CORESTA Joint Study Groups Meeting
Smoke Science / Product Technology
2011 - Graz, Austria

Using a Structural Model based on a Class of Generalized Covariance Criteria, to explore the generation process of smoke compounds

Xavier Bry[1], Patrick Redont[1], Thomas Verron[2], Xavier Cahours[2]

[1] I3M, Univ. Montpellier II - Route de Mende, 34199 Montpellier Cedex 5, France

[2]SEITA, Imperial Tobacco Group - 4, rue André Dessaux, 45404 Fleury-les-Aubrais, France

Outline

- Brief state of Art
  - From linear modeling to structural equation modeling

- Existing methods and limitations
  - PLSPM, SEM-ML, TC-PM, MB-PLS, GSCA, MCCRM, GLLAM, RGCCA

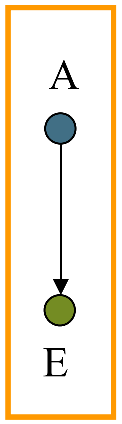
- New approach THEME-SEER
  - Global criterion, optimization program and properties

- Application to explore the generation process of smoke compounds

Modeling

A mathematical model usually describes a system by a set of **variables** and a **set of equations** that establish **relationships** between the variables (**explanatory variables** and **dependant variables)**.

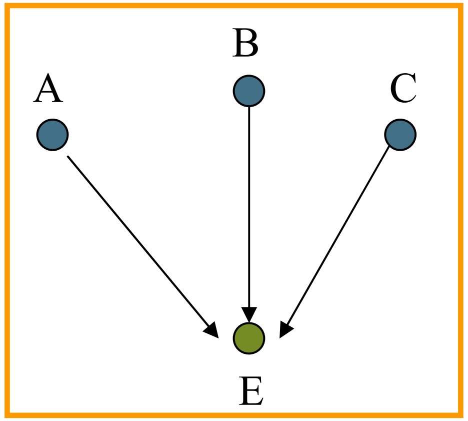
Single relationship



The only identified source of variation in E is A

OLS

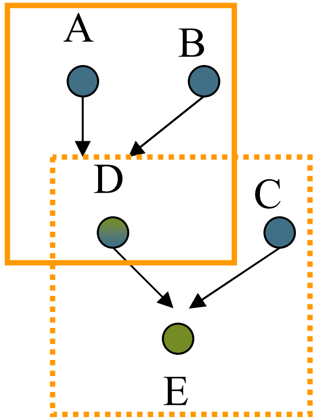
Multiple relationships



The only identified sources of variation in E are A, B and C

MLR

Multiple relationships and equations



The only identified sources of variation in:
E are D and C, and in
D are A and B

Path Modeling

The multiple equations is the most realistic approach but we must take into account the fact that the dimensions (variables) are not always totally identified.

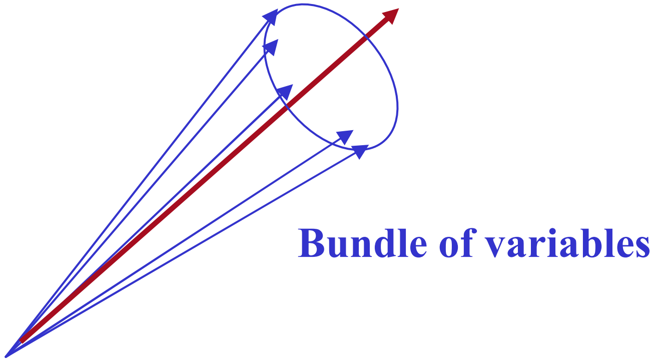
Unclear dimensions

The predictive variables (precursors) of a dependent variable (compound):

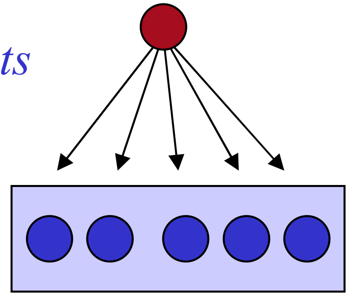
- can be unknown or not totally known (exploratory phase)

- can be difficult or impossible to measure (for example retention, combustibility **are not observed directly**)

*To get round these difficulties, we can replace the **unobserved characteristic** by **several variables** related to it*



Measurable (easily)
Selected by the experts



– Need to use component-based modeling to reduce the dimension and extract the relevant information

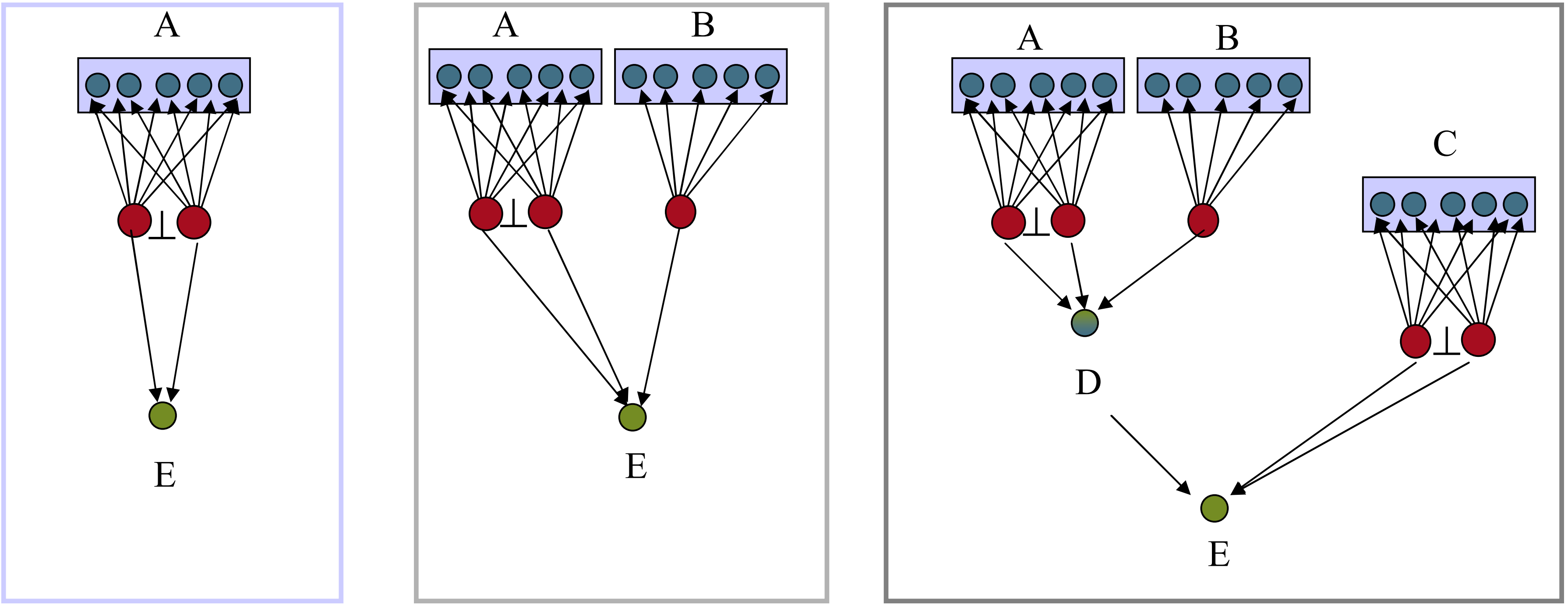
Component-based modeling (1)

Reduce the dimension and extract the relevant information of a group of variables and measure the importance of the relation between **a dependent variable** and some explanatory variables.

Partial least squares

Multiblock analysis

Structural Equation Modeling



PLS

SEER

THEME-SEER

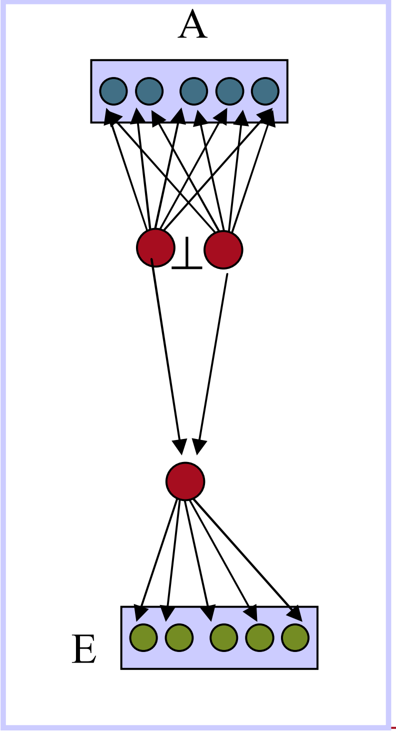
Component-based modeling (2)

Reduce the dimension and extract the relevant information of a group of variables and measure the importance of the relation between dependent variables and some explanatory variables

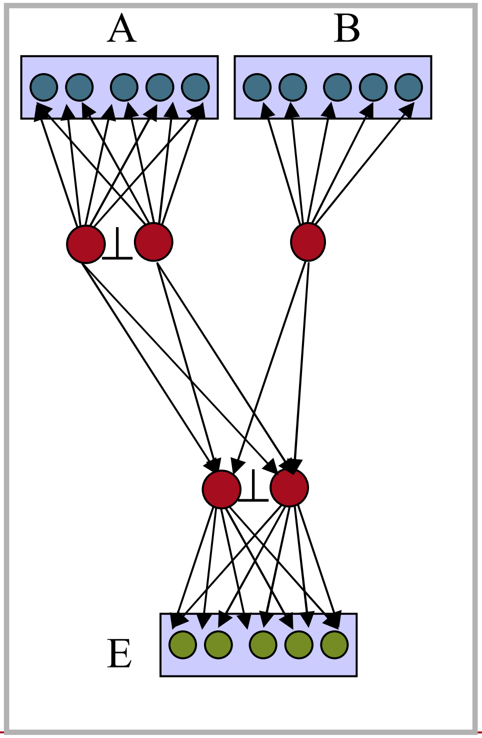
Partial least squares

Multiblock analysis

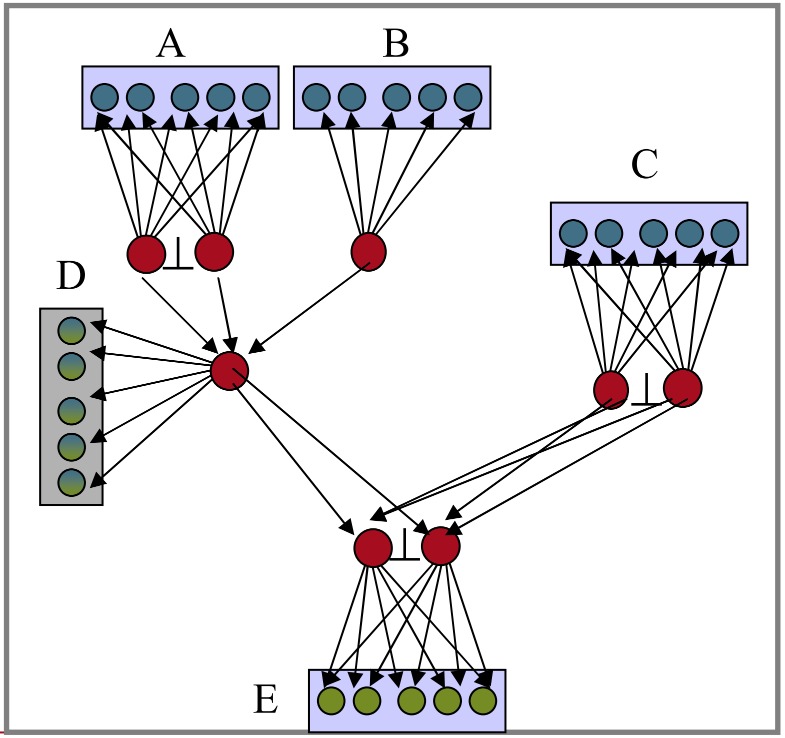
Structural Equation Modeling



PLS2



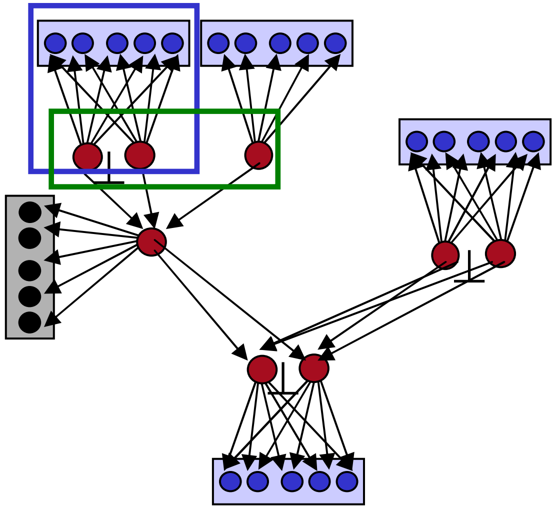
SEER



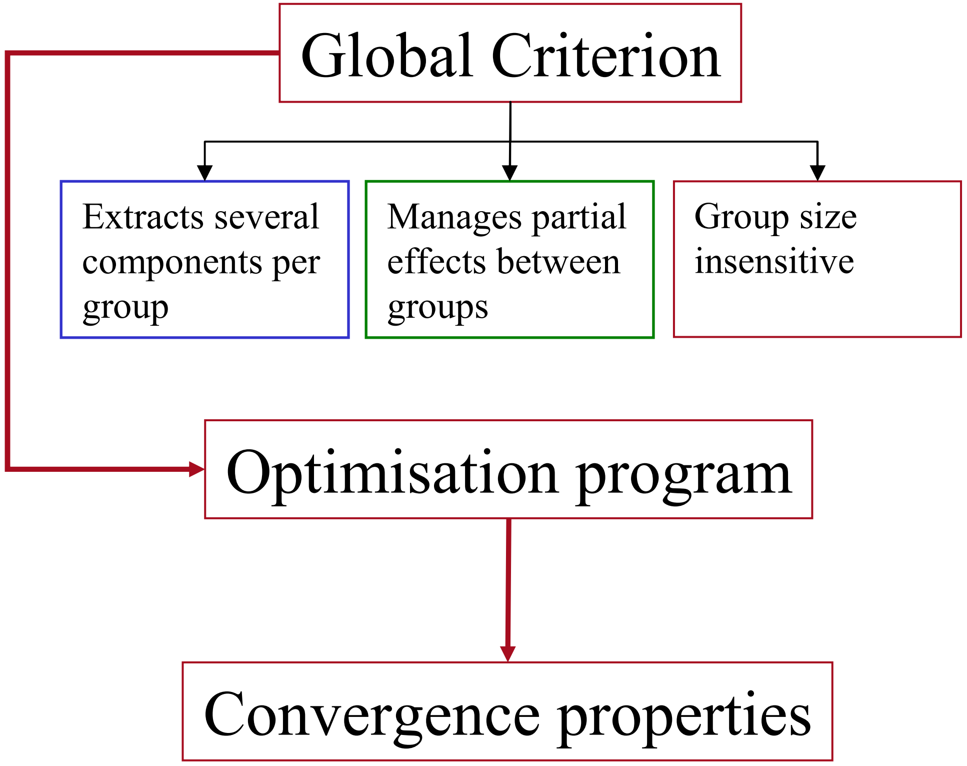
THEME-SEER

Statistical modelling

Thematic Scheme



Mathematical Strategy



Global Criterion

Extracts several components per group

Manages partial effects between groups

Group size insensitive

Optimisation program

Convergence properties

Structural Equation Methods

	SEM Method							
	PLSPM	TC-PM	MB-PLS	SEM-ML	GLLAM	RGCCA	GSCA	MCCRM
Global criterion	✗	✗	✗	✓	✓	✓	✓	✓
Criterion optimization type	✗	✗	✗	1	1	3	2	2
Manages partial effects between groups	✗	✓	✗	✓	✓	✗	✓	✓
No probabilistics assumption	✓	✓	✓	✗	✗	✓	✓	✓
Convergence of criterion	✗	✗	✗	?	?	✓	?	?
Extracts several components / group	✗	✗	✓	✗	✗	✗	✗	✓
Group size insensitive	✓	✓	✓	✗	✗	✗	✗	✗

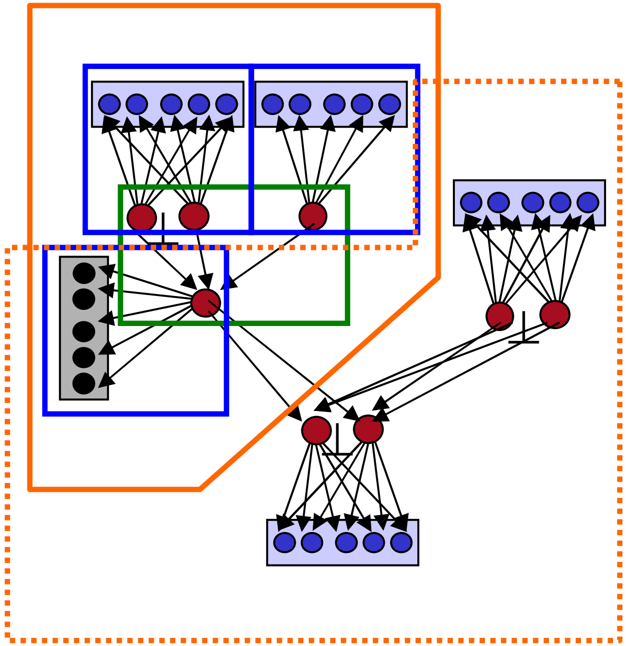
1 Max Likelihood

3 Max Compound Bivariate Covariance

2 (Alternated) Least Squares

THEME-SEER

Thematic Scheme

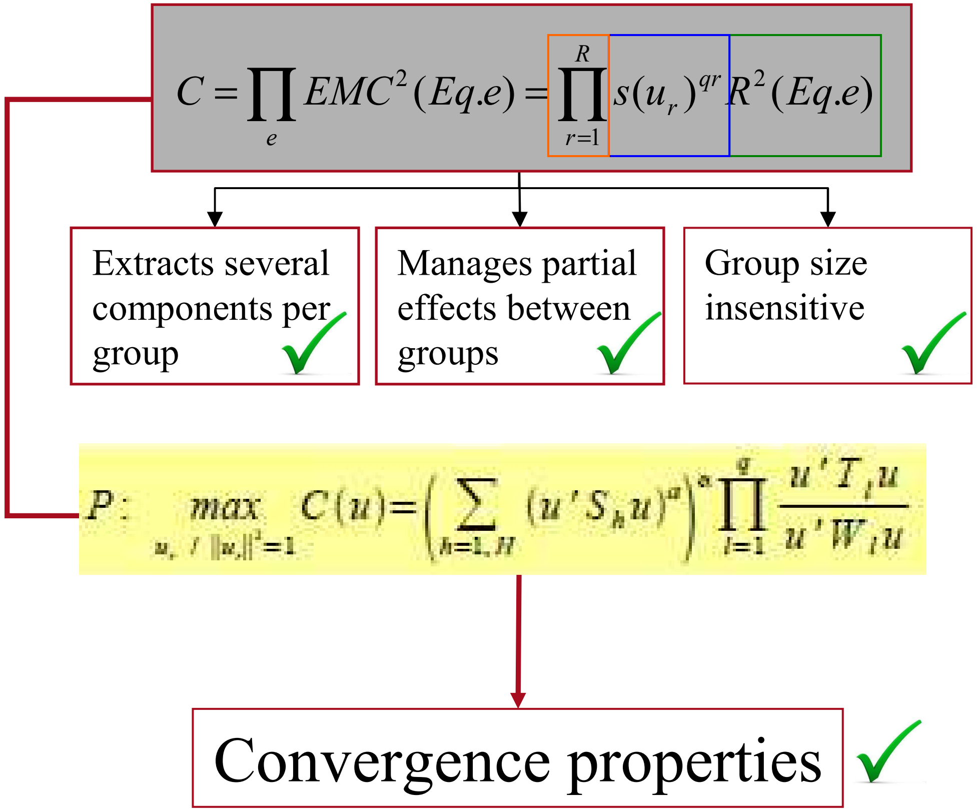


Product of all variances

Linear model fit

For each equation

Mathematical Strategy



$$C = \prod_{e} EMC^2(Eq.e) = \prod_{r=1}^{R} s(u_r)^{qr} R^2(Eq.e)$$

Extracts several components per group ✓

Manages partial effects between groups ✓

Group size insensitive ✓

$$P: \max_{u_r \,/\, \|u_r\|^2=1} C(u) = \left(\sum_{h=1,H} (u' S_h u)^{a} \right)^{s} \prod_{l=1}^{q} \frac{u' T_l u}{u' W_l u}$$

Convergence properties ✓

EMC = Extended Multiple Covariance

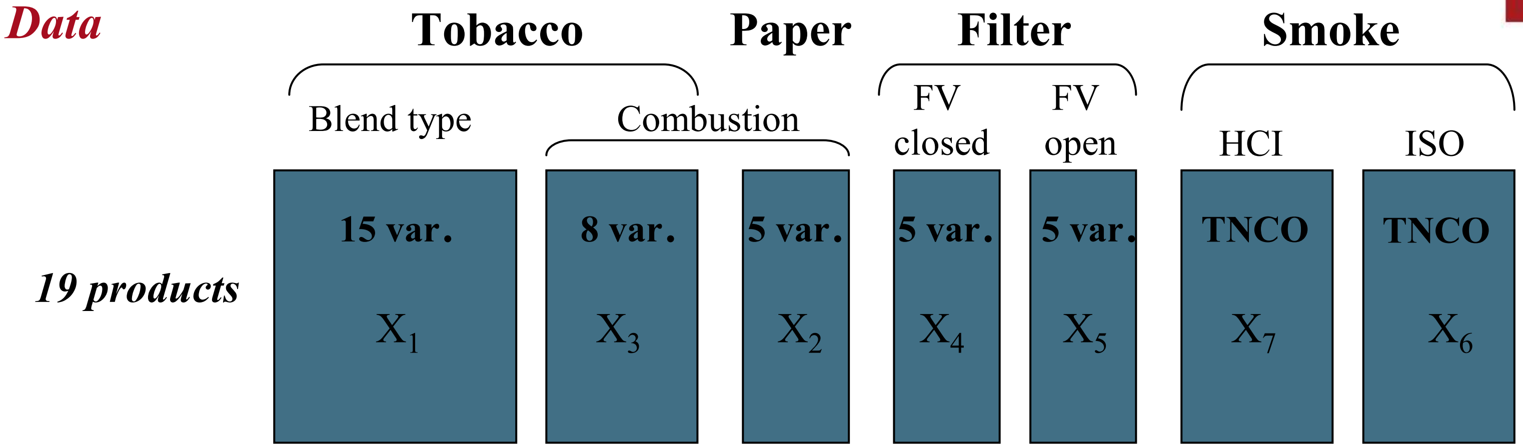
THEME-SEER

	SEM Method								
	PLSPM	TC-PM	MB-PLS	SEM-ML	GLLAM	RGCCA	GSCA	MCCRM	THEME-SEER
Global criterion	✗	✗	✗	✓	✓	✓	✓	✓	✓
Criterion optimization type	✗	✗	✗	1	1	3	2	2	4
Manages partial effects between groups	✗	✓	✗	✓	✓	✗	✓	✓	✓
No probabilistics assumption	✓	✓	✓	✗	✗	✓	✓	✓	✓
Convergence of criterion	✗	✗	✗	?	?	✓	?	?	✓
Extracts several components / group	✗	✗	✓	✗	✗	✗	✗	✓	✓
Group size insensitive	✓	✓	✓	✗	✗	✗	✗	✗	✓

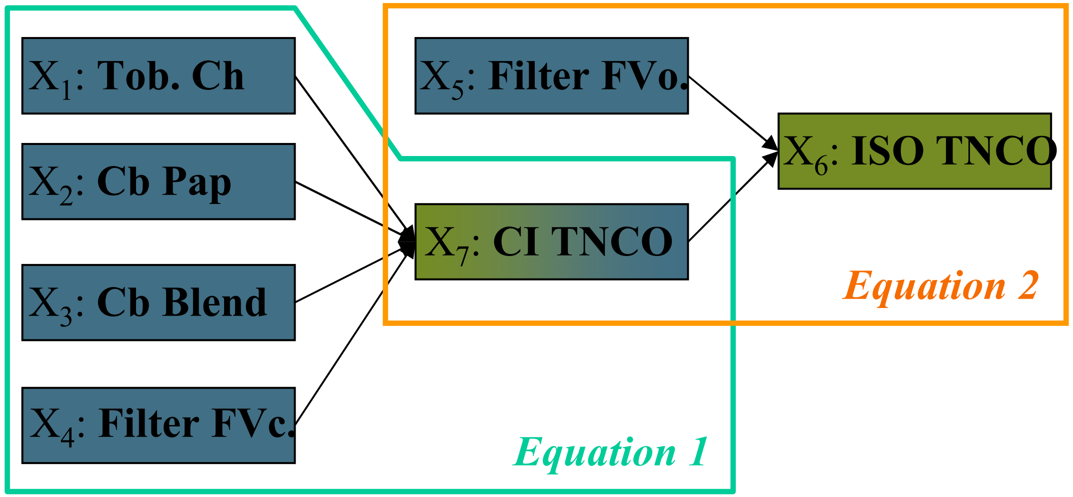
1	Max Likelihood	3	Max Compound Bivariate Covariance
2	(Alternated) Least Squares	4	Max Extended Multiple Covariance

Application : data & thematic concept

Data



Thematic conceptual model



Model design motivations

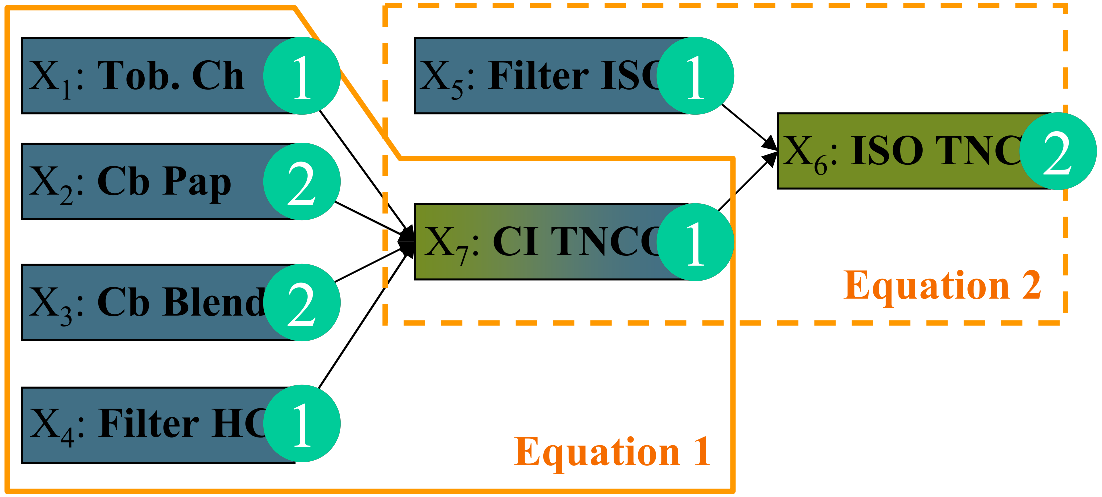
Equation 1:

Smoke compounds are generated / transferred to smoke through combustion. Filter only plays a *retention* role (Filter ventilation blocked in intense mode)

Equation 2:

Final output of smoke compounds is conditioned by other filter properties, as ventilation/dilution.

Application: number of components



X_1: Tob. Ch 1

X_2: Cb Pap 2

X_3: Cb Blend 2

X_4: Filter HC 1

X_5: Filter ISO 1

X_6: ISO TNC 2

X_7: CI TNCO 1

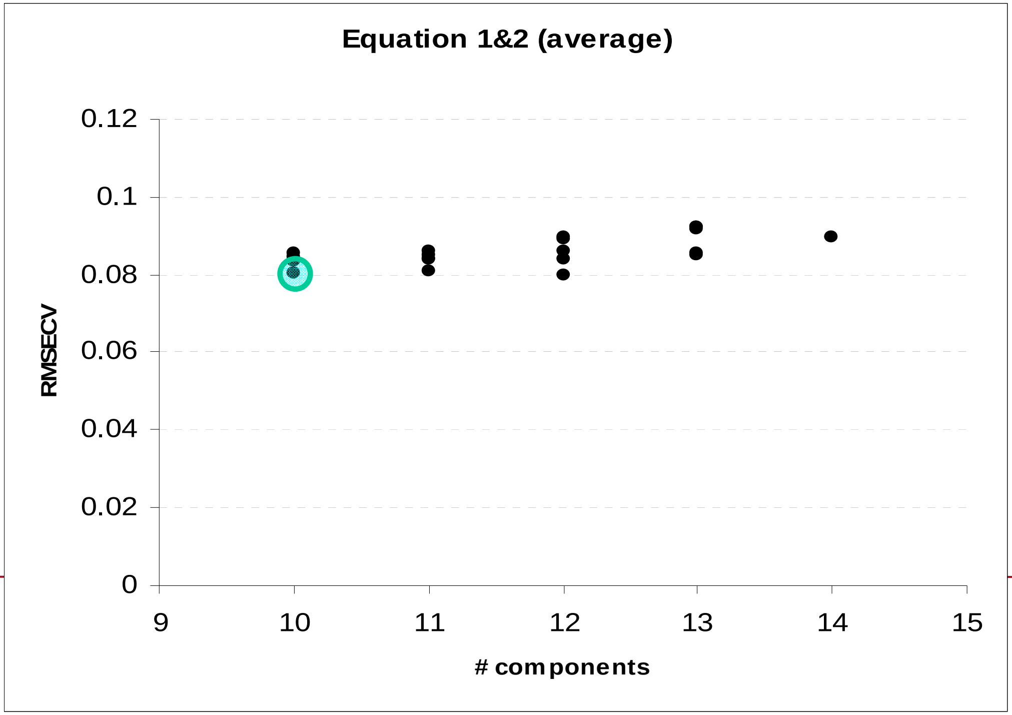
Equation 1

Equation 2

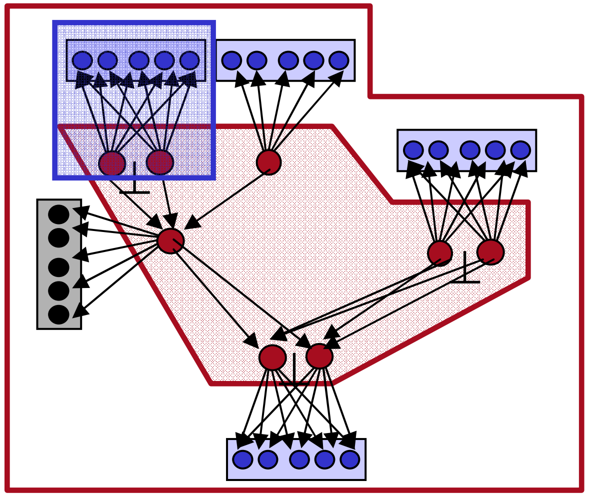
- Initially: $K = 2$ components *per* group (total=14 components)

- Remove rank K_r component alternately in each group X_r
  - → 6 « shrunk » models
  - → Evaluated *via* cross-validation
  - → Best model selected

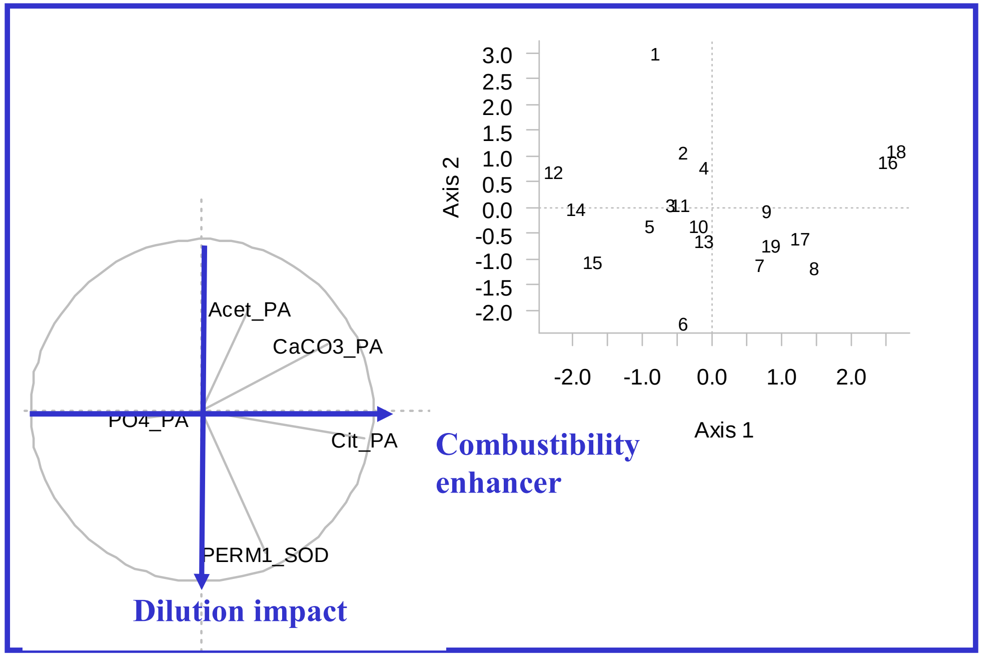
- Resume with selected model



Application: Interpretation rules



Factorial plans: *Paper parameters*



Combustibility enhancer

Dilution impact

Coefficients

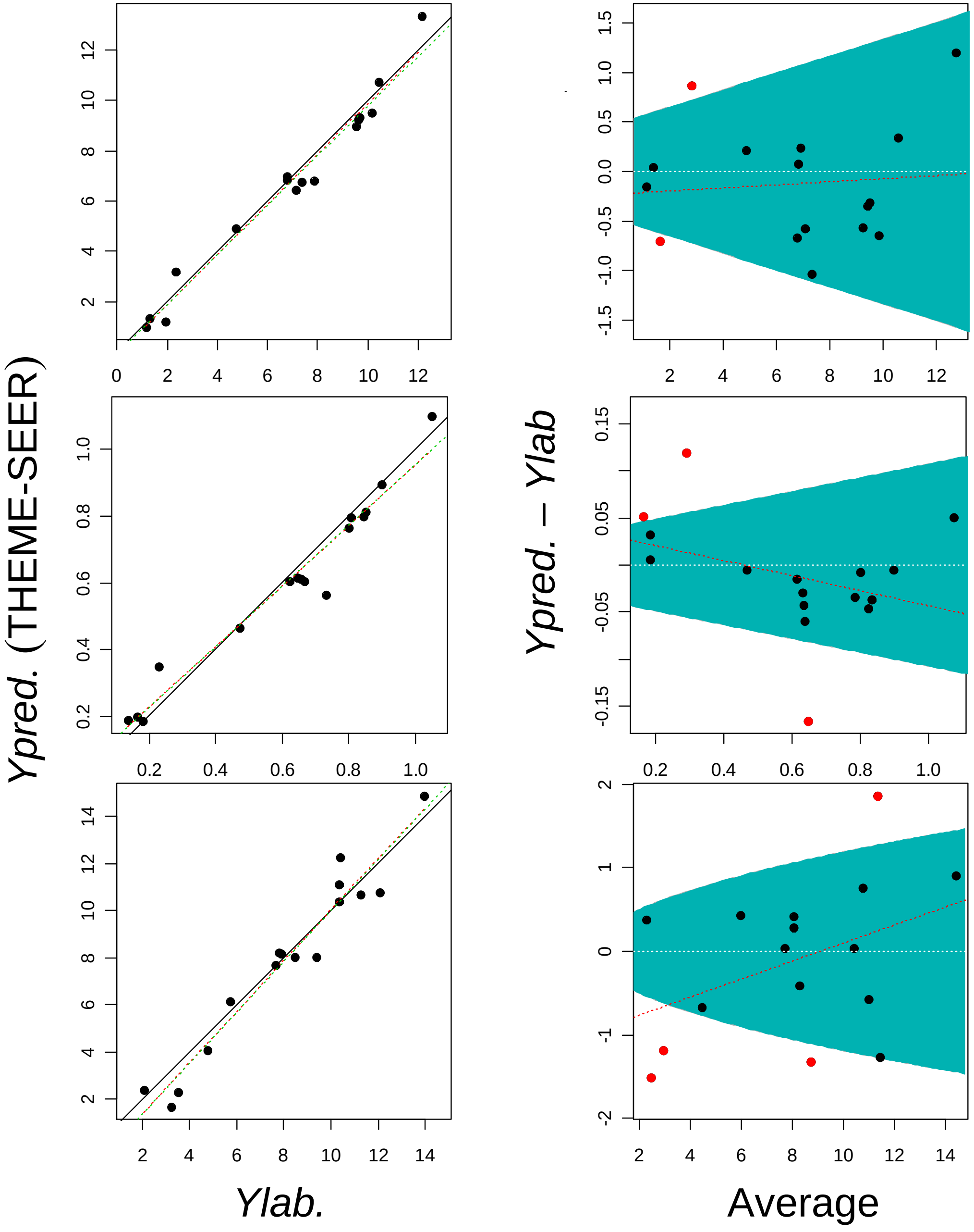
Equation 1

		NFDPM	NICO	CO
Group1	F1	-0.02	-0.11*	0.15*
	C	0.0049	0.0019	-0.0179
	Mal	-0.0040	-0.0016	0.0146
	N	0.0200	0.0078	-0.0727
	PP	-0.0168	-0.0065	0.0609
	MV	0.0000	0.0000	0.0000
	Asp	0.0782	0.0303	-0.2839
	Cit	0.0164	0.0063	-0.0594
	NO3	0.0197	0.0076	-0.0713
	Alka	0.0083	0.0032	-0.0302
	GFS	-0.0022	-0.0008	0.0078
	NH3	0.1589	0.0616	-0.5766
	NAB	0.0047	0.0018	-0.0171
	NAT	0.0002	0.0001	-0.0007
	NNK	0.0003	0.0001	-0.0012
	NNN	0.0001	0.0000	-0.0004
Group2	F1	-0.15**	-0.075	-0.136
	F2	0.07	0.030	0.22*
	Cit	-1.616	-0.064	-0.976
	PO4	6.834	0.280	2.770
	Acet	0.443	0.005	2.543
	CaCO3	-0.200	-0.009	0.008
	PERM1	-0.036	-0.001	-0.042
Group3	F1	0.01	0.38***	-0.49**
	F2	-0.12	-0.28*	0.033
	Ca	0.0141	-0.0219	0.2285
	Mg	-0.6190	-0.2500	1.3063
	Cl	-0.1413	-0.0738	0.4547
	PO4	-0.6483	-0.2215	0.9899
	K_pc	-0.0458	-0.0311	0.2144
	Hg	0.0001	0.0000	0.0002
	Pb	0.0008	0.0000	0.0008
	Cd	0.0003	-0.0001	0.0014
	NO3	0.1135	-0.0050	0.2366
Group4	F1	0.60***	1.08***	-0.289
	FL	-0.502	-0.075	0.137
	FDENSC	0.095	0.014	-0.026
	PDEF	-0.047	-0.007	0.013
	Tria	0.156	0.023	-0.042
	DIAM	-20.383	-3.055	5.558
	Weight NTM	-0.050	-0.008	0.014

Equation 2

		NFDPM	NICO	CO
Group5	F1	0.41***	0.44*	0.44***
	FV	-0.049	-0.004	-0.055
	PD	0.040	0.003	0.046
	PDFNE	-0.072	-0.006	-0.082
Group6	F1	0.27**	0.26	0.24*
	NFDPM_INT	0.118	0.008	0.110
	NICO_INT	1.268	0.081	1.184
	CO_INT	0.154	0.010	0.144

ISO Nicotine prediction quality



NFDPM

Ypred. = 0.97Ylab + 0.32

R^2 = 0.97

NICOTINE

Ypred. = 1.06 Ylab - 0.02,

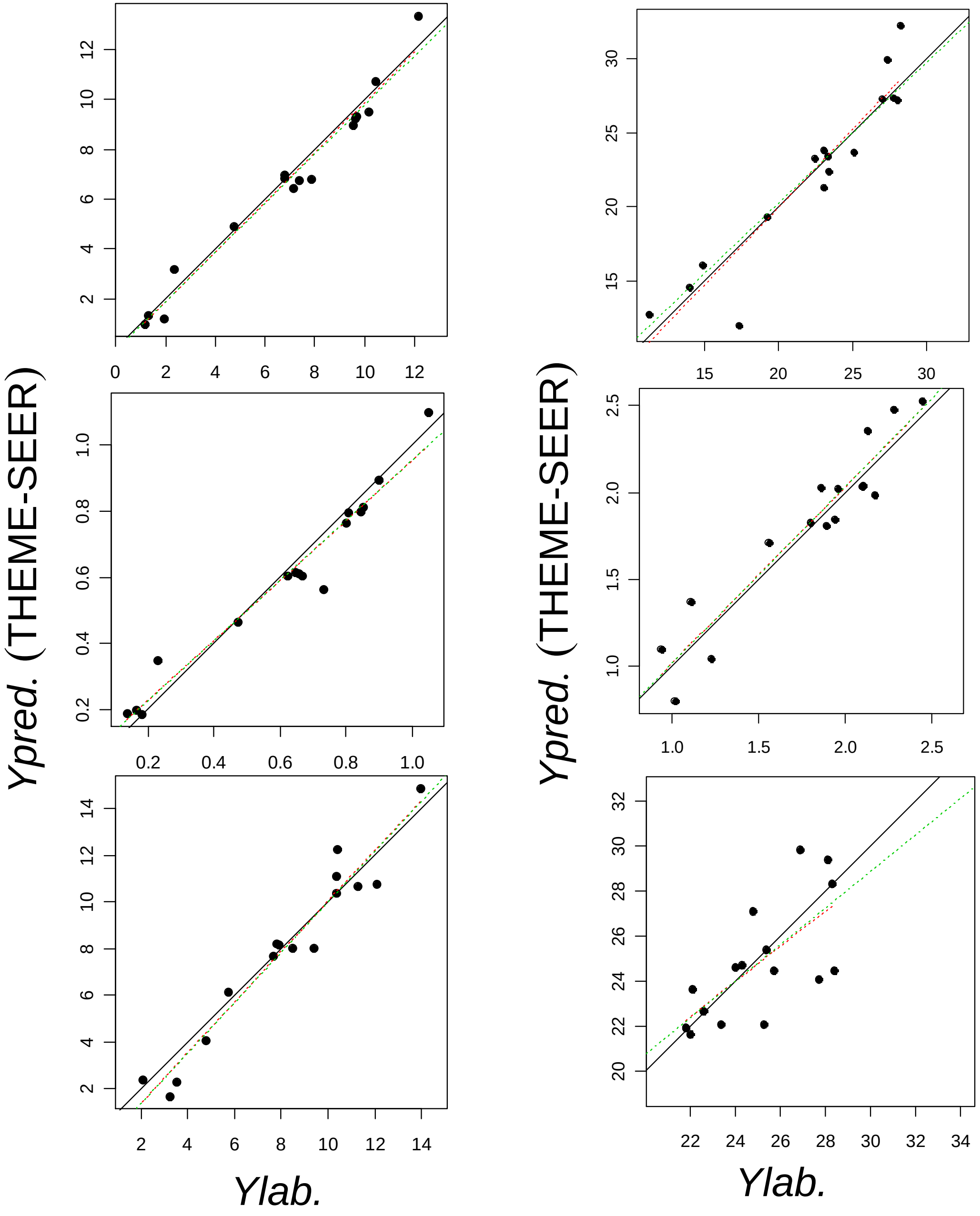
R^2 = 0.96

CO

Ypred. = 0.87 Ylab + 1.11,

R^2 = 0.94

ISO & Intense Nicotine prediction quality



NFDPM

Ypred. = 0.97Ylab + 0.32

R^2 = 0.97

Ypred. = 0.84Ylab + 0.46

R^2 = 0.88

NICOTINE

Ypred. = 1.06 Ylab - 0.02,

R^2 = 0.96

Ypred. = 0.89 Ylab - 0.17,

R^2 = 0.90

CO

Ypred. = 0.87 Ylab + 1.11,

R^2 = 0.94

Ypred. = 0.60 Ylab + 10,

R^2 = 0.47

Conclusions

Theory

- Thematic partitioning allows to interpret components conceptually, and also to analyze the complementarities of thematic aspects. Compared to other multi-group techniques, THEME-SEER:
  - ➢ solves the problem of group-weighting;
  - ➢ extends PLSR (Extended Multiple Covariance criterion);
  - ➢ allows various measures of component structural strength.

Application

- From the ***explanatory*** point of view, THEME-SEER allowed to separate the **complementary roles**, on smoke Compounds, of:
  - ➢ Tobacco type (Burley, Flue Cured, Oriental, Virginia)
  - ➢ Combustion chemical enhancers or inhibitors related to tobacco or paper
  - ➢ Filter retention power.
  - ➢ Filter ventilation power

- From the ***predictive*** point of view,

  THEME-SEER gave out a complete and robust model having accuracy within reproducibility limits (ISO regime)