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

Xavier Bry\textsuperscript{1}, Patrick Redont\textsuperscript{1}, Thomas Verron\textsuperscript{2}, Xavier Cahours\textsuperscript{2}

\textsuperscript{1}I3M, Univ. Montpellier II - Route de Mende, 34199 Montpellier Cedex 5, France
\textsuperscript{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
  - PLS-PLM, 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).

<table>
<thead>
<tr>
<th>Single relationship</th>
<th>Multiple relationships</th>
<th>Multiple relationships and equations</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Single relationship diagram" /></td>
<td><img src="image" alt="Multiple relationships diagram" /></td>
<td><img src="image" alt="Multiple relationships and equations diagram" /></td>
</tr>
</tbody>
</table>

- **OLS**
  - The only identified source of variation in E is A
  - The only identified sources of variation in E are A, B and C

- **MLR**
  - 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

- Bundle of variables

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.
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.
Statistical modelling

Global Criterion

- Extracts several components per group
- Manages partial effects between groups
- Group size insensitive

Optimisation program

Convergence properties

Thematic Scheme

Mathematical Strategy
### Structural Equation Methods

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

1. Max Likelihood
2. (Alternated) Least Squares
3. Max Compound Bivariate Covariance
Manages partial effects between groups

**THEME-SEER**

GlobalSCriteria

Extracts several components per group

Group size insensitive

Convergence properties

Mathematical Strategy

\[ C = \prod_{e} EMC^2 (Eq.e) = \prod_{r=1}^{R} s(u_r)^{q_r} R^2 (Eq.e) \]

- Extracts several components per group
- Manages partial effects between groups
- Group size insensitive

For each equation

\[ C(u) = \left( \sum_{h=1}^{H} (u'S_h u)^a \right) \prod_{l=1}^{q} \frac{u'I_l u}{u'W_l u} \]

**EMC = Extended Multiple Covariance**
<table>
<thead>
<tr>
<th>SEM Method</th>
<th>PLSPM</th>
<th>TC-PM</th>
<th>MB-PLS</th>
<th>SEM-ML</th>
<th>GLLAM</th>
<th>RGCCA</th>
<th>GSCA</th>
<th>MCCRM</th>
<th>THEME-SEER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global criterion</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Criterion optimization type</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Manages partial effects between groups</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>No probabilistic assumption</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Convergence of criterion</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>?</td>
<td>?</td>
<td>✓</td>
<td>?</td>
<td>?</td>
<td>✓</td>
</tr>
<tr>
<td>Extracts several components / group</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Group size insensitive</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>

1. Max Likelihood
2. (Alternated) Least Squares
3. Max Compound Bivariate Covariance
4. Max Extended Multiple Covariance
Application: data & thematic concept

**Data**

<table>
<thead>
<tr>
<th>Tobacco</th>
<th>Paper</th>
<th>Filter</th>
<th>Smoke</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blend type</td>
<td>Combustion</td>
<td>FV closed</td>
<td>FV open</td>
</tr>
<tr>
<td>19 products</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 var.</td>
<td>8 var.</td>
<td>5 var.</td>
<td>5 var.</td>
</tr>
<tr>
<td>X₁</td>
<td>X₂</td>
<td>X₃</td>
<td>X₄</td>
</tr>
</tbody>
</table>

**Thematic conceptual model**

- **X₁**: Tob. Ch
- **X₂**: Cb Pap
- **X₃**: Cb Blend
- **X₄**: Filter FVc
- **X₅**: Filter FVo
- **X₆**: ISO TNCO
- **X₇**: CI TNCO

**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

- Initially: $K = 2$ components per group (total=14 components)
- Remove rank $K_r$ component alternately in each group $X_r$
  - $6 \llbracket \text{shrunk} \rrbracket$ models
  - Evaluated via cross-validation
  - Best model selected
- Resume with selected model

Equation 1

1. $X_1$: Tob. Ch
2. $X_2$: Cb Pap
3. $X_3$: Cb Blend
4. $X_4$: Filter HC
5. $X_5$: Filter ISC
6. $X_6$: ISO TNC
7. $X_7$: CI TNCC

Equation 2

Equation 1&2 (average)
Equation 1

\[\begin{align*}
NFDPM & = 0.0049 - 0.0040 + 0.0200 - 0.0168 + 0.0433 + 0.0002 + 0.0003 + 0.0001 + 0.0000 + 0.0200 + 0.0078 - 0.0727 + 0.0669 + 0.0000 + 0.0001 + 0.0001 - 0.0012 + 0.0000 - 0.0004 \\
NICO & = -0.02 - 0.11 + 0.0200 + 0.0019 + 0.0146 - 0.0001 + 0.0018 - 0.0179 + 0.0019 + 0.0002 - 0.0016 + 0.0000 + 0.0001 + 0.0001 - 0.0007 + 0.0003 + 0.0000 - 0.0012 \\
CO & = 0.15 - 0.22 + 0.01 + 0.27 + 0.15 + 0.03 + 0.11 + 0.04 - 0.02 + 0.22 \\
\end{align*}\]

Equation 2

\[\begin{align*}
NFDPM & = 0.0049 - 0.0040 + 0.0200 - 0.0168 + 0.0433 + 0.0002 + 0.0003 + 0.0001 + 0.0000 + 0.0200 + 0.0078 - 0.0727 + 0.0669 + 0.0000 + 0.0001 + 0.0001 - 0.0012 + 0.0000 - 0.0004 \\
NICO & = -0.02 - 0.11 + 0.0200 + 0.0019 + 0.0146 - 0.0001 + 0.0018 - 0.0179 + 0.0019 + 0.0018 - 0.0016 + 0.0000 + 0.0001 + 0.0001 - 0.0007 + 0.0003 + 0.0000 - 0.0012 \\
CO & = 0.15 - 0.22 + 0.01 + 0.27 + 0.15 + 0.03 + 0.11 + 0.04 - 0.02 + 0.22 \\
\end{align*}\]

Coefficients

<table>
<thead>
<tr>
<th>Group1</th>
<th>NFDPM</th>
<th>NICO</th>
<th>CO</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>-0.02</td>
<td>0.11</td>
<td>0.15</td>
</tr>
<tr>
<td>Group2</td>
<td>Cit</td>
<td>PO4</td>
<td>Acet</td>
</tr>
<tr>
<td>F1</td>
<td>-1.61</td>
<td>0.28</td>
<td>2.77</td>
</tr>
<tr>
<td>Group3</td>
<td>Ca</td>
<td>Mg</td>
<td>Cl</td>
</tr>
<tr>
<td>F1</td>
<td>0.01</td>
<td>-0.21</td>
<td>0.25</td>
</tr>
<tr>
<td>Group4</td>
<td>FL</td>
<td>FDEF</td>
<td>Tria</td>
</tr>
<tr>
<td>F1</td>
<td>0.60</td>
<td>-0.07</td>
<td>0.13</td>
</tr>
<tr>
<td>Group5</td>
<td>FV</td>
<td>PD</td>
<td>PDFNE</td>
</tr>
<tr>
<td>F1</td>
<td>0.41</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Group6</td>
<td>F1</td>
<td>FV</td>
<td>PD</td>
</tr>
<tr>
<td>F1</td>
<td>0.27</td>
<td>0.11</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Application: Interpretation rules

Factorial plans: Paper parameters

Combustibility enhancer

Dilution impact
ISO Nicotine prediction quality

NFDPM
\[ Y_{pred.} = 0.97Y_{lab} + 0.32 \]
\[ R^2 = 0.97 \]

NICOTINE
\[ Y_{pred.} = 1.06Y_{lab} - 0.02, \quad R^2 = 0.96 \]

CO
\[ Y_{pred.} = 0.87Y_{lab} + 1.11, \quad R^2 = 0.94 \]
ISO & Intense Nicotine prediction quality

NFDPM

Y_{pred.} = 0.97Y_{lab} + 0.32
R^2 = 0.97

Y_{pred.} = 0.84Y_{lab} + 0.46
R^2 = 0.88

NICOTINE

Y_{pred.} = 1.06Y_{lab} - 0.02,
R^2 = 0.96

Y_{pred.} = 0.89Y_{lab} - 0.17,
R^2 = 0.90

CO

Y_{pred.} = 0.87Y_{lab} + 1.11,
R^2 = 0.94

Y_{pred.} = 0.60Y_{lab} + 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)