Fifth International Workshop on

Compositional Data Analysis

CoDaWork - 2013 - Vorau

June 3-7, 2013, Vorau, Austria

http://www.codawork2013.com/

Organized by
Vienna University of Technology
Sponsors:

Austrian Statistical Society (ÖSG)

International Association for Mathematical Geology (IAMG)

United Nations Industrial Development Organization (UNIDO)

data-analysis OG

Vienna University of Technology
CoDaWork 2013

The workshop is held from June 4-7, 2013, in the Education Center of Vorau, Austria. On June 3 and 4, 2013, an introductory course to compositional data analysis is offered. CoDaWork 2013 intends to bring together specialist researchers, data analysts, postgraduate students, as well as those with a general interest in the field, to summarize and share their contributions and recent developments.

Invited Speakers:
Morris L. Eaton (University of Minnesota, Minneapolis, USA)
Chris Glasbey (Biomathematics & Statistics Scotland, Edinburgh, UK)
Clemens Reimann (Geological Survey of Norway, Trondheim, Norway)
Raimon Tolosana-Delgado (Helmholtz-Institute Freiberg for Resources Technology, Germany)

Scientific Committee:
Peter Filzmoser (Chair, Vienna University of Technology)
John Bacon-Shone (The University of Hong Kong)
Karl Gerald van den Boogaart (Technical University Bergakademie Freiberg)
Antonella Buccianti (University of Florence)
Juan José Egozcue (Technical University of Catalonia)
Michael Greenacre (Pompeu Fabra University of Barcelona)
Karel Hron (Palacký University Olomouc)
Josep Antonio Martín-Fernández (University of Girona)

Local Organizers:
Peter Filzmoser (Chair, Vienna University of Technology)
Karel Hron (Palacký University of Olomouc)
Matthias Templ (Vienna University of Technology)
Alexander Kowarik (Statistics Austria)
Contact:

Prof. Peter Filzmoser
Department of Statistics and Probability Theory
Vienna University of Technology
Wiedner Hauptstr. 8-10
A-1040 Vienna
E-mail: codawork2013@gmail.com
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoDaWork 2013</td>
<td>1</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>3</td>
</tr>
<tr>
<td>List of Participants</td>
<td>4</td>
</tr>
<tr>
<td>Program <em>Introductory Course</em></td>
<td>6</td>
</tr>
<tr>
<td>Program <em>CoDaWork</em></td>
<td>7</td>
</tr>
<tr>
<td><em>Abstracts</em></td>
<td>13</td>
</tr>
<tr>
<td>General Information about Vorau</td>
<td>67</td>
</tr>
<tr>
<td>Telephone Numbers</td>
<td>69</td>
</tr>
<tr>
<td>Social Program</td>
<td>70</td>
</tr>
<tr>
<td>Location Map of the Abbey</td>
<td>71</td>
</tr>
<tr>
<td><em>Time Table</em></td>
<td>72</td>
</tr>
</tbody>
</table>
## List of Participants

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacon-Shone, John</td>
<td>The University of Hong Kong</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>Bajorski, Peter</td>
<td>Rochester Institute of Technology</td>
<td>USA</td>
</tr>
<tr>
<td>Barceló-Vidal, Carles</td>
<td>Universitat de Girona</td>
<td>Spain</td>
</tr>
<tr>
<td>Brandl, Michael</td>
<td>Austrian Academy of Sciences</td>
<td>Austria</td>
</tr>
<tr>
<td>Braumann, Alexander</td>
<td>Vienna University of Technology</td>
<td>Austria</td>
</tr>
<tr>
<td>Buccianti, Antonella</td>
<td>University of Florence</td>
<td>Italy</td>
</tr>
<tr>
<td>Comas-Cufí, Marc</td>
<td>Universitat de Girona</td>
<td>Spain</td>
</tr>
<tr>
<td>Cuadras, Carles</td>
<td>University of Barcelona</td>
<td>Spain</td>
</tr>
<tr>
<td>de la Rosa de Sáa, Sara</td>
<td>Universidad de Oviedo</td>
<td>Spain</td>
</tr>
<tr>
<td>de Mulder, Eduardo</td>
<td>Planet Earth Institute</td>
<td>The Netherlands</td>
</tr>
<tr>
<td>de Rooij, Mark</td>
<td>Leiden University</td>
<td>The Netherlands</td>
</tr>
<tr>
<td>Dietze, Elisabeth</td>
<td>GFZ German Research Centre for Geosciences</td>
<td>Germany</td>
</tr>
<tr>
<td>Donevska, Sandra</td>
<td>Palacký University Olomouc</td>
<td>Czech Republic</td>
</tr>
<tr>
<td>Eaton, Morris</td>
<td>University of Minnesota</td>
<td>USA</td>
</tr>
<tr>
<td>Egozcue, Juan José</td>
<td>Universidad Politécnica de Catalunya</td>
<td>Spain</td>
</tr>
<tr>
<td>Ellefsen, Karl</td>
<td>US Geological Survey</td>
<td>USA</td>
</tr>
<tr>
<td>Fačevicová, Kamila</td>
<td>Palacký University Olomouc</td>
<td>Czech Republic</td>
</tr>
<tr>
<td>Filzmoser, Peter</td>
<td>Vienna University of Technology</td>
<td>Austria</td>
</tr>
<tr>
<td>Fišerová, Eva</td>
<td>Palacký University Olomouc</td>
<td>Czech Republic</td>
</tr>
<tr>
<td>Gallo, Michele</td>
<td>University of Naples</td>
<td>Italy</td>
</tr>
<tr>
<td>Garrido Fernández, Antonio</td>
<td>Instituto de la Grasa (CSIC)</td>
<td>Spain</td>
</tr>
<tr>
<td>Glasbey, Chris</td>
<td>Biomathematics &amp; Statistics Scotland</td>
<td>UK</td>
</tr>
<tr>
<td>Graf, Monique</td>
<td>Université de Neuchâtel</td>
<td>Switzerland</td>
</tr>
<tr>
<td>Graffelman, Jan</td>
<td>Universitat de Girona</td>
<td>Spain</td>
</tr>
<tr>
<td>Greenacre, Michael</td>
<td>Universitat Pompeu Fabra</td>
<td>Spain</td>
</tr>
<tr>
<td>Hron, Karel</td>
<td>Palacký University Olomouc</td>
<td>Czech Republic</td>
</tr>
<tr>
<td>Hrizová, Klara</td>
<td>Palacký University Olomouc</td>
<td>Czech Republic</td>
</tr>
<tr>
<td>Jarauta-Bragulat, Eusebi</td>
<td>Universitat Politècnica de Catalunya (UPC)</td>
<td>Spain</td>
</tr>
<tr>
<td>Jelínková, Marcela</td>
<td>Centre of Applied Research of Vegetables and</td>
<td>Czech Republic</td>
</tr>
<tr>
<td></td>
<td>Special Crops of the Crop Research Institute</td>
<td></td>
</tr>
<tr>
<td>Johnson, Daniel</td>
<td>Lancaster University</td>
<td>UK</td>
</tr>
<tr>
<td>Kalivodová, Alżbeta</td>
<td>Palacký University Olomouc</td>
<td>Czech Republic</td>
</tr>
<tr>
<td>Kynělová, Petra</td>
<td>Vienna University of Technology</td>
<td>Austria</td>
</tr>
<tr>
<td>Leung, Tak Ching</td>
<td>University of Hong Kong</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>Lin-Ye, Jue</td>
<td>Universidad Politécnica de Catalunya</td>
<td>Spain</td>
</tr>
<tr>
<td>Lovell, David</td>
<td>CSIRO</td>
<td>Australia</td>
</tr>
<tr>
<td>Martín-Fernández, Josep Antoni</td>
<td>Universitat de Girona</td>
<td>Spain</td>
</tr>
<tr>
<td>Mateu-Figueras, Gloria</td>
<td>Universitat de Girona</td>
<td>Spain</td>
</tr>
</tbody>
</table>
Menafoglio, Alessandra  Politecnico di Milano  Italy
Mert, Can  Vienna University of Technology  Austria
Monti, Gianna  University of Milano Bicocca  Italy
Musolas, Antoni  Universitat Politècnica de Catalunya  Spain
Ortego, Maribel  Universitat Politècnica de Catalunya  Spain
Palarea-Albaladejo, Javier  Biomathematics & Statistics Scotland  UK
Parent, Leon Etienne  Université Laval  Canada
Parent, Serge-Étienne  Université Laval  Canada
Pawlowsky-Glahn, Vera  Universitat de Girona  Spain
Reimann, Clemens  Geological Survey of Norway  Norway
Reitner, Heinz  Geological Survey of Austria  Austria
Rosteck, Veronika  Springer  Germany
Roth, Georg  Cologne University  Germany
Sajdak, Marcin  Institute for Chemical Processing of Coal  Poland
Schroeder, Fabian  Vienna University of Technology  Austria
Templ, Matthias  Vienna University of Technology / Statistics Austria
Thió-Henestrosa, Santiago  Universitat de Girona  Spain
Todorov, Valentin  UNIDO  Austria
Tolosana-Delgado, Raimon  Helmholtz-Institut-Freiberg for Resources Technology  Germany
van den Boogaart, K. Gerald  Helmholtz-Institut-Freiberg for Resources Technology  Germany
Venieri, Catia  Università d’Annunzio of Chieti-Pescara  Italy
Verma, Surendra  Universidad Nacional Autónoma de México  Mexico
Walczak, Beata  University of Silesia  Poland
Zehetgruber, Judith  Vienna University of Technology  Austria
### Introductory Course — Tutorial on Compositional Data Analysis
**June 3-4, 2013, Vorau, Austria**

#### Monday, June 3, 2013:

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.00-10.45</td>
<td><strong>V. Pawlowsky-Glahn and J.J. Egozcue</strong>: <em>Introduction to CoDa</em></td>
</tr>
<tr>
<td>10.45-11.15</td>
<td><strong>COFFEE BREAK</strong></td>
</tr>
<tr>
<td>11.15-12.00</td>
<td><strong>V. Pawlowsky-Glahn and J.J. Egozcue</strong>: <em>Introduction to CoDa</em></td>
</tr>
<tr>
<td>12.00-13.00</td>
<td><strong>J.A. Martín-Fernández and M. Comas-Cufí</strong>: <em>CoDaPack, basic operations, transformations, biplot and CoDa-dendrogram</em></td>
</tr>
<tr>
<td>13.00-14.00</td>
<td><strong>LUNCH (Stiftstaverne)</strong></td>
</tr>
<tr>
<td>14.00-15.45</td>
<td><strong>P. Filzmoser</strong>: <em>Robustness for CoDa, outlier detection</em></td>
</tr>
<tr>
<td>15.45-16.15</td>
<td><strong>COFFEE BREAK</strong></td>
</tr>
<tr>
<td>16.15-18.00</td>
<td><strong>K. Hron</strong>: <em>Robust multivariate methods (PCA, LDA, regression, etc.)</em></td>
</tr>
<tr>
<td>19.00</td>
<td><strong>DINNER (Stiftstaverne)</strong></td>
</tr>
</tbody>
</table>

#### Tuesday, June 4, 2013:

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.00-10.45</td>
<td><strong>J.A. Martín-Fernández and M. Comas-Cufí</strong>: <em>Data preprocessing (missings, zeros), software “CoDaPack”, applications</em></td>
</tr>
<tr>
<td>10.45-11.15</td>
<td><strong>COFFEE BREAK</strong></td>
</tr>
<tr>
<td>11.15-13.00</td>
<td><strong>M. Templ, K. Hron, and P. Filzmoser</strong>: <em>Software “compositionsGUI”, applications</em></td>
</tr>
<tr>
<td>13.00-15.00</td>
<td><strong>LUNCH (Stiftstaverne)</strong></td>
</tr>
</tbody>
</table>
CoDaWork — Workshop on Compositional Data Analysis  
June 4-7, 2013, Vorau, Austria

Tuesday, June 4, 2013:

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 15.00-15.50 | **Keynote presentation** *(Chair: K.G. van den Boogaart)*  
**M.L. Eaton**: *A vector space approach to aspects of multivariate analysis* |
| 15.50-16.10 | **Geometrical aspects and CoDa methods** *(Chair: K.G. van den Boogaart)*  
**V. Pawlowsky-Glahn, J.J. Egozcue, and D. Lovell**: *The product space $T$ (tools for compositional data with a total)* |
| 16.10-16.30 | **G. Mateu-Figueras, J. Daunis-i-Estadella, and J.A. Martín-Fernández**: *Analysing grouped CoDa: Manova and Post-hoc contrasts* |
| 16.30-17.00 | **COFFEE BREAK** |
| 17.00-17.20 | **Counts and tables** *(Chair: J.J. Egozcue)*  
**J. Bacon-Shone**: *The risk ratios versus odds ratios argument revisited from a compositional data analysis perspective* |
| 17.20-17.40 | **C. Barceló-Vidal, M.E.R. Pierotti, and J.A. Martín-Fernández**: *Individual specialization analysis for count data* |
| 17.40-18.00 | **V. Todorov, K. Fačevicová, K. Hron, D. Guo, and M. Templ**: *Statistical analysis of compositional $2 \times 2$ tables with an economic application* |
| 18.10-19.15 | **GUIDED TOUR THROUGH THE MONASTERY** |
| 19.15-20.30 | **DINNER (Stiftstaverne)** |

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 20.30-21.00 | **Compositional data and R** *(Chair: P. Filzmoser)*  
**K.G. van den Boogaart and R. Tolosana-Delgado**: *The R package “compositions”* |
**Wednesday, June 5, 2013:**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 9.00-9.50  | **Keynote presentation**  
> (Chair: J.A. Martín-Fernández)  
> C. Glasbey, D. Allcroft, and A. Butler: *Tobit models for multivariate, spatio-temporal and compositional data* |
| 9.50-10.10 | **Zero problems**  
> (Chair: J.A. Martín-Fernández)  
> M. Templ, K. Hron, P. Filzmoser, and G.S. Monti: *Methods to detect outliers in compositional data with structural zeros*  
> T.C. Leung and J. Bacon-Shone: *Compositional Data Analysis and the zero problem: comparison of additive and multiplicative replacements with interval censoring* |
| 10.30-11.00| **COFFEE BREAK**                          |
| 11.00-11.20| **High-dimensional compositions and chemometrics**  
> (Chair: C. Glasbey)  
> D. Lovell, V. Pawlowsky-Glahn, and J.J. Egozcue: *Have you got things in proportion? A practical strategy for exploring association in high-dimensional compositions*  
> C. Mert, P. Filzmoser, and K. Hron: *Sparse principal balances* |
| 11.20-11.40| M. Greenacre: *Detecting inliers in compositional data* |
| 11.40-12.00| M. Greenacre: *Detecting inliers in compositional data* |
| 12.00-12.20| J. Palarea-Albaladejo, K. McLean, F. Wright, and D.G.E. Smith: *On the use of CODA methods for bacterial identification from proteomic mass spectra*  
> M. Sajdak and S. Stelmach: *Application of chemometric analysis for determination of contamination degree of materials obtained by thermal conversion of biomass* |
| 12.40-14.00| **LUNCH (Stiftstaverne)**                  |
| 14.00-14.20| **Distributional issues**  
> (Chair: G. Mateu-Figueras)  
> M. Graf and D. Nedyalkova: *Compositional analysis of a mixture distribution with application to categorical modelling*  
> M. Comas-Cufi, G. Mateu-Figueras, and J.A. Martín-Fernández: *Model-based clustering via Gaussian mixture models: a compositional sensitivity analysis* |
<p>| 14.40-15.00| M.I. Ortego and J.J. Egozcue: <em>Spurious copulas</em> |
| 15.00-15.20| J. Graffelman: <em>On the asymptotic distribution of isometric log-ratio coordinates</em> |</p>
<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.20-16.50:</td>
<td><strong>POSTER PRESENTATIONS</strong></td>
<td></td>
</tr>
<tr>
<td>P. Bajorski</td>
<td>Target detection in constrained models in hyperspectral images</td>
<td></td>
</tr>
<tr>
<td>M. Brandl, P. Filzmoser, C. Hauzenberger, M.M. Martinez, and W. Postl</td>
<td><em>Northern Alpine versus Carpathian radiolarites – a case study from the Upper Palaeolithic Krems-Wachtberg site (Lower Austria)</em></td>
<td></td>
</tr>
<tr>
<td>C.M. Cuadras</td>
<td>Some quirks in simple, multiple and canonical correlation</td>
<td></td>
</tr>
<tr>
<td>K. Fačevicová and K. Hron</td>
<td>Statistical analysis of $2 \times 2$ compositional tables using coordinates</td>
<td></td>
</tr>
<tr>
<td>A. Cortés-Delgado, A. López López, and A. Garrido Fernández</td>
<td>Effect of processing on the Manzanilla and Hojiblanca green Spanish-style table olive fat as assessed by compositional data analysis</td>
<td></td>
</tr>
<tr>
<td>K. Hrůzová and K. Hron</td>
<td>Compositional analysis of stock exchange data</td>
<td></td>
</tr>
<tr>
<td>M. Jelínková, K. Hron, I. Doležalová, J. Peč, and K. Dušek</td>
<td>Statistical analysis using the logratio approach for plant genetic resources evaluation</td>
<td></td>
</tr>
<tr>
<td>A. Kalivodová, K. Hron, M. Župková, and D. Friedecký</td>
<td>Partial least squares for compositional data used in metabolomics</td>
<td></td>
</tr>
<tr>
<td>P. Kynčlová, P. Filzmoser, and K. Hron</td>
<td>Compositional time series: The VAR model</td>
<td></td>
</tr>
<tr>
<td>J. Lin-Ye</td>
<td>Performance analysis of wastewater treatment in constructed wetlands</td>
<td></td>
</tr>
<tr>
<td>A. Menafoglio and P. Secchi</td>
<td>Spatial prediction of probability density functions: a kriging approach based on Aitchison geometry</td>
<td></td>
</tr>
<tr>
<td>G.S. Monti, K. Hron, K. Hrůzová, and E. Fišerová</td>
<td>Dose-response modeling using the logratio approach</td>
<td></td>
</tr>
<tr>
<td>A. Musolas and J.J. Egozcue</td>
<td>Vulnerability analysis in nuclear power plant containment building</td>
<td></td>
</tr>
<tr>
<td>M. Sajdak and S. Stelmach</td>
<td>Application of chemometric analysis in the classification and the confirmation of bio-char origin</td>
<td></td>
</tr>
<tr>
<td>S.P. Verma, L. Díaz-Gonzáles, and J.R. García-Giles</td>
<td>Monte Carlo comparison of mean with median and standard deviation for statistically contaminated samples of size 5-20 and application for handling trace element concentrations in international geochemical reference materials</td>
<td></td>
</tr>
<tr>
<td>B. Walczak, A. Smolinska, and F.-J. van Schooten</td>
<td>The effect of normalization on chromatographic data</td>
<td></td>
</tr>
<tr>
<td>J. Zehetgruber, P. Filzmoser, and K. Hron</td>
<td>Robust multivariate regression with compositional data</td>
<td></td>
</tr>
</tbody>
</table>
### Compositional methodology (Chair: V. Pawlowsky-Glahn)

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.50-17.10</td>
<td><strong>E. Jarauta-Bragulat</strong> and J.J. Egozcue</td>
<td>Modeling compositional change with simplicial linear ordinary differential equations</td>
</tr>
<tr>
<td>17.10-17.30</td>
<td>M. Vives-Mestres, J. Daunis-i-Estadella, and <strong>J.A. Martín-Fernández</strong></td>
<td>Interpretation of out-of-control signals in a compositional $T^2$ control chart</td>
</tr>
<tr>
<td>17.30-17.50</td>
<td><strong>M. Gallo</strong>, V. Simonacci, and M.A. Di Palma</td>
<td>Joint biplots for CoDa</td>
</tr>
</tbody>
</table>

18.15: Walk to Mostschank Kuchlbauer (DINNER)

### Thursday, June 6, 2013:

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.00-9.50</td>
<td>Keynote presentation (Chair: P. Filzmoser)</td>
<td><strong>C. Reimann</strong>, P. Filzmoser, and K. Hron: Challenges for CoDa in geochemical practice</td>
</tr>
</tbody>
</table>

### Variable selection (Chair: P. Filzmoser)

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.50-10.10</td>
<td><strong>S. Donevska</strong>, E. Fišerová, P. Filzmoser, and K. Hron</td>
<td>Covariance-based variable selection for compositional data</td>
</tr>
<tr>
<td>10.10-10.30</td>
<td><strong>F. Schroeder</strong>, <strong>A. Braumann</strong>, P. Filzmoser, and K. Hron</td>
<td>Robust variable selection in linear regression with compositional explanatory variables</td>
</tr>
</tbody>
</table>

10.30-11.00: COFFEE BREAK

### Applications (Chair: C. Reimann)

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.00-11.20</td>
<td><strong>A. Buccianti</strong></td>
<td>Geochemistry of Archean rocks from the perspective of compositional data analysis: an overview</td>
</tr>
<tr>
<td>11.20-11.40</td>
<td><strong>E. Dietze</strong>, K. Hartmann, and M. Dietze</td>
<td>Robust end-member modelling analysis for interpreting grain size data in palaeoenvironmental science</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td><strong>K.J. Ellefsen</strong>, D.B. Smith, and J.D. Horton</td>
<td>A modified procedure for multivariate clustering of soil geochemical data and its application</td>
</tr>
<tr>
<td>12.20-12.40</td>
<td><strong>D.E. Johnson</strong>, C. Llamas, and D. Watt</td>
<td>A compositional approach to sociophonetic variation</td>
</tr>
</tbody>
</table>

12.40-13.30: LUNCH (Stiftstaverne)

13.30-18.30: EXCURSION

19.15: CONFERENCE DINNER (Kutscherwirt)
### Friday, June 7, 2013:

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 9.00-9.50  | **Keynote presentation** *(Chair: M. Templ)*
R. Tolosana-Delgado and K.G. van den Boogaart: *Regression between compositional data sets* |
| 9.50-10.10 | **Supporting tools** *(Chair: M. Templ)*
M. de Rooij and P. Eilers: *Ternary diagrams based on a probabilistic ideal point model* |
| 10.10-10.30| S. Thió-Henestrosa and M. Comas-Cufí: *CoDaPack: still more userfriendly?*                       |
| 10.30-11.00| **COFFEE BREAK**                                                                                |
| 11.00-11.20| **Regression** *(Chair: R. Tolosana-Delgado)*
J.J. Egozcue, D. Lovell, and V. Pawlowsky-Glahn: *Testing compositional association*          |
| 11.20-11.40| E. Fišerová, S. Donevska, and K. Hron: *On calibration between parts of compositions*           |
| 11.40-12.00| C. Venieri, M. Di Marzio, and A. Panzera: *Local regression for compositional data*              |
| 12.00-12.20| K.G. van den Boogaart, R. Tolosana-Delgado, K. Hron, M. Templ, and P. Filzmoser: *Compositional regression with unobserved components or below detection limit values* |
| 12.20-13.40| **LUNCH (Stiftstaverne)**                                                                       |
| 14.00      | **BUS TRANSFER TO VIENNA**                                                                       |
The risk ratios versus odds ratios argument revisited from a compositional data analysis perspective

J. BACON-SHONE

Social Sciences Research Centre, The University of Hong Kong, Hong Kong, johnbs@hku.hk

The argument amongst epidemiologists and statisticians about whether to use risk ratios or odds ratios has raged for more than 30 years (Cummings, 2009). Those favouring risk ratios highlight the ease of interpretation by clinicians and that the risk ratio is not affected when adjustment is made by a variable that is not a confounder. Those favouring odds ratios point out that odds ratios are symmetrical with respect to both the outcome and risk variables, which is consistent with the likelihood ratio principle, unlike risk ratios. From a Bayesian perspective (Aitchison and Bacon-Shone, 1981), it is obvious that the model should be expressed in odds ratio terms.

This presentation revisits the argument from the perspective of compositional data analysis and concludes that log binomial models used by those who favour risk ratios are irrational and that it is not surprising that log binomial models fail to fit many real datasets. This can be most easily seen when the risk variables are continuous.

References


Target Detection in Constrained Models in Hyperspectral Images

P. BAJORSKI

1 Graduate Statistics Department, Rochester Institute of Technology, USA, Peter.Bajorski@rit.edu

In hyperspectral images, each image pixel is described by a spectral curve, which can be represented as a \( p \)-dimensional vector, where \( p \) is the number of spectral bands. In this paper, we discuss the task of pixel-level, or subpixel, target detection, where we assume knowledge of the target spectrum. In evaluating performance of detectors, it is often assumed that the model under which the detector is constructed is the correct model. However, in practice, the ability to identify all elements (called endmembers) of the model might be limited. Consequently, the detector would use only a subset of all endmembers. In this paper, we consider consequences of using such incomplete information. To this end, we contrast between two detectors, one that uses all endmembers and another one that uses only some endmembers. The remaining endmembers might be unknown or deliberately discarded if we believe that this would improve the detector performance. We provide formulas for detection power of the two detectors. We also calculate their relative efficiency, which is a useful tool for comparison of detectors without directly calculating the detection power. We then show some theorems that help us in understanding relationships between the power detection of the two detectors. We show when the detector using more information is more powerful, but we also show examples of the opposite to be true. For the purpose of subpixel target detection, the following model is often used in literature (Manolakis and Shaw (2002)):

\[
\mathbf{r} = \mathbf{d} \cdot \mathbf{\theta} + \mathbf{W} \cdot \mathbf{\tau} + \mathbf{\epsilon}
\]

where \( \mathbf{r} \) is a \( p \)-dimensional vector (of reflectance or radiance, for example) of a pixel spectrum, \( \mathbf{d} \) is a fixed known vector of the target spectrum, \( \mathbf{W} \) is a matrix of background endmembers as columns, and \( \mathbf{\theta}, \mathbf{\tau} \) are the unknown target scalar abundance and the vector of background abundances, respectively. The error term \( \mathbf{\epsilon} \) is assumed to follow the multivariate Gaussian (normal) distribution \( \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \). We assume that the matrix \( \mathbf{W} \) consists of \( k \) known (or estimated) background endmembers written as columns of the matrix \( \mathbf{U} \) and \( r \) unknown background endmembers written as columns of the matrix \( \mathbf{Q} \). We can write that

\[
\mathbf{r} = \mathbf{d} \cdot \mathbf{\theta} + \mathbf{U} \cdot \mathbf{\gamma} + \mathbf{Q} \cdot \mathbf{\eta} + \mathbf{\epsilon}
\]

where \( \mathbf{\tau}^T = [\mathbf{\gamma}^T, \mathbf{\eta}^T] \), that is, \( \mathbf{\gamma} \) and \( \mathbf{\eta} \) are the unknown vectors of abundances of the known and unknown background endmembers, respectively. In this setup, the target detection-testing problem can be formulated as the hypothesis-testing problem

\[
H_0 : \mathbf{r} = \mathbf{d} \cdot \mathbf{\theta} + \mathbf{U} \cdot \mathbf{\gamma}_0 + \mathbf{Q} \cdot \mathbf{\eta}_0 + \mathbf{\epsilon}
\]

\[
H_1 : \mathbf{r} = \mathbf{d} \cdot \mathbf{\theta} + \mathbf{U} \cdot \mathbf{\gamma}_1 + \mathbf{Q} \cdot \mathbf{\eta}_1 + \mathbf{\epsilon}, \quad (\mathbf{\theta} > 0)
\]

Here is one of the theorems that we use in order to find a lower bound on the power \( P_{1,\alpha} \) of the detector \( D_1(\mathbf{r}) = \mathbf{d}^T \mathbf{P}_U^+ \mathbf{r} \), where \( \mathbf{P}_U^+ = \mathbf{I} - \mathbf{U} (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T \).

**Theorem.** Under the hypotheses (3), and the fully constrained case \( (\mathbf{\theta} + \sum_j \gamma_{1j} + \sum_j \eta_{0j} = 1, \sum_j \gamma_{0j} + \sum_j \eta_{0j} = 1) \), where \( \gamma_{1j}, \gamma_{0j}, \eta_{1j}, \) and \( \eta_{0j} \) are non-negative coordinates of the vectors \( \mathbf{\gamma}_1, \mathbf{\gamma}_0, \mathbf{\eta}_1, \) and \( \mathbf{\eta}_0 \), respectively),

\[
\min_{\gamma_{1j}, \eta_{0j}} P_{1,\alpha} = 1 - \Phi \left( z_{1,\alpha} - \sqrt{\mathbf{d}^T \mathbf{P}_U^+ \mathbf{d}} \left( \mathbf{\theta} - \max_i a_i \right) / \sigma \right)
\]

for \( \alpha \in (0,1) \), where \( a_i \)'s are coordinates of the vector \( \mathbf{a}^T = \left( \mathbf{d}^T \mathbf{P}_U^+ \mathbf{d} \right)^{-1} \mathbf{d} \mathbf{P}_U^+ \mathbf{Q} \).

**Reference**

Individual specialization analysis for count data

C. BARCELÓ-VIDAL¹, M. E. R. PIEROTTI² and J. A. MARTÍN-FERNÁNDEZ¹

¹ Department of Computer Science, Applied Mathematics and Statistics - University of Girona, Spain
carles.barcelo@udg.edu

²Department of Biology, East Carolina University, North Carolina Center of Biodiversity, Greenville NC, USA

discover different patterns in their resource use. Many apparently generalist populations are in fact composed of a number of specialists, that is, individuals that use only small subsets of the population’s niche. It has been showed (e.g., Bolnick et al. (2011)) that among-individual variation in ecological traits (e.g., prey preference) may affect the dynamics of a population, species, and communities. Consequently, it is crucial to develop appropriate measures of individual specialization in relation to common methodological concerns and these complex ecological scenarios.

Martín-Fernández et al. (2011), from a compositional point of view, review four of the most relevant indices applied in ecology. Although more questions are raised than answers, it is possible to conclude from their work that traditional indices do not verify the main principles of compositional data analysis. Importantly, these workers establish that any appropriate index of individual-level of specialization (ILS) should be based on two statistics: distance and variance. The distance is needed to measure the difference between one individual and a representative resource, and the variance is required to evaluate the variation among-individuals of the same group, population, species, etc.

In this contribution we focus on the definition of an appropriate index of ILS for count data and show its properties. We discuss important methodological considerations, like the effect of the categorising resources. Note that the definition of the resource categories are related to the simplicial operations amalgamation and subcomposition. Moreover, using Monte Carlo resampling methods, we calculate the null model, used in inferential studies. For example, this model is needed to test the null hypothesis “All individuals select resource with the probability distribution given by a representative resource use”, i.e. all individuals in a population are generalist in relation to this representative resource use. Finally, we discuss how cluster analysis methods may be helpful to analyze the structure of a population, as far as resource use is concerned. In other words, how individuals could be grouped in relation to its pattern in their preferences. To illustrate above contributions we analyze a dataset from a real population harbouring individual variation in diet.

References


Northern Alpine versus Carpathian radiolarites – a case study from the Upper Palaeolithic Krems-Wachtberg site (Lower Austria)

M. Brandl1, P. Filzmoser2, C. Hauzenberger3, M.M. Martinez4 and W. Postl5
1Austrian Academy of Sciences, Prehistoric Commission, Vienna, Austria; 2Vienna University of Technology, Department of Statistics and Probability Theory; 3Department of Earth Sciences, University of Graz, Austria; 4University of Texas at Austin, Department for Anthropology; 5c/o Department of Mineralogy, Universalmuseum Joanneum, Graz, Austria.

The presence of high quality radiolarites at Palaeolithic sites in Austria has triggered an enduring discussion concerning their origin. Because the Carpathian Mountains contain such raw materials, there is a debate as to the role these sources play in Palaeolithic raw material procurement in Lower Austria. Additionally, most researchers assert that such high quality radiolarites were not available in this region during Palaeolithic times. For the current study, 10 radiolarite sources (four from the Northern Calcareous Alps, two from the St. Veit Klippen Belt and four within the Carpathian Mountains; see Figure 1) and 10 artifacts from the Krems - Wachtberg site were investigated macroscopically, microscopically and geochemically applying LA-ICP-MS. Geochemical data were further analysed with Fisher's linear discriminant analysis based on the ilr-transformed data (Filzmoser et al., 2012). Figure 1 shows the resulting plot of the first and second Fisher scores. Nine of the Krems - Wachtberg artifacts could thus be assigned to a northern alpine provenance, deriving from different sources (most likely collected from river gravels at the Danube or nearby rivers). One artifact however displayed a markedly distinct chemical composition and could clearly be assigned to the Carpathian Mountains.

![Figure 1. Fisher scores plot of the radiolarite sources.](image)

References

Geochemistry of Archean rocks from the perspective of compositional data analysis: an overview

A. BUCCIANTI

1Department Earth Sciences – University of Florence, Italy, antonella.buccianti@unifi.it

The term Archean means “ancient” and was originally used to refer to the oldest known rocks. These rocks contain the earliest fossils of life on Earth, were formed very long ago—between 2.5 to 4.0 billion years ago and have been covered by younger sediments, eroded, or subducted into the Earth's mantle. Nevertheless, some Archean strata survived in the central parts of continents. These Archean “shields” lie in the heart of Canada, Australia, Africa, India, and Siberia. Rocks of this Era consist mainly of metamorphic gneisses and granitic intrusions representing early protocontinents that assembled to produce larger continents (Precambrian cratons). Between protocontinents are belts of metamorphosed oceanic crust (greenstone belts) and sedimentary cover representing the margins of protocontinents that were deformed during collisions (Rollinson, 2007).

The geochemistry of Archean materials has been widely investigated by using traditional graphical and numerical tools. In this research chemical data were analysed from the compositional data analysis perspective, with the aim to highlight the geochemical behaviour of the elements in this important phase of Earth formation and to verify if general laws governed their partition worldwide. To achieve this aim the database GEOROC (Geochemistry of Rocks of the Oceans and Continents) maintained by the Max Planck Institute for Chemistry in Mainz was used. It contains a comprehensive collection of published analyses of Archean rocks from different geological settings worldwide. This research wish to contribute to tracks the changes in the Earth from its primordial state to the time at which modern processes were established, about 2.5 billion year ago, or earlier. The changes required significant interactions between all the major reservoirs of the Earth system and compositional data analysis was here used as a new probe for their understanding.

Figure 1. Distribution of Archean rocks worldwide (Rollinson, 2007).

References

Model-based clustering via Gaussian mixture models: a compositional sensitivity analysis

Marc Comas-Cufi\textsuperscript{1}, Gloria Mateu-Figueras\textsuperscript{2}, Josep Antoni Martin-Fernandez\textsuperscript{2}

\textsuperscript{1}Informàtica, Matemàtica Aplica i Estadística - Universitat de Girona, Spain mcomas@ima.udg.edu

\textsuperscript{2}Informàtica, Matemàtica Aplica i Estadística - Universitat de Girona, Spain

Compositional data (CoDa) appear naturally in many applied fields. Any statistical analysis involving CoDa should fulfill two main principles: scale invariance and subcompositional coherence. In this work the subcompositional coherence principle is studied when model-based clustering methods are applied to CoDa set. In addition, to Gaussian mixture models, a sensitivity analysis is done for two important parameters: the mixture proportions and the posterior cluster probabilities. This analysis shows that both parameters can be considered as compositional random vectors.
Effect of processing on the Manzanilla and Hojiblanca green Spanish-style table olive fat as assessed by compositional data analysis

A. CORTÉS-DELGADO\(^2\), A. LÓPEZ LÓPEZ\(^2\), A. GARRIDO FERNÁNDEZ\(^1\)  
\(^1\)Food Biotechnology Department – Instituto de la Grasa (CSIC), Spain, garfer@cica.es  
\(^2\)Food Biotechnology Department Instituto de la Grasa (CSIC), Spain,

There are only a few studies related to the changes in the fat composition during table olive processing. A first approach, using conventional multivariate analysis, showed that the oil from ripe olives could suffer diverse transformations during processing, mainly over the storage phase. Usually, fatty acid composition is expressed as percentages, being then a clear case of compositional data (Aitchison, 1986). However, the effect of treatments, diverse origins, and olive classes is still studied by traditional multivariate techniques. In this work, the application of compositional analysis is applied by the first time to study the effect of processing on the fatty acid composition of Manzanilla and Gordal cultivars elaborated as green Spanish-style table olives. The largest variance for \(\ln(X_i/X_j)\) was observed for \(\ln(C_{18:3\text{n-6}}/X_j)\) and ranged from 0.0256 to 0.0920. The largest \(\ln(X_i/X_j)\) was observed for \(\ln(C_{18:1\text{c}}/X_j)\) and ranged between 1.5368 (\(X_j=C_{16:0}\)) and 8.956 (\(X_j=C_{15:1}\)). Exploratory data analysis by biplot (Pawlowsky-Glahn et al., 2011) as well as discriminant analysis (Templ et al., 2011) were used to disclose changes due to processing.

Both graphs showed that there were more differences between cultivars than between treatments within cultivar. In conclusion, Spanish-style processing only introduces very limited changes in the fatty acid composition of Manzanilla and Hojiblanca table olive fat.

References


Some quirks in simple, multiple and canonical correlation

C. M. CUADRAS

Department of Statistics - University of Barcelona, Spain  
ccuadras@ub.edu

We present three quirks and counterexamples in correlation. First, given two correlated random variables $X, Y$, we may find a paradox (correlation coefficient greater than one) if we try to obtain a i.i.d. sample of $X$ (Cuadras, 2007). This paradox often disarranges probabilists and statisticians and have an explanation if the sample is exchangeable. Second, if $Y$ is correlated with $X_1, \ldots, X_k$, we expect that increasing the simple correlations with $Y$ may increase the multiple correlation. We show that this is not always true and give a principal component interpretation (Cuadras, 1993). Third, the singular values which arise in correspondence analysis can be interpreted as canonical correlations. In general, any bivariate density has a diagonal expansion where the coefficients constitute a finite or countable set $C$ of canonical correlations (Lancaster, 1958). We give a counterexample with a bivariate distribution such that $C$ is a continuous set rather than countable (Cuadras, 2005).

References


The ternary diagram is a familiar and useful display of triples of probabilities that sum to one. The scales of the diagram are linear, and so small probabilities lead to dots close to the boundary or even in the corners. Details are hard to judge then.

We propose a transformation, inspired by the probabilistic ideal point model of De Rooij (2009). In a plane an object $i$ with probabilities $(p_{i1}, p_{i2}, p_{i3})$ is represented by a point with coordinates $(x_i, y_i)$, such that $p_{ij} = c \exp(-d_{ij}^2)$, where $d_{ij}^2 = (x_i - u_j)^2 + (y_i - v_j)^2$ is the squared Euclidean distance to an “anchor point” $j$, with coordinates $(u_j, v_j)$. These anchor points are defined by the user, and will generally form an equilateral triangle.

The proposed display has several interesting properties. Triples with very small probabilities can be represented well. Equal log-odds of pairs of probabilities corresponds to straight perpendicular to the line connecting two anchor points. Equal log-odds of a single probability against the two others are given by smooth curves.

An example diagram is shown in the figure below. The data are simulated using the R package compositions. The left panel shows the standard linear diagram; The right panel shows the newly proposed diagram.

References

Robust end-member modelling analysis for interpreting grain size data in palaeoenvironmental science

Elisabeth Dietze¹, Kai Hartmann², Michael Dietze³

¹GFZ German Research Centre for Geosciences, Section 5.2 Climate Dynamics and Landscape Evolution, Potsdam, Germany, edietze@gfz-potsdam.de
²Freie Universität Berlin, Institute of Geographical Sciences, Germany
³Technische Universität Dresden, Institute of Geography, Germany

Interpreting geomorphological and sedimentological processes from grain size data in sediment archives typically runs into problems when source- and process-related grain size distributions become mixed during deposition. Classical approaches to interpret grain size data rely usually, e.g., on the methods of moments and simple means of fixed grain size classes (Folk and Ward, 1957), which are biased when applied to multimodal distributions. A powerful approach to overcome this ambiguity is to statistically “unmix” the samples, considering the inherent compositional data constraints (after Aitchison, 1986; Weltje, 1997).

This contribution presents a package for the statistical software R and its application on lacustrine grain size data sets (Dietze et al., 2013) from a user-point of view. The R package EMMAgeo originated from a recently presented Matlab-based end-member modelling algorithm (Dietze et al., 2012) that relies on eigenspace analysis, dimension reduction and the weight transformation after Miesch (1976). The package contains several extensions, added functionality and a series of user-friendly functions to perform different kinds of data tests, preparation, modelling of grain-size end-member loadings and scores along with several measures of model quality. It further provides pre-processing and visualisation tools, allows to model data sets with artificial or user-defined end-member loadings, and helps to define the optimal as well as the (system-related) most robust end-members from a large range of similarly-likely end-member model runs.

The contribution will illustrate how large data sets of artificial and natural, multi-modal grain size distributions from different, mainly lacustrine depositional environments can be analysed to infer quantitative process-related palaeoenvironmental proxies.

References


Covariance-based variable selection for compositional data

S. DONEVSKA¹, P. FILZMOSER², E. FIŠEROVÁ¹ and K. HRON¹

¹Department of Mathematical Analysis and Applications of Mathematics - Palacký University, Czech Republic
²Department of Statistics and Probability Theory - Vienna University of Technology, Austria

Often in practice we have to make decision which parts of the composition should be selected for the multivariate analysis. The reduction of the compositional parts leads to consequent facilitation of the analysis and simultaneously to simplification of the interpretation of the results of multivariate statistical analysis. Nevertheless, one has to be careful in the selection of compositional parts because the results of the multivariate statistical analysis of the subcomposition can substantially differ from those from the original full composition. For example, if we perform principal component analysis or if we construct the compositional biplot (Aitchison and Greenacre, 2002), both the interpretation of loadings and scores of the remaining subcomposition is affected.

The aim of the contribution is to present a stepwise procedure with which the reduction of parts of the original composition will result in a subcomposition that retains the important information of the original multivariate data structure. We propose stop criterion based on a statistic derived by Hron and Kubáček (2011). Real world examples are demonstrated in order to show the usefulness of this procedure.

References


A vector space approach to aspects of multivariate analysis

Morris L. Eaton

1School of Statistics - University of Minnesota, USA eaton@stat.umn.edu

Multivariate analysis, including linear model theory, is often presented in one of two ways. The classical development is presented using the standard coordinate representations of vectors and matrices. An alternative approach is to use inner product spaces, including the induced geometry, thus providing a theory not tied to traditional coordinate representations.

In this talk, we will present a third approach using vector space theory as developed in Halmos’s classical text *Finite Dimensional Vector Spaces* first written in 1946 and appearing in hard cover in 1958. It is shown that a variety of statistical notions can be defined and studied using the theory of finite dimensional real vector spaces, without reference to coordinates or inner products. In this context, the statistical notions discussed include mean vectors, covariances, the normal distribution, characteristic functions and the Gauss-Markov Theorem. Our presentation is based in part on material in Eaton(1985) referenced below.

In the context of compositional data analysis, the k-simplex is a k-1 dimensional real vector space. Thus the above discussion applies directly to a variety of matters that arise in the treatment of statistical issues involving compositional data.

References

Testing compositional association

J. J. EGOZCUE\textsuperscript{1}, D. LOVELL\textsuperscript{2} and V. PAWLOWSKY-GLAHN\textsuperscript{3}

\textsuperscript{1}Dept. Applied Mathematics III, U. Politècnica de Catalunya, Barcelona, Spain, juan.jose.egozcue@upc.edu
\textsuperscript{2}CSIRO Mathematics, Informatics, and Statistics, Canberra, Australia
\textsuperscript{3}Dept. Informatics and Applied Mathematics, U. de Girona, Spain

Proportionality is a meaningful measure of association for parts of compositions: when two parts
are perfectly proportional, the variance of their log-ratio is zero. However, when they are not perfectly
proportional, it is not immediately clear how to interpret the positive variance of their log-ratio. Two
methods for testing compositional association of two parts are presented. They check whether an
observed positive variance of a log-ratio is significantly different from what one would see if the parts
were approximately proportional. As a first step a normalisation of the variation array is proposed.
Normalisation consists in computing the proportion of total variance of the sample explained by
a variance of a simple log-ratio, which is comparable to other log-ratios of the same composition.
However, testing techniques are more involved. The main difficulty comes from the fact that the
natural null hypothesis would be the null variance of the log-ratio. Under this null hypothesis, the log-
ratio has no variability and any sample variability implies rejection of the null hypothesis. Alternative
testing consists in comparing variability of the simple log-ratio against other sources of variability.

Two approaches have been successfully used. The first one considers the linear regression

\[
b_{i1} = \beta_0 + \sum_{j=1}^{D} \beta_j b_{ij} + \epsilon_i, \ i = 1, 2, \ldots, n,
\]

where \(b_{i1}\) is the sample balances involving the two reference parts, and \(b_{ij}\) are the sample balances
orthogonal to \(b_{i1}\). A non-significant \(F\)-test suggests that the balance \(b_i\) is approximately constant up
to residuals \(\epsilon_i\). The main shortcomings of this procedure are: the source of variability is restricted to
the sample, and the null hypothesis is only a necessary, but not sufficient, condition for proportionality
of the reference parts. Alternatively, consider the regression

\[
clr_1(x_i) = \beta_0 + \beta_1 clr_2(x_i) + \epsilon_i, \ i = 1, 2, \ldots, n,
\]

where \(x_i\) are the sample compositions and subscripts 1, 2 refer to reference parts. It can be tested
for \(\beta_1 = 1\) using a \(t\)-statistics, due to the fact that \(\beta_1 = 1\) implies proportionality of parts 1, 2. The
fact that this regression is not symmetric in the two parts, and total least squares does not work in
this case, leads to consider the symmetrical regressions and consider the probit-average of the two
\(p\)-values. Again, non-significant \(p\)-values are expected for well associated parts.

The normalised variation array, accompanied with the two proposed \(p\)-values, appears to be an
exploratory tool assessing compositional association, able to replace the largely claimed correlation
matrix of raw compositional data, known to be spurious.
A modified procedure for multivariate clustering of soil geochemical data and its application

K.J. Ellefsen,1 D.B. Smith,2 and J.D. Horton2

1Crustal Geophysics and Geochemistry Science Center, U.S. Geological Survey, U.S.A., ellefsen@usgs.gov
2Central Mineral and Environmental Resources Science Center, U.S. Geological Survey, U.S.A.

Interpretation of regional-scale geochemical data may be aided by multivariate clustering of the samples. To perform the clustering, the element concentrations are transformed with the isometric log-ratio transformation, using singular value decomposition (Egozcue and Pawlowsky-Glahn, 2011) because this decomposition facilitates reduction of the dimension (with principal components) and reduction of the noise. The transformed concentrations are clustered using a model-based algorithm called MCLUST (Fraley and Raftery, 2002).

This procedure was used to cluster 960 soil samples that were collected across the state of Colorado (United States of America) (Smith et al., 2010). Thirty-one element concentrations were used for clustering. After the isometric log-ratio transformation, the dimension was reduced to 15 components, which accounted for 94% of the total variance. It is assumed that this reduction also eliminated noise. The clustering algorithm was used to find 15 clusters, and an example of a cluster is shown in Figure 1a. The cluster is spatially associated with the Pikes Peak Batholith, the Dawson Arkose, and their erosional sediments. This cluster has relatively high concentrations of those elements characteristic of silicic rocks; conversely, it has relatively low concentrations of those elements characteristic of mafic and mineralized rocks (Figure 1b).

Figure 1. (a) Samples constituting one cluster (black dots). The samples outside this cluster are represented by gray dots. The red polygon represents the exposure area of both the Pikes Peak Batholith and the Dawson Arkose. (b) Compositionally-standardized element concentrations for this cluster. The horizontal green line represents the barycenter.

References


Statistical analysis of $2 \times 2$ compositional tables using coordinates

K. FAČEVICOVÁ$^1$, K. HRON$^{12}$

$^1$Department of Mathematical Analysis and Applications of Mathematics - Palacký University, Czech Republic

$^2$Department of Geoinformatics - Palacký University, Czech Republic

A $2 \times 2$ compositional table is a special case of four-part compositional data (Aitchison, 1986), which describes relative relationships between two factors listed in rows and columns. Since only proportions between parts of composition are relevant, the analysis of such table does not depend on any normalization of cells to proportions or percentages (although these representations are popular in the practice). A $2 \times 2$ compositional table follows the Aitchison geometry on the simplex and the analysis of relationships between two factors could be performed through decomposition of the table onto its independent and interaction parts (Egozcue et al., 2008). When the goal of this analysis is test independence of factors, the interaction table needs to be transformed isometrically to real space with the standard Euclidean geometry. Concretely, cells of the interaction table are expressed in proper orthonormal coordinates (Egozcue and Pawlowsky-Glahn, 2005) and these coordinates (balances) are consequently tested on zero mean value. Since the dimension of the subspace of interaction tables is equal to one, only one coordinate can be assigned and the testing simplifies to the well-known t-test. As an alternative also principal component analysis could be applied to obtain the only nonzero coordinate of interaction table. When the hypothesis of independence between factors is rejected, the relationship between them could be further analyzed using cross- and cell-contrasts (Egozcue et al., 2008). The case of $2 \times 2$ compositional tables also points out some interesting features of statistical analysis in coordinates, possible to be applied to general compositional tables.

References


On calibration between parts of compositions

E. FIŠEROVÁ1, S. DONEVSKA2 and K. HRON2

1Department of Mathematical Analysis and Applications of Mathematics - Palacký University, Czech Republic
eva.fiserova@upol.cz

2Department of Mathematical Analysis and Applications of Mathematics - Palacký University, Czech Republic

The aim of calibration is usually to express linear relationship between the errorless measurement results obtained by measuring the same object on two different measuring methods (or, alternatively, from two different measuring devices), see e.g. Brown (1993). When the measurement results are compositional data, multivariate observations carrying only relative information (proportions, percentages), a special treatment is necessary. Because all the relevant information in compositional data is contained in log-ratios, the multivariate analysis can be decomposed into a set of univariate problems, where the calibration is performed for the corresponding coordinate and can be advantageously analyzed using linear regression modelling. Algorithms for fitting a calibration line and making statistical inference in case of three-part compositions are derived in Fišerová and Hron (2010, 2012). For D-part compositions, the calibration problem can be partitioned into $D(D - 1)/2$ partial calibration problems, performed on log-ratios of all possible two-part subcompositions separately.

In the contribution we will present tests for conformity between two measuring methods. We propose multiple tests for identification of a significant systematic difference between results from both methods and for verification that they measure with the same predetermined precision. As a consequence, we will demonstrate an analogy between the compositional variation array and the matrices of the predicted averages and residual variances, respectively, resulting from the estimation of calibration lines, that is useful for descriptive statistics.

Theoretical results will be applied to a real-world example with motivation from biochemistry.

References


Joint Biplots for CoDa

M. GALLO1, V. SIMONACCI, M.A. DI PALMA1

1Department of Human and Social Sciences – University of Naples - L’Orientale, Italy, mgallo@unior.it

Compositional data (CoDa) consist of vectors of positive values summing to a unit, or in general, to some fixed constant for all vectors. They appear as proportions, percentages, concentrations, absolute and relative frequencies. Sometimes, compositions arise from non-negative data (such as counts, area, weights, volume) that have been scaled by the total of the components because the analyst is not interested in the total sum of the vector.

The multidimensional analysis of this kind of data requires a careful consideration because the sample space for CoDa is the simplex. The first consistent methodological proposal to deal with CoDa was proposed by Aitchison (1986) when he introduced the log-ratio approach. Basically, the idea that this approach conveys is to move from the simplex space to the real space by using log.ratio transformations, applying standard statistical methods, and finally, by means of an inverse log-ratio transformation, to interpret the results in the simplex space. Starting from this paper, pairwise, centered, additive and isometric log-ratio transformations, in short plr, clr, alr (Aitchison, 1986) and ilr respectively, are proposed in literature (Egozucue et al., 2003).

In the context of dimension-reducing techniques, Aitchison (1983) proposed applying principal component analysis (PCA) after having applied a centered log-ratio (clr) transformation to CoDa. Aitchison and Greenacre (2002) suggested an adaptation of the biplot to CoDa. The biplot is a well established graphical aid in other branches of statistical analysis and can prove to be a useful exploratory and expository tool for compositions.

In literature many papers on dimensional-reduction techniques for CoDa are proposed. Based on log-ratio strategy, Gallo (2012a, 2012b, 2013) recently proposed to use three-mode analysis of compositional data. Starting from Gallo (2012b), we propose using of plr and clr joint biplots. Where in some cases the plr joint biplot is the only ones that show clearly the correlations.

References


Tobit models for multivariate, spatio-temporal and compositional data

C. GLASBEY, D. ALLCROFT, and A. BUTLER
Biomathematics & Statistics Scotland

Variables whose distributions have a singularity at zero pose a challenge for statistical modelling, particularly when data are also multivariate, spatial, temporal or compositional in structure. Tobin (1958) proposed a form of latent Gaussian model for such data: if a Gaussian variable is positive then it is observed, but otherwise it is censored and a zero is observed. Motivated by applications involving crop lodging, food intake, rainfall disaggregation and food composition, we demonstrate the effectiveness of Tobit models, and consider issues of estimation and validation.

References

Compositional analysis of a mixture distribution with application to categorical modelling

Monique Graf\textsuperscript{1} and Desislava Nedyalkova\textsuperscript{2}

\textsuperscript{1}Institut de Statistique - Université de Neuchâtel, Switzerland
and Elpacos Statistics, Switzerland monique.p.n.graf@bluewin.ch
\textsuperscript{2}Statistical Methods Unit - Federal Statistical Office - Switzerland

Many probability distributions can be represented as compound distributions, see e.g. Kleiber and Kotz (2003): Consider some vector of parameters $\theta$ as random with a distribution $h(\theta)$. Let $y$ be the vector of observations. Specify a conditional distribution for the observations $g(y|\theta)$. The compound distribution is the marginal distribution of $y$. We shall consider a positive range for the components of $\theta$. Of course $f$, $g$ and $h$ depend on some other deterministic parameters.

If we define a partition of the domain of definition of $\theta$ into $L$ parts $D_1, \ldots, D_L$, we can represent the above density as a finite mixture of densities:

$$f(y) = \sum_{\ell=1}^{L} \int_{D_{\ell}} g(y|\theta)h(\theta) \, d\theta = \sum_{\ell=1}^{L} p_{\ell} f(y|\theta \in D_{\ell}). \quad (1)$$

The densities $f(y|\theta \in D_{\ell})$ will be called component densities in the mixture representation (1) of $f(y)$. The weights $p_{\ell}$ of the component densities form a composition $p$. Graf et al. (2011) introduce this representation after fitting the income distribution with a GB2 distribution. Keeping the component densities fixed, a linear model predicting the alr representation of $p$ is introduced, thus modifying the original model. The necessary computations are available in Graf and Nedyalkova (2012) and the performance of the procedure is investigated in Hulliger et al. (2011).

Our example uses the European Union Statistics on Income and Living Conditions (EU-SILC) data. For each country, we fit model (1) in which $p$ is predicted with household categories. Comparisons across countries are processed using robust principal component analysis for compositional data as implemented in Templ et al. (2011). Interesting balances (Egozcue and Pawlowsky-Glahn, 2005) will be presented.

References


On the asymptotic distribution of isometric log-ratio coordinates

J. GRAFFELMAN\textsuperscript{1}

\textsuperscript{1}Department of Statistics and Operations Research - Universitat Politècnica de Catalunya , Spain
jan.graffelman@upc.edu

In compositional data analysis log-ratio transformations are widely used as a tool to obtain new variables that are no longer subject to the annoying unit sum constraint of the original compositions (Aitchison, 1986). A standard statistical analysis of the transformed variables usually follows. Statistical inference in compositional data analysis is hampered by the fact that the distribution of the log-ratio coordinates is unknown.

In previous work (Graffelman and Egozcue, 2011), we initially used bootstrap methods for inference with isometric log-ratio coordinates (Egozcue et al., 2003). Later (Graffelman, 2011), we obtained the asymptotic distribution of the isometric log-ratio coordinates for three-way compositions. The isometric log-ratio coordinates have the advantage that they are expressed with respect to an orthogonal basis. In brief, we part from a count vector with three elements having a multinomial distribution with parameter vector $p = (p_1, p_2, p_3)$. By applying the delta method, we derived the asymptotic distribution of the isometric log-ratio coordinates $\hat{y} = ilr(\hat{p})$. In particular, $\sqrt{n}(\hat{y} - y)$ was shown to be approximately multivariate normal $N(0, \Sigma)$, with covariance matrix

$$
\Sigma = \begin{bmatrix}
\frac{1}{2} \left( \frac{1}{p_1} + \frac{1}{p_3} \right) & \frac{1}{2\sqrt{3}} \left( \frac{1}{p_1} - \frac{1}{p_3} \right) & \frac{1}{6} \left( \frac{1}{p_1} + \frac{4}{p_2} + \frac{1}{p_3} \right) \\
\frac{1}{2\sqrt{3}} \left( \frac{1}{p_1} - \frac{1}{p_3} \right) & \frac{1}{2} \left( \frac{1}{p_1} + \frac{1}{p_3} \right) & \frac{1}{6} \left( \frac{1}{p_1} + \frac{1}{p_2} + \frac{1}{p_3} \right) \\
\frac{1}{6} \left( \frac{1}{p_1} + \frac{4}{p_2} + \frac{1}{p_3} \right) & \frac{1}{6} \left( \frac{1}{p_1} + \frac{1}{p_2} + \frac{1}{p_3} \right) & \frac{1}{3} \left( \frac{1}{p_1} + \frac{1}{p_2} + \frac{1}{p_3} \right)
\end{bmatrix} \quad (1)
$$

In this contribution, we extend this result to compositions of $k$ elements, again under the assumption that the elements of the composition are counts obtained from a multinomial distribution with $k$ categories.

Isometric log-ratio coordinates are useful in the field of genetics, where they can be used to quantify to what extent a genetic marker deviates from Hardy-Weinberg equilibrium. Equation (1) is useful in this respect for bi-allelic genetic markers that have only 3 genotypes, as is the case of single nucleotide polymorphisms (SNPs). Microsatellites, also called short tandem repeats (STRs), constitute another type of highly polymorphic genetic markers for which typically more than two alleles are observed. An example of the use of our asymptotic result will be provided, where we test microsatellite loci with multiple alleles for Hardy-Weinberg equilibrium.

References


Graffelman, J. (2011) Statistical inference for Hardy-Weinberg equilibrium using log-ratio coordinates. 4\textsuperscript{th} International Workshop on Compositional Data Analysis, Sant Feliu de Guíxols, Girona, 9-13 of May 2011.


Detecting inliers in compositional data

MICHAEL GREENACRE

1Department of Economics and Business – Universitat Pompeu Fabra, Catalonia, Spain, michael.greenacre@upf.edu

The problem of outliers, univariate or multivariate observations that are extreme compared to the others, is well-known. By contrast, inliers are points that are unexpectedly inside the data range. Inliers in univariate data are clearly difficult, if not possible, to detect. For multivariate data, however, we can capitalize on the inter-relationships between variables to detect data values that are close to the centre of their respective variable’s distribution whereas we would have expected them to be more extreme. Multivariate inliers are an issue in multivariate calibration in chemometrics (for example, Jouan-Rimbaud et al., 1999) and in fraud detection in biomedical research (for example, Evans, 2001). In those studies it is the whole observation vector that is considered as a potential inlier, whereas the present study is more directed towards identifying inlying values in observation vectors.

Some examples of inlier detection, based on imputation strategies using the singular value decomposition (Greenacre, 2010), are presented for the cases of continuous and frequency variables (Greenacre & Ayhan, 2013). Finally, I will consider how this methodology may be applied to the special case of compositional data where the data for each sampling unit are closed.

References


Compositional data naturally occur in many fields, including economical applications (Fry et al., 2000; Fry, 2011). This is due to relative character of many economical processes; reliance upon absolute information by their statistical analysis usually results in misleading conclusions (Aitchison, 1986). Similarly, in many real-world problems from stock exchange markets one can find data that describe quantitatively parts of a whole, carrying exclusively relative information. A typical example are base instruments of exchange indices, where not the absolute values of market capitalization, but rather shares of market capitalization (expressed popularly in percentages) are of primary interest. Another problem, that frequently occurs in practice, is optimization of structure of a portfolio. Finally, also distributions of changes in a stock portfolio calls for a specific treatment respecting their relative character. Reasonable processing of compositions using the logratio approach enables to apply a variety of multivariate methods for statistical analysis of stock exchange data sets like outlier detection, principal component analysis, classification issues and time series. In particular, proper statistical treatment of change of relative structure in time can form an important support for future decisions.

The aim of this contribution is to provide an overview of stock exchange data analysis using the logratio approach. A special care will be devoted to problems concerning proper choice and interpretation of orthonormal coordinates, used for the statistical analysis (Egozcue and Pawlowsky-Glahn, 2005; Fišerová and Hron, 2011). Selected statistical problem will be illustrated also with real-world data together with interpretation of the results.

References


Modeling compositional change
with simplicial linear ordinary differential equations

E. JARAUTA-BRAGULAT\textsuperscript{1} and J. J. EGOZCUE\textsuperscript{1}

\textsuperscript{1}\textsuperscript{1}Dept. Applied Mathematics III, U. Politècnica de Catalunya, Barcelona, Spain, eusebi.jarauta@upc.edu

In many fields, and particularly in economical and social sciences, compositional data evolving in time or space appear quite frequently. For instance, shares of a market or a portfolio, proportions of social groups from the point of view of the production structure, proportions of GDP by countries or geographic areas, are compositional data changing in time and they can be represented by evolutionary compositions. Their prediction and analysis using suitable models is an important goal.

There is a lack of models in which the different proportions are treated jointly and satisfying the principles of compositional data: scale invariance and subcompositional coherence. Scale invariance requires analyses to be invariant under change of units; subcompositional coherence demands that ratios between parts do not change after reduction of the number of components. First order simplicial linear differential equations satisfy the mentioned requirements and provide flexible enough models for low frequency evolutions. These kind of models can be fitted to data using least squares techniques on coordinates of the simplex. The matrix of the differential equation is interpretable, thus providing a powerful analytical tool.

The evolution of the Spanish population is studied as illustrative example; population is divided into four classes (roughly speaking, children, employees, non-employees, retired) and its evolution in the period 1976-2011 is analysed. A forecast of these population groups is provided. The fitted model reveals a remarkable instability even using only years previous to the present crisis. Evolution of GDP by geographic areas is also analysed.
Statistical analysis using the logratio approach for plant genetic resources evaluation

M. JELÍNKOVÁ1, K. HRON2, I. DOLEŽALOVÁ1, J. PEČ1 and K. DUŠEK1

1 Centre of Applied Research of Vegetables and Special Crops of the Crop Research Institute, Department of Genetic Resources for Vegetables, Medicinal and Special Plants, Czech Republic, jelinkova@genobanka.cz
2Department of Mathematical Analysis and Applications of Mathematics, Palacky University, Czech Republic

Standard statistical methods are most common techniques for analyzing chemical or biological data. However, this standard way is not able to take into account the nature and character of data carrying relative information (compositional data). To overcome this inappropriate approach that often leads to useless results, logratio transformations for compositional data are applied and, consequently, data are processed by standard statistical methods (Aitchison, 1986; Egozcue et al., 2003).

Plant genetic resources maintained in gene banks are irreplaceable natural source of biodiversity for future generations. Genetic resources information system involves passport data (data of origin) and descriptive ones as well (Dotlačil et al., 2009). Descriptive data (results on characterization and evaluation) represent morphological and phenological characteristics, yield parameters, data on resistance, biochemical and molecular data. Plants are, in general, complex organisms where intraspecific connections among various types of data are expected and as such must be considered. Processing of data using appropriate multivariate statistical methods provides for a complex view and a deeper understanding of plant genetic resources variability.

References


A compositional approach to sociophonetic variation

D.E. JOHNSON\textsuperscript{1}, C. LLAMAS\textsuperscript{2}, and D. WATT\textsuperscript{2}

\textsuperscript{1}Department of Linguistics and English Language, Lancaster University, UK d.e.johnson@lancaster.ac.uk
\textsuperscript{2}Department of Language and Linguistic Science, University of York, UK

Since the 1970’s, the bulk of quantitative research in the field of sociolinguistics has used logistic regression to estimate the social (between-speaker) and linguistic (within-speaker) constraints on the realization of many types of binary dependent variables (Cedergren and Sankoff, 1974). Linear regression has also been applied to continuous responses, mainly in phonetics (particularly vowel measurements). In both cases, mixed-effects models have recently become the method of choice, due to the unbalanced and grouped structure of linguistic data sets consisting of multiple observations drawn from multiple speakers (Gorman and Johnson, 2013).

Though linguistic alternations with three or more categories are common, especially with consonantal variables, they have been relatively neglected. Some researchers have used multinomial logistic regression (or equivalent log-linear models), although the extension to mixed multinomial models has not yet been made, casting doubt on any conclusions about the significance of between-speaker variables (Johnson, 2009). A single well-known study has employed compositional data analysis (Stuart-Smith et al., 2007). But while aggregating each speaker’s variants into a unit-sum composition largely solves the problems of imbalance and grouping, it also makes it impossible to include within-speaker covariates that may themselves be of interest, or act as confounding or at least nuisance variables.

This study uses compositional data analysis to investigate the social constraints on the realisation of coda /r/ in a balanced data set collected from 160 speakers in the AISEB (Accents and Identity on the Scottish-English Border) project (Llamas et al., 2008). Because each speaker produced the same 158 tokens of /r/ in a reading passage, the linguistic factors omitted from the compositional analysis are nuisance variables at worst. The demographic factors of interest are locality (two on each side of the border), age group (2), social class (2), and gender (2); the sample is balanced with respect to these variables, with 5 speakers in each of the 32 cells. Most importantly, there are covariates representing the degree of identification of the speaker as English, Scottish, British, Borderer, or local.

The three main realizations of /r/ – tap/trill, approximant, and zero – are already known to be affected by all the demographic factors, but the CoDA analysis allows a proper evaluation and comparison of their effects. We see that locality is the most important predictor: the tap/trill is stereotypically Scottish, and zero /r/ in the syllable coda is typically English. The effect of the other three factors interacts heavily with locality; for example, the use of coda /r/ is slowly increasing among young speakers in the town of Eyemouth, on the east coast of the Scottish Borders, but it is rapidly decreasing in Gretna, on the west coast. But the most important question is whether (self-reported) identity has an effect on language use once these other correlated factors have been taken into account. And despite the current popularity of identity as an explanatory factor in sociolinguistics, the AISEB data shows little significant effect of identity on the use of /r/.

References


Partial least squares for compositional data used in metabolomics

A. KALIVODOVÁ¹, K. HRON², M. ŽUPKOVÁ³ and D. FRIEDECKÝ³

¹Department of Mathematical Analysis and Applications of Mathematics - Palacký University, Czech Republic
²Department of Mathematical Analysis and Applications of Mathematics - Palacký University, Czech Republic
³Laboratory of Metabolomics - Institute of Molecular and Translational Medicine, Palacký University, Czech Republic

The metabolome is the collection of small molecular mass organic compounds which are in a given biological material. It includes all organic substances naturally occurring from the metabolism of the studied living organism. Metabolomics is the analysis of metabolome in a given condition (Roux et al., 2011).

When quantifying information in metabolomics, the results are often expressed in the form of data carrying only relative information. Vectors of these data have positive components and the only relevant information is contained in the ratios between their parts. They are called compositional data (Aitchison, 1986). In such a case, the sum of the variables (parts) can be rescaled to a prescribed constant.

Partial Least Squares (PLS) is a class of methods for modeling relations between sets of observed variables by means of latent variables (Rosipal and Kramer, 2006). It is a combination of principal component analysis and multiple regression. PLS is the most widely used method in chemometrics for multivariate calibration (Varmuza and Filzmoser, 2009). The aim of PLS is to predict or analyze a set of dependent variables from a set of independent variables or predictors. Prediction is made by extracting a set of orthogonal factors from the predictors which are called latent variables. The goal is to maximize the covariance between different sets of variables (Rosipal and Kramer, 2006).

PLS for compositional data needs a specific treatment for regression of estimation parameters. Concretely a particular choice of balances (Egozcue and Pawlowsky-Glahn, 2005) is used to construct orthonormal coordinates with respect to each of latent variables and consequently the standard PLS regression is applied.

Theoretical developments concerning PLS for compositional data are illustrated by real data from metabolomics. Data reflect metabolic profile and metabolic changes of model organism. Approach of targeted metabolomics was used to the analysis detecting a lot of metabolites. The compositional biplot is applied to PLS regression results.

References


Compositional Time Series: The VAR model

P. KYNČLOVÁ1, P. FILZMOSER1, and K. HRON2

1Department of Statistics and Probability Theory, Vienna University of Technology, Austria kynelova.petra@gmail.com
2Department of Mathematical Analysis and Applications of Mathematics, Palacký University Olomouc, Czech Republic

Compositional time series can be represented as time series with a constant sum at each time point \( t \) (frequently the unit constant is considered). This assumption constitutes the main problem of modelling compositional time series by standard multivariate time series methods in the sense of obtaining inadequate results. Multivariate time series of compositions, mainly proportions, occur frequently. Many economic factors are generally measured in absolute numbers. However, in some cases it is more informative and appropriate to work with proportions than directly with the absolute values.

The principal approach of compositional time series modelling is based on the use of logratio transformations (Bergman, 2008; Barcelo-Vidal et al., 2011). The procedure consists of transforming given time series to coordinates which are in the real vector space with the Euclidean structure. Subsequently, the standard modelling methods for the transformed time series can be applied, and standard tests for the parameters can be used.

In this paper an isometric logratio transformation is used for given compositional time series followed by applying a suitable VAR model (Lütkepohl, 2005). It is shown that the resulting final model and predictions do not depend on the specific ilr transformation used. In addition, particular care was devoted to the proper choice of balances (Egozcue and Pawlowsky-Glahn, 2005) which makes the corresponding statistical inference like hypothesis testing easily interpretable. The application to real data shows that the traditional non-compositional approach can result in non-sense predictions.

References


Compositional Data Analysis and the zero problem: comparison of additive and multiplicative replacements with interval censoring

T.C.LEUNG and J. BACON-SHONE

Social Sciences Research Centre, The University of Hong Kong, Hong Kong, b8894118@hku.hk

Compositional data analysis (Aitchison, 1986) has been widely applied to geological data, and in biomedical, educational, social, psychological and political disciplines.

However, there’s always one difficulty: the presence of zeros. Various methods to handle the zeros such as combining data, replacing the zeros with predicted non-zero values had been proposed with its interpretation analysed by scholars. (Martin-Fernandez, Barcelo-Vidal and Pawlowsky-Glahn, 2003; Aitchison and Kay, 2003; Bacon-Shone, 2003; Butler and Glasbey, 2008).

In this analysis, additive and multiplicative replacements on the Halimba Bauxite Deposit data (Martin-Fernandez, Barcelo-Vidal and Pawlowsky-Glahn, 2003) have first been compared. An MSD plot had been made to enable comparison with the original data, showing a lower MSD for multiplicative replacement of zero items when compared with that of additive replacement, which further supports the use of multiplicative replacement as a simple means to deal with zeros in log-ratio analysis of compositional data. The same methods are also applied to the foraminiferal data set found in Aitchison (1986).

In this presentation, Bayesian analysis assuming interval censoring on the original scale will be compared with the additive and multiplicative replacements, using the above two data sets. This analysis has the advantage of reflecting the data recording process, which necessarily involves interval censoring and may avoid the need for special treatment of zeros in a compositional dataset.

References


Performance analysis of wastewater treatment in constructed wetlands

J. LIN-YE

Graduate student - UPC Barcelona Tech, Spain jl.iccp@gmail.com

The use of horizontal subsurface flow constructed wetlands (SSF) for the treatment of municipal wastewater in small communities is becoming a commendable alternative to conventional wastewater management in a society led to the sustainable living. From 2001 to 2003, the Environmental Engineering Division of the Hydraulics Department of the BarcelonaTECH had carried out an experiment on SSF in the small town of Les Franqueses. The SSF’s were connected in parallel to a wastewater source so that each SSF received the same discharge at time $t$. Three perforated tubes were uniformly distributed along the wetland bed. The SSF’s were designed with different aspect ratios: A(1:1), B(1.5:1), C(2:1) and D(2.5:1). Moreover, in type A SSF, water depth was 0.5 m; in type B SSF water depth was 0.5 m in 2001 and 2002 and changed to 0.27 m in 2003; in type C SSF, water depth was 0.5 m and for SSF D water depth was 0.27 m. For each of the SSF groups described, there were 2 subgroups with different substrate medium sizes and porosities: subgroup 1 had a substrate medium size of 10 mm and a porosity of 39%, while subgroup 2 had a substrate medium size of 3.5 mm and a porosity of 40%. Additionally, a variety of hydraulic load rates (HLR) were applied in order to test HLR effects on SSF water depuration: 20, 27, 36 and 45 mm/day. In summary, there were five factors characterizing the SSF: aspect ratio, water depth, substrate medium size, porosity and HLR. According to the way that the factor levels are chosen, the number of treatments used in the campaign was 80. Concentrations of affluent and effluent flow were measured in 824 measurements. This data-set was analysed by J. García et al. (2005) using conventional statistics.

This study is aimed at reviewing the previous results using compositional techniques on COD, BOD$_5$, and NH$_3$ and H$^+$ concentrations (517 measurements remaining). An exploratory analysis of affluent, effluent flows and their perturbation-difference is conducted. CoDa-MANOVA is used to detect significant differences between designs. The result is that the most suitable subsurface flow constructed wetland is that of aspect ratio of 1:2.5, water depth of 0.27 meters which is significantly different from other SSF designs. Longer SSF’s (aspect ratio of 1:2.5) are clearly more efficient than SSF’s with other aspect ratios in BOD$_5$ reduction. While the most suitable depth is 0.27 meters, which helps reducing BOD$_5$ because aerobic environments are more efficient in organic material removal.

**Reference**

Have you got things in proportion? A practical strategy for exploring association in high-dimensional compositions

D. LOVELL\textsuperscript{1}, V. PAWLOWSKY-GLAHN\textsuperscript{2} and J. J. EGOZCUE\textsuperscript{3}

\textsuperscript{1} CSIRO Mathematics, Informatics, and Statistics, Canberra, Australia, David.Lovell@csiro.au
\textsuperscript{2} Dept. Informatics and Applied Mathematics, U. de Girona, Spain
\textsuperscript{3} Dept. Applied Mathematics III, U. Politècnica de Catalunya, Barcelona, Spain

This paper is motivated by a desire to apply compositional data analysis to molecular bioscience where small numbers of observations of large numbers of components are typical. Awareness is gradually growing among bioscientists that most of their measurement processes yield only relative information and that these data require appropriate analysis and interpretation. We are keen to ensure that practical compositional methods exist and can be applied as well-principled alternatives to the potentially misleading correlation-based methods that are generally used in bioscience at the moment.

Our strategy combines the principal balances approach to partitioning the variation array, with recently proposed methods for testing the hypothesis that pairs of components are proportional. Principal balances are defined as a sequence of orthonormal balances which maximize successively the explained variance in a data set. This sequence forms a binary tree. By testing the hypothesis that the two child branches of balances at any given node are proportional, we have a data-driven means of determining whether to group them.

The decision to group at a particular principal balance can be seen as identifying components that carry the same information. In bioscience, the practical impact of this is to identify components that are strongly related, giving clues about the biological mechanisms at play. We see this partitioning-plus-association-testing strategy as a sound alternative to popular techniques such correlation-based clustering (often visualized using heatmaps) which are generally applied without considering the compositional nature of the underlying data.

We explore the usefulness, strengths and limitations of this approach by using it to find proportional sets of components in yeast gene expression data where the levels of over 3000 messenger RNAs were observed in a 16-point time course.
Multivariate analysis of variance (Manova) is a powerful tool for analysing mean differences across different groups. This technique is widely used in many disciplines (e.g., Aebischer et al, 1993, Howel 2007, Kacjan-Marsic et al, 2010) in order to analyse with inferential tools the structure of grouped data.

The typical Manova technique, designed for data in the Euclidean space, is obviously not suitable for compositional data. However, the particular Euclidean geometric structure of the simplex and the log-ratio approach (Aitchison, 1986) are often not considered in the literature. Interestingly, some authors realize that compositions require a particular treatment because they apply other transformations like the arcsin squared-root (e.g., Bolnick et al 2010, Smith et al 2011).

In this communication, first we show the drawbacks and effects of applying typical Manova in an inappropriate space. Secondly, we establish basic rules to apply Manova in a compositional context and we adapt Manova methodology to this specific case. In particular we study the effect of the log-ratio transformation applied and the orthonormal basis selected. Finally, we discuss the post-hoc contrasts used when we obtain significant differences. We apply the proposed methodology, as well as appropriate descriptive tools to highlight differences between groups, to real data set.

References


Sparse principal balances

C. MERT\(^1\), P. FILZMOSER\(^1\), and K. HRON\(^2\)

\(^1\)Department of Statistics and Probability Theory, Vienna University of Technology, Austria

\(^2\)Department of Mathematical Analysis and Applications of Mathematics, Palacký University Olomouc, Czech Republic

Principal component analysis (PCA) transforms the variables of a dataset into new variables, called principal components (PCs) which do not correlate with each other and subsequently maximize the explained variance. It is thus frequently used for dimension reduction. Applying PCA to clr-transformed compositional data results in loadings that form an orthonormal basis of the simplex, and in scores that are ilr coordinates. The ilr coordinates are usually difficult to interpret since they involve all compositional parts to some extent. This has been solved by the concept of balances (Pawlowsky-Glahn and Egozcue, 2005), orthonormal coordinates consisting of two non-overlapping groups of compositional parts. Balances can be constructed by a sequential binary partition (Pawlowsky-Glahn and Egozcue, 2005), but this construction does not consider a criterion like maximizing the explained variance.

For this reason, Pawlowsky-Glahn et al. (2011) defined principal balances (PBs) which are balances that are subsequently constructed to maximize the explained (remaining) variance. The computation of PBs requires an exhaustive search among all possible sets of orthogonal balances, which is computationally infeasible especially for high-dimensional data. Pawlowsky-Glahn et al. (2011) thus introduced different approximative algorithms using the PCs of the clr-transformed data, and based on hierarchical clustering. In an example with 21 compositional parts they have demonstrated the usefulness of this approach.

In this paper we want to go further. We propose an algorithm to compute PBs for data sets with hundreds or even thousands of parts, as it is typical in chemometrics (e.g. mass spectral data) or biology (e.g. gene expression data). The basic assumption is that only a small subset of parts contributes to the main data variability. This requirement is typically met with microarray data, where only few genes are responsible for group separation, causing the essential data variability.

The proposed algorithm is based on sparse PCA (SPCA) (e.g. Croux et al., 2013) applied to the clr-transformed data. SPCA finds components that maximize the variance, with an additional penalty term to take sparseness into account. The result is a sparse loading matrix containing many zeros, which makes the interpretation easier. The penalty parameter, say \(\lambda\), regulates the priority between variance maximization and sparseness, and it can be determined by a BIC-type criterion (Croux et al., 2013). The resulting loadings of the first \(k\) sparse PCs still need to be converted to balances. We will call them sparse principal balances. We suggest a sequential procedure: The signs of the first sparse PC loadings are used in the usual way to form the first balance. Since for a balance we always need at least one positive and one negative sign, the penalty parameter \(\lambda\) is optimized not only according to the BIC, but also to include this condition. From subsequent sparse PCs we only use those signs for defining a balance which were not used for previously.

It turns out that for high-dimensional data this approach leads to a more efficient dimensions reduction—in terms of explained variance—than the algorithms of Pawlowsky-Glahn et al. (2011). Our method is illustrated with simulated and real high-dimensional data.

References


In ecotoxicology, dose-response models are developed to relate the variation in the effect on an exposed organism caused by changing the dose or concentration level of a particular compound. An appropriate choice of the model used to describe this relationship is of fundamental importance in the ecological risk assessment (Ritz, 2010), i.e. in the development of a new chemicals, or in risk assessment of existing chemicals. The magnitude of the response, i.e. probability of death or probability of inhibition in living organisms, is measured against the logarithm transformed doses or concentrations. The concentration-response curve is the result of the best-fit selected curve among different non-linear regression models, such as log-logistic, probit or Weibull models (Scholze et al., 2001). Some extrapolation factors, which correspond to certain concentration values associated to a fixed level of inhibition, could be estimated using the fitted model. In particular, the analysis is focused on estimation of the $EC_p$, concentrations that give $100p\%$ of the total achievable effect, $(0 < p < 1)$, i.e. the concentrating showing $p\%$ effect.

Since the response consists in percentage of inhibited organisms exposed to different levels of concentrations, a reformulation of the basic task is proposed by performing simplicial regression of the observed proportion of inhibitions to the log-concentration values using the log-ratio methodology (Aitchison, 1986; Egozcue et al., 2012). This approach, leading to an easy interpretable regression model, enables to preserve the relative scale of the response as well as to perform the consequent statistical inference like confidence and prediction bounds. Moreover, the non-linear regression problem is moved via the logratio approach to the case of standard linear regression model. The advances of the logratio approach will be compared to the standard regression modeling. An application to a real-world data set is also provided, consisting of toxicity data on chemicals related to the luminescent bacterium *Vibrio fischeri* (Vighi et al., 2009).

**References**


When supervising a nuclear power plant, the containment building is crucial. Its functions are guaranteeing structural integrity and avoiding leaks in case of accident. Both events are considered of high risk. Once a given over-pressure is registered inside the containment building, three possible outputs are considered: serviceability, breakdown, and collapse. The current aim is the study of vulnerability. The vulnerability of the containment building under over-pressure is described by the conditional probability of the three mentioned outputs.

The study consists of three steps: (a) modelling the containment building using a finite element software; (b) given an over-pressure, simulating uncertain parameters related to material constitutive equations in order to obtain the corresponding output; (c) performing a simplicial regression to get a meaningful vulnerability model. The simulation provides proportions of outputs under the over-pressure conditions. The vulnerability model can be obtained by a simplicial regression of those proportions, as a response variable, on the over-pressure, as explanatory variable.

Some difficulties appear in step (c). For any given over-pressure, some outputs are very improbable, producing zeros in the corresponding proportion. In order to reduce the presence of zero proportions, importance re-sampling is applied in the simulation. Importance re-sampling allows the simulation of favourable conditions to get the desired output. Then, importance of simulated data is taken into account when fitting the vulnerability model in step (c). The obtained vulnerability model is similar to previous results in nuclear power plant safety designs.
Modeling dependence between two or more variables is a common issue in statistical applications. The Pearson correlation coefficient is often used to measure dependence, although it only captures linear dependence. Other dependence coefficients, such as Kendall’s $\tau$ or Spearman’s $\rho$, among others (Schweizer and Wolff, 1981; Scarsini, 1984) model dependencies other than linear. The use of this other coefficients of dependence is directly linked to the use of copula functions to model the multivariate variables at hand (Nelsen, 1999; Genest and Favre, 2007). The use of copula functions allows to treat separately the marginal distributions of the variables and the dependence structure among them. Sklar’s theorem (Sklar, 1959), ensures that under continuity hypothesis, the bivariate (multivariate) distribution between two ($n$) variables can be expressed in a unique manner as a combination of the marginal distributions and their copula. Copulas seem then a useful tool to represent dependence between variables.

In a compositional framework, dealing with dependence is a key problem. In fact, the detection of spurious Pearson correlation can be considered as the spark that boosted the interest for compositional data and its suitable treatment. Some researchers may expect that Kendall’s $\tau$, Spearman’s $\rho$ or the copula between parts of a composition are not spurious thus suggesting a way to circumvent log-ratio treatment of compositional data. In this work, these dependence coefficients and copula functions are used to measure the dependence between parts of a composition and its subcompositions. It is shown that if the compositional structure of data is ignored, the use of these measures of dependence also presents great shortcomings, as they are subcompositionally incoherent and, consequently, they are spurious measures of dependence between parts of a composition.

References

Genest, C. and A.-C. Favre (2007). Everything you always wanted to know about copula modeling but were afraid to ask. Journal of Hydrologic Engineering 12(4), 347–368.


On the use of CODA methods for bacterial identification from proteomic mass spectra

J. PALAREA-ALBALADEJO¹, KEVIN McLEAN², FRANK WRIGHT¹ and DAVID G.E. SMITH²,³

¹Biomathematics & Statistics Scotland, Scotland, UK javier@bioss.ac.uk
²Moredun Proteomics Facility - Moredun Research Institute, Scotland, UK
³Institute of Infection, Immunity and Inflammation - College of Medical, Veterinary and Life Sciences - University of Glasgow, Scotland, UK

The complexity of the proteome demands high-throughput methods for acquiring useful information from biological material. Mass spectrometry (MS) using a matrix-assisted laser desorption/ionization (MALDI) ion source and a time-of-flight (TOF) detection system has become a powerful tool for fast extraction of proteomic profiles. MALDI-TOF mass spectra represent so-called intensity over a range of mass to charge (m/z) values, typically between 10.000 and 100.000 data points (Fig. 1). The masses are computed from the observed time of flight by a quadratic transformation. Ideally the intensity scale would reflect the amount of ions hitting the detector within a fixed time interval (mapped to a m/z value afterwards), but unfortunately there is substantial variability in the MALDI-TOF MS measurements. The process of plating samples is irreproducible and, among others, there are changes in laser intensity and fluctuations in the detector response that alter noise levels, baseline and peak intensities in a collection of spectra; even when the same sample is analysed repeatedly. In consequence, the correlation between raw signal intensity and protein abundance is poor and absolute peak values are useless. However, the relative signal intensities within a mass spectrum are remarkably stable. Proper analysis of mass spectrometry data necessarily involves several preprocessing stages, including normalisation of spectra (typically based on the total ion current) to allow comparison of different spectra by making the data independent of variation in protein abundances and detector sensitivities. These features suggest that compositional methods may be a suitable approach to analysing proteomic patterns obtained by MALDI-TOF MS.

![Figure 1. Example of MALDI-TOF MS proteomic profile.](image)

In this work we will explore the use of CODA methods for analysing MS data, and will discuss some challenges involving their application in such a context. This includes high-dimensionality issues and the existence of features not present in all spectra. In particular, we will focus in data analysis for the identification/differentiation of bacterial species of value to diagnostics.
Nutrient balance ionomics: case study with mango (Mangifera Indica)

S.-É. PARENT\textsuperscript{1}, L.E. PARENT\textsuperscript{1}, D.E. ROZANE\textsuperscript{2} and W. NATALE\textsuperscript{3}

\textsuperscript{1} ERSAM, Department of Soils and Agrifood Engineering, Université Laval, Québec, Canada, email: serge-etienne.parent.1@ulaval.ca
\textsuperscript{2} Departamento de Agronomia, Unesp, Universidade Estadual Paulista, Brazil
\textsuperscript{3} Departamento de Solos e Adubos, Unesp, Universidade Estadual Paulista, Brazil

A plant ionome is the compositional vector of nutrients and trace elements in plant tissues. Ionomes and soil nutrients are commonly diagnosed in agronomy using concentration and nutrient ratio ranges. However, both diagnoses are biased by redundancy, scale dependency and non-normal distribution inherent to compositional data, potentially leading to conflicting results and wrong inferences. Our objective was to present an unbiased statistical approach of plant nutrient diagnosis using a balance concept and mango (Mangifera indica) as test crop. Ionomes were represented by balances computed as isometric log ratios (ilr). We collected foliar samples at flowering stage in 175 ‘Tommy’, ‘Palmer’, ‘Espada’ and ‘Haden’ mango orchards in São Paulo state, Brazil. The ionomes comprised 11 nutrients (S, N, P, K, Ca, Mg, B, Cu, Zn, Mn, Fe). Soil fertility attributes (pH and bioavailable nutrients) were analyzed to reflect soil nutrient supply. Significant (p < 0.05) cultivar effect was confounded with soil effect in discriminant analysis and interpreted as high phenotypic plasticity of the species. A customized receiver operating characteristic (ROC) iterative procedure was developed to classify observation between true/false negative or positive and high/low-yielders. The ROC partitioning procedure showed that the critical Mahalanobis distance of 4.08 separating balanced from imbalanced specimens about yield cut-off of 128.5 kg fruit tree\textsuperscript{-1} proved to be a fairly informative test (area under curve = 0.84-0.92). Traditional multivariate methods were found to be numerically biased as shown by their deviation from the Mahalanobis distance of ilrs. We propose using a coherent pan balance diagnostic method with median ilr values of top yielders centered at fulcrums of a mobile and the critical Mahalanobis distance as a guide for global nutrient balance. Nutrient concentrations in weighing pans assist appreciating nutrients as relative shortage, adequacy or excess in balances.
The analysis of compositional data deals with relative information between parts. The total (abundances, mass, amount, ...) is in general not known, or not informative. Occasionally, interest lies in analysing a composition for which the total is known and of interest. Original data are in the real positive orthant, $\mathbb{R}^D_+$, which has a vector space structure (in fact a Euclidean structure) taking the logarithm of each component. The corresponding composition is obtained by closure of the original data and lies in the simplex $S^D$, where the Aitchison geometry is assumed to be the appropriate one. The total can be described by a positive function of the original components with a relative scale in $\mathbb{R}_+$. Jointly, the total function and the composition are elements of the product space $\mathbb{R}_+^D \times S^D = T$. In order to carry out statistical analyses in $T$, a Euclidean structure allows to use standard multivariate techniques. However, the vector space structure in $T$ should be compatible with the previously assumed vector space structure in $\mathbb{R}_+^D$ and $S^D$. Not any total function satisfies such compatibility. Particularly, the total, conceived as the sum of all original components, leads to incompatibilities. A typical inconsistency is found when computing the mean value in $T$ and $\mathbb{R}_+^D$. Compatibility conditions are fulfilled by total functions which are powers of the products of the original components, e.g. the geometric mean. In practice, multivariate techniques in $T$ such as MANOVA, principal component and cluster analyses would require to use a compatible total function in order to preserve the properties of the Aitchison geometry of the simplex. Suitable procedures are illustrated with real data about total abundances of phytoplankton in a river.
Characterization of Sediments of the Carnuntum Wine Growing Area, Austria, by Exploratory Analysis of Compositional Data

Heinz REITNER1, Andreas BAUMGARTEN2, Josef EITZINGER3, Thomas GERERSDORFER3, Johann GRAßL4, Maria HEINRICH1, Gerhard HOBIGER1, Wolfgang LAUBE3, Erwin MURER5, Herbert PIRKL6, Julia RABEDER1, Heide SPIEGEL2, Ingeborg WIMMER-FREY1

1Geological Survey of Austria, Neulinggasse 38, A-1030 Vienna, Austria, heinz.reitner@geologie.ac.at
2Austrian Agency for Health and Food Safety, Institute for Soil Health and Plant Nutrition, Spargelfeldstr. 191, A-1226 Vienna, Austria
3University of Natural Resources and Applied Life Sciences, Institute of Meteorology, Peter Jordan Str. 82, A-1190 Vienna, Austria
4Die Rubin Carnuntum Weingüter, Carnuntum Wine Region Cooperation, Fischamenderstr. 12/3, A-2460 Bruck an der Leitha, Austria
5Federal Agency for Water Management, Institute for Land and Water Management Research Pollnbergstraße 1, A-3252 Petzenkirchen, Austria
6Geological Office, Plenergasse 5/27, A-1180 Wien, Austria

An interdisciplinary study of the vineyards of the Austrian wine district Carnuntum was carried out over a three-year period. The study aimed at the investigation of the basic conditions of the viticulture by describing the physiogeographic properties of the region.

Climatology studies, soil mapping, geological compilation, mapping of the quaternary loess and loam cover of the region, hydrologic, geophysical and geomorphological investigations were performed. Extensive mapping of the thematic layers utilized a geographic information system (Reitner et al., 2011).

To evaluate the vertical and lateral variations of rocks and soils in the study area more than 200 samples taken by drilling or at sampling pits were analysed for grainsize distribution, clay mineral and bulk mineral content and whole rock geochemistry. As the analysed datasets were considered as compositional data (Aitchison, 1986 reprint 2003), exploratory data analysis of the sample data made use of statistical methods which are targeted at the requirements of closed data (Pawlowsky-Glahn et al., 2011). The parameters were compared for varying regional areas and lithostratigraphic units with descriptive statistics and graphs, making use of software tools for processing of compositional data (van den Boogaart et al., 2012, Templ et al., 2013, Thió-Henestrosa and Comas, 2011). The results of the exploratory data analysis contribute to the characterization of the various stratigraphic units and soil properties of the study region.

References


Challenges for CoDa in geochemical practice

C. REIMANN¹, P. FILZMOSER² and K. HRON³

¹Norges geologiske undersøkelse (NGU), Norway, clemens.reimann@ngu.no
²Department of Statistics and Probability Theory, Vienna University of Technology, Austria
³Department of Mathematical Analysis and Applications of Mathematics, Palacky University, Czech Republic

Geochemists almost exclusively work with compositional data and have to deal with large multi-element datasets. Although problems with correlation analysis for compositional data were recognized as early as 1897 (Pearson, 1897) and a comprehensive book on the special geometrical properties of compositional data was published in 1986 (Aitchison, 1986), the vast majority of practitioners continue to use statistical tools that are not ideally suited for the data at hand. This has probably several reasons: (1) geochemists are actually interested in the variation of the absolute concentrations of every single analysed element in space (i.e. in univariate data analysis), (2) classical data analysis appears to deliver interpretable results, (3) most geoscientists are not sufficiently trained in mathematics and statistics to understand the formalism behind log-ratio transformations, (4) geochemical data are of very varying quality and many people may be reluctant to mix good and bad data in a log-ratio transformation to then perform multivariate data analysis.

Data from a large geochemical mapping project covering a large part of Europe (the GEMAS project) and univariate regional distribution maps highlighting why geochemists are interested in absolute concentrations and univariate data analysis even when having a multivariate multi-element dataset at hands will be shown. Absolute concentrations are compared to relative information (clr-transformed variable). Results can be quite similar and many geochemists might argue „why then use more complicated techniques?“ In the end, both, absolute concentration maps and maps of the clr-transformed variables will provide important information. Entering multivariate data analysis (e.g. via a PCA) with absolute information (log-transformed variables) will often provide the expected results, a summary of the elements with high or low concentrations in certain parts of the survey area. Due to the effects of data closure mis-interpretations can easily happen. Using the openend data (log-ratio transformed variables) will provide the opportunity to detect new processes in the data.

To encourage geoscientists to use CoDa methods, convincing practical examples, including solutions of how to work in the univariate and bivariate case are needed.

References


Application of chemometric analysis in the classification and the confirmation of bio-char origin

M. SAJDAK and S. STELMACH

Institute for Chemical Processing of Coal, Poland, msajdak@ichpw.zabrze.pl

World primary energy demand in 2011 increased by 2.5% compared to the previous year. According to market analysts, global energy demand will increase by 30% over 2035 (DOE, 2012). Due to the rising prices of fossil fuels it is expected that renewable energy sources (RES) by 2035 will become the world's second largest source of electricity generation (BP, 2012). In Poland, the main source of energy is still hard coal, which in 2011 was used for production of 62% of total energy generated. In line with the global trend, also in Poland there is increasing interest in renewable energy sources, among which the most popular is biomass. In 2011 it was recorded the highest (13.7%) increase in the amount of energy production from renewable sources compared to the previous year (GUS, 2012). With increasing use of biomass for production of primary energy (by combustion or co-combustion with coal), more frequently problems with the physicochemical properties of biomass are observed, e.g.:

- wide range of moisture content causing the instability of the combustion process,
- different chemical composition of the ash compared to coal,
- low-density resulting in difficulties in transport and storage,
- low energy density,
- high content of volatile matter.

These problems may be eliminated (or at least limited) by biomass pre-processing into a fuels with greater stability, characterized by better physicochemical properties (e.g. calorific value). The methods of thermal biomass upgrading (such as torrefaction or pyrolysis) are well known. Using these processes it is possible to obtain a fuel with the quality parameters comparable with hard coal (e.g. high calorific value up to about 30MJ/kg). These parameters are dependent on the process conditions, as well as the properties of the material used. The key to wide application of bio-chars as alternative renewable fuels for energy generation is the type of material used in the thermal pre-processing. To consider bio-char as a renewable fuel it must be fully produced from biomass (it should contain no any additives). Unfortunately, bio-char or torrified biomass entering the market may be knowingly contaminated by chars obtained from coals or plastics pyrolysis. It is therefore necessary to build a suitable (and relatively cheap) method of analysis allowing for the correct classification of the bio-material subjected to energy production. Such a methods, based on chemometric analysis and data mining methods, are under elaboration in Institute for Chemical Processing of Coal (Zabrze, Poland). This paper presents the current results of application of chemometric analysis in the classification and the confirmation of bio-char origin.

References


Application of chemometric analysis for determination of contamination degree of materials obtained by thermal conversion of biomass.

M. SAJDAK and S. STELMACH
Institute for Chemical Processing of Coal, Poland, msajdak@ichpw.zabrze.pl

Biomass is the oldest and the third largest in the world source of renewable energy used nowadays (European Commission, 2000). Due to the rising price of fossil fuels it is expected that the renewable energy sources (RES), including biomass, become the world's second largest source of electricity generation by 2035. According to the report of International Energy Agency (IEA, 2012), the world's bio-energy sources are quite sufficient to ensure the supply of biomass and bio-fuels for energy purposes, without the need to compete with food production. The growing demand for biomass and products of its thermal conversion (e.g. torrefied biomass or bio-chars) carries a risk of knowingly raising the calorific value of natural renewable energy sources. For example, fossil fuels, polymers and wastes from furniture industry (which are not classified as biomass) can be added to that kind of green energy source. The latest fuels contain significant amounts of different resins (polyester, alkyd, polyurethane) and glues (e.g. on the basis of urea or polyvinyl acetate). Determination of the contents of such pollutants in bio-char used for energy production would be usually time-consuming and expensive, especially taking into account the fact that the most reliable method to confirm the origin of biomass is currently the analysis of the 14C content. In order to meet the requirements of energy producers on the quality of renewable fuels, combination of several chemometric methods with standard analysis of solid fuels (elemental analysis, heat of combustion, oxides content in the ash) for simple and fast verification of biomass and its thermal conversion products in terms of its purity has been proposed (Sajdak, 2012, Sajdak and Piotrowski, 2012). Chemometric methods were initially used in the classification of biomass (Tao at al., 2012 a, b), but their application to check the purity of the products of biomass thermal conversion bring a new challenge.

In the study performed the Principal Component Analysis combined the Classification and Regression Trees has been used. In addition, both methods are combined with regression analysis which allows for the quantitative determination of material contamination degree. The Principal Component Analysis task was to detect the variables that can significantly indicate the type of material being tested and describe the existing relationship between the variables and their groups. Classification and Regression Trees method made it possible to build a classification algorithm of studied materials, like torrefied biomass or bio-char. In the case of bio-material with artificial additives, regression analysis allows one to estimate contamination degree of the sample. The method described in the paper is currently under development in order to be optimized.

References


Robust variable selection in linear regression with compositional explanatory variables

F. SCHROEDER¹, A. BRAUMANN¹, P. FILZMOSER¹, K. HRON²

¹Department of Statistics and Probability Theory, Vienna University of Technology, Austria
²Department of Mathematical Analysis and Applications of Mathematics, Palacký University Olomouc, Czech Republic

This paper considers robust variable selection for a multiple linear regression model, where the regressors are compositional and the response is a real variable. As raw compositional data cannot be used directly in this setting, we use an approach based on the isometric log-ratio (ilr) transformation, see Hron et al. (2012). The special choice of the ilr transformation proposed in Hron et al. (2012) allows to test for the contribution of all the relative information concerning one compositional part to explaining the response. This idea is used for forward and backward stepwise selection algorithms. For model selection we propose the Mallows’ $C_p$ and the Akaike information criterion, as well as robust versions thereof, see Ronchetti (1985); Ronchetti and Staudte (1994). Simulation results confirm that the method can successfully select good models. In presence of outliers, classical model selection breaks down, while the robust methods still show excellent behavior. The approach was furthermore applied to a real data set.

References


Methods to detect outliers in compositional data with structural zeros

M. TEMPL, K. HRON, P. FILZMOSER, G.S. MONTI

1Department of Statistics and Probability Theory - Vienna University of Technology, Austria
matthias.templ@gmail.com
2Department of Mathematical Analysis and Applications of Mathematics - Palacký University, Czech Republic
3Department of Geoinformatics - Palacký University, Czech Republic
4Methods Unit, Statistics Austria
data-analyis OG, Austria
5Department of Economics, Quantitative Methods and Business Strategy - University of Milano-Bicocca, Italy

In compositional data analysis the ratios between parts are of interest. However, in many data sets structural zeros (values that are truly observed and are zero) are present in the data, sometimes with concentration of more than 50% of the values. Examples of structural zeros are plant species that are not able to survive in a given soil type or climate, a political party that has no candidates in a region, or teetotal households that do not have expenditures on alcohol and tobacco. The presence of structural zeros causes serious difficulties for the analysis and standard transformations, and methods from a log-ratio approach to compositional data analysis can not be used.

In this contribution we focus on outlier detection for compositional data and enhance the work from Filzmoser and Hron (2008) and Filzmoser et al. (2012) by allowing structural zeros in the data. We also employ the Mahalanobis distance approach for outlier detection, and thus one goal is the robust estimation of the covariance matrix.

Two new methods are proposed and compared with a method that uses zero replacement (Fry et al., 2000). The first method is based on first imputing the zeros. After this imputation step, the covariance is not derived directly from the imputed data set, but we use the isometric logratio transformation (Egozcue et al., 2003), concretely the particular one from Hron et al. (2010), to construct an isometric logratio representation for each zero pattern rearranging the parts according to their zero structure. In other words, to obtain a covariance estimate and Mahalanobis distances, only the information on the non-zero part of the data set is used. Observations with Mahalanobis distances higher than a suitable threshold are then flagged as outliers. The second approach is based on the variation matrix which is estimated pairwise and robustly from their non-zero (bivariate) parts. The covariance matrix is then obtained by multiplication of the variation matrix with an certain orthonormal basis.

In contrast to existing methodology (Fry et al., 2000), the imputation based method shows an excellent performance in real-world applications and simulation studies that have been carried out.

Our approach is not only suitable for outlier detection, it can be used for all covariance based methods like principal component analysis or discriminant analysis and is not only restricted to structural zeros - it can be applied also for data with missing values.

References


CoDaPack: still more userfriendly?

S. THIÓ-HENESTROSA and M. COMAS-CUFÍ

1Departament d’informàtica, matemàtica aplicada i estadística- Universitat de Girona, Spain,
mallto:santiago.thio@udg.edu
2 Departament d’informàtica, matemàtica aplicada i estadistica- Universitat de Girona, Spain

From the early times of CoDaPack it has intended to be a software of Compositional Data with an easy and intuitive way of use. On last CoDaWork (Comas and Thió-Henestrosa, 2001) we introduced the new version of CoDaPack made in Java that runs on windows and all unix based plataforms.

CoDaPack 2.0 main window (Figure 1) has four main areas: Menus, variables, alphanumerical results and data. Also graphical outputs appear in independent windows.

![CoDaPack 2.0 structure](image)

Figure 1. CoDaPack 2.0 structure: Menus on very top, variables on left side, alphanumerical results on right top and data on left bottom. Graphical outputs appear in independent windows.

The data part contains variables organized in columns. The window has an Excel like appearance but until now it can’t be edited. The new version of CoDaPack to be presented on CoDaWork’13 will incorporate edition features as: 1) to modify values of cells, 2) to delete variables and 3) to add new variables.

This new version of CoDaPack also incorporates new customize capabilities of the graphical outputs like change of colour, width and shape of the different lines, points and legends of the graph. Also with the new version using the mouse it is possible to identify which observation corresponds to a single point of the graphs. All under the philosophy of what you see is what you get.

Finally CaDaPack has a new capability. The user could draw straight lines and ellipses inside the Ternary Diagram.

References

A 2 × 2 table that describes joint distribution between two (row and column) factors can be considered as a special case of four-part compositional data (Aitchison, 1986), if its cells (represented by positive values) contain quantitatively expressed relative contributions on a whole, carrying exclusive relative information (Egozcue et al., 2008). The nature of that tables can be expressed by the Aitchison geometry. In particular, for those tables any normalization of cells (like normalization to probabilities with cell sum equal 1) should not alter the only relevant information, formed by ratios between cells. It enables to analyze relations between the corresponding row/column factors through decomposition of the original table to its independent and interaction parts. Independence analysis is also in focus when using log-linear and correspondence analysis. However, these methods do not reflect the structural characteristics of the joint distribution as the marginals (obtained summing columns or rows of the table) or conditionals (formed by single row/columns of the table).

Although a general theory for structural analysis of compositional tables on the simplex exists (Egozcue et al., 2008), the 2 × 2 case enables to simplify some key ideas and explain illustratively key features of the model. Moreover, in the sample case it is possible to visualize the real data structure as well as to perform statistical inference like hypotheses testing, if the decomposed tables are expressed using a proper choice of balances - the orthonormal coordinates with respect to the Aitchison geometry (see Egozcue and Pawlowsky-Glahn, 2005).

In reality, the degree of globalization draws our attention to applications of the proposed configuration to real world problems. The increasing trend of globalization has deepened the interaction degree among countries, which results in a country’s increasing share of economic flow originated from internationals. The decomposition of economic flow into independent and interaction parts would be able to trace the economic linkages between domestic and international perspective for a country, which in return will provide an insight into the cause and effect of globalization. To perform this analysis we use the recently developed and publicly available World Input-Output Database (Timmer et al., 2012) which covers 27 EU countries and 13 other major countries, as well as an additional region - the rest of the world, for the period 1995 to 2009. The computations are performed using the R package rwiot which will be soon available at CRAN.

References


Regression between compositional data sets

R. TOLOSANA-DELGADO¹ and K.G. VAN DEN BOOGAART¹,²

¹Department of Modelling and Valuation - Helmholtz Institute Freiberg for Resources Technology, Germany
²Institute for Stochastics - Technical University Bergakademie Freiberg, Germany
r.tolosana@hzdr.de

Linear regression is one of the grounding techniques of statistics, aimed at detecting and quantifying the degree of dependence between two (sets of) variables, one considered explanatory and the other considered as dependent or explained. In this contribution we focus on the case where both the dependent and independent variables are compositions. Thus a compositional data set is explained as a function of another compositional data set. This contribution presents graphical and quantitative tools to visualize and assess the degree of dependence, and to isolate that dependence to certain subcompositions. This is illustrated with two examples, one from mining, and one linking educational level to political preferences.

The logratio approach and the principle of working on coordinates tell us that a model may be estimated in the following steps:

1. transform both data set to logratios, each with its own alr or ilr transformation;
2. build a multiple multivariate linear regression model between the two sets of logratios; and
3. back-transform predicted logratios (or estimated coefficients) to predicted compositions (or estimated coefficient compositions) with the adequate inverse ilr/alr.

Let us denote with $X = [x_{ni}]$ and $Y = [y_{nj}]$ the two data sets, with $n = 1, \ldots, N$ observations and respectively $i = 1, \ldots, P$ and $j = 1, \ldots, Q$ components. We work then with a model

$$\text{ilr}_y(Y) = 1_N \cdot b_0 + \text{ilr}_x(X) \cdot B,$$

Assuming that the data are jointly normally distributed, classical multivariate regression methods provide in this way estimates of each coefficient $\hat{B}$, as well as their variance-covariance matrix $\hat{\Sigma}_B$.

The resulting estimates of the coefficients $\hat{B}$ can be backtransformed to the compositional space and interpreted in a compositional context, since the results of the multivariate regression are in fact representations of compositional objects (a homomorphism) in the chosen ilr or alr bases.

The estimation variance matrix $\hat{\Sigma}_B$ can also be seen as a particular representation of an object with an intrinsic compositional nature. It allows to compute compositional confidence areas for the parameters and a compositional predictive areas for model predictions. These regions can be plotted as compositional ellipses in various standard plots for compositional data (e.g. ternary diagrams, clr-scatterplot, etc.). Via a singular value decomposition of the regression coefficient matrix is also possible to identify pairs of projection directions on the simplex for the dependent and the simplex for the independent variable showing an univariate regression dependence. A Scheffé-type confidence region can be used to test simultaneously which of these pairs show significant influences and to test for a priori suspected directions of dependence. Note that all these results are not dependent on the logratio basis actually used to obtain the coefficients or the variance matrix.
Compositional Regression with unobserved components or below detection limit values

K.G. VAN DEN BOOGAART1,2, R. TOLOSANA-DELGADO1, K. HRON3, M. TEMPL4 and P. FILZMOSER4

1Department of Modelling and Valuation - Helmholtz Institute Freiberg for Resources Technology, Germany
2Institute for Stochastics - Technical University Bergakademie Freiberg, Germany
3Faculty of Science, Palacký University Olomouc, Czech Republic
4Department of Statistics and Probability Theory; Vienna University of Technology, Vienna, Austria
boogaart@hzdr.de

The regression of a compositional variable against a set of covariables is an important tool to analyse the dependence of compositions to external influences. It can essentially be treated with the principle of working in coordinates, by using a multivariate regression model for the additive or isometric log ratio transform, using available standard software (e.g. R). This approach is based on Maximum Likelihood (ML).

A problem arises when some of the components in the composition are missing, since missings in compositions do not simply correspond to missing values in the logratio transform. This renders compositional regression unavailable for many existing datasets, like e.g. containing incomplete analysis data or very low concentrations in individual components. The state-of-the-art way out of this is to impute each missing with some value prior to its analysis. However, if the imputation is unaware of the dependence, then estimates will be substantially influenced by the imputation rule.

We propose a more rigorous approach, aimed at including in the ML procedure the information that some values were missing. The two following improvements are grounded on assumptions of joint additive logistic normality for the compositional dataset.

For values missing completely at random, like e.g. perhaps in the case of unobserved components, it is possible to restrict the likelihood to the observed components by replacing the observed ilr by a projection onto the ilr subspace linked to the observed components. This leads to likelihoods and ML-equations which are slightly more complicated but still recognizable as such.

Observations below detection limit (BDL) are more complicated. Based on on the meaning of the detection limit from van den Boogaart et al. (2011) we developed a Bayesian model for this situation, which can be used to estimated the parameters in an MCMC procedure. The key idea is that the distribution of a datum with a BDL is not anymore a "pure" additive logistic normal, but it is polluted with some additive noise. The resulting distribution is the one that must be taken into account in the likelihood calculations.

Combining both strategies as needed, we can fit compositional regression models efficiently in case of missing data. The approach also provides a complete "imputed" dataset, distributed according to the true conditional distribution (given all the data) of the unobserved true compositions, filtering the measurement error out.

References

Local regression for compositional data

C. VENIERI\textsuperscript{1} and M. DI MARZIO\textsuperscript{1} and A. PANZERA\textsuperscript{1}

\textsuperscript{1}DMQTE - University “d’Annunzio” of Chieti-Pescara, Italy c.venieri@unich.it

Regression for compositional data has been considered only from a parametric point of view. We introduce a nonparametric estimator for the regression function when both the response variable and the predictor can be defined on the simplex, and derive its asymptotic properties. To this end, we use the isometric log-ratio transformation along with properly defined kernel functions. The performances of our estimator are compared to those of a parametric model using a real data set.
Monte Carlo comparison of mean with median and standard deviation with median absolute deviation for statistically contaminated samples of size 5-20 and application for handling trace element concentrations in international geochemical reference materials

SURENDRA P. VERMA1, LORENA DÍAZ-GONZÁLEZ2, and JOSÉ RAÚL GARCÍA-GILES3

1 Departamento de Sistemas Energéticos, Centro de Investigación en Energía, Universidad Nacional Autónoma de México, Priv. Xochicalco s/no., Col. Centro, Apdo. 34, Temixco 62580, Mexico, spv@cie.unam.mx

2 Departamento de Computación, 3 Licenciatura en Ciencias, Facultad de Ciencias, Universidad Autónoma de Estado de Morelos, Av. Universidad 1001, Chamilpa, Cuernavaca, Mor., 62209, Mexico.

Following our precise and accurate procedure (Verma and Quiroz-Ruiz, 2006), Monte Carlo simulations of 10,000,000 replications for sample sizes of five to twenty were used to objectively compare two outlier-based with two robust statistical parameters.

Our simulations show that both mean and standard deviation must be computed only after the application of discordancy tests (Barnett and Lewis, 1994; Verma, 1997) along with the newly simulated precise and accurate critical values, which is fully consistent with the statistical theory (viz., mean and standard deviation are highly affected by discordant outliers).

However, we also observed for the first time an unexpected result that mean and standard deviation (outlier-based estimators) after proper application of discordancy tests generally provide a better estimate of the population than median and MAD (robust estimators), respectively. A measurable parameter $\tau$ (Normalized distance of the outlier from the censored mean) was formulated for interpretations. Complex $\tau$ dependent correction factors are required for any given sample size rather than the constant correction factors envisioned thus far in the literature. Therefore, further simulation work to cover more complex asymmetric statistical contamination and larger sample sizes is urgently needed.

We emphasize that this is the first time that any one has carried out precise and accurate simulation work along with the application of discordancy tests to achieve an objective comparison of four frequently used statistical parameters (two robust and two outlier-based). Reasons for the better performance of the outlier-based methods as compared to the robust methods are also suggested.

We then applied these simulation results and recommendations for correctly handling compositional data (Aitchison, 1986; Egozcue et al., 2003) to selected cases of international geochemical reference materials. Finally, an objective comparison of crude compositional data with additive as well as isometric log-ratio transformations is also reported.

References


Interpretation of out-of-control signals in a compositional $T^2$ control chart

M. VIVES-MESTRES$^1$, J. DAUNIS-I-ESTADELLA$^1$ and J. A. MARTÍN-FERNÁNDEZ$^1$

$^1$Department of Computer Science, Applied Mathematics and Statistics – University of Girona, Spain, josepantoni.martin@udg.edu

One of the main goals in Statistical Process Control (SPC) is the detection of special-cause variation in order to control and monitor a process. This is achieved by using a statistical control chart and interpreting the output. When a signal is detected, the causes of the anomaly must be identified in order to apply appropriate remedial measures.

In industry it is common to measure process performance by a set of interrelated continuous variables. In those cases, a widely used charting statistic is the Hotelling’s $T^2$ which takes into account both the univariate and the interrelationship effects between variables. One of the drawbacks with multivariate process monitoring is in interpreting signals because in addition to changes in location and variation, there may also be changes in the interrelationships between variables. One efficient procedure for interpreting signals in a $T^2$ control chart is by decomposing the $T^2$ statistic into orthogonal components directly interpretable (Mason et al., 1995), known as MYT decomposition.

Typical $T^2$ control chart is not suitable for compositional data (CoDa) due to the special geometry of the simplex. A new control chart for compositional data ($T^2_c$ control chart) has been proposed by Vives-Mestres et al. (submitted) which is plenty consistent with CoDa nature. It is based on the log-ratio methodology and it is shown to be equivalent to applying the typical methodology to the compositions expressed in coordinates with respect to an orthonormal basis.

The difficulty arises again when the causes of the signal in the $T^2_c$ control chart have to be identified. In this work, we present a method for interpreting such signals based on the MYT decomposition method.

References


Gas chromatography coupled with Mass spectrometry (GC-MS) is routinely used in many areas of chemistry and biology. Analysis of raw GC-MS data requires several successive steps, the first one being preprocessing. One of the objectives of data preprocessing is to remove experimental, inter-sample variations which usually obscure the biologically meaningful changes and/or lower the statistical power. This is traditionally achieved by normalizing areas under curves of the individual peaks in the chromatogram to the total area. The individual intensities (i.e. peaks) are divided by the sum of all the measured intensities. This type of normalization is generally referred to as “integral normalization”. This approach is valid for data with certain simple statistical structure but it is not sure if this approach is valid and efficient for general use. It has been already shown that for Nuclear Magnetic Resonance spectra integral normalization has several drawbacks. For instance minor perturbation in high abundant compounds affects the scaling of others regardless of their actual amount. It ought to be pointed out that normalization changes the variance in the data. In data analysis, variance is often considered as an approximation of information; therefore one could state that the normalization affects the information contained in the chromatograms. If the information of interest is located in low abundant compounds, total area normalization might change the correlation with Y vector (i.e. class vector).

Using different datasets of breath samples delivered from GC-MS analysis we tested the histogram normalization, integral normalization and robust orthogonal regression, and compare their performance. To do so, two real examples are used. The first dataset come from research on patients suffering from Inflammatory Bowel Disease. The second study used here consists of elderly individuals exposed to air pollution from dense traffic. By comparing the different normalization methods we will show that integral normalization might negatively influence the final outcomes.
Robust multivariate regression with compositional data

J. ZEHETGRUBER\textsuperscript{1}, P. FILZMOSER\textsuperscript{1}, K. HRON\textsuperscript{2}, M. TEMPL\textsuperscript{1,3}, R. TOLOSANA-DELGADO\textsuperscript{4} and K.G. VAN DEN BOOGAART\textsuperscript{4,5}

\textsuperscript{1}Department of Statistics and Probability Theory, Vienna University of Technology, Austria
\textsuperscript{2}Department of Mathematical Analysis and Applications of Mathematics, Palacký University Olomouc, Czech Republic
\textsuperscript{3}Statistics Austria, Austria
\textsuperscript{4}Department of Modelling and Valuation - Helmholtz Institute Freiberg for Resources Technology, Germany
\textsuperscript{5}Institute for Stochastics - Technical University Bergakademie Freiberg, Germany

Linear regression analysis belongs to the most important tools in statistics for analysing the relations between two sets of variables. In the multivariate context, both the explanatory and the response part consist of multivariate observations. In the following we will focus on this situation. Moreover, we will consider the cases, where the responses are compositional data and the explanatory variables are non-compositional data, as well as on the reverse situation.

Traditionally, the estimation of the regression coefficients is based on the least-squares method. In presence of outliers it is well known that the least-squares estimation may lead to distorted results, being heavily influenced by the outlying observations. There are two approaches to robust multivariate linear regression: Rousseeuw et al. (2004) suggested to robustly estimate the joint covariance matrix of the explanatory and response variables, and to use this for constructing a robust equivalent to the least-squares solution. The method of Agulló et al. (2008) is a more direct approach, it generalises the Least Trimmed Squares (LTS) estimator (Rousseeuw, 1984) from the multiple to the multivariate regression case.

We employ the isometric logratio transformation for the compositional part in the regression model, and perform robust multivariate regression according to both robust approaches. We will discuss advantages and disadvantages of the methods, also with respect to inference statistics and interpretation of the results. Finally, the usefulness of the methods is demonstrated using real-world data.

References


GENERAL INFORMATION ABOUT VORAU

Geographical Location

Vorau is located in the northeastern region of Steiermark (Styria), the so-called “green heart” of Austria. It is surrounded by beautiful scenery, with mountains (up to 1700 m), alps, and forests. The hilly landscape changes into the Hungarian lowland a few kilometers to the east. Graz, the capital of Steiermark, is approximately 80 km south of Vorau, and Vienna, the capital of Austria, is 120 km north of Vorau. The Hungarian border is 60 km east of Vorau.

Climate

The weather can be unstable with temperatures of about 10-25°C. At night, the temperatures cool to 5-10°C. Rainy weather for several days is possible, although this region belongs to the areas in Austria with the most days of sunshine.

Healthy Region

Vorau is far away from industry, traffic, and noise pollution. The unique and charming scenery around Vorau is inviting for relaxation and recovery. Several villages in the neighbourhood call themselves “strength-giving-villages”. The balancing of body, soul, and spirit in a harmonious environment should give new strength for everyday life. Vorau and the neighbouring communities are integrated in the “healthy region Vorau”. Most of the farmers in this region have biological cultivation and keep animals in a natural way. They produce all basic foods such as milk products, sausages, and bread.

The Village

Vorau is a small village with about 1500 inhabitants. At the moment, 60 small enterprises provide some jobs and short distance supply. However, about 70% of the working population have to commute, a lot of them even to Graz or Vienna.

The Augustinian monastery is definitely the cultural centre of the village. But the village has charm as well, with its old houses, the parish hall (more than 400 years old), and the parish church St. Agidius (a Romanesque building).

The Abbey

The Augustinian monastery of Vorau was founded in the year 1163 by Margrave Ottokar III for gratitude of the birth of a long-desired son. In this year, the Abbey of Vorau is thus celebrating its 850 year’s anniversary. The Abbey is well-known for its valuable handwritings of the 12th century and the frescos of Johann Cyriak Hackhofer.

The library contains more than 18,000 books, and another 22,000 volumes are kept in fire proof safes. Vorau has 416 parchments and 215 incunabula, that are prints made before 1450. The baroque church which was built between 1660 and 1662, is decorated with many frescos and statues. The famous mural painting in the sacristy is certainly worth seeing. The mural is titled “The Last Judgement” and originates from the Styrian painter Hackhofer. There are still far more objects of interest which can be discovered during a guided tour through the Abbey.

The Abbey consists of many buildings which are used for different purposes. Besides the ecclesiastical use, there are private appartments, a domestic science school, certainly a tavern, and a large part of the Abbey is used by the education center. A picturesque court and lecture halls in baroque style make the Abbey of Vorau one of the most beautiful in Austria.
Accommodation
Vorau has several guest-houses and private boarding-houses. Most of the rooms are situated in the Abbey. All rooms are very cheerful and are furnished with a toilet and a shower. Of course, the comfort of a 5-star congress hotel should not be expected. The luxury which is probably missing, will surely be compensated by the very kind and informal atmosphere.

The payment of the hotel costs has to be done directly by the workshop participants (except for keynote speakers). Payment with credit card is possible.

Leisure Offers
The beautiful environment entices the visitor into hiking. Close to the center there are two tennis-courts and a heated public swimming pool. 5km away from Vorau there is the possibility to go horseback riding. A few meters from the Abbey one finds an open-air museum of traditional handicraft and ancient farmhouses of this region.

Shopping
The majority of all shops are for food and clothing. Most of the essential things can be purchased. Attention: There are no computer shops in Vorau.

Banking
Changing money is possible in each of the four banks. The bancomats accept the usual credit cards.

Medical Care
Vorau hosts a general hospital with 100 beds. Furthermore, there are three physicians, a dentist, a pharmacy and the red cross organization.

Security
Vorau is very safe, like most of Austria.

Public Transportation
Vorau is quite isolated from public transportation. The nearest train station is in Rohrbach (12km from Vorau). In the same village there is a fast (but irregular) bus connection to Vienna. Vorau has two small taxi companies. During the workshop there will also be the opportunity to rent a car (reservation is requested).
The public bus and train schedule is available in the conference office.

Other Services
The workshop participants have free internet access in the Education Center.
Fax transmission and a copy service is available at the office of the Education Center (1st floor).
TELEPHONE NUMBERS

<table>
<thead>
<tr>
<th>SOS:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fire department</td>
<td>122</td>
</tr>
<tr>
<td>Rural police</td>
<td>133 (without dialing code)</td>
</tr>
<tr>
<td>Emergency service</td>
<td>144</td>
</tr>
</tbody>
</table>

Poisoning central office: 01/4064343-0

The following numbers are phone numbers in Vorau (except numbers of mobile phones, starting with 06**). Use the dialing code 03337 if you call from another place in Austria. For dialing from outside of Austria use +43 3337 or +43 6** to mobile phones.

SOS:
- Rural police: 2233
- Emergency service: 2244
- Hospital: 2254, 2255

Physicians:
- Dr. Kryza-Gersch: 2371
- Dr. Lux: 2170
- Dr. Sauer: 3125

Dentists:
- Albrechtsberger: 2212
- Dr. Sauer: 3124

Pharmacy:
- Apotheke: 4150
- Mobile: 0664/4452080

Taxi:
- Pötz Ewald: 2322 or 0664/1205979
- Pötz Josef: 2286 or 0664/3554885
- Vienna airport taxi: 01/1731 (~Euro 50,- airport–city)

Accommodation and Conference Site:
- Education Center: 2815-0
- Abbey (doorman): 2351-31
- Gasthof Kutscherwirt: 2211
- Gasthof Vorauerhof: 2329

People with local knowledge:
- Filzmoser Peter: 0650/3702177 or 2155
- Matthias Templ: 0660/1220236
- Education Center (Abbey): 2815-0

In case of any problem, please contact Peter Filzmoser: +43 650 3702177
SOCIAL PROGRAM

June 4, 18.10: Guided tour through the monastery
The monastery of Vorau is celebrating its 850th year anniversary. The church has been renovated now since more than two years. The new altar and a new organ are consecrated just on June 2, 2013. The monastery can be considered as a real “jewel” in our country.  
Meeting place: in front of the library, ground floor

June 5, 18.15: Walk to Mostschank Kuchlbauer
We will walk about 40 minutes to a “Mostschank”, a farm offering their own products. A shuttle bus is available on request.  
Meeting-place: in front of the main entrance of the monastery, or on the way through the market place of the village

June 6, 13.30: Excursion
We will take a bus to the nearby village Pöllauberg, famous for its wonderful floral decoration, but also for its pilgrimage church. Then we walk down through scenic nature to visit a distillery of Austrians famous ”Schnaps”. We then continue to the next village Pöllau to visit the famous market place and the impressive church.  
Meeting-place: in front of the main entrance of the monastery

June 6, 19.15: Conference Dinner
We will walk 250m to the restaurant Kutscherwirt in the village. Traditional food will be offered. Everybody is welcome to contribute with music and singing.  
Meeting-place: in front of the main entrance of the monastery

The expenses of the social program are included in the conference fee. We hope to see all of you at these events!
## TIME TABLE

### Sunday, June 2, 2013:

- **19.30:** Dinner (Stiftstaverne)
- **20.30–21.30:** Registration

### Monday, June 3, 2013:

- **9.00–18.00:** Introductory Course
- **19.00:** Dinner (Stiftstaverne)
- **20.00–21.30:** Registration

### Tuesday, June 4, 2013:

- **9.00–13.00:** Introductory Course
- **13.00–15.00:** Lunch (Stiftstaverne)
- **14.00–15.00:** Registration
- **15.00–18.00:** CoDaWork
- **18.10–19.15:** Guided tour through the monastery
  *(meeting-place in front of the library at ground floor)*
- **19.15–20.30:** Dinner (Stiftstaverne)
- **20.30–21.30:** CoDa and R

### Wednesday, June 5, 2013:

- **9.00–12.40:** CoDaWork
- **12.40–14.00:** Lunch (Stiftstaverne)
- **14.00–18.10:** CoDaWork
  - **18.30:** Walk to Mostschank Kuchlbauer (Dinner)
    *(meeting point at the main entrance of monastery)*

### Thursday, June 6, 2013:

- **9.00–12.40:** CoDaWork
- **12.40–13.30:** Lunch (Stiftstaverne)
- **13.30–18.30:** Excursion
  *(meeting point at the main entrance of monastery)*
  - **19.15:** Conference Dinner (Kutscherwirt)
    *(meeting point at the main entrance of monastery)*

### Friday, June 7, 2013:

- **9.00–12.20:** CoDaWork
- **12.20–13.40:** Lunch (Stiftstaverne)
- **14.00:** Bus transfer to Vienna
  *(in front of the main entrance of monastery)*