

# A cluster approach to analyze preference data: Choice of the number of clusters

*Karin Sahmer, Evelyne Vigneau  
and El Mostafa Qannari*



*Unité de Sensométrie et de Chimiométrie  
Nantes, France*

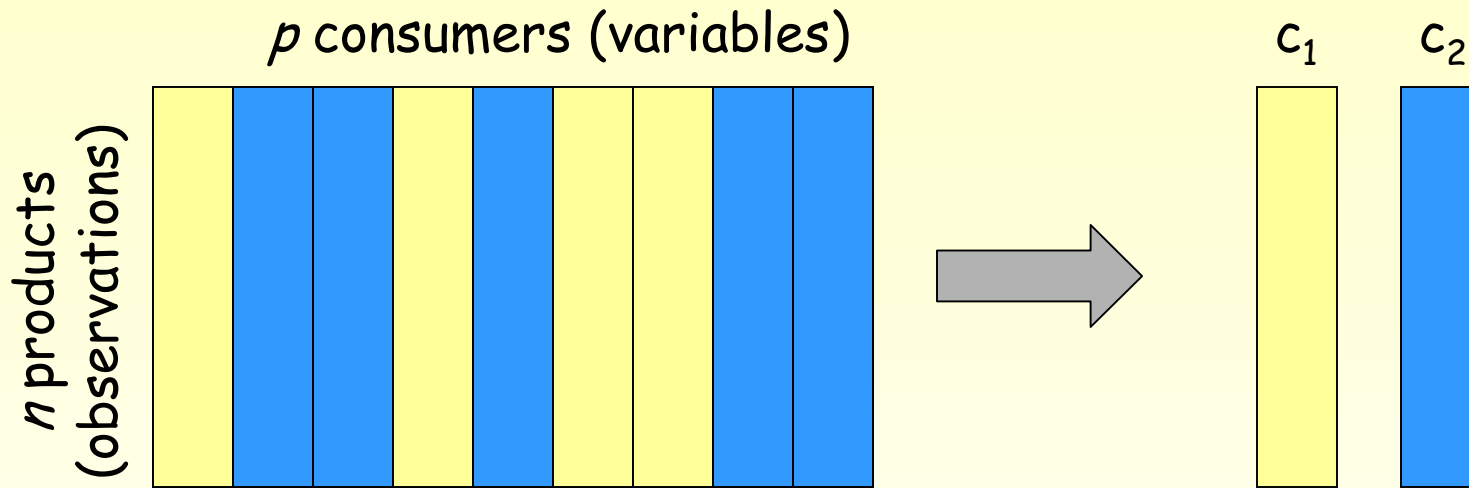


# Overview

- Clustering of variables approach
- Cluster tendency test:  
Are there different clusters of consumers?
- Illustration: Consumer preferences for coffees
- Cluster validity test:  
How many clusters are there?
- Illustration: Segmentation of consumers according to their preferences for coffees

**Segmentation  
of a panel of consumers:  
A cluster  
of variables approach**

# Clustering of variables approach



$$\text{Minimize } Q = \frac{1}{n} \sum_{k=1}^K \sum_{j=1}^p \delta_{jk} \|z_j - c_k\|^2$$

$z_j$  the standardized scores of consumer  $j$

$\delta_{jk} = 1$ , if consumer  $j$  belongs to cluster  $k$   
 $= 0$ , otherwise

# Clustering of variables approach

Assumed model:

Each consumer belongs to one cluster.

If consumer  $j$  belongs to cluster  $k$ ,  
his score for product  $i$  is given by

$$z_{ij} = c_{ik} + \varepsilon_{ij}$$

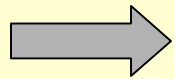
$c_k$ : latent variable of cluster  $k$

The consumers use the given scale in a different manner.

So we observe:

$$x_{ij} = a_j + b_j (c_{ik} + \varepsilon_{ij})$$

# Clustering of variables approach



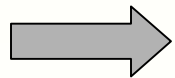
We want to determine

- the number  $K$  of clusters
- the partition into  $K$  clusters
- the latent variable  $c_k$  of each cluster



Number of clusters?

Which variable belongs to which cluster?



To minimize  $Q$ , we have to choose  $c_k = \bar{\mathbf{z}}_k$

# Clustering of variables approach

Cluster tendency test:  
Is there more than one cluster?

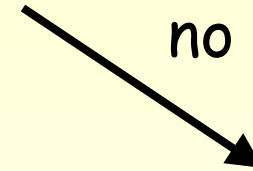


yes

Hierarchical clustering  
and cluster validity tests



Partitioning algorithm  
for the chosen number  $K$   
of clusters



no

We don't need  
any clustering.

**Cluster tendency test:  
Are there different  
clusters of consumers?**

# Cluster tendency test: The hypotheses

$$H_0: K = 1$$

$$H_1: K > 1$$

If there is just one cluster,  
the score of consumer  $j$  for product  $i$  is given by

$$x_{ij} = a_j + b_j (c_i + \varepsilon_{ij})$$

$$z_{ij} = \frac{1}{b_j} (x_{ij} - a_j) = c_i + \varepsilon_{ij}$$

Residuals:  $E = (z_1, z_2, \dots, z_p) - (c, c, \dots, c)$

$H_0$ : E is just formed by noise.

$H_1$ : E is structured.

# Cluster tendency test: The test statistic

$\lambda_1, \lambda_2, \dots, \lambda_r$  : the first  $r$  eigenvalues of  $E'E$

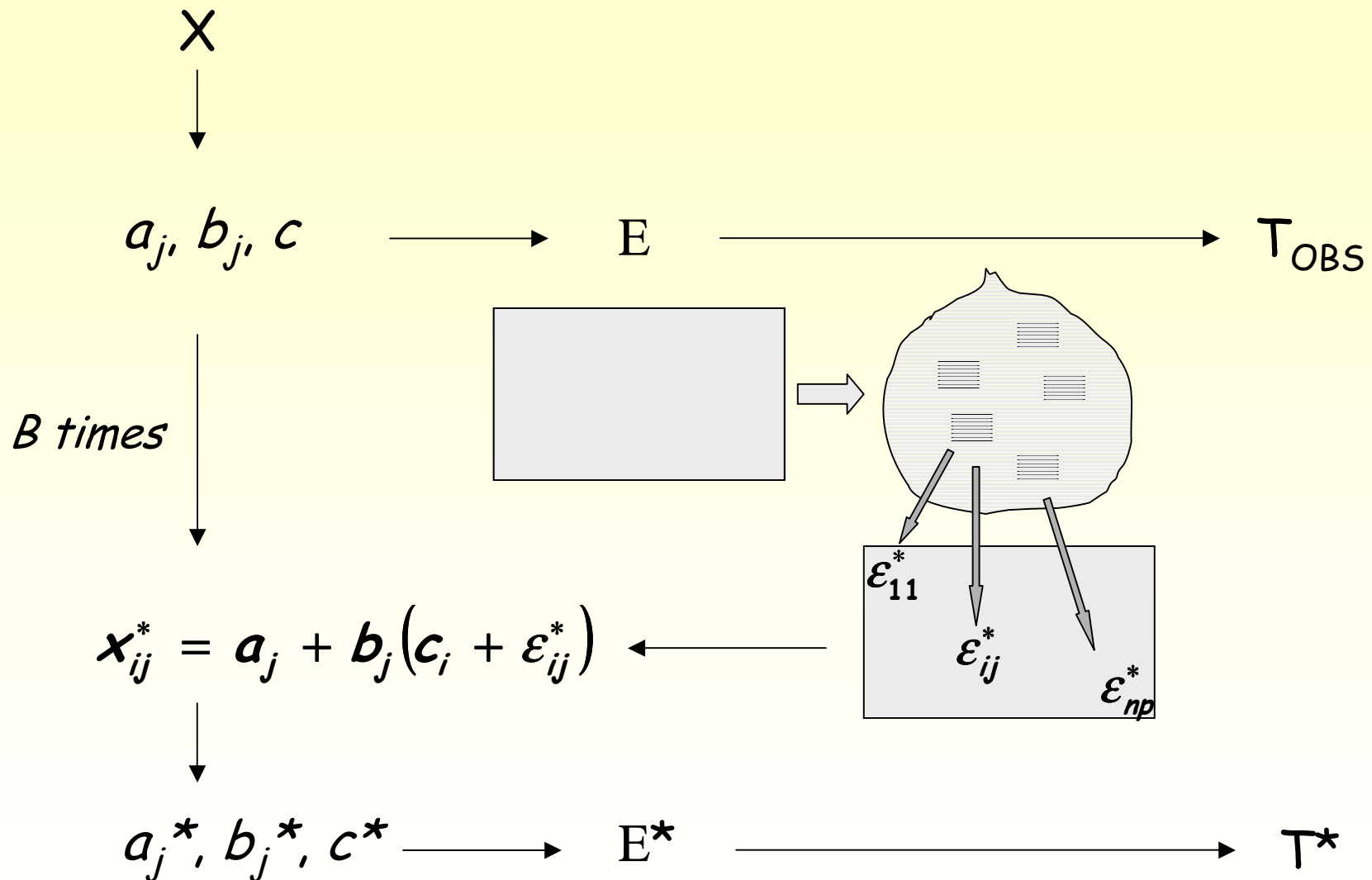
$$r = \min(n - 1, p - 1)$$

$$T = \frac{\left( \prod_{i=1}^r \lambda_i \right)^{1/r}}{\frac{1}{r} \sum_{i=1}^r \lambda_i}$$

$$0 \leq T \leq 1$$

$T$  "near" 1  $\rightarrow H_0$

# Cluster tendency test: A bootstrap procedure

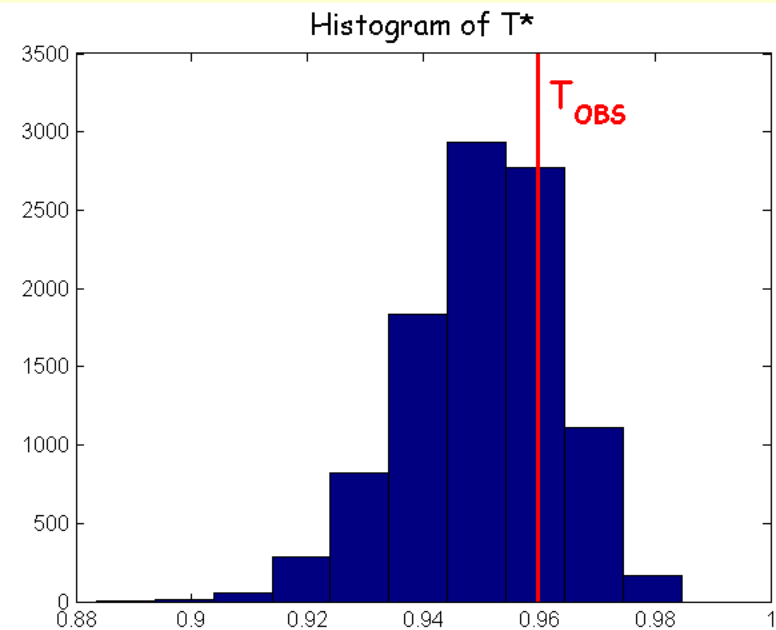
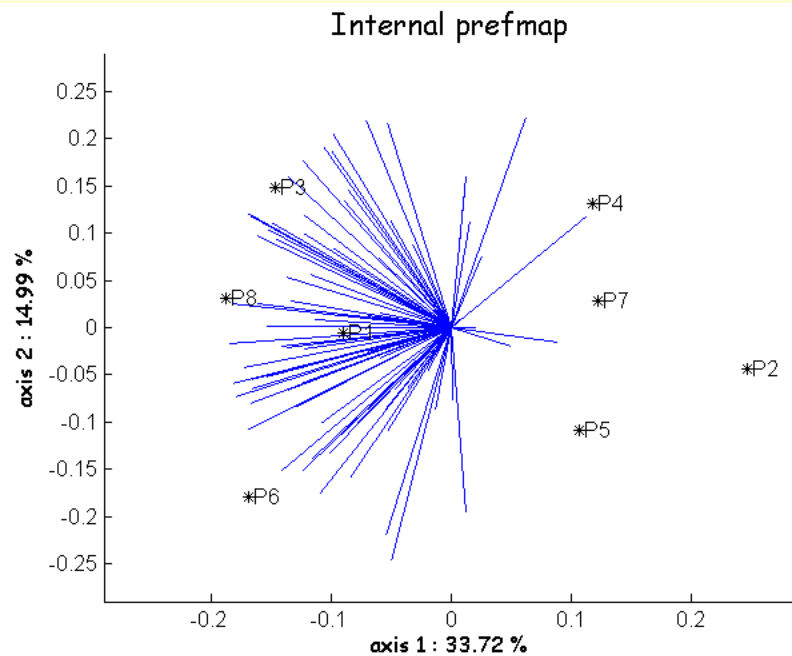


$$\text{p-value: } \#(T^* \leq T_{OBS}) / B$$

**Illustration:  
Consumer preferences  
for coffees**

# Coffee experiment: French Consumers

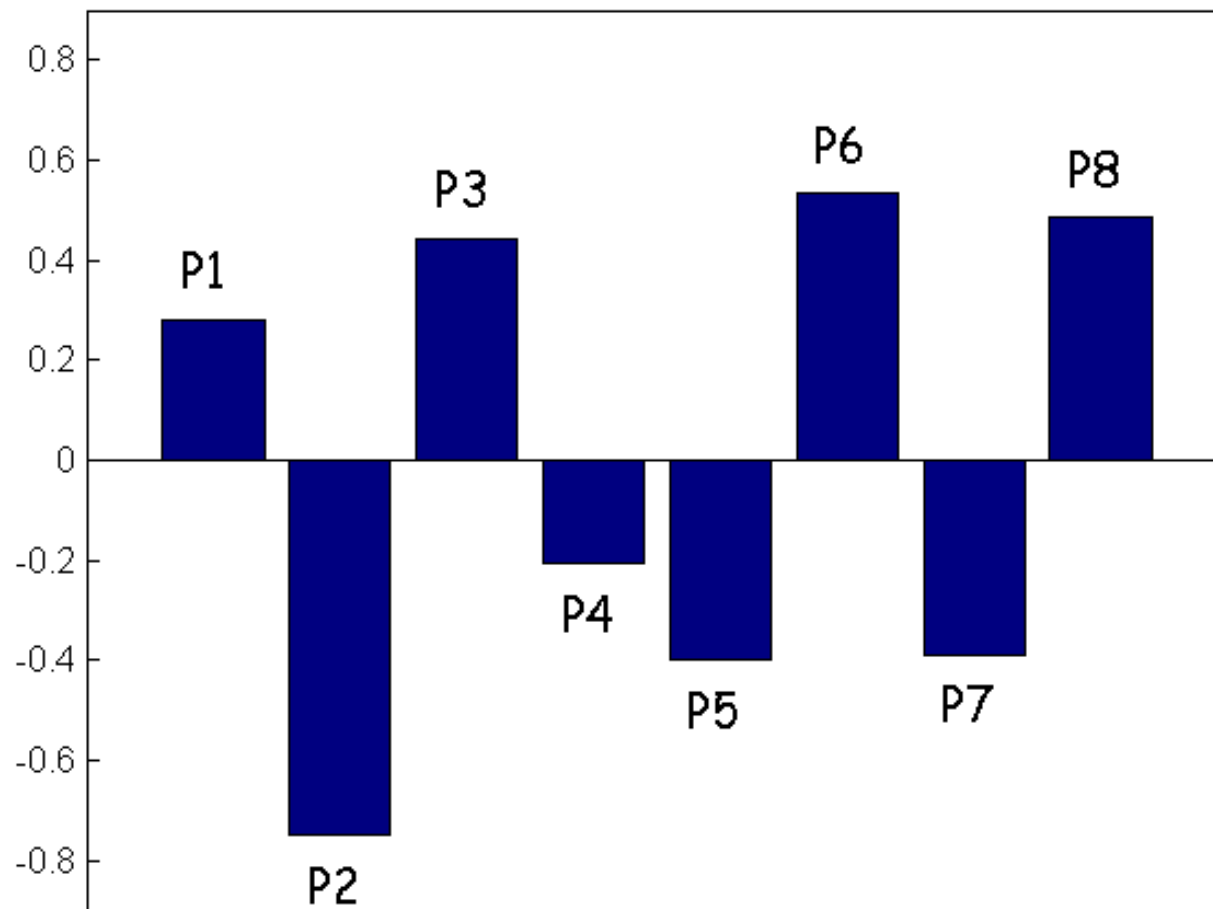
Overall liking scores for 8 coffees  
80 consumers from France



p-value: 0.76  
→ one cluster

Data: ESN (1996). *A European Sensory and Consumer Study: a Case Study on Coffee*. Published by European Sensory Network

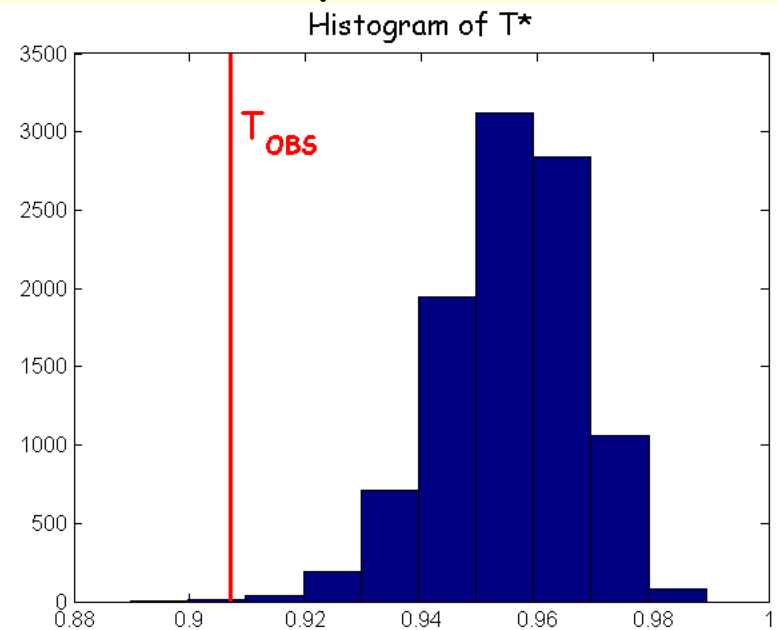
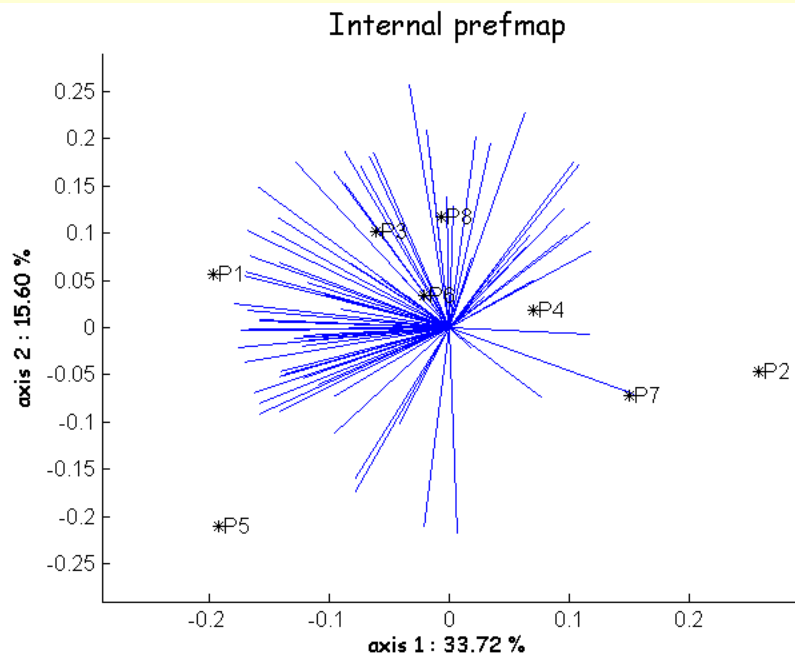
# Coffee experiment: French Consumers



Latent variable of the French Consumers

# Coffee experiment: Norwegian Consumers

Overall liking scores for 8 coffees  
79 consumers from Norway



p-value:  $< 0.01$   
→ more than one cluster

Data: ESN (1996). *A European Sensory and Consumer Study: a Case Study on Coffee*. Published by European Sensory Network

**Cluster validity tests:  
How many clusters  
are there?**

# Hierarchical clustering

$$Q_i = \frac{1}{n} \sum_{k=1}^{K_i} \sum_{j=1}^p \delta_{jk} \|\mathbf{z}_j - \bar{\mathbf{z}}_k\|^2 = p - \sum_{k=1}^{K_i} p_k \text{Var}(\bar{\mathbf{z}}_k)$$

Step 1 : Each consumer forms a cluster by himself:  $Q_1 = 0$ .

Step  $i$  : Two clusters are merged into a new cluster.

If clusters A and B are merged,

$Q$  increases by

$$\begin{aligned} \Delta Q_i &= Q_i - Q_{i-1} \\ &= p_A \text{Var}(\bar{\mathbf{z}}_A) + p_B \text{Var}(\bar{\mathbf{z}}_B) - (p_A + p_B) \text{Var}(\bar{\mathbf{z}}_{A \cup B}) \end{aligned}$$

➔ Merge the two clusters with the minimum  $\Delta Q_i$ .

# Cluster validity test

Hypotheses:  $H_0: \Delta Q_i = 0 \rightarrow$  merge without loss  $\rightarrow K \leq K_i$

$H_1: \Delta Q_i > 0 \rightarrow$  don't merge  $\rightarrow K > K_i$

Test statistic: 
$$D = \frac{\Delta Q}{p_A \text{Var}(\bar{z}_A) + p_B \text{Var}(\bar{z}_B)}$$

Reference sets under  $H_0$ : Random selection of  $p_A + p_B$  variables of the  $p$  initial variables.  
 $\Rightarrow D^*$

p-value:  $\# (D^* \geq D_{OBS}) / B$

# Testing procedure

Test 2

$H_0: K = 2$

$H_1: K > 2$

$H_0$  rejected

$H_0$  accepted

two clusters



Test  $g$

$H_0: K = g$

$H_1: K > g$

$H_0$  rejected

$H_0$  accepted

$g$  clusters



*Continue testing*

**Illustration:  
Segmentation  
of the consumers  
of Norway**

# Coffee experiment: Consumers of Norway

Cluster tendency test: more than one cluster



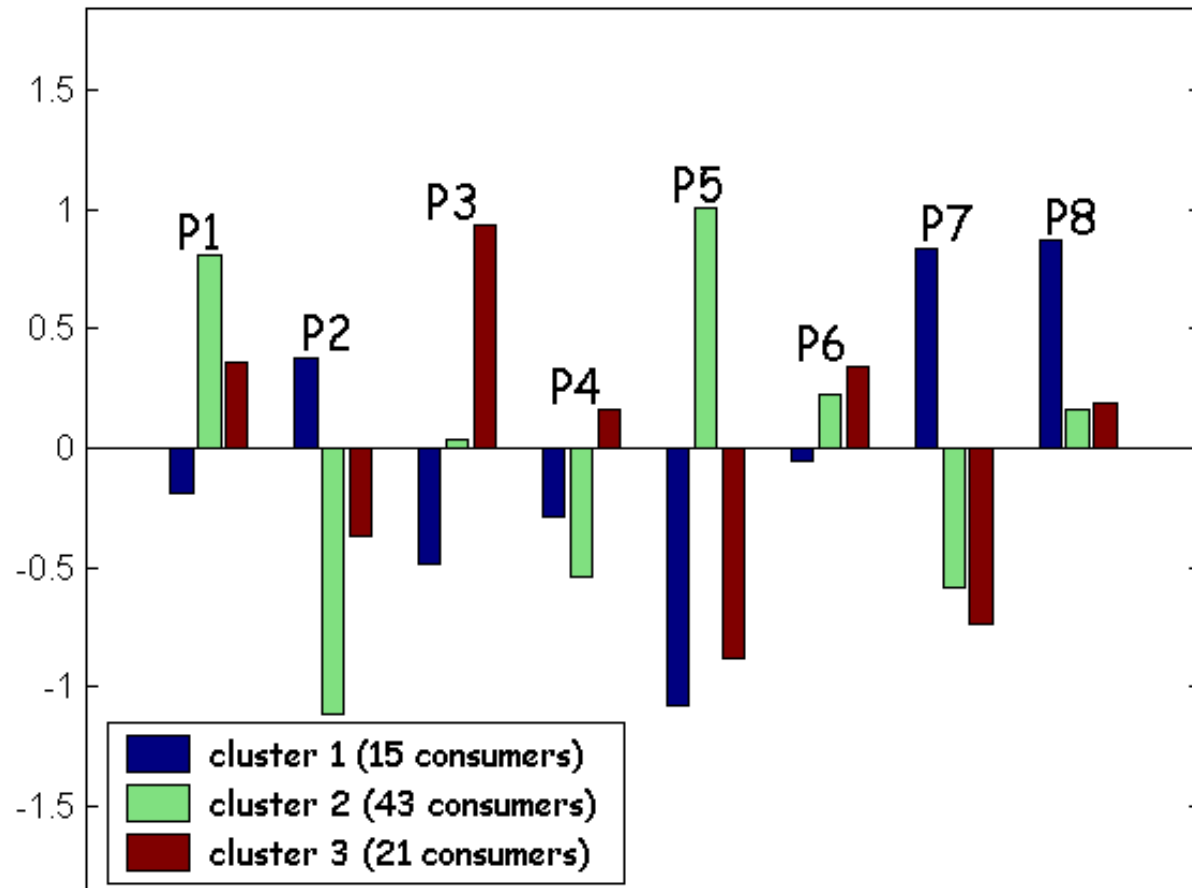
Cluster validity tests

$H_0$	$H_1$	p-value
2 clusters	3 or more clusters	< 0.01
3 clusters	4 or more clusters	0.12



**Decision for three clusters**

# Coffee experiment: Consumers of Norway



Latent variables of the three clusters

# Conclusion

**Cluster approach to analyze preference data which allows**

- to determine if there are different clusters of consumers
- to determine the number of clusters
- to associate each consumer to one cluster
- to summarize the preferences of each cluster

## **Perspective**

- taking account of external data
- other contexts:  
cluster analysis of variables around principal components