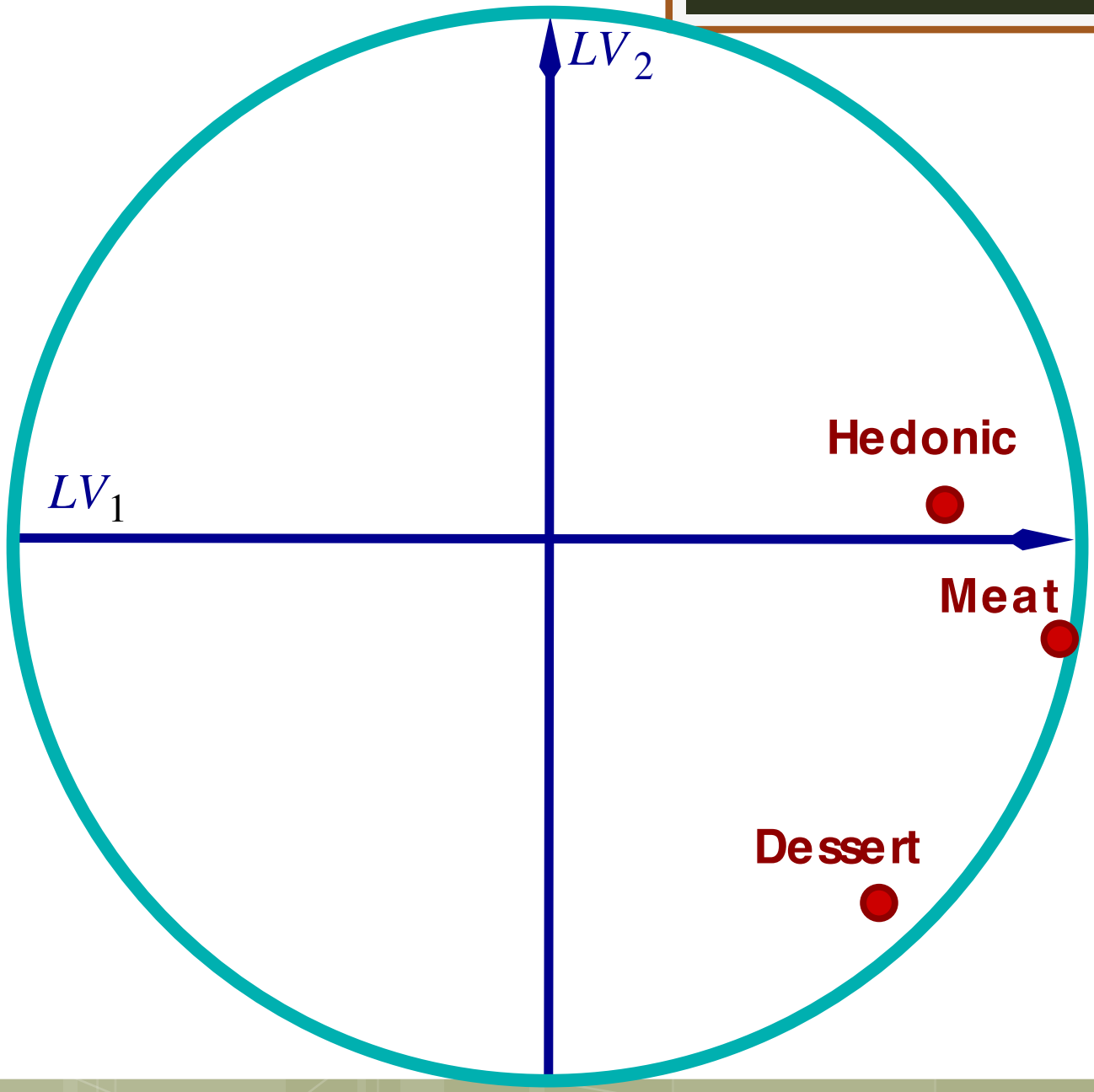
Wine	c_1	c_2	c_3
Hedonic	.61	.05	.97
Meat	.70	–.27	–.22
Dessert	.37	–.96	–.13

What about some pictures

Plot T and W biplot



Y Correlation Circle



The Prediction With 3 latent variables

		Y		
		Hedonic	Meat	Dessert
	Intercept	48.50	−8.92	−3.85
	Price	−1.00	−0.03	0.04
×	Sugar	0.75	0.28	0.59
	Alcohol	−4.00	1.00	0.50
	Acidity	2.75	0.18	0.09

A Wine costing \$10, with 5 g of sugar per litre, 10 degree of alcohol and an acidity of 5 (whatever) will have a hedonic rating of:

$$\hat{y}_1 = 48.50 - 1 \times 10 + .75 \times 5 - 4 \times 10 + 2.75 \times 5 = 16$$

THE PREDICTION WITH 2 LATENT VARIABLES

		Y		
		Hedonic	Meat	Dessert
×	Intercept	–3.28	–3.38	–1.39
	Price	–0.26	–0.11	0.01
	Sugar	0.14	0.34	0.62
	Alcohol	–0.81	0.49	0.27
	Acidity	0.69	0.40	0.19

A Wine costing \$10, with 5 g of sugar per litre, 10 degree of alcohol and an acidity of 5 (whatever) will have a hedonic rating of:

$$\hat{y}_1 = -2.28 - .26 \times 10 + .14 \times 5 - .81 \times 10 + 0.69 \times 5 = 6.38$$

How many

Big Question:

How many latent variables?

What's give the best prediction?

Risk of over-fitting

So 3 sets. Learning, tuning, testing

Another example: Crunch crucnh

OBJECTIVE: To study the relationships between the 3 texture attributes: *croustillant* (crispy), *craquant* (crackly), *croquant* (crunchy) and 9 attributes describing the biting sounds and other textural characteristics.

Descriptive Analysis

Panel: 8 trained assessors

Attributes

	+
crispy crunchy crackly	hard crumbly puffy brittle sticky duration
eating noises	loudness high-pitched low-pitched

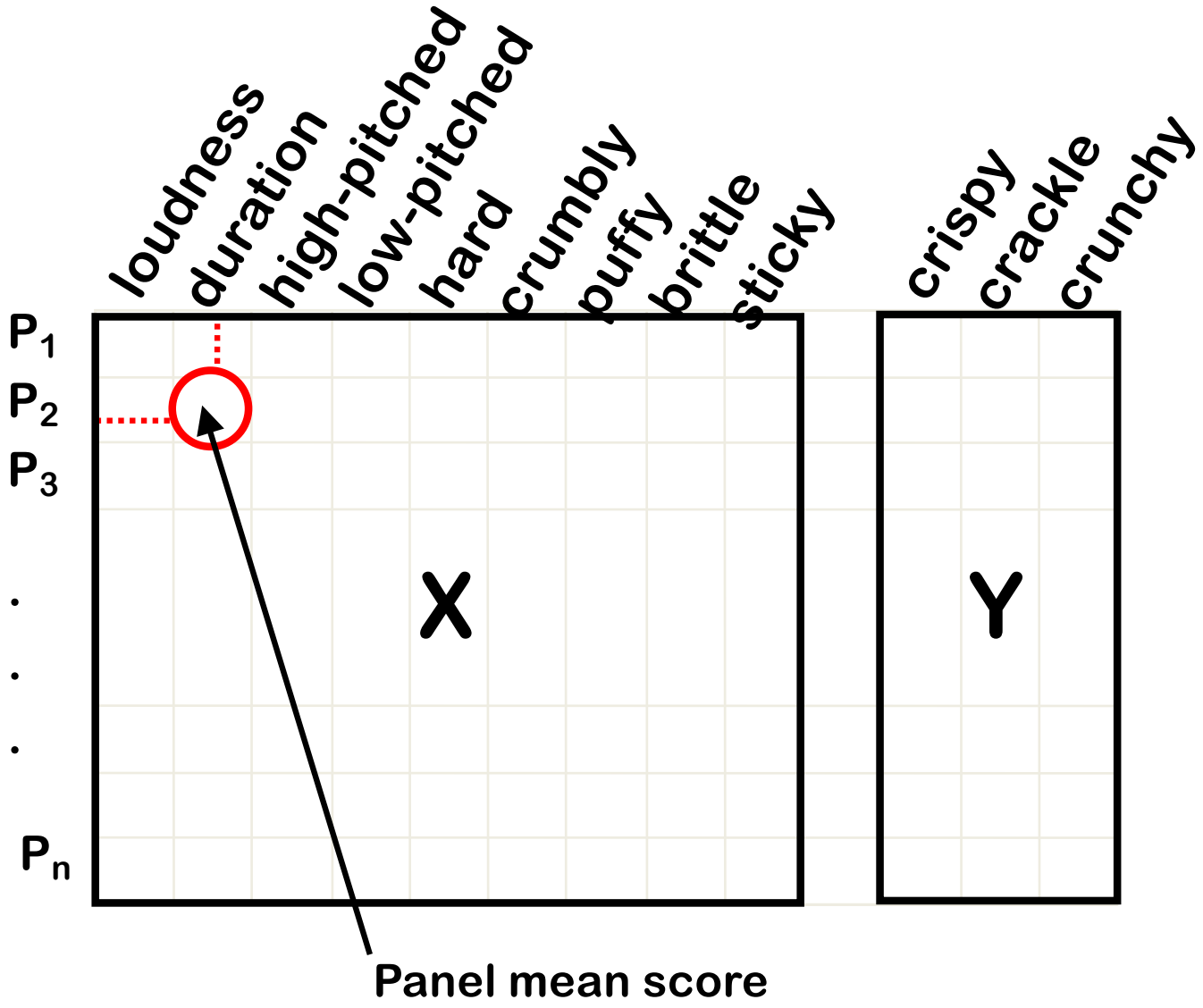
35 Products

extruded bread & snacks
dry biscuits, cold cereals
chips, biscotte, chocolate
raw vegetables & salad
dry fruits & seeds

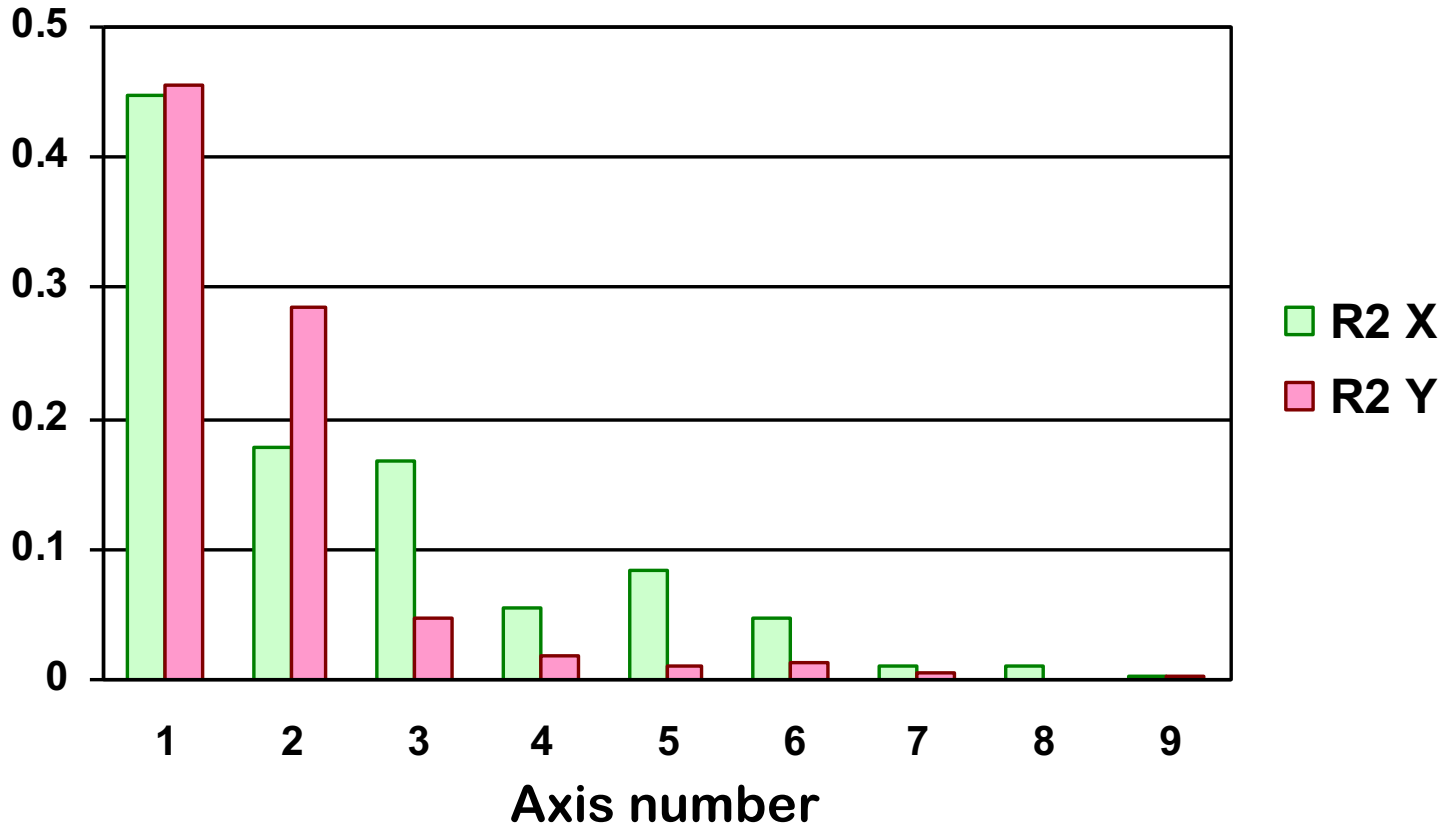
Evaluation

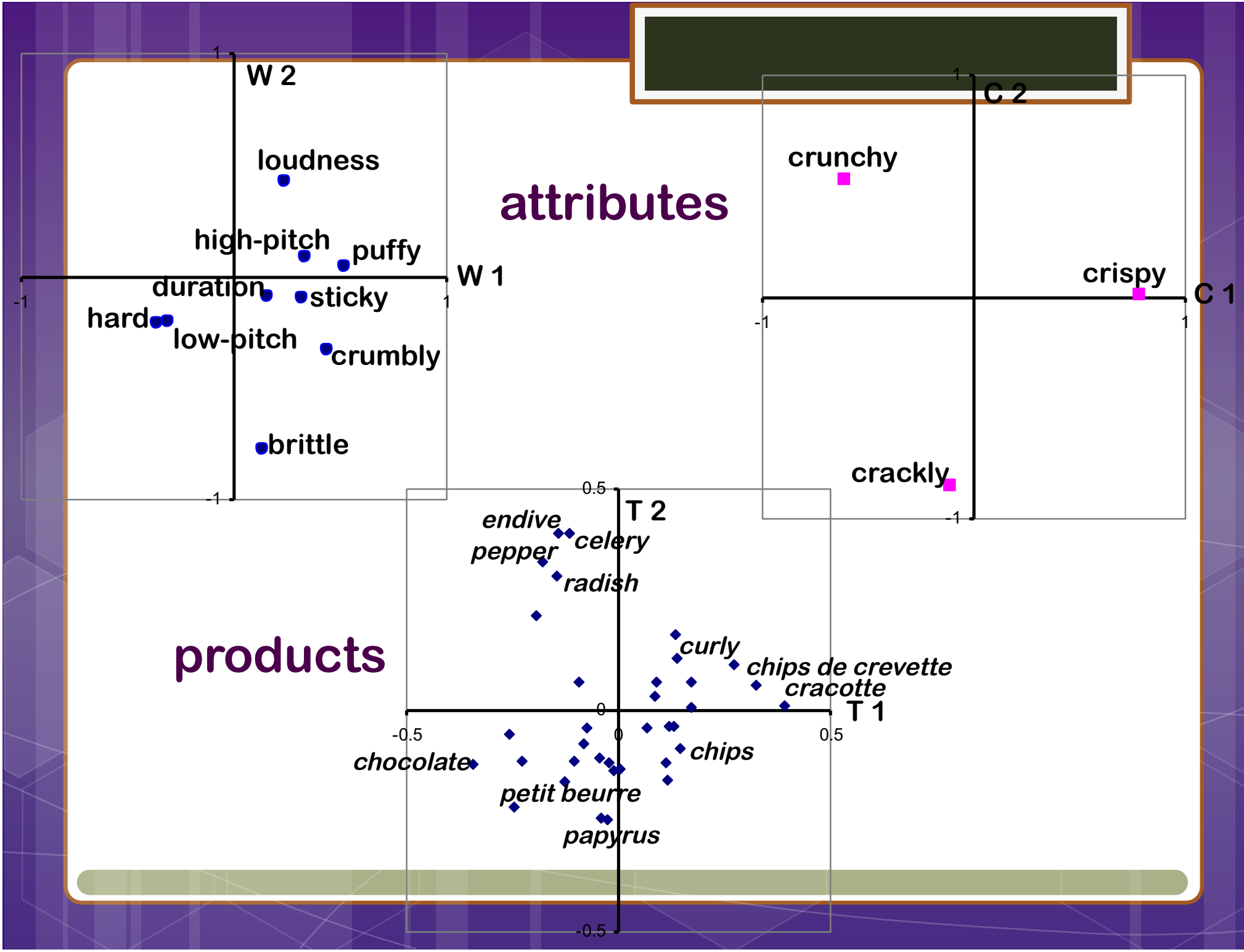
- in duplicate
- linear scale
- scores : 0 → 10

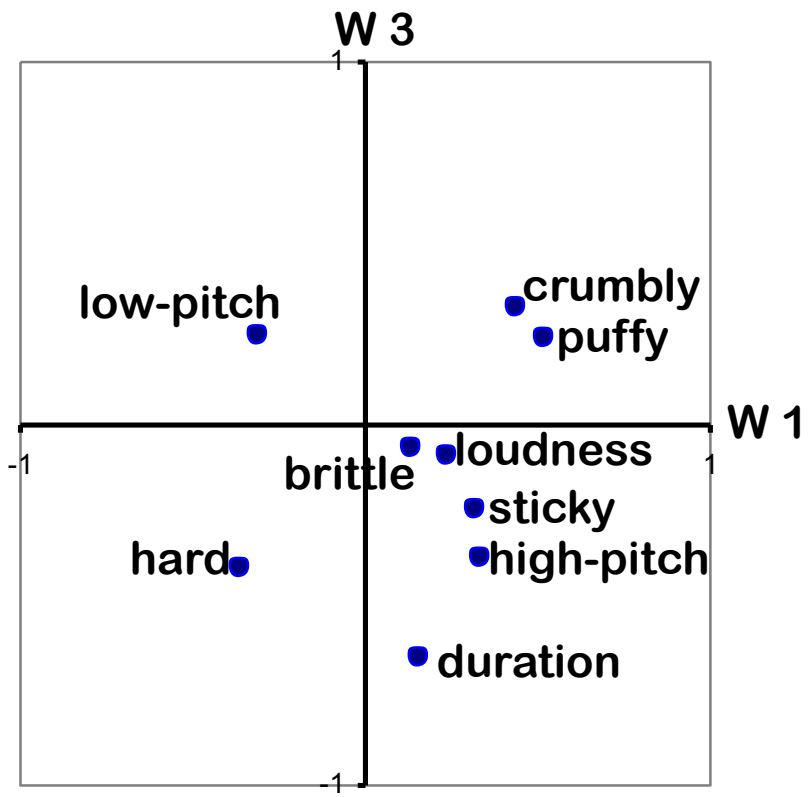
Data

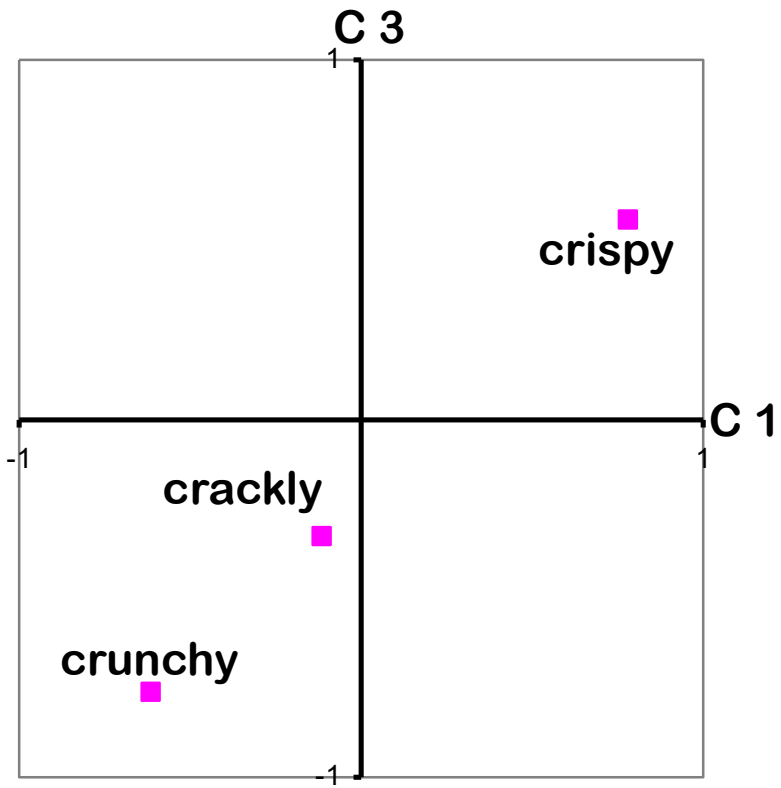

loudness
duration
high-pitched
low-pitched
hard
crumbly
puffy
brittle
sticky
crispy
crackle
crunchy
P1
P2
P3
.
.
.
Pn
X
Y
Panel mean score

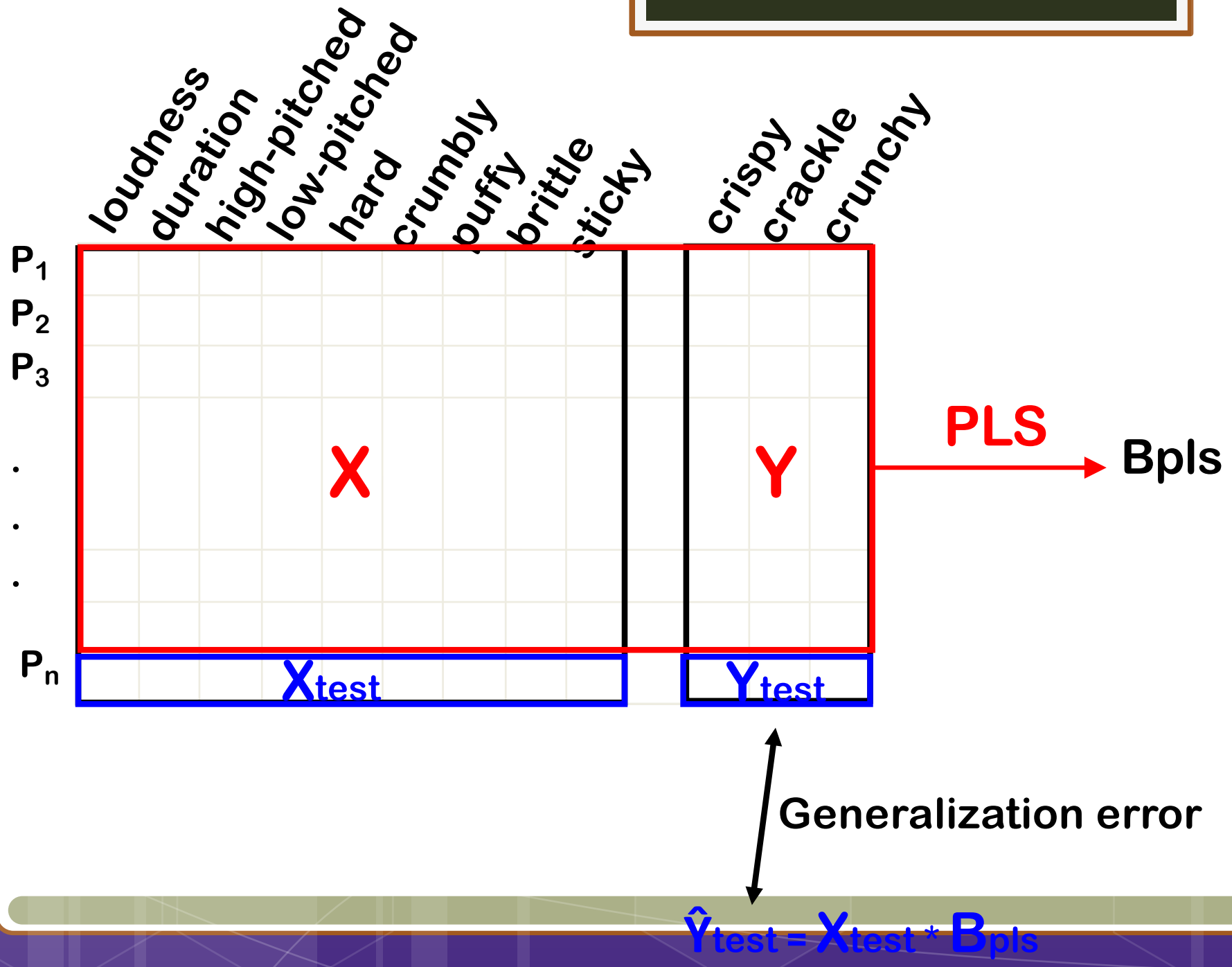
Error




W 2
1
-1
loudness
high-pitch
puffy
W 1
duration
sticky
hard
low-pitch
crumbly
brittle
attributes
C 2
crunchy
crispy
C 1
crackly
products
T 2
0.5
-0.5
endive
celery
pepper
radish
curly
chips de crevette
cracotte
T 1
0
chocolate
chips
petit beurre
papyrus

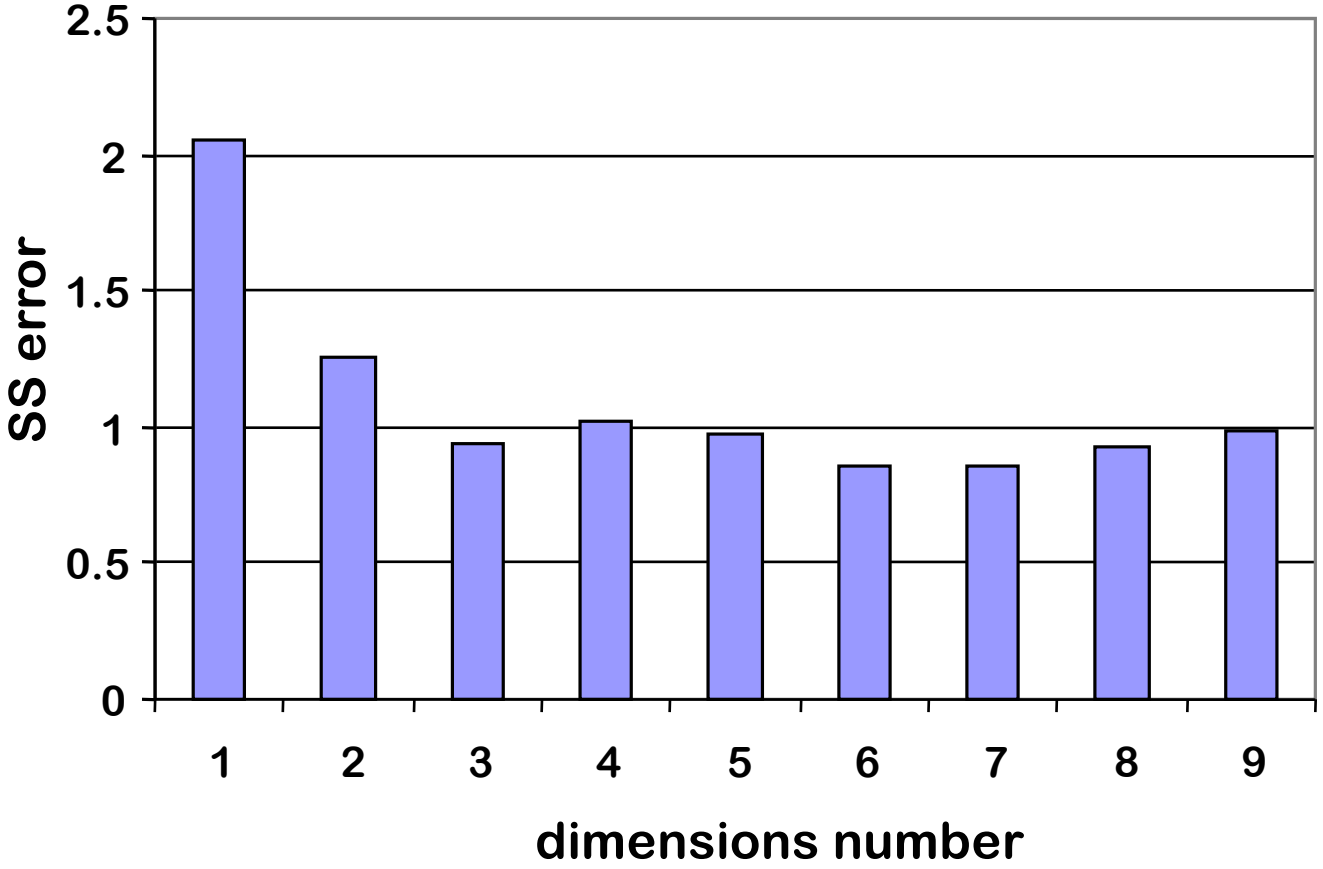

W 3
W 1
1
-1
1
-1
low-pitch
crumbly
puffy
loudness
brittle
sticky
hard
high-pitch
duration


C 3
C 1
1
-1
1
-1
crispy
crackly
crunchy

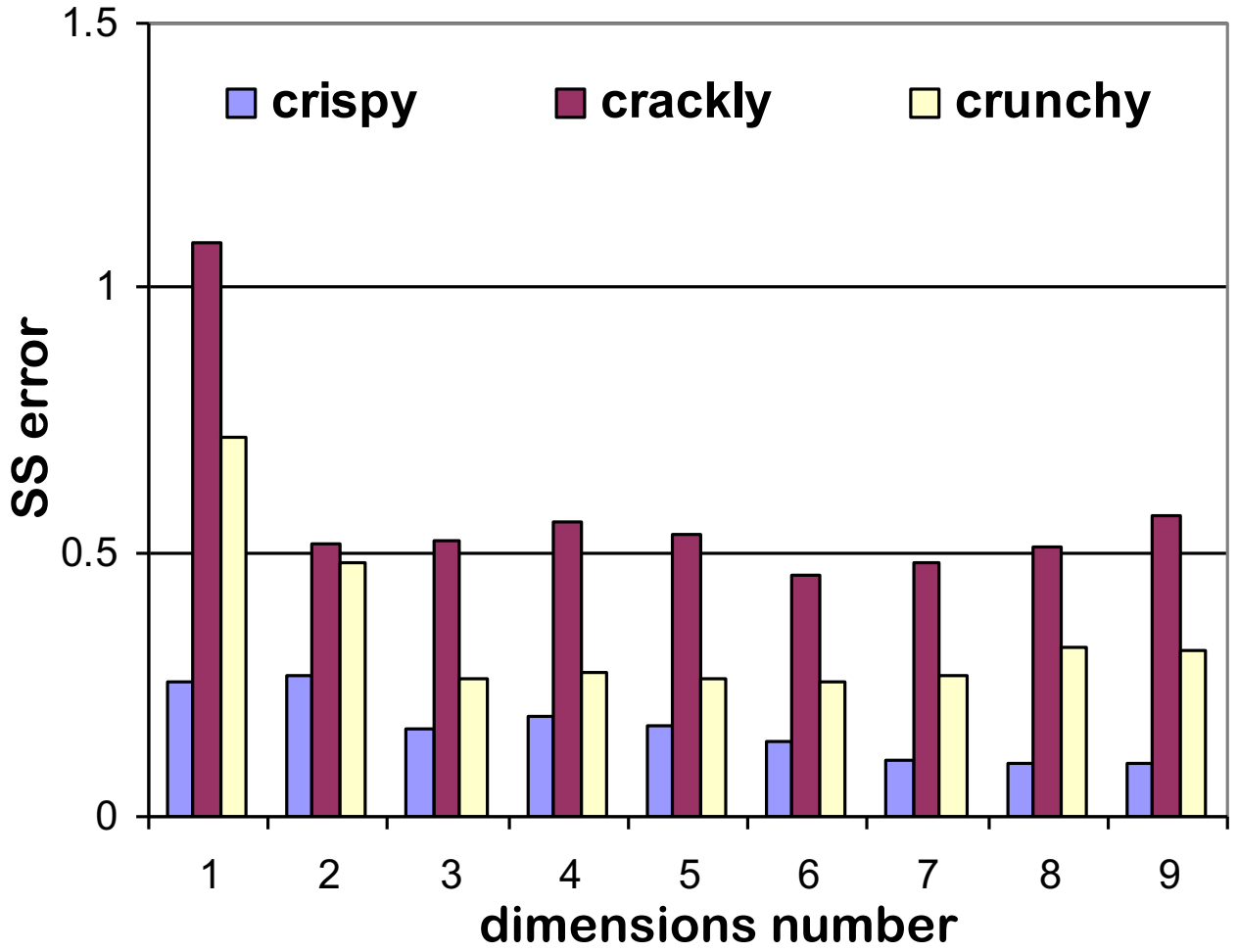

loudness
duration
high-pitched
low-pitched
hard
crumbly
puffy
brittle
sticky
crispy
crackle
crunchy
P1
P2
P3
.
.
.
Pn
X
Y
PLS
Bpls
Xtest
Ytest
Generalization error
Ŷtest = Xtest * Bpls

Generalisation Error

Learning set: 34 products ; test set: 1product
35 runs



Generalisation Error



Developing PLS

- PLS for Distances
  - DISPLS Correlation
  - DISPLS Regression
- PLS for Qualitative Data
  - PLS Correspondence Analysis
- PLS for multiple data tables
  - Multi-Block PLS, DISPLS

Thank You