8th International Conference on
PARTIAL LEAST SQUARES
AND RELATED METHODS
26-28 May 2014 - Paris, France

BOOK OF ABSTRACTS
The 8th International Conference on Partial Least Squares and Related Methods (PLS’14) takes place in Paris (France) on May 26th-28th, 2014. PLS’14 follows the path started 15 years ago in Jouy-en-Josas (PLS’99, France), then in Anacapri (PLS’01, Italy), Lisbon (PLS’03, Portugal), Barcelona (PLS’05, Spain), Ås (PLS’07, Norway), Beijing (PLS’09, People’s Republic of China) and Houston (PLS’12, USA).

PLS’14 is jointly organized by ESSEC Business School and the Conservatoire National des Arts et Métiers (CNAM). It is organized with the scientific sponsorship of two sections of the International Statistical Institute (ISI), namely the International Society for Business and Industrial Statistics (ISBIS) and the International Association for Statistical Computing (IASC), The International Federation of Classification Societies (IFCS), the Société Française de Statistique (SFdS), the Società Italiana di Statistica (SIS), the Association Française du Marketing (AFM) and the Association Internationale Francophone d’Extraction et de Gestion des Connaissances (EGC).

The PLS’14 Organizing Committee members are grateful to all the sponsors who provide financial support and thus make this event possible, namely La Région Île-de-France, Addinsoft-XLSTAT, Camo-The Unscrambler, Ipsos France, JMP, Médiamétrie and Sigma Plus. A word of thanks also goes to Springer and Chapman & Hall publishing houses for their support and material.

PLS’14 is a unique event where the most outstanding academic researchers and expert professionals in PLS and related methods come from different countries and regions to meet, communicate and debate in the heart of Paris!

During the three days, the conference delegates present the state-of-art on the PLS methods as well as the most recent developments - conceptual, methodological or mathematical - as well as the most challenging advances demanded for the future. The PLS-based techniques are compared to other methods and several applications to real data are shown in different disciplines.

This eighth edition focuses on original methodological contribution in both PLS Regression, PLS Path Modeling and their related methods with application in areas and disciplines related to Management, Social Sciences, Chemometrics, Sensory Analysis, Industry and Life Sciences including Genomics.

The scientific programme of PLS’14 boasts 3 keynote speakers, 11 organized specialized sessions, 1 round table, 12 sessions of contributed papers as well as a rich poster session and a session in Honor of Carlo Lauro, thus totalling more than 100 presentations!

The keynote speakers talk about accuracy estimation for PLS and related methods via resampling-based procedures, inhomogeneous large-scale data, thinking multi-block.

The specialized sessions cover the following themes: advances in regression modeling, a focus on the use of PLS in Marketing in France (organized by the Association Française du Marketing), applied latent variable models, Genomics and brain imaging, high-dimensional models for genomic prediction, Kronecker product models, multi-group analysis, multi-way component analysis, Partial Least Squares Path Modeling applications in Information Systems research, recent advances on functional PLS and applications, and recent developments in Generalized Structured Component Analysis. The round table focuses on “Formative versus reflective measurement: fallacy or reality.”
The contributed papers are organized in sessions around the following topics: methodological extensions, multi-group analysis, measurement scale heterogeneity and model testing in PLS-PM, multivariate models for longitudinal data, PLS-PM and PLS regression applications in Marketing and Customer Relationship Management, Industry, Health Sciences, Genetics, Ecology and Spectroscopy, multi-block models for classification, methodological and mathematical contributions to PLS regression, variable selection and sparse PLS regression.

The work done by the Scientific Programme Committee, together with the qualified effort of the Authors of both keynote lectures, specialized sessions and contributed papers, ensure a high scientific level to PLS’14 that constitutes a rich and interesting balance between state-of-art and up-to-date research achievements on PLS and related methods. All delegates are invited to witness recent trends and applications in this field but also to put forward the most challenging theoretical and practical advances demanded for the future.

The attendance to PLS’14 shows a high number of participants (about 150) coming from more than 15 different countries (almost all European countries are represented as well as very far countries such as Korea, China, Thailand, Japan, USA and Canada) and including people from academy, industry and private companies operating in several fields.

Finally, we want to acknowledge the huge work of Organizing Committee members whose enthusiasm, professionalism and abnegation have created the most favourable atmosphere for scientific and social exchanges and actually made PLS’14 possible.

To them, as well as to all the sponsors that have financially and technically supported PLS’14 goes our deepest gratitude and that of the PLS’14 participants.

Paris, May 2014.

Vincenzo Esposito Vinzi
Gilbert Saporta
PLS’14 Conference co-chairs
Contents

Scientific Programme .............................................................................................................................................. 8

Keynotes .................................................................................................................................................................. 11

• Anne-Laure Boulesteix
  Accuracy estimation for PLS and related methods via resampling-based procedures ................................. 13

• Peter Bühlmann
  Inhomogeneous large-scale data: maximin effects ................................................................................................. 15

• Mohamed Hanafi and Gaston Sanchez
  Thinking Multiblock ............................................................................................................................................... 17

Oral presentations ................................................................................................................................................ 19

• M. Carmen Aguilera-Morillo and Ana M. Aguilera
  Using PLS regression in the noisy functional data context .................................................................................. 21

• Tomas Aluja, Giuseppe Lamberti and Antonio Ciampi
  Extension to the PATHMOX approach to detect which constructs differentiate segments and to test factor invariance: application to Mental Health Data .............................................................. 23

• Simona Balbi, Vincenzo Esposito Vinzi, Giuseppe Giordano, Franco Mola, Francesco Palumbo, Germana Scepi, Roberta Siciliano, Rosanna Verde
  Non-Symmetrical Data Analysis: the Neapolitan School ...................................................................................... 25

• Derek Beaton, Anne Bernard, Herve Abdi and Gilbert Saporta
  An integration of partial least squares and sparsified multiple correspondence analysis applied to genetic data .................................................................................................................................................. 27

• Derek Beaton, Joseph Dunlop, Francesca Filbey and Hervé Abdi
  CONCATPLS: PLSC for one continuous data set and one categorical data set .................................................. 29

• Alessandra Biancolillo, Ingrid Mäge and Tormod Næs
  Multiblock Classification: SO-PLS-DA .............................................................................................................. 31

• Silvia Bianconcini, Silvia Cagnone and Dimitris Rizopoulos
  Approximate likelihood inference in latent variable models with application to the Health and Retirement Study .................................................................................................................................................. 33

• Carole Binard
  Study of the variance term in the quadratic risk of the PLS1 estimator ............................................................. 35

• Marianna Bolla
  General latent variable models of PLS enhanced with spectral techniques and representations in Hilbert spaces .................................................................................................................................................. 37

• Stephanie Bougeard, Aida Eslami, El Mostafa Qannari and Stephanie Legleye
  Multigroup analyses to study Cannabis Abuse Screening Test ............................................................................. 39

• Mohamed Amine Bouhlel, Nathalie Bartoli, Joseph Mortier and Abdelkader Otsmane
  Surrogate models for optimization in high dimension using a mixed Kriging/PLS method .............................. 41

• Anne-Laure Boulesteix
  A regression-based approach to relate datasets’ characteristics to relative prediction accuracy of algorithms, with applications to PLS for omics data .................................................................................................................................................. 43

• Xavier Bry, Catherine Trottier, Frederic Mortier and Thomas Verron
  Extending SCGLR to multiple regressor-groups: the Theme-SCGLR method .................................................. 45

• Silvia Cagnone and Cinzia Vircoli
  The analysis of heterogeneity and cognitive structure of dementia by means of a factor mixture model .......... 47

• Furio Camillo and Valentina Adorno
  PLS-PM in opinion surveys when respondent minds generate size-effect (involvement axis) in the data .......... 49

• Gabriele Cantaluppi and Giuseppe Boari
  A partial least squares algorithm handling ordinal variables ............................................................................... 51
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal Covariates Clusterwise Regression</td>
<td>53</td>
</tr>
<tr>
<td>PLS Path Modelling of the molecular origins of gene expression noise</td>
<td>55</td>
</tr>
<tr>
<td>Time-Variant Effect on PLS Regression Models of Complex Industrial Systems</td>
<td>57</td>
</tr>
<tr>
<td>Using PLS for Variable Selection: A Simulation Study</td>
<td>65</td>
</tr>
<tr>
<td>Estimating and correcting optimism bias in multivariate PLS regression: application to the study of the association between Single Nucleotide Polymorphisms (SNPs) and multivariate traits in Attention Deficit Hyperactivity Disorder (ADHD)</td>
<td>67</td>
</tr>
<tr>
<td>Quantile PLS Path Modeling</td>
<td>69</td>
</tr>
<tr>
<td>A Markov-switching regression model with non Gaussian errors for investigating stock market behavior</td>
<td>71</td>
</tr>
<tr>
<td>Identifying common and distinctive processes underlying multiset data</td>
<td>73</td>
</tr>
<tr>
<td>Component-based Redundancy Path Modelling</td>
<td>75</td>
</tr>
<tr>
<td>Component-based Redundancy Path Modelling, Carlo Lauro and Alfred Olivier Hero Kronecker Products within NIPALS algorithms</td>
<td>77</td>
</tr>
<tr>
<td>Multigroup and multiblock PLS regression</td>
<td>79</td>
</tr>
<tr>
<td>Tree-structured Partial Least Squares with an Application to Orthodontics Dat</td>
<td>81</td>
</tr>
<tr>
<td>A least squares approach to latent variables extraction in formative-reflective models</td>
<td>83</td>
</tr>
<tr>
<td>Probabilistic Disjoint Principal Component Analysis</td>
<td>85</td>
</tr>
<tr>
<td>A mixed method approach to help demonstrate saturation in qualitative research: applying Partial Least Square regression to qualitative data</td>
<td>87</td>
</tr>
<tr>
<td>Handling Unobserved Heterogeneity in PLS Path Modeling Using REBUS: Case of TAM Apply to Social Media Adoption</td>
<td>89</td>
</tr>
<tr>
<td>Sparse Partial Least Squares for Macroeconomic Forecasting</td>
<td>87</td>
</tr>
<tr>
<td>Seemingly unrelated linear regression models based on multivariate mixtures</td>
<td>93</td>
</tr>
<tr>
<td>Model Based Composite Indicators. Why to use the PLS Path Modeling approach</td>
<td>97</td>
</tr>
<tr>
<td>Robust Kronecker Product PCA for Spatio-Temporal Covariance Estimation</td>
<td>99</td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Boris Hejblum, Robin Genuer and Rodolphe Thiebaut</td>
<td>101</td>
</tr>
<tr>
<td>Variable selection in high-dimensional dataset: comparison of sPLS with other approaches in an HIV vaccine trial</td>
<td></td>
</tr>
<tr>
<td>Mounia N. Hocine, Giorgio Russolillo and Gilbert Saporta</td>
<td>103</td>
</tr>
<tr>
<td>A PLS Path Model for Predicting Impact of Job Characteristics on Work-Related Stress</td>
<td></td>
</tr>
<tr>
<td>Karel Hron, Alzbeta Kalivodova, Peter Filzmoser and Lukas Najdekr</td>
<td>105</td>
</tr>
<tr>
<td>PLS-DA for compositional data with application to metabolomics</td>
<td></td>
</tr>
<tr>
<td>Heungsun Hwang</td>
<td></td>
</tr>
<tr>
<td>Three-way Generalized Structured Component Analysis</td>
<td>107</td>
</tr>
<tr>
<td>Emmanuel Jakobowicz</td>
<td>109</td>
</tr>
<tr>
<td>A software for Generalized Structured Component Analysis</td>
<td></td>
</tr>
<tr>
<td>Kwanghee Jung and Hyun Jung Park</td>
<td>111</td>
</tr>
<tr>
<td>Multilevel Dynamic Generalized Structured Component Analysis</td>
<td></td>
</tr>
<tr>
<td>Aurélie Kessous, Fanny Magnoni and Pierre Valette-Florence</td>
<td>113</td>
</tr>
<tr>
<td>Brand Nostalgia and Consumers’ Relationships to Luxury Brands: a Continuous and Categorical Moderated Mediation Approach</td>
<td></td>
</tr>
<tr>
<td>Henk Kiers</td>
<td></td>
</tr>
<tr>
<td>Treating Similar Aspects Differently, and Different Aspects Similarly</td>
<td>115</td>
</tr>
<tr>
<td>Athanasios Kondylis and Joe Whittaker</td>
<td></td>
</tr>
<tr>
<td>Preconditioning methods for PLS regression. Towards functional alternatives</td>
<td></td>
</tr>
<tr>
<td>Lesley Land, Natasha Kamalesanandb, Wynne Chin and R. Ryan Nelson</td>
<td>119</td>
</tr>
<tr>
<td>E-Health Website Adoption: A Test of the Theory of Trying</td>
<td></td>
</tr>
<tr>
<td>Ludovic Lebart</td>
<td></td>
</tr>
<tr>
<td>Analysis of structured data: All roads lead to Naples</td>
<td>121</td>
</tr>
<tr>
<td>Gisela Lechuga, Laurent Le Brusquet, Vincent Perlberg, Louis Puybasset, Damien Galanaud and Arthur Tenenhaus</td>
<td>123</td>
</tr>
<tr>
<td>Discriminant Analysis for Multi-way Data</td>
<td></td>
</tr>
<tr>
<td>Ying Li and Dietrich von Rosen</td>
<td>125</td>
</tr>
<tr>
<td>PLS-inspired multivariate linear modeling</td>
<td></td>
</tr>
<tr>
<td>Tommy Löfstedt, Fouad Hadj-Selem, Vincent Guillemtot, Cathy Philippe, Edouard Duchesney, Vincent Frouin and Arthur Tenenhaus</td>
<td>127</td>
</tr>
<tr>
<td>Structured variable selection for generalized canonical correlation analysis</td>
<td></td>
</tr>
<tr>
<td>Cindy Lombart and Didier Louis</td>
<td></td>
</tr>
<tr>
<td>A study of the impact of Corporate Social Responsibility and price image on retailer personality and consumers’ reactions (satisfaction, trust and loyalty to the retailer)</td>
<td>129</td>
</tr>
<tr>
<td>George A. Marcoulides, Andrew Hardin and Jerry Chang</td>
<td></td>
</tr>
<tr>
<td>Formative versus reflective measurement: Fallacy or reality</td>
<td>131</td>
</tr>
<tr>
<td>Francesca Martella, Donatella Vicari and Maurizio Vichi</td>
<td></td>
</tr>
<tr>
<td>Multivariate regression modeling with clustered correlated predictors</td>
<td>133</td>
</tr>
<tr>
<td>Harald Martens</td>
<td></td>
</tr>
<tr>
<td>Sparse nominal-level PLSR with interactions</td>
<td>135</td>
</tr>
<tr>
<td>Elena Menichelli, Trygve Almey, Oliver Tomic, Nina V. Olsen and Tormod Næs</td>
<td>137</td>
</tr>
<tr>
<td>SO-PLS as an exploratory tool for path modelling</td>
<td></td>
</tr>
<tr>
<td>Simona C. Minotti, Antonio Ciampi, Giuseppe Lamberti and Tomas Aluja-Banet</td>
<td></td>
</tr>
<tr>
<td>Statistical properties of coefficient estimates in PLS-PM: reflective versus formative measurement model</td>
<td>139</td>
</tr>
<tr>
<td>Joanna Orzel and Michał Daszykowski</td>
<td></td>
</tr>
<tr>
<td>PLS and n-PLS for modelling the total antioxidant capacity of food extracts using excitation emission fluorescence spectroscopy</td>
<td>141</td>
</tr>
<tr>
<td>Francesco Palumbo and Rosaria Romano</td>
<td></td>
</tr>
<tr>
<td>Partial Possibilistic Regression Path Modeling</td>
<td>143</td>
</tr>
<tr>
<td>Heewon Park and Sadanori Konishi</td>
<td></td>
</tr>
<tr>
<td>Robust penalized logistic regression modeling and tuning parameter selection</td>
<td>145</td>
</tr>
</tbody>
</table>
• Leila Peltier-Ben Aoun and Anne Dubrocard
  Eco-innovation and performance: what is the story in Luxembourg? ....................................................... 147

• Cathy Philippe, Arthur Tenenhaus, Vincent Frouin and Vincent Guillemot
  Multi-block Logistic Regression for High Dimensional Data........................................................................... 149

• Karine Picot-Coupey and Julien Troville - The Use of PLS-SEM in Marketing Research: Advantages, Limits and Recommendations for Rigorous Applications- «Session AFM» ........................................... 151

• Andrew Prattley, Ronald Ting Tai Chan and Erik van Voorthuysen
  Applications of Mode C analysis to understanding team performance....................................................... 153

• Cristian Preda and Gilbert Saporta
  PLS for multivariate functional data .......................................................................................................... 155

• Marco Reis and Tiago Rato
  Fault detection in the Tennessee Eastman process using dynamic principal components analysis with decorrelated residuals ........................................................................................................... 157

• Giorgio Russolillo
  Extending the Non-Metric PLS approach to inwards directed path models............................................. 159

• Ahmed Shuhaiber and Hans Lehmann
  Embedding PLS analysis into the theory building process – the case of consumer trust in mobile payments ............................................................................................................................... 161

• Marco Singer, Tatjana Krivobokova, Axel Munkand Bert de Groot
  Partial Least Squares for Dependent Data .................................................................................................. 163

• Hye Won Suk and Heungsun Hwang
  Functional Generalized Structured Component Analysis ........................................................................... 165

• Arthur Tenenhaus and Michel Tenenhaus
  Multigroup or multiblock data analysis with RGCCA .................................................................................. 167

• Vanesa Tennant, Annette Mills and Wynne Chin

• Arun Venkitaraman and Magnus Jansson
  On Estimation of Covariance Matrices Modeled as a Sum of Kronecker Products...................................... 171

• Huiwen Wang and Lele Huang
  Functional Linear Regression Analysis With Partial Least Squares And Its Application................................ 173

• Sebastian Wolfrum, Norman Siebrecht, Susanne Papaja-Hülsbergen, Gisela Lüscher, Philippe Jeanneret, Michaela Arndorfer, Jürgen K. Friedel, Maurizi G. Paoletti, Maximilian Kainz, Kurt-Jürgen Hülsbergen und Felix Herzog
  Complex indicators for earthworm (Lumbricus terrestris) population viability – a PLS-PM path model approach ............................................................................................................................... 175

• Genane Youness and Ahmad Ramadan
  Use of PLS path modeling to study customer satisfaction of two Lebanese mobile operators...................... 177

• Lili Zheng, Michel Plaisent and Prosper Bernard
  Comparability Regarding Cross-Cultural Research: Establishing Data Equivalence Using Amos Multi-Group Analysis ..................................................................................................................... 179

• Yi Zhou and Siu-Wai Leung
  Partial least square component as frailty in survival analysis on gene expression data................................ 181

Posters ...........................................................................................................................................  183

• Nada Assi, Vivian Viallon, Anne Fages, Bénédicte Elena-Herrmann, Mazda Jenab and Pietro Ferrari
  Challenges in applying PLS to un-targeted serum NMR metabolomic data from a nested case-control study in the EPIC study ................................................................. 185

• Philippe Bastien, Frédéric Bertrand, Nicolas Meyer and Myriam Maumy-Bertrand
  Deviance residuals based sparse PLS and sparse kernel PLS regression for censored data ....................... 187

• Marta Bevilacqua and Federico Marini
  Local classification by Locally Weighted Partial Least Squares-Discriminant Analysis (LW-PLSDA) .......... 189

• Mélanie Blazère, Fabrice Gamboa and Jean-Michel Loubes
  A new statistical insight on PLS through orthogonal polynomials ................................................................ 191
• Susana Faria, Shirin Shahriari and A. Manuela Gonçalves  
A Comparison of variable selection methods in high-dimensional regression ........................................... 193

• Giuseppe Giordano, Stefania Spina and Laura Trinchera  
PLS-Path model for Social Network Data .................................................................................................. 195

• Robson Godinho, Mauricio Santos and Ronei Poppi - Raman Spectroscopy and Multivariate  
Calibration as a New Technique for Fragrances Flash Points Determination .......................................... 197

• Tomoko Hatta, Tamaki Takaya, Tatsuya Takagi, Norihiro Kawashita and Kousuke Okamoto  
Development of Combinatorial Regression Method for Avoiding Chance Correlations ............................. 199

• Tae-Joo Kang, Byoung-Su Kang, Shin-Jae Lee and Tae-II Kim  
A better method of predicting face changes after cosmetic surgery: the partial least squares regression ...... 201

• Barbara Krakowska, Joanna Orzel, Michal Daszykowski, Ireneusz Grabowski, Grzegorz Zaleszczyk and Miroslaw Sznajder  
Partial Least Squares for detection of diesel oil adulteration ..................................................................... 203

• Jean-Pierre Labuschagne  
Exploring the impact of scaling techniques on classification accuracy of NMR metabolomic data using PLS-DA .............................................................................................................. 205

• Kristian H. Liland, Achim Kohler and Volha Shapaval  
Hot PLS – a framework for Hierarchically Ordered Taxonomic classification by Partial Least Squares ...... 207

• Jérémy Magnanensi, Frédéric Bertrand, Myriam Maumy-Bertrand and Nicolas Meyer  
A new bootstrap based stopping criterion in PLS components construction ............................................. 209

• Panagis Moschopoulos  
Testing in Two-Way Analysis of Variance with Random Cell Sizes .......................................................... 211

• Mauro Nogueira, Danniela Faleiro, Fernando Da Costa, Reto Brun, Marcel Kaiser and Thomas Schmidt  
Linking plants metabolite profile and antitypanosoma activity data with PLS analysis:  
An approach to discover new natural active compounds ........................................................................... 213

• Alba Martinez-Ruiz and Cristina Montañola-Sales  
A Review on Strategies to Factor Matrices and Parallelize Algorithms .................................................... 215

• Sirli Seisonen, Keili Kajava, Rain Kuldjärv and Kristel Vene  
- Correlation of gas chromatography-olfactometry and sensory descriptive analysis of oregano samples by PLSR ................. 217

• V. Tobar, L. A. Sarabia and M.C. Ortiz  
PLS calibration model based on UV-visible signals to quantify and confirm the presence of formaldehyde after the migration testing from melamine kitchenware ............................................................. 219

• M. Villalba, L.A. Sarabia and M.C. Ortiz  
Fast and inexpensive method for the determination of bisphenol A in a migration test from polycarbonate tableware using fluorescence spectroscopy with multiway techniques ...................... 221

• Genane Youness and Wassim Nassar  
An empirical comparison between PLS, Lasso, Elasticnet and other models for highly correlated data.... 223

• Andi Zheng and Zhihong Zou  
Estimation of NBA Rookies Performances Using Partial Least Squares Regression Model...................... 225
### Scientific Programme

#### Monday, May 26

<table>
<thead>
<tr>
<th>Time</th>
<th>Amphi Robert Faure (Z)</th>
<th>Amphi Paul Painlevé (PP)</th>
<th>Amphi J.B. Say (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:00</td>
<td>Registration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>09:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>09:30</td>
<td></td>
<td>KN1: P. Bühlmann</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inhomogeneous large-scale data: maximin effect</td>
<td>Chair: G. Saporta</td>
</tr>
<tr>
<td>10:30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:00</td>
<td>Coffee break</td>
<td>CS1: Measurement scale heterogeneity in PLS-PM</td>
<td>OS1: Recent advances on functional PLS and applications</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chair: V. Esposito Vinzi</td>
<td>Chair: A. M. Aguilera</td>
</tr>
<tr>
<td>12:30</td>
<td></td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td>CS2: Multivariate models for longitudinal data</td>
<td>OS2: Recent developments in Generalized Structured Component Analysis</td>
<td>OS3: AFM Session: Focus on the use of PLS in marketing in France</td>
</tr>
<tr>
<td></td>
<td>Chair: A. Kondylis</td>
<td>Chair: H. Hwang</td>
<td>Chair: M. Korchia</td>
</tr>
<tr>
<td>15:30</td>
<td>Coffee break &amp; Poster Session 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td>CS3: PLS applications in health sciences</td>
<td>OS4: Kronecker product models</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chair: C. Guinot</td>
<td>Chair: A. Hero</td>
<td></td>
</tr>
<tr>
<td>17:15</td>
<td>CS4: PLS applications in ecology &amp; spectroscopy</td>
<td>OS5: High-dimensional models for genomic prediction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chair: E.M. Qannari</td>
<td>Chair: A.L. Boulesteix</td>
<td></td>
</tr>
<tr>
<td>18:00</td>
<td>Free time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:30</td>
<td>Visit of CNAM Museum &amp; Welcome Reception at the Chapel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22:00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Scientific Programme

### Tuesday, May 27

<table>
<thead>
<tr>
<th>Times</th>
<th>Amphí Robert Faure (Z)</th>
<th>Amphí Paul Painlevé (PP)</th>
<th>Amphí J.B. Say (Y)</th>
</tr>
</thead>
</table>
| 09:00  | CS5: PLS-PM methodological extensions  
Chair: N. Niang Keita  
OS6: Genomics and brain imaging  
Chair: H. Abdi | | |
| 10:30  | Coffee break & Poster Session 1 | | |
| 11:00  | CS6: Multigroup analysis in PLS Path Modeling  
Chair: P. Valette-Florence  
OS7: Multiway component analysis  
Chair: C. Derquenne  
CS7: Multiblock models for classification  
Chair: S. Bougeard | | |
| 12:15  | Lunch | | |
| 13:45  | Round Table: Formative versus reflective measurement: fallacy or reality  
Chair: G. Marcoulides | | |
| 14:45  | | OS8: Multigroup analysis  
Chair: M. Tenenhaus | |
| 16:00  | Coffee break & Poster Session 2 | | |
| 16:30  | CS8: Methodological and mathematical contributions to PLS Regression  
Chair: C. Preda | OS9: PLS Path Modeling applications in information systems research  
Chair: W. Chin | |
| 17:30  | | CS9: PLS-PM applications in marketing and Customer Relationship Management  
Chair: F. Camillo | |
| 18:15  | | | |

### Wednesday, May 28

<table>
<thead>
<tr>
<th>Times</th>
<th>Amphí Robert Faure (Z)</th>
<th>Amphí Paul Painlevé (PP)</th>
</tr>
</thead>
</table>
| 09:00  | CS10: Variable selection and sparse PLS Regression  
Chair: L. Trinchera | OS10: Applied latent variable models  
Chair: A. Montanari | |
| 10:30  | Coffee break & Poster Session 2 | | |
| 11:00  | CS11: Applications in genetics  
Chair: A. Tenenhaus | Session in Honour of Carlo Lauro  
Chair: V. Esposito Vinzi | |
| 12:30  | Lunch | | |
| 13:00  | | | |
| 14:30  | KN3: M. Hanafi  
Thinking multiblock  
Chair: H. Abdi | | |
| 15:30  | CS12: Applications in CRM and industry  
Chair: T. Aluja-Banet | OS11: Advances in regression modeling  
Chair: M. Vichi | |
| 16:30  | | XLSTAT Award & Closing Ceremony | |
| 17:15  | | | |
KEYNOTES
Accuracy estimation for PLS and related methods via resampling-based procedures

Anne-Laure Boulesteix

"Ludwig-Maximilians-University of Munich, Munich, Germany

Keywords: benchmarking, comparison studies, real data, test, parameter tuning, cross-validation

Introduction

Resampling-based methods such as, e.g., k-fold cross-validation or repeated splitting into training and test sets are routinely used in the context of supervised statistical learning to assess the prediction performance of prediction methods, including PLS-based regression procedures. In this talk, I discuss two important issues related to the use of such methods: the design of resampling-based benchmark experiments from the perspective of statistical testing [1] and the stability of resampling-based procedures for the choice of tuning parameters [2].

1 Statistical framework for resampling-based benchmarking experiments

The first part of the talk deals with benchmark experiments that aim at comparing the performance of different algorithms and presents a statistical framework for hypothesis testing in real data comparison studies [1]. In computational sciences, most abstracts of articles presenting new supervised learning methods end with a sentence like "our method performed better than existing methods on real data sets", e.g. in terms of error rate. However, these claims are often not based on proper statistical inference and, if statistical hypothesis tests are performed, the tested hypothesis is not clearly defined and poor attention is paid to the type I and type II error.

We propose a proper statistical framework for hypothesis tests comparing the performance of supervised learning methods based on several real data sets with unknown underlying distributions [1]. After giving a statistical interpretation of ad-hoc tests commonly performed in practice, we devote special attention to power issues and outline a simple method to determine the number of data sets to be included in a comparison study to reach an adequate power.

This approach clearly suggests that most comparison studies presented in the literature are considerably underpowered. Comparison studies presented within research or review articles often include a very limited number of datasets as demonstrated in a special case by our recent literature survey of papers on supervised classification published in computational statistics, bioinformatics and machine learning journals [3]. In light of our new statistical framework, it becomes clear that these applications should be considered as illustrative rather than representative of the considered area of application [4].

This part of the talk is illustrated by concrete power calculations for typical values of the effect size as identified from example studies from the literature and through a large benchmark study based on 50 (high-dimensional) microarray data sets with binary response variable. The investigated algorithms include linear discriminant analysis performed on the PLS components, a method that has already shown good performance in similar settings [5].

As an outlook, I will also briefly present results due to Fuchs et al [6] on the practical implication of the recent identification of estimators of the error rate and its variance as U-statistics. In particular, I will present a hypothesis test comparing the performance of several algorithms based on a single dataset [6].

*Corresponding author. E-mail: boulesteix@ibe.med.uni-muenchen.de.
2 Resampling-based methods for parameter tuning

The second part of the talk is devoted to the problem of the choice of tuning parameters—for instance the number of PLS components—based on resampling methods. Cross-validation or related methods are commonly applied within the training dataset to internally assess candidate values of tuning parameters. The value yielding the best internal performance is then chosen to fit the final prediction rule using the whole training dataset.

In this part of the talk I will show results obtained with linear discriminant analysis performed on PLS components [5] and sparse PLS [7]. These results empirically demonstrate the instability induced by the random character of resampling-based procedures both in terms of variable selection and prediction performance [2]. As suggested by this study and substantiated by theoretical results [6], we recommend to perform a large number of resampling iterations whenever computationally possible to yield a better stability.

References


Inhomogeneous large-scale data: a new approach

Peter Bühlmann

(joint work with Nicolai Meinshausen)

ETH Zurich, Zurich, Switzerland

Keywords: big data, linear program, mixture model, regularized estimation, robustness

Introduction

Large-scale or “big” data usually refers to scenarios with very many variables (dimension $p$) and very large sample size $n$. Such data is most often of “inhomogeneous” nature, i.e., not being i.i.d. realizations from a distribution or not being realizations from a stationary distribution. Such heterogeneity also implies that the amount of information for statistical accuracy might not be as large as naively believed: if the data is “homogeneous” for small sub-groups of the entire sample only, we are still confronted with the problem of high-dimensional inference, namely that the number of samples in a homogeneous sub-group is substantially smaller than the dimension $p$.

We propose a new methodology for some class of large-scale inhomogeneous data: the advocated procedure is efficient with respect to computation and we provide corresponding statistical accuracy guarantees.

1 A mixture model

For simplicity, we consider here a mixture of regressions model:

$$Y_i = X_i^T B_i + \varepsilon_i \quad (i = 1, \ldots, n),$$

where $Y_i \in \mathbb{Y} \subseteq \mathbb{R}$, $X_i \in \mathbb{X} \subseteq \mathbb{R}^p$, and the regression coefficient vector $B_i \in \mathbb{R}^p$ is allowed to change for every subject $i = 1, \ldots, n$. The $B_i$’s are random variables from a distribution. We assume that the $\varepsilon_i$’s are uncorrelated from the $X_i$’s and the $X_i$’s are independent from the $B_i$’s so that there is no information from $X$ to the mixture regression coefficients $B$. The following examples fit to the framework.

Known groups of observations. There are groups $C_1, \ldots, C_G$ which build a partition of the indices of the entire sample $\{1, \ldots, n\}$. Every subject $i$ corresponds to a group $C_g \subseteq \{1, \ldots, n\}$, for some $g = 1, \ldots, G$. In every group $C_g$, we assume that $B_i \equiv b_g$ for all $i \in C_g$. Thus, the model in (1) is a finite mixture of $G$ regression coefficients with known mixture components.

Correlated $B_i$’s. When having strong positive correlation among neighboring $B_i$’s (neighboring w.r.t. the sample ordering $i$), we have a smooth trend for the regression coefficients $B_i$’s (w.r.t. $i$). The model in (1) is then a random coefficient linear model with smooth trend.

Contaminated samples. Consider the case where a large fraction $(1 - \delta)$ of the $B_i$’s assumes a fixed value $b$, and there is a smaller fraction $\delta$ of outliers where the $B_i$’s can take other values in $\mathbb{R}^p$. The model (1) can then be viewed from the viewpoint of robust statistics for guarding against some outliers where as usual, the (groups of) outliers are unknown.

1.1 Maximin effects without fitting the entire mixture model

Fitting a mixture model with say maximum likelihood estimation can be very cumbersome when the dimension $p$ and the sample size $n$ are large. We propose in [1] instead a new “maximin” parameter $b_{\text{maximin}} \in \mathbb{R}^p$ which is an...
important “summary quantity” of all the possible values \( B_i \) so that prediction and interpretation in heterogeneous models remains powerful (prediction is based on a linear model \( X_{new}^T \hat{b}_{\text{maximin}} \) with the estimated maximin effects parameter). The main idea and motivation of the definition of the maximin effects is to optimize on the worst possible case which then enables “robustness” (against all sorts of inhomogeneities) in the estimation step. In particular, estimation of the maximin effects does not require fitting the entire mixture model in (1).

In [1], the following main issues are covered: establishing estimation rates and accuracy for the “maximin” parameter \( b_{\text{maximin}} \), covering also the high-dimensional setting where the dimension \( p \) might be much larger than the sample size of the “homogeneous subgroups”; and showing that the estimator can be computed, under some circumstances, with a very efficient linear program allowing for very large scales. Some empirical examples demonstrate the usefulness of the proposed approach and corresponding estimation method.

2 Conclusions

We propose a novel approach for large-scale data analysis. The procedure allows to deal with some form of inhomogeneities, it has good statistical estimation properties, and it can be efficiently computed even when both the dimension and sample size are very large.

References

Thinking multiblock

Mohamed Hanafi and Gaston Sanchez
Oniris, Unité de Sensométrie et Chimiométrie, Site de la Géraudière, Nantes, France
mohamed.hanafi@oniris-nantes.fr, gaston.stat@gmail.com

Keywords: Multiblock data analysis, Matrix algebra, R

Introduction

A challenging problem in multivariate statistics is the study of relationships between several blocks of data, also known as “multi-block data analysis”. Multiblock methods [1] are used in a variety of applications ranging from chemometrics [2,3] to ecology [4].

After twenty years of involvement around this methodology, and nearly a century of literature in this field, it appears that: the framework provided by matrix algebra and matrix factorization principles [5] are not enough to describe the whole richness and the complexity of block matrices, their manipulation, or the factorization strategies behind multiblock methods.

The present talk proposes an algebraic framework for data analysis involving block data structures. This framework can be considered as a new form of thinking about multiblock data structures with new ways of manipulation and statistical treatment. More precisely, we extend the classical matrix algebra to the case of block matrices and show some practical interests of this new framework in order to prototype algorithms for multi-block data analysis. An experimental R package for storage and manipulation of block matrix is also presented.

1 Block Matrix and Blok dimension

Definition. A partition of an integer \( n \) is a decomposition of \( n \) as sum of \( k \) integers all strictly positive (called parts). \( P(n,k) \) denotes the partition of the integer \( n \) in \( k \) parts.

Partition for pair of integers. A partition of a pair of integers \((n,p)\) is a pair of partitions \((P(n,k), P(p,l))\), with \( P(n,k) \) is a partition of the integer \( n \) in \( k \) parts (respectively, \( P(p,l) \) is a partition of the integer \( p \) in \( l \) parts).

Definition. A block matrix (partitioned matrix) is a couple \((X, (P(n,k), P(p,l)))\) where \( X \) is a matrix of dimension \((n,p)\) and \((P(n,k), P(p,l))\) is a partition of its dimension \((n,p)\).

Definition. By analogy with the dimension of a matrix \( X \), the pair of partitions \((P(n,k), P(p,l))\) is called here the block dimension of \( X \) considered as a block matrix.

2 Block Matrix multiplications

Block matrix multiplications are very important and complex. The issues have to do with specifying which blocks are to be multiplied and how the block result should be computed. This problem is approached by introducing 12 new products and notations, 3 of them are reported in table 1.
Table 1. Names and notations of some new block matrix are given in columns 1 and 2. The block definition of $X$ considered as block matrix is given in column 3 where $(*u)$ denotes the well known usual matrix product, $(*u)$ denotes matrix Hadamard product, $(*s)$ denotes scalar by matrix product.

<table>
<thead>
<tr>
<th>Names</th>
<th>Notations</th>
<th>block definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(u,u) product</td>
<td>$X = A^{*_{(u,u)}} B$</td>
<td>$X_{ij} = \sum_{k,l} A_{ik} *<em>{u} B</em>{kj}$</td>
</tr>
<tr>
<td>(u,h) product</td>
<td>$X = A^{*_{(u,h)}} B$</td>
<td>$X_{ij} = \sum_{k,l} A_{ik} *<em>{h} B</em>{kj}$</td>
</tr>
<tr>
<td>(u,s) product</td>
<td>$X = A^{*_{(u,s)}} B$</td>
<td>$X_{ij} = \sum_{k,l} A_{ik} *<em>{s} B</em>{kj}$</td>
</tr>
</tbody>
</table>

3 An experimental R package for block matrix algebra

From a practical point of view, the concepts introduced in previous sections allow to implement some R classes for block matrix and generic methods for manipulations that can be used for fast algorithm prototyping. These classes extend the functionality of R’s to block matrix by supporting additional operations such as block matrix multiplication. The matrix as block matrix supports the block dimension update of a matrix, i.e., the conversion of a matrix to a block matrix (and vice versa), and modify the block structure of a matrix which is a commonly used operation in many algorithms. We describe all of these classes and then demonstrate their use by showing how to implement several block matrix algorithms that have appeared in the literature.

Acknowledgements

“Région Pays de Loire” is greatly acknowledged for its financial support of the present work. This work was carried out under the regional project “Approches Intégratives du déterminisme structural, génétique et écophysiologique de la qualité des fruits”.

References


Using PLS regression in the noisy functional data context

M. Carmen Aguilera-Morillo\textsuperscript{a*} and Ana M. Aguilera\textsuperscript{b}

\textsuperscript{a}University Carlos III of Madrid, Spain
\textsuperscript{b}University of Granada, Spain

Keywords: Functional data, P-splines, Partial Least Squares, Linear Discriminant Analysis.

Introduction

The scope of this work is the improvement of the estimation and prediction ability in regression when we are working with functional data in the presence of noise. Saying noise is to make emphasis in situations such as measurement errors provided by a machine (for example an spectrometer used in chemometric studies) during the collection of the sample data. It is a very important problem which must be taken into account in the data modeling process. Important applications of functional data analysis (FDA) in chemometric can be seen in [1].

In a general way, it can be said that a functional variable is one whose values depend on a continuous scale as for example time or wavelength. So, let us consider \( \{x_1(t), x_2(t), \ldots, x_n(t)\} \) the sample information of a functional variable \( X \). The sample curves can be considered observations of a centered second order stochastic process \( X = \{X(t) : t \in T\} \) whose sample functions belong to the Hilbert space \( L^2(T) \) of square integrable functions with the usual scalar product given by \( \langle f, g \rangle = \int_T f(t) g(t) \, dt, \forall f, g \in L^2(T) \).

In practice, sample curves are usually observed at a finite set of sampling points that could be unequally spaced and different among the sample units. Moreover, in this work we are focus on smooth data observed with some error or noise. So, the first step is to reconstruct the functional form of each sample curve from a finite set of discrete observations by using a B-spline basis representation [2].

In this work we are focus on the functional linear model (FLM) given by

\[
Y = \beta_0 + \int_T X(t) \hat{\beta}(t) \, dt + \varepsilon, \tag{1}
\]

with \( Y \) being a scalar response variable and \( \hat{\beta}(t) \) being the parameter function to be estimated. Our aim is to solve the main problems associated with this model (high dimension, multicollinearity, lack of smoothness in the estimated parameter function) by means of penalized versions of functional Partial Least Squares regression (PLSR) using different types of P-spline penalty. The PLSR was extended to the functional case by [3]. In [4] the estimation of this model with B-splines basis was developed. Working with noisy data these algorithms do not control the degree of smoothness in the estimated parameter function. So, in order to get good estimations different P-spline penalties [5] are introduced in the functional PLSR algorithm.

On the other hand, functional PLSR can be applied in different contexts as functional linear discriminant analysis (LDA) as we can see in [6, 7].

Finally, a real data set will be studied by using the proposed methods based on penalized functional PLS regression.

1 P-spline approaches for functional PLS regression

In order to solve the high dimension and multicollinearity problems associated with the FLM, functional PLSR or functional Principal Component regression (PCR) could be considered. Both methods were compared in [4].

\*Corresponding author. E-mail: maguiler@est-econ.uc3m.es.
with PLSR giving better estimations of the parameter function and similar predictions. Once the high dimension and multicollinearity problems are solved, the problem comes from the lack of smoothness in the estimated parameters. Penalized approaches for functional PLSR are proposed as solution. In [8] penalized versions for functional PCR and PLSR were proposed without taking into account the B-spline approximation of the original sample curves.

P-spline penalties will be considered in this work. P-splines proposed by [5] are an extension of regression splines with a discrete penalty based on differences between coefficients of adjacent B-splines. It is well known that penalized splines have very good properties [9], which make them a good choice for penalizing the lack of smoothness in the estimators of the main functional regression models.

Two different ways of introducing the P-spline penalty in the PLS algorithm will be considered. The first one is based on smoothing the sample curves by P-splines. After that, the non-penalized functional PLSR of the response variable on the smoothed sample curves is carried out. The second version follows the idea developed in [10] for functional Principal Component Analysis (PCA) and introduce the discrete penalty in the definition of the norm of PLS component weight functions. Different penalized approaches for functional PCA have been recently introduced in the estimation of the functional logit model in [11].

References


Extension to the PATHMOX approach to detect which constructs differentiate segments and to test factor invariance: application to Mental Health Data

Tomas Aluja-Banet \textsuperscript{a*} and Giuseppe Lamberti \textsuperscript{b} and Antonio Ciampi \textsuperscript{c}

\textsuperscript{a}Universitat Politecnica de Catalunya, Barcelona Tech, Spain
\textsuperscript{b}Universitat Politecnica de Catalunya, Barcelona Tech, Spain
\textsuperscript{c}McGill University, Montreal (Quebec), Canada

\textbf{Keywords:} PLS-PM, Segmentation, PATHMOX, models comparison, Fisher’s $F$

\section*{Introduction}

In 2009 Gastón Sánchez proposed PATHMOX \cite{Sanchez2009}, a tree-growing algorithm which aims to identify and resolve heterogeneity in the context of Partial Least Square Path Modeling (PLS-PM). This algorithm is indicated in the following situation: 1) We are interested in modeling the correlation structure of a given set of variables that we assume to be adequately represented by a Path Model (PM). 2) We have reason to believe that the coefficients of the PM vary across subgroups of the population. 3) We also believe that these subgroups can be identified by values of another set of variables, the segmentation variables. In a specific application of PATHMOX, all PMs considered have the same topology, so that “different PMs” means PMs with the same topology but different values of the coefficients.

PATHMOX constructs a tree by recursively partitioning a given data set into segments, so that each node of the tree is associated to a specific Path Model (PM) and the leaves (terminal nodes) represent homogeneous subpopulations, i.e. subpopulations which are adequately described by a unique PM. Segmentation is performed by choosing at each node the optimal monothetic split, i.e. the binary partition induced by a single segmentation variable, which best distinguished between the PM’s associated to the subsets of the partition (children nodes). The criterion defining optimality used in PATHMOX is an adaptation of the Fisher’s $F$ statistic for testing the equality of two regression models \cite{Fisher1925} and \cite{Kendall1938}.

In this work we present an extension of PATHMOX which includes these statistics. The advantages of the new version are demonstrated by an application to mental health data.

\section{Extended PATHMOX}

The limitation of the $F$ statistics used in the standard PATHMOX \cite{Sanchez2009}, is that it provides only a global comparison of two PMs: it tests whether or not all the path coefficients for two structural models are equal, but, when it detects a difference, it does not detect which path coefficients are responsible for it. In previous works we have discussed two additional $F$-statistics, one to test equality of coefficients of the inner model \cite{Aluja-Banet2011}, the other to test factor invariance, i.e. equality of the coefficients of the measurement models \cite{Aluja-Banet2012}. In our extension of PATHMOX we have introduced these two statistics in the tree construction and in the interpretation of the PATHMOX tree.

\*Corresponding author. E-mail: Campus Nord UPC. C5-204. Jordi Girona 1-3, 08034 Barcelona, Spain. tomas.aluja@upc.edu.
2 The mental health data-set

Using the new PATHMOX, we analyze data from a survey carried by a team of the St. Mary’s Hospital Research Centre between July 2005 and January 2007. Our aim is to study the relationship between three constructs representing mental disorders that are common in elderly populations: dementia, delirium and depression. The data were collected on 138 patients from seven distinct Quebec hospitals; it consists of a total of 43 variables divided in two groups; one group is formed by 37 indicator variables, which define the PM; and the other is formed by 6 segmentation variables: hospital, patients’ pain, patients gender, Charlson scale (number of associated diseases), number of hospitalizations, patients’ age. The topology of the PM is shown in Figure 1.

![Figure 1. Path Diagram for Mental Health data](image)

In Figure 1, the inner model is represented by the three constructs (latent variables) Dementia, Depression and Delirium; each construct is linked by arrows to specific indicator variables (measurement models).

In our extended PATHMOX, at each step of the construction we also compare the two segments in terms of their inner models. For example, if gender determines a split as in Figure 2, the new version of PATHMOX also provides indications on whether or not the coefficients of the inner model are the same for the children nodes. For examples, we can test whether or not the relationship between Dementia and Delirium has the same intensity for Male and Female patients.

![Figure 2. PATHMOX tree diagram: Mental Health data model partitioned in two terminal nodes by the segmentation variable “gender”](image)

Moreover (not in the picture), indications are obtained about the invariance of the definition of the constructs, e.g. whether or not the relationship between Delirium and its indicator variables is the same for Males and Females.

References


Non-Symmetrical Data Analysis: the Neapolitan School

S. Balbi\textsuperscript{a}, V. Esposito Vinzi\textsuperscript{b} G. Giordano\textsuperscript{c}, F. Mola\textsuperscript{d}, F. Palumbo\textsuperscript{a}\textsuperscript{*}, G. Scepi\textsuperscript{a} R. Siciliano\textsuperscript{a} R. Verde\textsuperscript{e}

\textsuperscript{a}University of Naples Federico II, Naples, Italy \textsuperscript{b} ESSEC Business School, Cergy, France \textsuperscript{c} University of Salerno, Salerno, Italy \textsuperscript{d} University of Cagliari, Cagliari, Italy \textsuperscript{e} Second University of Naples, San Leucio, Italy

Keywords: Non-Symmetrical Correspondence Analysis, Principal Component Analysis, Reference Subspace

Introduction

Most often called in French ‘Analyse des Données’ to stress the direct link with the French school and with its founder J.P. Benzécri, who is famous for the development of the Correspondence Analysis (CA) (probably the most used statistical technique for analyzing contingency tables), Data Analysis has represented and still represents one of the most fruitful statistical domains. Many applications of Statistics, in particular in the behavioral, cognitive and social sciences, can be referred to Data Analysis. However, when the variables have different roles in the analysis, we talk of Non Symmetrical Data Analysis (NSDA).

The thirtieth birthday of the \textit{L’analyse non symétrique des correspondances} (1984) \cite{2} offers an excellent opportunity for drawing a balance of the effect and of the fall-out of the Neapolitan School contribution to the Non-Symmetrical Data Analysis (NSDA) approach in Multivariate Statistics.

Before going into the details of the methods presented in this paper, we are pleased to spend a few words about the Neapolitan School of Data Analysis and in particular a few words of gratitude to its founder. At the beginning of the 80’s the charismatic personality of Carlo Lauro, his devotion to the scientific research and his dedication to an ever larger number of students made possible a significant growth up of Statistics in Naples and made the Department of Mathematics and Statistics of the University of Naples one of the most important poles for the Statistical studies in Data Analysis all around the World. He catalyzed the formation of a team of young researchers on the Data Analysis domain. Most developments that will be mentioned in this paper have been produced during the last thirty years in Data Analysis domain from the Neapolitan school. This paper wants to honor and to thank Carlo Lauro: his time, his devotion and the energy spent for and with us have made all these developments possible.

This presentation has a twofold aim: (i) to outline the state of the art, including the most recent developments, of the NSDA focusing the attention on the Neapolitan School contribution in the context of the data analysis in general; (ii) to identify a (likely non exhaustive) set of domains and fields of application where the NSDA methods, which can be directly or indirectly referred to the Neapolitan statistical school, have affirmed themselves as reference methods.

A (very) short glance at Non-symmetrical Data Analysis

The role played by the variables in statistical analysis can be somehow different and can depend upon the objectives of the researcher. Therefore, there is a strong demand for a variety of methodologies able to cope with different types of data as well as to answer different questions. In this framework, the paper intends to

\footnote{Corresponding author: F. Palumbo - Dept of Political Science, Via Rodinò 22, 80138 Naples, e-mail: fpalumbo@unina.it}
systematize the methodological contributes and the related applications that can be referred to the Neapolitan school of Data Analysis.

For sake of brevity we just cite only the two fundamental papers that have probably contributed to inspire more than hundreds of papers, tens of PhD and MSc theses and are currently used in many application fields in the NSDA framework.

**Numerical variables domain (PCAR)**

In the numerical variables domains the paper entitled ‘Analisi in componenti principali in rapporto ad un sottospazio di riferimento’ \(^1\) \cite{1} can be assumed as the reference approach for many exploratory and confirmatory statistical multivariate methods developed to study the relationships between two groups of variables playing different roles.

Let \(Y\) be a \(n \times q\) matrix, where \(Y_1, Y_2, \ldots, Y_q\) represent the main variables, observed on \(n\) statistical units. Analogously, the \(n \times p\) matrix \(X\) refers to the \(X_1, X_2, \ldots, X_p\) instrumental variables, observed on the same units. Both the variable sets are standardized and divided by \(\sqrt{n}\) so that \(XY\) defines the cross-correlation matrix.

The aim of PCAR is to find a lower-dimensional matrix \(Z = UX\) that maximizes the predictive efficiency of \(Z\) for \(Y\). The solution to this problem can be found by solving the following characteristic equation:

\[
X(X'X)^{-1}XYU = P_XYU = AU
\]

where \(P_X = X(X'X)^{-1}X'\) indicates an orthogonal projector of \(Y\) onto \(X\). It is equivalent to perform a PCA on the matrix \(Y'X(X'X)^{-1}X'Y\). Developments or applications that can be referred to the PCAR are: Conjoint Analysis, Multivariate Control Charts, Principal Component Analysis for Interval-valued Data, Sensory Analysis, Textual Analysis with External Information.

**Categorical variables domain (NSCA)**

Dealing with categorical data, let \(X\) and \(Y\) be a pair of categorical variables with \(K\) and \(J\) categories respectively and observed on the same \(n\) units. The indicator matrices \(X\) and \(Y\) such that \(x_{ik} = y_{ij} = 1\) if the \(i\)th unit belongs to the \(k\)th category of \(X\) and to the \(j\)th category of \(Y\), \(x_{ik} = y_{ij} = 0\) otherwise \((i = 1, \ldots, n; k = 1, \ldots, K; j = 1, \ldots, J)\). Let us now introduce the matrices \(F\) and \(D\), where: \(F = n^{-1}X'Y\) has order \(K \times J\), with general term \(f_{kj}\) with \(k = 1, \ldots, K\) and \(j = 1, \ldots, J\); \(D = n^{-1}XY = \text{diag}(r)\) and \(r\) has general term \(r_k = \sum_{j=1}^{J} f_{kj}\).

Non Symmetrical Correspondence Analysis (NSCA) \cite{2} aims at evaluating the influence of the \(J\) categories of the column variable \(Y\) on the \(I\) categories of the row variable \(X\). Thus attention will be paid to the column profiles matrix \(FD^{-1}\), which represents the observed conditional distribution of \(X\) with respect to \(Y\). NSCA looks for the orthonormal basis which allows to account for the ‘largest’ part of inertia, so as to visualize the dependence structure of \(Y\) w.r.t. \(X\). From a mathematical viewpoint, the search leads to the eigenanalysis of the centered profiles matrix

\[
(F'F^{-1} - \frac{1}{n} Y'h'h'Y) = (Y'X(X'X)^{-1}X'Y - \frac{1}{n} Y'h'h'Y)
\]

i.e. to the singular value decomposition \((FD^{-1} F - \frac{1}{n} Y'h'h'Y) = V AW'\) under the usual orthonormality constraints \(V'V = WDW' = I\), where \(h\) is a \(n \times 1\) vector of ones.

NSCA has inspired many developments and applications of the following topics: Factorial Discriminant Analysis for Categorical Variables, Fuzzy Non-Symmetrical Correspondence Analysis, Factor Clustering for Binary Data, Non Symmetrical Lexical Correspondence Analysis, Classification Trees for Categorical Data, Latent Budget Models, Reduced-Rank Models.

**References**


\(^1\) Principal Component Analysis onto a Reference Subspace
An integration of partial least squares and sparsified multiple correspondence analysis applied to genetic data

Derek Beaton\textsuperscript{a}*, Anne Bernard\textsuperscript{b}, Hervé Abdi\textsuperscript{a}, and Gilbert Saporta\textsuperscript{c}

\textsuperscript{a}The University of Texas at Dallas, USA
\textsuperscript{b} University of Queensland, Australia
\textsuperscript{c} CNAM Paris, France

Keywords: Genetics, behavioral data, sparsification, multiple correspondence analysis, group sparsification

Abstract

Genetic data are now common in many domains. Typically, these genetic studies try to associate genetics with single phenotypes, behaviors, or diagnostic criteria. However, many of these studies include multiple behavioral variables and very large genetic data sets. The analysis of these data sets faces two particular challenges: 1) How to integrate many behavioral and genetic variables when 2) only a small number of variables are interpretable. To address these issues, we propose the integration of partial least squares and a sparsified approach to multiple correspondence analysis.

1 Introduction

The singular value decomposition (SVD)—and its weighted least-squares extension, the generalized SVD (GSVD)—are the core of many multivariate techniques such as principal components analysis (PCA), partial least squares (PLS), and multiple correspondence analysis (MCA). All of these techniques produce components (a.k.a. latent variables) which are obtained as linear combinations of the original variables. However, when the sample size is small and the data large (i.e., $n << p$) many variables will have non-zero loadings, a pattern that makes it difficult to interpret the results. A number of regularization and sparsification techniques have been proposed in order to produce only few non-zero loadings. Some of these methods include SCoTLASS \cite{1}, SPCA \cite{2}, SPCA-rSVD\cite{3}, and rPCA\cite{4}.

However, with large genetics data there are still some problems not easily addressed by sparsification alone. In fact, using behavioral data in conjunction with large genetics data (such as genome-wide data) may increase power to detect genetic effects \cite{5}. Therefore, it would be advantageous to use a PLS method designed for behavioral and genetic data. One such method is PLSCA \cite{6}. But, PLSCA would still produce many non-zero loadings. Because most genetic data (such as SNPs) are categorical they are naturally structured by blocks (see Table 3). A recent extension of SPCA-rSVD\cite{3} has been adapted as group-sparse PCA and extended to MCA, called sparse MCA (SMCA) \cite{7}.

In this paper we present a solution to the analysis of genetic data that integrates PLSCA with SMCA. This approach can associate SNPs—typically large and noisy data—to various behavioral markers—typically well-defined instruments—while producing as few non-zero loadings as possible.
Table 3: Example of nominal (left) and disjunctive (right) coding of illustrative SNPs (SNP 1 and 2)

<table>
<thead>
<tr>
<th>SNP 1</th>
<th>SNP 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subj.1</td>
<td>AG</td>
</tr>
<tr>
<td>Subj.2</td>
<td>...</td>
</tr>
<tr>
<td>Subj.3</td>
<td>GG</td>
</tr>
</tbody>
</table>

(a) Nominal

<table>
<thead>
<tr>
<th>SNP 1</th>
<th>SNP 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subj.1</td>
<td>1</td>
</tr>
<tr>
<td>Subj.2</td>
<td>1</td>
</tr>
<tr>
<td>Subj.3</td>
<td>0</td>
</tr>
</tbody>
</table>

(b) Disjunctive

2 Asymmetric sparsification of genetic data

Call $Y$ a disjunctive data matrix that represents behavioral data (e.g., a survey). Call $X$ a disjunctive data matrix that represents genetic data (e.g., SNPs; see, e.g., Table 3). SNPs can be viewed as multiblock data, where each SNP represents a block of (usually) 3 columns. We would approach the analysis of these data sets with PLSCA [6] as such: $R = X^T Y$ where $R$ is a contingency table (behavioral data $\times$ genetic data). Correspondence analysis (CA) is a natural choice for analyzing very large contingency tables. Therefore, in PLSCA, we would preprocess $R$ as we normally would with CA, where the row and column weights are proportional to the row and column sums and stored in diagonal matrices ($W_Y$ and $W_X$, respectively). We then decompose $R$ with the GSVD where $R = P \Delta Q^T$, such that $PMP = QWQ = I$, where the latent variables are: $L_X = XW_X P \times r_{++}^{-1}$ and $L_Y = YW_Y Q \times r_{++}^{-1}$, where $r_{++}$ is the sum of the table $R$. The latent variables have maximal covariance (due to the properties of the GSVD) as: $L_X^T L_Y = \Delta$

However, there are typically many SNPs in data such as these, thus implying that $X$ has the following two properties: 1) a natural block structure and 2) is much larger than $Y$ and thus could produce components with many non-zero values. If $X$ has a natural block structure such that $X = [X_1, ..., X_b, ..., X_B]$ where $R = Y^T [X_1, ..., X_b, ..., X_B]$ and given the block structure of $R$, we have: $R = P \Delta Q^T = \Delta [Q_1, ..., Q_b, ..., Q_B]^T$. The blocks of $Q$ represent the group coding of SNPs, and hence, we would want to (group) sparsify the SNPs (genetic) data. Sparsification can be achieved with an extension of group sparse PCA to MCA: sparse multiple correspondence analysis [7].

We will present the details of the method illustrated with a small and a realistic example.

References


*Corresponding author. E-mail: derekbeaton@utdallas.edu.*
**CONCATPLS: PLSC for one continuous data set and one categorical data set**

Derek Beaton\textsuperscript{a}*, Joseph Dunlop\textsuperscript{a,b}, Francesca Filbey\textsuperscript{a,b}, and Hervé Abdi\textsuperscript{a}

\textsuperscript{a}School for Behavioral and Brain Sciences
\textsuperscript{b}Center for BrainHealth
The University of Texas at Dallas, Richardson, Texas USA

**Keywords:** Mixed-data, neuroimaging, genetics, principal components analysis, multiple correspondence analysis

**Abstract**

We present an extension of PLS correlation—called continuous-categorical partial least squares (CONCATPLS)—that analyzes two data sets, one comprising continuous variables and the other comprising nominal variables. CONCATPLS combines features of PLSC (analyzing the information common to two tables), principal components analysis, and multiple correspondence analysis. We also present inferential techniques for CONCATPLS such as bootstrap, permutation, and χ\textsuperscript{2} omnibus tests. We illustrate CONCATPLS with a large example: a functional neuroimaging data set (continuous) and a genomic (categorical) data set.

**1 Introduction**

Imaging Genetics seeks the genetic correlates of neural structures and functions related to behaviors or clinical diagnoses. Imaging and genetic data sets tend to be very large (e.g., more than one million variables). Partial least squares correlation (Partial Least Squares Correlation)—a popular method in neuroimaging—can analyze the information common to two data sets [1]. As such, it should be well suited to identify common relationships between, say, voxels (i.e., continuous brain variables) and SNPs (i.e., nominal genetic variables). However, PLSC works only with quantitative data. While imaging data are quantitative, the most commonly used genetic data are categorical (e.g., SNPs exist as three categories: a major homozygote, a minor homozygote, and a heterozygote). Here, we show how to simultaneously analyze a quantitative data set (fMRI) and a categorical data set (GWAS).

We introduce a new variant of PLSC—called continuous-categorical PLS (CONCATPLS)—designed for imaging genetics analyses. CONCATPLS is a compromise between PLSC [1] and PLSC-correspondence analysis [2]. Like PLSC, CONCATPLS computes pairs of latent variables associated to each data set such that paired latent variables have maximum covariance, while unpaired latent variables are orthogonal. To do so, CONCATPLS integrates features of PCA (for quantitative data) and multiple correspondence analysis (MCA; for categorical data). Some features include pre-processing from PCA and MCA—for their respective data types—and the computation of weights for variables (from MCA), such as χ\textsuperscript{2} weights for SNPs (to identify rare variants). CONCATPLS uses bootstrap and permutation tests to (1) assess the generalizability of the results, and (2) provide a sparse and simplified representation of the results. Permutation tests identify significant latent variables whereas the bootstrap identify SNPs and voxels that significantly contribute to the results.

*Corresponding author. E-mail: derekheaton@utdallas.edu.
Table 3: Example of nominal (left) and disjunctive (right) coding of illustrative SNPs (SNP 1 and 2)

<table>
<thead>
<tr>
<th></th>
<th>SNP 1</th>
<th>SNP 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subj.1</td>
<td>AG</td>
<td>CA</td>
</tr>
<tr>
<td>Subj.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj.1</td>
<td>GG</td>
<td>AA</td>
</tr>
</tbody>
</table>

(a) Nominal

<table>
<thead>
<tr>
<th></th>
<th>SNP 1</th>
<th>SNP 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subj.1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Subj.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj.1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Subj.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) Disjunctive

2 Formalization of CONCAT PLS

CONCatPLS reveals the common information between two data matrices: one continuous data set and one categorical data set. The generalized singular value decomposition (GSVD)—a weighted least squares version of the SVD—is the core of CONCatPLS, (as well as PCA, MCA, PLS, and many other techniques). CONCatPLS can be formalized in terms of PCA and MCA.

Call Y a quantitative matrix with continuous data (e.g., fMRI). A PCA is a suitable choice to analyze this table. We call ZY the centered, standardized version of Y. PCA is computed from the GSVD of ZY as

\[ Z_Y = P \Delta Q^T \quad \text{with} \quad P^T M P = Q^T W Y Q = I, \tag{1} \]

where M and WY are diagonal matrices whose diagonal elements sum to 1.

Call X a disjunctive data matrix (see, e.g., Table 3) that represents a categorical data set (e.g., SNPs). MCA is a suitable choice to analyze this table. MCA can be viewed as a weighted, non-standardized version of PCA [3]. First, we denote the center (column means) of X as x. Next, compute \( Z_X = (X - [1 1 1] \odot x) \times N^{-1} \). MCA is computed as

\[ Z_X = U \Sigma V^T \quad \text{with} \quad U^T M U = V^T W_X V = I, \tag{2} \]

where \( W_X = \text{diag}\{x^{-1}\} \).

We can now formalize CONCatPLS. First, compute \( R = Z_X^T Z_Y \), which is then decomposed by the GSVD as:

\[ R = S_X \Psi S_Y \quad \text{with} \quad S_X^T W_X S_X = S_Y^T W_Y S_Y = I. \tag{3} \]

The original data are expressed as latent variables with respect to their singular vectors (S):

\[ L_X = Z_X W_X S_X \quad \text{and} \quad L_Y = Z_Y W_Y S_Y. \tag{4} \]

They provides us with a score for each of the I individuals in the analysis with respect to the X and Y matrices, where

\[ \text{cov}(l_{X,i}, l_{Y,j}) \propto l_{X,i}^T l_{Y,j} = \max \tag{6} \]

(as follows from Eq. 3). Given the properties of the GSVD, when \( \ell = 1 \), we have the largest possible covariance between \( Z_X \) and \( Z_Y \); when \( \ell = 2 \), we have the next largest possible covariance. This property holds for all subsequent pairs of latent variables with orthogonality between all unpaired latent variables.

References


Multiblock Classification: SO-PLS-DA

Alessandra Biancolillo\textsuperscript{a*} Ingrid Måge\textsuperscript{a} Tormod Næs\textsuperscript{a}

Nofima, Norwegian Institute of Food, Fisheries and Aquaculture Research, Ås, Norway\textsuperscript{a*}

Keywords: SO-PLS, Linear Discriminant Analysis, Multiblock

Introduction

In recent years, the advancement of new analytical technologies has generated an increased need for interpreting large and complex data sets and also the relationship between them. Often, variables can be separated into conceptually meaningful blocks of data which can represent for instance different steps of an industrial process, or measurements taken by different instruments. This kind of multi-block data can be found in various fields, such as microbiology, industrial processes, consumer and sensory data, -omics, or medical protocols. It is therefore crucial to develop and implement new methods which can process this large amount of data and that allow for combining multiple blocks of data from different experimental observations.

A number of methods are available for analyzing multiple data blocks. Many of these, such as MB-PLS, SO-PLS, PO-PLS and OnPLS are widely used in all fields as they give excellent results. When linking data sets there are a number of different aspects that are of interest, for instance looking for common information and unique information in the data and using some of the data sets for predicting others.

The aim of the present work is to extend the SO-PLS regression method, usually used for continuous output, to situations where classification is the main purpose. For this reason SO-PLS discriminant analysis will be compared with other commonly used techniques such as PLS-DA and MBPLS-DA. In particular we will focus on how multiblock strategies can give better discrimination than by analyzing the individual blocks. We will also show that SO-PLS discriminant analysis yields some valuable interpretation tools that give additional insight into the data. We will introduce some new ways to graphically represent the information, taking into account both interpretation and predictive aspects.

1 Theory and Method

1.1 Sequentially orthogonalized PLS (SO-PLS)

The SO-PLS model can be written as $Y = X\beta_1 + Z\beta_2 + E$, where $Y$ are the response variables, $X$ and $Z$ are blocks of predictor variables, and $E$ are residuals. SO-PLS extracts information sequentially from each data block. The SOPLS algorithm starts by fitting $Y$ to $X$ by PLS regression. After that, the $Z$ block is orthogonalized with respect to the scores of the previous PLS. Therefore, the orthogonalized $Z$ is fitted to the deflated $Y$. In this way, it is possible to extract further information from $Z$ that explains the variance in $Y$. At the same time, since the space spanned by $X,Z$ and $X,Z_{orth}$ is the same, the orthogonalization does not represent any loss of information. For how the procedure is conducted, it is clear that the order of the blocks is important. In the last step, the full predictive model can also be written as function of the original measures, and it can be computed as explained in [1]. The number of latent variables used is decided independently for each block, usually by cross-validation. The combination of components is selected by evaluating the cross-validated prediction error in a so-called Måge plot.[2]

1.2 Linear Discriminant Analysis (LDA)

In classification, the objective is to assign objects to a number of different categories. Methods that require a training set where the categories are known in advance are called “supervised methods” or discriminant analysis. Linear Discriminant Analysis (LDA) is one of these discriminant analysis methods [2]. This methodology is based on the assumption that the probability that a sample belongs to a specific class follows a Gaussian distribution. The main limitation of LDA is that it requires a well-conditioned covariance matrix. This means that the method cannot be used when the number of variables exceeds the number of samples, or...
when the variables themselves are highly correlated. To overcome this limitation, it is possible to compress the original data into a few latent variables before applying LDA.

### 1.3 SO-PLS-LDA

This method involves the combination between LDA and SO-PLS. The first step in the proposed procedure is to calculate SO-PLS using a dummy Y-matrix and then classify by LDA. The strength of coupling these two techniques, in addition of what is already explained, is that the orthogonalization allows performing classification separately on each step of SOPLS.

By defining \( t \) the concatenated scores vector for a sample:

\[
t = [t_X \ t_{Z_{orth}}] \tag{1}
\]

the SOPLS-LDA classification criterion could be stated as:

\[
\max_g \left[ -1/2 \left( \mathbf{t} - \bar{\mathbf{t}}_g \right)^T \mathbf{S}^{-1} \left( \mathbf{t} - \bar{\mathbf{t}}_g \right) \right] \quad g=1,\ldots,G \tag{2}
\]

where \( g \) is the number of categories.

Accordingly to Equation (2), since the orthogonality of the two blocks results in the inverse covariance matrix being block-diagonal, the previous equation can be rewritten as:

\[
\max_g \left[ -1/2 \left( \mathbf{t}_X - \bar{\mathbf{t}}_g \right)^T \mathbf{S}_{X}^{-1} \left( \mathbf{t}_X - \bar{\mathbf{t}}_g \right) - 1/2 \left( \mathbf{t}_{Z_{orth}} - \bar{\mathbf{t}}_{Z_{orth}} \right)^T \mathbf{S}_{Z_{orth}}^{-1} \left( \mathbf{t}_{Z_{orth}} - \bar{\mathbf{t}}_{Z_{orth}} \right) \right] \tag{3}
\]

This is important for interpretative purposes. Indeed it implies that one can naturally look at the blocks separately and together for interpretation of where for instance the classification goes wrong or right. The comparison between the results of LDA obtained both overall and for individual samples, allows making several considerations on the interpretation of the dataset. So one can examine classification results obtained by each singular block and compare them with the aim of interpreting agreements/dissimilarities on responses. After that, looking at data in the space of canonical variates allows further interpretations.

### 2 Dataset

The dataset analyzed is composed by 60 samples of 42 different types of beer. Samples were analyzed by Mid Infra-Red, Near Infra-Red, Ultraviolet and Visible spectroscopies and by Thermogravimetry. Beers are both industrial and artisanal; produced in different brewery with different processes, of different geographical areas, and distributed both in bottle and in cans. Samples were subjected to some physical-chemical pretreatment before being analyzed by Thermogravimetry and UV spectroscopy.

### 3 Results

Samples are considered belonging to two classes, the first one, constituted by all the samples of a specific type of beer (called Class “Reale”), and the second class, with all the others (called Class “Others”). SOPLS models were created on the pre-treated data matrices. Of the different signals coming from the various analytical techniques, only three blocks at a time were used for the construction of each SOPLS model: one as Y, one of spectroscopic signals as X and one as Z.

The number of latent variables chosen for each model was selected as the one that gave the minimum error in cross validation. Predicted samples were then classified by LDA. SOPLS-LDA was able to predict with great accuracy classes of the beers under exam giving very low classification errors. The same dataset has been classified also by PLS-DA and MB-PLS-DA; often SO-PLS-LDA has given rise to higher percentages of correct classification.

### 4 References


Approximate likelihood inference in latent variable models
with application to the Health and Retirement Study

Silvia Bianconcini\textsuperscript{a,*}, Silvia Cagnone \textsuperscript{a} and Dimitris Rizopoulos\textsuperscript{b}

\textsuperscript{a}Department of Statistical Sciences, University of Bologna, Italy.
\textsuperscript{b}Department of Biostatistics, Erasmus University Medical Center, The Netherlands

\textbf{Keywords:} Laplace approximation, adaptive quadrature, dimension reduction method, generalized linear latent variable models.

1 Introduction

Generalized Linear Latent Variable Models (GLLVM) \cite{1} represent a useful tool in different fields of research
in which the constructs of interest are not directly observable, so that one or more latent variables are required
to reduce the complexity of the data. The relationship between a set of \( p \) responses or items, denoted for each
individual by the \( p \)-dimensional vector \( \mathbf{y}_l, l = 1, \ldots, n \), and a set of \( q \) latent variables or/and random effects \( \mathbf{b}_l \),
that are fewer in number than the observed variables \( q < p \), is specified through the probability associated to the
individual response pattern as follows

\[
    f(\mathbf{y}_l) = \int_{\mathbb{R}^q} g(\mathbf{y}_l | \mathbf{b}_l) h(\mathbf{b}_l) d\mathbf{b}_l
\]

(1)

where \( h(\mathbf{b}_l) \) is usually referred as the structural part of the model, generally assumed to be normal with null mean
and covariance/correlation matrix equal to \( \Psi \), and \( g(\mathbf{y}_l | \mathbf{b}_l) \) is usually referred to as the measurement part of the
model. The latter is taken from the exponential family (with canonical link function), and for the specific case of
binary items treated in this study it results

\[
    g(\mathbf{y}_l | \mathbf{b}_l) = \exp \left[ \mathbf{y}_l^T \eta_l - 1^T \log(1 + \exp(\eta_l)) \right]
\]

(2)

where \( \eta_l \) represents the systematic component of the model defined as \( \eta_l = \alpha_0 + \alpha^T \mathbf{b}_l \) with \( \alpha_0 \), being a \( p \)-
dimensional vector of item specific intercepts and \( \alpha \) is the factor loading matrix associated to the latent variables.

An analytical solution for the integral in eq. (1) does not exist. The aim of this paper is to propose an alternative
approach based on a reduction in the dimensionality of the integrals involved in the computations that allows
for significant computational savings. This approach is based on a fundamental theorem by \cite{2} that provides a
convenient way to represent the Taylor series expansion of the integrand up to a specific order without involving
any partial derivative. We study and discuss how to apply the method in generalized linear latent variable models
with application to the University of Michigan Health and Retirement Study (HRS).

2 Model estimation

Maximum Likelihood (ML) estimates in the GLLVM framework are typically obtained by using either the EM or
the Newton-Raphson algorithms. The key component for applying both the algorithms is the score vector of the

\*Corresponding author. E-mail: silvia.bianconcini@unibo.it.
observed data log-likelihood function. For a random sample of size $n$, the latter is defined as

$$\ell(\theta) = \sum_{i=1}^{n} \log f(y_i; \theta) \tag{3}$$

$$= -\frac{nq}{2} \log(2\pi) - \frac{n}{2} \log |\Psi| + \sum_{i=1}^{n} \log \int_{\mathbb{R}^q} \exp \left\{ y_i^T \eta_i - 1^T \log[1 + \exp(\eta_i)] - \frac{1}{2} b_i^T \Psi^{-1} b_i \right\} db_i,$$

where $\theta$ denotes the vector of model parameters, that is $\theta = [\alpha_0, \text{vec}(\alpha), \text{vec}(\Psi)]$.

### 2.1 The dimension reduction method

In this study, the Dimension Reduction Method (DRM) proposed by [2] is applied to approximate the log-likelihood function (3), that is the integral (1). Defining $L(b_i) = g(y_i | b_i) + h(b_i)$, and considering its Taylor expansion around the mode $\hat{b}_i = \arg\max_{b_i \in \mathbb{R}} L(b_i)$, the integral (1) results

$$\int_{\mathbb{R}^q} \exp[L(b_i)] db_i = (2\pi)^{q/2} |\hat{\Sigma}_e|^{1/2} \exp[L(\hat{b}_i)] \int_{\mathbb{R}^q} \exp[v(b_i)] \phi(b_i; \hat{\Sigma}_e, \hat{\Sigma}_e) db_i \tag{4}$$

where $v(b_i)$ includes all the terms of the expansion of higher order than the second one and $\phi(b_i; \hat{\Sigma}_e, \hat{\Sigma}_e)$ is the normal density function whose mean vector is given by the mode $\hat{b}_i$ and its covariance matrix is minus the inverse of the Hessian matrix of $L(b_i)$ evaluated at its mode. The DRM is then applied to approximate the integral in eq. (4). In the latter, the leading term in the RHS is exactly the Laplace approximation of the integral (1). It follows that the DRM approximation provides more accurate estimates than the classical Laplace approximation, due to the inclusion of higher than two order terms in the Taylor series expansion of $L(b_i)$.

To integrate on the domain of standardized variables, we perform the Cholesky decomposition of the covariance matrix $\hat{\Sigma}_e$, that is $\hat{\Sigma}_e = \hat{C}_T \hat{C}_T'$, and the standardized variables $b^*_i$ are derived as $b^*_i = \hat{C}_T^{-1}(b_i - \hat{b}_i)$. Based on DRM, the approximation of the likelihood function results

$$L \approx \prod_{i=1}^{n} (2\pi)^{q/2} |\hat{\Sigma}_e|^{1/2} \exp[L(\hat{b}_i)] \tag{5}$$

$$\sum_{i=0}^{s} (-1)^i \left( \begin{array}{c} q - s + i - 1 \\ i \\ \end{array} \right) \int_{\mathbb{R}^q} \sum_{k_1 < \cdots < k_{i-s}} v_{1x-i}(b^*_i) \phi(b^*_i) \cdots \phi(b^*_i) db^*_i \cdots db^*_i$$

where each integral depends on $s, s - 1, \cdots, 1$ variables, respectively, and can be easily approximated by using the classical Gauss-Hermite quadrature.

### 3 Simulation study and application to real data

The performance of DRM is evaluated by means of a Monte Carlo simulation study for the GLLVM for binary data. Two different scenarios, corresponding to models with 2 and 4 correlated latent variables, are considered in order to compare the performance of DRM, under different conditions, with other widely applied approximation techniques, that is Adaptive Gauss-Hermite (AGH) quadrature and Laplace approximation. Furthermore, the performance of the proposed method is evaluated on real data coming from the University of Michigan Health and Retirement Study (HRS). The latter is a longitudinal panel study that surveys a representative sample of more than 26,000 Americans over the age of 50 every two years. Supported by the National Institute on Aging and the Social Security Administration, the HRS explores the changes in labor force participation and the health transitions that individuals undergo toward the end of their work lives and in the years that follow.

### References


Study of the variance term in the quadratic risk of the PLS1 estimator

Carole Binard

" Université de Nice Sophia-Antipolis, Laboratoire J.A. Dieudonné, Nice, France

Keywords: PLS algorithm, PLS1 regression, quadratic risk, variance, Singular Value Decomposition

Introduction

The pioneering work in Partial Least Squares (PLS) was done in the late sixties by H. Wold in the field of econometrics. In the late seventies PLS method was pioneered by S. Wold and H. Martens for chemical applications. That algorithm is constructed on the properties of the NIPALS (Non-Linear Iterative Partial Least Squares) one [1]. It iteratively builds an orthogonal set of score vectors, called latent variables, by maximising the covariance between different sets of variables. That algorithm can be used to solve regression problems when the classical Least Squares method cannot be applied. In this case the exogenous $X$ and endogenous $Y$ variables are each considered as a set of variables. From these blocks, PLS extracts the latent variables considering as new predictors and regresses the endogenous variables on those new exogenous ones.

For the user of that algorithm, it is important to know what kind of properties to expect from it. Thus one important challenge is to identify the risk of the PLS estimator. In the following study we are interested in controlling the variance term in the quadratic risk of the PLS estimator when the dependent variable $Y$ is a vector.

1 First result

We consider the following model

$$Y = X\beta + \epsilon,$$  

(1)

with $Y$ a centred and scaled endogenous random variable of size $n$, $X$ a centred and scaled deterministic exogenous matrix of size $n \times p$ and $\epsilon$ a random vector of size $n$ such that $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. The rank of the matrix $X$ is denoted by $r$.

Thanks to the PLS1 algorithm in [2, Chapters 7 and 8] we compute $r$ latent variables $t_1, t_2, ..., t_r$. We denote by $\Pi_r$ the orthogonal projection onto the subspace $T$ generated by the set $\{t_1, t_2, ..., t_r\}$ and $\Pi_r Y = X\hat{\beta}_{PLS}$ an estimation of $Y$ by the PLS1 algorithm. Let us denote by $f$ the quantity $X\beta$ and by $\hat{f}$ the quantity $X\hat{\beta}_{PLS}$. Thus from the equation (1) we can compute the quadratic risk of the PLS1 estimator $\hat{\beta}_{PLS}$ after $r$ steps by

$$\mathbb{E} \left[ \|f - \hat{f}\|^2 \right] = \mathbb{E} \left[ \|f - \Pi_r Y\|^2 \right] = \mathbb{E} \left[ \|f - \Pi_r f\|^2 \right] + \mathbb{E} \left[ \|\Pi_r \epsilon\|^2 \right],$$  

(2)

Thanks to the orthogonality of the latent variables we rewrite the variance term in a sort of waterfall

$$\mathbb{E} \left[ \|\Pi_r \epsilon\|^2 \right] = \mathbb{E} \left[ \sum_{i=1}^{r} \|\Pi_i \epsilon\|^2 \right] = \sum_{i=1}^{r} \mathbb{E} \left[ \|\Pi_i \epsilon\|^2 \right].$$  

(3)

*E-mail: binard@unice.fr
We study the case where there is no missing data. By applying a Singular Value Decomposition (SVD) on the matrix $X$, we succeed in finding an upper bound for the variance term at step one

$$\mathbb{E} \left[ \| \Pi_t \epsilon \|^2 \right] \leq r \sigma^2. \quad (4)$$

2 Attempt of generalization, numerical results and perspectives

2.1 Attempt of a theoretical generalization

Let us denote by $X_0$ the previous $X$ matrix and $\Pi_{t_1}$ the orthogonal projection on the vector $t_1$. We compute the residual matrix $X_1$ of the linear regression of $X_0$ on $t_1$ by $X_1 = X_0 - \Pi_{t_1} X_0$. Then by applying a SVD on $X_1$ of rank $r - 1$ we obtain an upper bound for the second intermediate term of variance

$$\mathbb{E} \left[ \| \Pi_{t_2} \epsilon \|^2 \right] \leq (r - 1) \sigma^2.$$

So we define the deflated matrix at step $i$ by $X_i = X_{i-1} - \Pi_{t_i} X_{i-1}$, for $i = 1, ..., r$. And if we apply a SVD on those matrices we finally have the relationship

$$B_1 > B_2 > ... > B_r,$$

where $B_i$ is the upper bound of $\mathbb{E} \left[ \| \Pi_{t_i} \epsilon \|^2 \right]$ such that $B_i = (r - i + 1) \sigma^2$.

If we combine those results with equation (3), we get an upper bound of order $r^2 \sigma^2$. However we know that when we can perform the Least Squares method, the PLS estimate at step $r$ and the Least Squares estimate are the same [4, Chapter 9]. Thus the upper bound of order $r^2 \sigma^2$ seems to be huge. That result can be explained by the fact we did not take the nature of $X$ into account.

2.2 Numerical results

We study some specific cases where $X$ is a $n \times p$ matrix such that $n < p$ and whose columns are highly correlated. From simulations three main results appeared:

1. $\mathbb{E} \left[ \| \Pi_{t_i} \epsilon \|^2 \right]$ has an upper bound of order $\sigma^2$,

2. we observe a hierarchic order between the intermediate terms of variance until the step $h$

$$\mathbb{E} \left[ \| \Pi_{t_1} \epsilon \|^2 \right] > \mathbb{E} \left[ \| \Pi_{t_2} \epsilon \|^2 \right] > ... > \mathbb{E} \left[ \| \Pi_{t_h} \epsilon \|^2 \right], \text{ with } h \leq r,$$

3. the total variance $\mathbb{E} \left[ \| \Pi_T \epsilon \|^2 \right]$ has an upper bound of order $r \sigma^2$.

We note first that when we study particular matrices $X$ ($n < p$ and high correlations between columns) we confirm our result (4) and our theoretical remark written at the end of the subsection 2.1. Then, in some cases, we observe that the step $h$ from our second numerical result corresponds on the stopping step obtained by cross-validation.

2.3 Perspectives

Our next work will be focused on improving previous theoretical results to reach the hierarchy observed between intermediate terms of variance. Also we will study the bias term of (2) in order to finally obtain a stopping criterion different from those already developed (for instance [5]).

References

General latent variable models of PLS enhanced with spectral techniques and representations in Hilbert spaces

Marianna Bolla\textsuperscript{a}\textsuperscript{*}

\textsuperscript{a}Institute of Mathematics, Budapest University of Technology and Economics, Hungary

Keywords: Conditional expectation, singular value decomposition, low-dimensional representation, correspondence analysis, compromise factors

Introduction

A general (usually not linear) statistical model of the latent variable approach of PLS is stated in terms of joint distributions, and solved by the spectral or singular value decomposition (SD or SVD) of the Hilbert–Schmidt operator taking conditional expectation between the margins. In the possession of an i.i.d. sample from a multivariate distribution, the components of which are divided into two parts (in a symmetric or asymmetric way), we make inference via the empirical covariance and cross-covariance matrices (in the multivariate Gaussian case) or the normalized contingency table (in the case of categorical data). Compromise factors of several data sets are also looked for, based on a novel algorithm for maximizing the sum of heterogeneous quadratic forms.

1 Representation of joint distributions

Let \((\xi, \eta)\) be a pair of random variables defined over the product space \(X \times Y\) and having joint distribution \(W\) with margins \(P\) and \(Q\), respectively. Assume that the dependence between \(\xi\) and \(\eta\) is regular, i.e., their joint distribution \(W\) is absolutely continuous with respect to the product measure \(P \times Q\), \(w\) denotes the Radon–Nikodym derivative (notation of A. Rényi\textsuperscript{1}, see [1]). In the spirit of [2], \(H = L^2(\xi)\) and \(H' = L^2(\eta)\) denote the sets of random variables which are functions of \(\xi\) and \(\eta\), and have zero expectation and finite variance with respect to \(P\) and \(Q\). Both \(H\) and \(H'\) are Hilbert spaces with the covariance as inner product; further, they are embedded as subspaces into the \(L^2\) space defined likewise by the \((\xi, \eta)\) pair over the product space.

Let \(P_X : H' \to H\) and \(P_Y : H \to H'\) be the integral operators taking \textit{conditional expectation} between the two margins, \(P_X^\ast = P_Y\). Provided \(\int_X \int_Y w^2(x, y)Q(dy)P(dx) < \infty\), \(P_X\) and \(P_Y\) are Hilbert–Schmidt operators with SVD

\[
P_X = \sum_{i=1}^{\infty} s_i \langle \cdot, \phi_i \rangle_H \psi_i, \quad P_Y = \sum_{i=1}^{\infty} s_i \langle \cdot, \psi_i \rangle_H \phi_i, \quad P_X \phi_i = s_i \phi_i, \quad P_Y \psi_i = s_i \psi_i \quad (i = 1, 2, \ldots), \tag{1}
\]

where for the singular values \(1 > s_1 \geq s_2 \geq \cdots \geq 0\) holds, since the operators \(P_X\) and \(P_Y\) are in fact orthogonal projections from one margin onto the other, and \(\phi_i \in H, \psi_i \in H'\) are the corresponding function pairs.

\textsuperscript{*}E-mail: marib@math.bme.hu. Address: 1 Egry József, Bldg H5/2, 1111 Budapest, Hungary. Research supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project.

\textsuperscript{1}Alfréd Rényi, founder of the Mathematical Institute of the Hungarian Academy of Sciences, in 1959 published two seminal papers on the maximal correlation in the Acta Math. Acad. Sci. Hung. with titles “On measures of dependence” and “New version of the probabilistic generalization of the large sieve”.

37
1.1 When the role of the two spaces is symmetric

The pair \((X, Y)\) of \(k\)-dimensional random vectors with components in \(H\) and \(H'\), respectively, is called a \textit{k-dimensional representation} of the product space endowed with the measure \(\mathcal{W}\) if \(E_\mathcal{W}XX^T = I_k\) and \(E_\mathcal{W}YY^T = I_k\).

\textbf{Theorem 1.} With the notation of (1), the minimum of the cost \(E_\mathcal{W}\|X - Y\|^2\) of this representation is \(2\sum_{i=1}^k (1-s_i)\), and it is attained with the \(k\)-dimensional representation \(X' = (\psi_1, \ldots, \psi_k)\) and \(Y' = (\phi_1, \ldots, \phi_k)\).

If \(k = 1\), Theorem 1 gives the solution of the maximal correlation problem of A. Rényi: \(\max_{\psi \in H, \phi \in H'} \text{Corr}_{\mathcal{W}}(\psi, \phi) = \max_{i} |\psi_i| = 1 \text{ Cov}_{\mathcal{W}}(\psi, \phi) = s_1\) or equivalently, \(\min_{i} |\psi_i| = 1 \text{ \ and both are attained at the } \psi_1, \phi_1 \text{ pair. When } X \text{ and } Y' \text{ are finite sets, the solution corresponds to the SVD of the normalized contingency table } P^{-1/2}WQ^{-1/2}, \text{ and the representatives are obtained by the correspondence vector pairs. Since numerical algorithms are capable to find singular vector pairs orthogonal in Euclidean norm, when performing correspondence analysis, we have to normalize the underlying contingency table } W \text{ with the diagonal matrices } P \text{ and } Q \text{ containing the probabilities, corresponding to } I_P \text{ and } I_Q, \text{ in their main diagonals. If } \xi \text{ and } \eta \text{ are multivariate Gaussian, then their maximum correlation is the largest canonical correlation, realized by appropriate linear combinations of them. We can find canonical correlations successively; the procedure relies on the SVD of the matrix } C_{11}^{-1/2}C_{12}C_{22}^{-1/2}, \text{ where the covariance- and cross-covariance matrices } C_{11}, C_{22}, \text{ and } C_{12} \text{ are estimated from a sample.}

1.2 When the role of the two spaces is asymmetric

Let \(\psi\) be the response, and \(\phi\) the predictor, and only \(\|\psi\| = 1\) is assumed when \(\|\psi - \phi\|^2\) is minimized. Now the minimum is \(1 - \sum_{i} s_i^2\), attained with the \(\psi_i, s_i, \phi_i\) pair. This is the nonlinear regression problem, and based on a sample, the iteration of the ACE (Alternating Conditional Expectation) algorithm of [2] converges to the solution.

1.3 The case of a symmetric joint distribution

Here \(\mathcal{W}\) is a symmetric measure with margin \(P, H\) and \(H'\) are isomorphic, and \(P_X = \sum_{i=1}^k \lambda_i(\psi_i') \psi_i \psi_i'\), where \(\lambda_i \leq 1\), and \(P_X = \phi_i \phi_i'\), where \(\psi_i\) and \(\psi_i'\) are identically distributed \((i = 1, 2, \ldots)\).

\textbf{Theorem 2.} Assume that \(1 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0\). Then the minimum of \(E_\mathcal{W} \|X - X'\|^2\) subject to \(E_\mathcal{W} XX' = I_k\), where \(X\) and \(X'\) are identically distributed, is \(2\sum_{i=1}^k (1 - \lambda_i)\), and it is attained with \(X' = (\psi_1, \ldots, \psi_k)\).

A finite \(X\) belongs to the vertex-set of a weighted graph, with edge-weights \(w_{ii} = 0, w_{ij} = w_{ji} \geq 0 (i \neq j)\) summing to 1. The operator \(P_X\) corresponds to the normalized modularity matrix of the graph, and based on the low-dimensional representations of the vertices, spectral clustering techniques are to be used, see [1].

2 Compromise factors of independent samples and conclusions

Having \(k\) independent samples of underlying \(n\)-dimensional random vectors \(X_1, \ldots, X_k\) \((k \leq n)\), we are looking for compromise factors, i.e., linear combinations \(a_i^T X\) that maximize \(\text{Var}(\sum_{i=1}^k a_i^T X_i) = \sum_{i=1}^k a_i^T C_i a_i\) subject to \(a_i^T a_i = \delta_{ij} (i, j = 1, \ldots, k)\), where \(C_i\) is the covariance matrix of the \(i\)-th sample. In [3], we introduced a novel algorithm to find the compromise vectors \(a_i\)’s, the coordinates of which are loadings characterizing the samples in relation to the others; an application for finding compromise factors of three nephrotic stages will be presented.

To sum up, via the above representations we are modeling the relations between two or more sets of observed variables. Our theory extends to the non-Gaussian case, and the number of latent variables depends on the spectral properties of the underlying operators. Sometimes we use preliminary regularization to treat non-linearities in the data. This theory has applications in non-parametric regression (see [2]), correspondence analysis, and spectral clustering of social networks (see [1]). For the SD or SVD, fast numerical algorithms are at our disposal (e.g., the Lánczos method), which usually make use of the conjugate gradient method, a well-known PLS technique.

References


Multigroup analyses to study Cannabis Abuse Screening Test in thirteen European countries

Stéphanie Bougeard\textsuperscript{a} * Aida Eslami\textsuperscript{b} El Mostafa Qannari\textsuperscript{b} and Stéphane Legleye\textsuperscript{c}

\textsuperscript{a}French Agency for food, occupational and health safety (Anses), Ploufragan, France
\textsuperscript{b}Oniris, Nantes, France
\textsuperscript{c}National institute of demographic studies (Ined), Paris, France

\textbf{Keywords:} Multilevel data, multigroup PCA, multigroup PLS, cannabis consumption, social sciences.

\section*{Introduction}

We address the problem of describing multivariate datasets divided into groups of individuals, usually known as multilevel, hierarchical, clustered, nested or multigroup data. Standard multivariate methods such as PCA or PLS are based on the assumption that individuals come from homogeneous population and are independent from each other. In multigroup data, individuals from the same group are likely to be more similar than individuals from other groups and it is of paramount interest to take this structure into account. Multigroup methods can be divided into two categories depending on the aim of the study in regard with the group structure: the within-group methods where the group effect is discarded and the between-group methods related to the framework of discriminant analysis. We focus here on within-group analyses, the group effect being a structuring but not relevant effect in the analysis. The aims of within-group methods is to seek common parameters (e.g., common loadings across groups) as well as group parameters to understand the group specificity in comparison with the common structure. In this field, few methods are proposed and do not exactly match our aims. We focus herein on multigroup methods proposed by Eslami \cite{1,2}. In particular, the interest of two of them, namely multigroup Principal Component Analysis (mgPCA) and multigroup Partial Least Squares (mgPLS), is illustrated on the basis of a questionnaire aiming at studying the cannabis consumption among teenagers of thirteen countries.

\section{Multigroup methods}

Consider the multigroup setting where we have a single dataset $X$ involving $P$ variables and $N$ individuals \emph{a priori} divided into $M$ groups $X_m$ with $m = (1, \ldots, M)$. We assume that each dataset $X_m$ of dimensions $(N_m \times P)$ is column-centred. The aim of the analysis is to seek a common vector of loadings $a$ to study the relationships between the $P$ variables, these links being common to all the individuals. To better understand the group specificity in comparison with the common structure, the $P$ variables may also be viewed through their group vector of loadings $a_m$. Multigroup PCA consists in seeking a vector of loading $a$ common to all the groups, tightly linked to the $m$ group vectors of loadings $(a_1, \ldots, a_M)$ so as to maximize the criterion $\sum_{m=1}^{M} < a_m, a >^2$ with $a_m = X_m^T t_m$ and $\|t_m\| = ||a|| = 1$ \cite{1}.

Consider the multigroup context of regression where a dependent dataset $Y$ is predicted by an explanatory one $X$, both these datasets being \emph{a priori} partitioned into the same $M$ groups of individuals. The aim of the analysis is to seek common loadings to all the groups, namely $a$ and $b$, respectively for explanatory $X$ and dependent $Y$ datasets. Furthermore, to better understand the group specificity in comparison with the common structure, let $a_m$ and $b_m$ be the specific group vectors of loadings associated with $X_m$ and $Y_m$. Multigroup PLS seeks vectors of

\footnote{Corresponding author. E-mail: stephanie.bougeard@anses.fr}
loadings $a$ and $b$ common to all the groups such as their associated latent variables $u_m = Y_m b$ and $t_m = X_m a$ are tightly linked following the criterion to be maximized $\sum_{m=1}^{M} \text{cov}(Y_m b, X_m a)$ with $||a|| = ||b|| = 1$ [2]. The group loadings are then retrieved from $a_m = X_m' u_m$ and $b_m = Y_m' t_m$ with the constraints $||a_m|| = ||b_m|| = 1$.

2 Illustration

The interest of the multigroup methods is illustrated on the 2011 ESPAD European survey which aims at collecting data on alcohol and other drugs following the same protocol in various countries (www.espad.org). In 2011, one optional module in the questionnaire was the Cannabis Abuse Screening Test (CAST) [3] chosen by thirteen countries. The database consists of 5204 teenagers who reported having smoked cannabis in the last twelve months originated from these countries. The CAST explores various aspects of their consumption by means of six questions concerning the last 12 months. The core question of the ESPAD survey also comprises questions about various aspects of drug use and consumption context. The first aim is to investigate the relationships among the six variables from the CAST, common to all the countries. The second aim is to explain the CAST by the nine variables which describe drug use and consumption context. Multigroup PCA is performed to investigate the relationships among the six variables from the Cannabis Abuse Screening Test (CAST). As the six variables have different variances, data are globally centred and scaled to give the same weight to all the variables in the analysis. The country effect accounts for 20.9\% of the total variance (between-group part) and is removed from multigroup PCA thus focusing on the within-group part of data. As the countries have also different variances for the six CAST questions, data are centred and scaled by group to give the same weight to all the countries in the analysis. It follows that three dimensions could be retained since they explain 74.4\% of the total variance. The common loadings allows us to investigate the relationships between the six variables from CAST which are common to all the countries. Thereafter, the country specificity in regard with the common structure is given by means of the similarity index between common and group loadings. Multigroup PLS is conducted to explain cannabis consumption ($Y$ dependent dataset) from variables describing drug use and consumption context ($X$ explanatory dataset). For the reasons described beforehand, all the variables are globally centred and scaled as well as centred and scaled by group. The country effect accounts for 11.4\% of the total variance (between-group part) and is removed from multigroup PLS which focus on the within-group part of data. Only one dimension could be retained as it explains 96.7\% of the total variance.

3 Conclusion

In this paper, we propose to apply two innovative methods, called multigroup PCA and multigroup PLS, to the framework of multivariate datasets divided into groups of individuals. The proposed methods focus on the within-group part of the variance, the group effect being a structuring but non-relevant effect in the analysis. Both these methods are based on maximization criteria that reflect the objectives to be addressed. An appealing feature of these methods lies in the fact that they seek common and group loadings which may be useful for enhancing the interpretation of the outcomes or in graphical displays. The proposed methods and their associated interpretation tools are performed using code programs developed in the free software R. The methods mgPCA and mgPLS are powerful approaches to explore hierarchical data and may provide useful tools for exhibiting common structures as well as intuitively appealing graphical representations of similarities between groups of observations in situations such as cross-national surveys and international comparisons.

References


Surrogate models for optimization in high dimension using a mixed Kriging/PLS method

Bouhlel M. A.\textsuperscript{a*} and Bartoli N.\textsuperscript{a} and Morlier J.\textsuperscript{b} and Otsmane A.\textsuperscript{c}

\textsuperscript{a}ONERA, Toulouse, France
\textsuperscript{b}ISAE, Toulouse, France
\textsuperscript{c}SNECMA, Paris, France

\textbf{Keywords:} Kriging, PLS, design of experiment

\section*{Introduction}

Kriging \cite{1} is a popular metamodel used to replace a computationally expensive simulation model. However, with the development of modern scientific codes in the field of engineering, the number of variables increases in order to capture the physical behaviour and it is well known that Kriging performances are degraded in this condition. In order to overcome this problem, several methods for reducing the dimension are available, particularly the Partial Least Squares method (PLS) \cite{2}. PLS creates new variables (latent variables or principal components) by modelling the relationship between input and output variables while maintaining most of information in the input variables.

In this work, a new method is developed by combining Kriging and PLS. This method consists in using information from the PLS in order to build an efficient Kriging metamodel adapted to high dimension. Numerical experiments on both academic and industrial test cases show the efficiency of such method in terms of accuracy and CPU time up to dimension 60.

\section{Overview of Kriging and PLS}

From now, \(d\) is the dimension of the problem, \(n\) is the number of observations (design of experiments), \(h\) is the number of principal components to retain, \(x^{(i)} (i = 1, \ldots , n)\) is a point from the design of experiment and \(y\) is a vector containing true responses associated.

### 1.1 Kriging

A Kriging model is a generalized linear regression model since it accounts for the correlation in the residuals between the regression model and the observations. Correlation \(r_{uv}\) between two points \(u\) and \(v\) is given by

\[
r_{uv} = \prod_{j=1}^{d} \exp(-\theta_j |u_j - v_j|^2),
\]

where \(\theta_j (j = 1, \ldots , d)\) is estimated using maximum likelihood. The number of parameters \(\theta\) to determine equals the problem dimension \(d\). Thus, this estimation may be a difficult task if the number of variables is large.

The mathematical form of Kriging prediction at a point \(x\) is given by Equations (2) and (3)

\[
\hat{y}(x) = \hat{\mu} + r_x^T R^{-1} (y - 1\hat{\mu}),
\]

\textsuperscript{*}Corresponding author. E-mail: ONERA, France. mohamed.bouhlel@onera.fr.
with
\[ \hat{\mu} = \frac{1}{1^T R^{-1} 1} \] \[ R_{ij} = r_{x(i)x(j)}, \quad i, j = 1, \ldots, n, \quad x = (r_{xx(1)}, \ldots, r_{xx(n)})^T \] \[ 1 = (1, \ldots, 1)^T. \] 

### 1.2 Partial Least Squares

PLS is a statistical method that finds a linear relation between input variables and output variable (one considers only one output in this study) by projecting input variables to a new space, formed by new variables called principal components \( T \). PLS will try to find the best multidimensional direction in the \( X = (x(1), \ldots, x(n)) \) space that explains the behaviour of the output \( y \). Thus, principal components are given by \( T = XW \) with \( W \) the matrix containing in each column the loading vector for the construction of the principal component. PLS is suited when \( X \) has many variables (high dimension) to explain.

### 2 Kriging combined with PLS

This paper presents a new method -Kriging combined PLS- for computer simulation. This method builds an efficient and a fast Kriging. It uses informations (contained in \( W \)) extracted from PLS. The key point consists in substituting the spatial correlation function from the Eq. (1) to the new expression
\[ r_{u,v} = \prod_{j=1}^{h} \exp(-\theta d \sum_{i=1}^{d} |W_{i,j}| ||u_i - v_j||^2). \] 

The main advantage of this kernel is that the number of \( \theta \) to estimate will be reduced significantly from \( d \) to \( h \). Indeed the number of latent variables to retain is often less than 4. Also, this approach allows us to build an anisotropic spatial correlation matrix with few \( \theta_j \) to estimate.

### 3 Results

Academic problem (Griewank function) and two engineering problems are carried out. Relative error \( \left( \frac{||\hat{Y} - Y||}{||Y||} \right) \times 100 \), with \( Y \) and \( \hat{Y} \) are the vectors containing true responses and predicted responses of validation points, respectively) and CPU time (\( h \) refers to hour, \( mn \) refers to minutes and \( s \) to seconds) are reported in Table 1.

<table>
<thead>
<tr>
<th>Surrogate</th>
<th>Griewank function</th>
<th>Engineering problem</th>
<th>Engineering problem</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60D (800 sample points)</td>
<td>10D (1295 sample points)</td>
<td>24D (99 sample points)</td>
</tr>
<tr>
<td>error (%)</td>
<td>CPU time</td>
<td>error (%)</td>
<td>CPU time</td>
</tr>
<tr>
<td>Kriging</td>
<td>11.47</td>
<td>4 mn 53 s</td>
<td>5.37</td>
</tr>
<tr>
<td>Kriging/PLS (1 component)</td>
<td>7.4</td>
<td>6.88 s</td>
<td>5.07</td>
</tr>
<tr>
<td>Kriging/PLS (2 components)</td>
<td>6.04</td>
<td>12.57 s</td>
<td>5.02</td>
</tr>
<tr>
<td>Kriging/PLS (3 components)</td>
<td>5.23</td>
<td>16.82 s</td>
<td>5.34</td>
</tr>
</tbody>
</table>

The results of Kriging combined with PLS show a better accuracy and a faster CPU time than ordinary Kriging existing in literature. The main perspective is to couple this method with an efficient global optimizer (EGO algorithm) to carry out an optimization in high dimension.

### References


A regression-based approach to relate datasets’ characteristics to relative prediction accuracy of algorithms, with applications to PLS for omics data

Anne-Laure Boulesteix\textsuperscript{a}*

\textsuperscript{a}Ludwig-Maximilians-University of Munich, Munich, Germany

Keywords: benchmarking, meta-learning, statistical testing, high-dimensional data, PLS+LDA

It is well-known that the relative accuracy of prediction algorithms may depend on the dataset’s characteristics such as, e.g., size or structure. For example, statements such as, say, “method A performs particularly well in the case of small sample sizes, while method B copes well with highly correlated predictors” are common in the computational literature. However, if drawn from real data studies, these conclusions on the relation between datasets’ characteristics and relative prediction accuracy of different methods are often not statistically substantiated. They are typically based on a too small number of example datasets. In this talk I propose a regression-based statistical approach to this “meta-learning” issue that is otherwise typically addressed by computer scientists. The difference between the error rates of methods A and B is considered as a dependent variable and the datasets’ characteristics as independent variables. Simple power calculations within this framework clearly show that statements on the connection between datasets’ characteristics and relative performance of the methods require a large number of example datasets. These methods are illustrated through applications to binary classification based on high-dimensional omics data using linear discriminant analysis applied to PLS components.

*Corresponding author. E-mail: boulesteix@ibe.med.uni-muenchen.de.
Extending SCGLR to multiple regressor-groups: the Theme-SCGLR method

Bry X.\textsuperscript{a}, Trottier C.\textsuperscript{ab}, Mortier F.\textsuperscript{c}, Verron T.\textsuperscript{d}

\textsuperscript{a}Université Montpellier II, Montpellier, France \textsuperscript{b}Université Montpellier III, Montpellier, France \textsuperscript{c}Cirad, Montpellier, France \textsuperscript{d}ITG-SEITA, Fleury les Aubrais, France

Keywords: Components, GLM, Latent Variables, PLS Regression, SCGLR, SEER.

1 Data and model

We lately published a technique, Supervised Component Generalized Linear Regression (SCGLR) \cite{1} which, combining PLS regression with GLM estimation, extends the Iteratively Reweighted Partial Least Squares (IRPLS) method proposed by \cite{2} to a multivariate response scheme. We here propose an extension of SCGLR to multiple explanatory groups.

A set of $q$ random variables $Y = \{y_1, \ldots, y^q\}$ is assumed to depend on $R$ regressor groups $X_1, \ldots, X_R$, with: $\forall r, X_r = \{x^r_1, \ldots, x^r_p\}$, along with a group $T$ of additional regressors. Let $X = \{X_1, \ldots, X_R\}$. Every $X_r$ is assumed to contain an unknown set $F_r = \{f^r_1, \ldots, f^r_k\}$ of \textit{structurally strong} components $f^r_k := X_r u^r_k$ (i.e. that have a large variance under constraint $u^r_k M_r^{-1} u^r_k = 1$, $M_r$ being some s.d.p. matrix), that are useful to predict $Y$. Let $F = \{F_1, \ldots, F_R\}$. Each $y^k$ is modelled through a GLM taking $F \cup T$ as regressor set, with the assumption that $\{y_1, \ldots, y^q\}$ are independent conditional on $F \cup T$. All variables are measured on the same $n$ statistical units. The aim is to calculate $F$ and model $Y$ on it.

2 Estimating a GLM through the Fisher scoring algorithm (FSA)

2.1 Univariate GLM and FSA

For the GLM of some variable $y$, with regressor set $X$ and linear predictor $\eta = X \beta$, the log-likelihood associated with the $n$-sample is:

$$L(\delta, y) = \sum_{i=1}^n (y_i \delta_i - b(\delta_i) + c(y_i))$$

Then: $\mu_i = E(y_i) = \tilde{b}(\delta_i)$. Now, $\eta$ and $\mu$ are linked through link function $g: \forall i: \eta_i = g(\mu_i)$. The MLE of this model can be found through the FSA, that iteratively performs Weighted Least Squares on the following linearised model (LM) on iteration $t$:

$$M^{|t|}: z_{\beta^{|t|}} = X \beta + \zeta^{|t|}; E(\zeta^{|t|}) = 0, V(\zeta^{|t|}) = W_{\beta^{|t|}}$$

where: $z_{\beta^{|t|}} = X \beta + \left(\frac{\partial \eta}{\partial \mu}\right)^{|t|} (y_i - \mu_i^{|t|})$; \left(\frac{\partial \eta}{\partial \mu}\right)^{|t|} = \text{diag} \left(\frac{g' \left(\mu_i^{|t|}\right)}{u_i^{|t|}} \right)_{i=1,n}$

and: $W_{\beta^{|t|}} = \text{diag}(\sigma^2(\mu_i^{|t|}) b''((\mu_i^{|t|})))_{i=1,n}$.

\textsuperscript{a}Corresponding author. E-mail: Place Eugène Bataillon, 34095 Montpellier, France. xavier.bry@um2.fr.
2.2 Multivariate GLM with partially common predictor (MGLMPCP)

When \( y^1, \ldots, y^q \) are assumed to depend on linear predictors collinear to one another, the LM has to be altered to: \( \forall k = 1, \ldots, q; \ z^k = y_k f + \zeta_k \) with component \( f = Xu \) and \( u'M^{-1}u = 1 \), \( M \) being a given s.d.p. matrix, for identification. [3] altered the FSA in order to estimate this LM and showed that the estimate of \( u \) was the solution:

\[
Q : \max_{u \in \mathbb{R}^p, u'M^{-1}u = 1} C(u) \quad \text{where} \quad C(u) = \sum_k \cos^2 \theta_k (z_k; \{Xu\})
\]

3 Theme-SCGLR

3.1 Calculating components when \( \forall r = 1, \ldots, R : k_r = 1 \):

First, \( Q \) is replaced with an equivalent program: \( Q' : \max_{\forall r = 1, \ldots, R; \forall u_r \in \mathbb{R}^{pr}, u_r/M_r^{-1}u_r = 1} C(u_1, \ldots, u_R) \), where \( C(u_1, \ldots, u_R) = \sum_k \cos^2 \theta_k (z_k; \{X_1u_1, \ldots, XRu_R\}) \) and \( M_1, \ldots, M_R \) are given s.d.p. matrices. Then, we introduce the components' variances into it, getting:

\[
R : \max_{\forall r = 1, \ldots, R; \forall u_r \in \mathbb{R}^{pr}, u_r/M_r^{-1}u_r = 1} C(u_1, \ldots, u_R) \prod_{r=1}^R \|X_ru_r\|^2_W
\]

\( R \) is solved by solving in turn and iteratively programs:

\[
R_r : \max_{u_r \in \mathbb{R}^{pr}, u_r/M_r^{-1}u_r = 1} C(u_1, \ldots, u_R)\|X_ru_r\|^2_W
\]

We use a specific algorithm, proposed by [4], to solve \( R_r \).

3.2 Calculating several components per group

- We ensure zero-correlation of components by putting : \( f^m_r = X^m_r - u^m_r \) with \( X^0_r = X_r \) and \( X^m_r = \Pi_{f^m_r \perp} X^{m-1}_r \).
- Interpretable ranking of components is obtained through the local nesting principle (LocNes) presented in [4]. \( \forall r, k \), let \( F^k_r := \{f^1_r, \ldots, f^k_r\} \). LocNes states that component \( f^k_r \) must best complement all components currently available at that point, i.e. \( F^{k-1}_r \cup F^k_r \), together with \( T \), for predicting \( Y \). So, \( f^k_r \) will be found solving program \( R_r \) taking \( X^{m-1}_r \) instead of \( X_r \) and \( F^{m-1}_r \cup F^k_r \cup T \) instead of \( T \).

3.3 Backward component selection

Starting with an original model involving “too large” a number \( K_r \) of components in every \( X_r \):

1. Remove every \( f^k_r \) in turn, and compare the performances of all resulting models with those of the original model through cross-validation. Retain the best-performing model.
2. If the retained model is different from the original model, resume step 1. Else, stop.

References


The analysis of heterogeneity and cognitive structure of dementia by means of a factor mixture model

Cagnone S. $^a$ and Viroli C. $^b$

$^a$Department of Statistical Sciences, University of Bologna, Bologna, Italy.
$^b$Department of Statistical Sciences, University of Bologna, Bologna, Italy.

**Keywords:** Categorical data, Cognitive functioning, Latent variables, Mixture models.

**Introduction**

Generalized Linear Latent Variable Models (GLLVM) are a common way to model mixed data in the exponential family as function of latent variables and covariates [1]. In the GLLVM framework the association between the manifest variables is assumed to be fully explained by the latent variables through a generalized linear model where the link function depends linearly on the latent variables. The latent variables are typically assumed to be normally distributed. This assumption can be unrealistic in some cases, especially when data are heterogeneous and may be clustered into some groups, or when they describe characteristics that are extremely skewed distributed, as often occurs in social and psychological research.

Recently the statistical literature has witnessed an increasing interest on alternatives to normality for the latent variables in GLLVM or its submodels.

For instance, in the classical factor analysis, [2] investigated skew-normal factors with the aim to model the asymmetric distribution of student satisfaction towards university courses. [3] modeled the latent variables in confirmatory factor analysis by mixtures of Gaussians. Then, [4] and [5] extended the approach in an exploratory framework by proposing the so-called factor mixture model in which the factors are assumed to be distributed as a mixture model with varying component means but with class invariant loadings, intercepts and specific variances. [6] extended the factor mixture model by explicitly assuming that the factors underlying the set of continuous manifest variables follow a mixture of multivariate Gaussians with group-varying parameters.

In latent trait models for binary data [7] investigated the inadequacy of the normality assumption by using bootstrap combined with non-parametric maximum likelihood and, more recently, [8] assumed Gaussian mixture latent variables with the aim to detect the potential groups of observations.

In GLLVM [9] showed that the inappropriateness of the normality assumption can be due to non-linear dependence among the latent variables. [10] derived consistent semiparametric estimators of the latent variables without explicitly assuming a distributional model. They also showed how an inappropriate specification of the latent variables distribution can bias the estimates. More recently, [11] considered a semi-nonparametric specification for the density of latent variables which appears to be greatly flexible.

In this paper we present a GLLVM framework for mixed data (including binary, ordinal and count data) with a semi-parametric specification for the density of the latent variables given by mixtures of multivariate Gaussians. Thus, as in [11], the emphasis is on a more flexible specification of the density of the latent variables. But differently from it, our goal is to explicitly perform clustering of units in the latent space by a mixture model which will be cast in a ‘supervised’ framework to properly deal with known pre-existent classes of individuals.

The proposed approach is motivated by the following study. Every 2 years the Institute for Social Research at the University of Michigan conducts a nationally representative Health and Retirement Study (HRS) funded by
the National Institute on Aging. The study is designed to investigate the health, social and economic implications of the aging of the American population [12]. Between August 2001 and December 2003, a random sub-sample of 856 subjects was selected from the total sample frame of approximately 22,000 HRS individuals in order to participate to a more extensive study on Aging, Demographics and Memory Study (ADAMS), which represents the first population-based study of dementia in the United States [13]. The participants received a thorough in-home clinical and neuropsychological assessment leading to a diagnosis of normal, cognitive impairment but not demented (CIND), or dementia. Approximately 36% of subjects were diagnosed as normal, 28% were diagnosed with CIND, and 36% received a diagnosis of dementia. Our interest focuses on investigating the relation between some cognitive measurements taken by the HRS survey of 2002 and the diagnosis of the individuals. This issue can be achieved by a GLLVM with latent variables which may be interpreted as indirect measures of overall cognitive scores. Because of the heterogeneity of the participants into the three classes, it is not difficult to imagine that also their performance of cognitive activity may tend to cluster into different groups. This aspect can be described by assuming that the latent variables are distributed according to a mixture of multivariate Gaussians, which can be unsupervised or supervised depending if the known individual class membership is incorporated into the model. We will present a general formulation of the model, and then describe the more interesting supervised variant for the situation of known labels to analyze our real data example. Because the HRS is a very extensive study a wide range of health and demographic information has been also collected. We considered some of them (like the age, sex and education) as covariates in the GLLVM which may directly affect the observed cognitive measurements.

References


PLS-PM in opinion surveys when respondent minds generate size-effect (involvement axis) in the data

Furio Camillo and Valentina Adorno
University of Bologna, Bologna, Italy

Keywords: Size effect, pls-pm, scale perception, unidimensionality, involvement axis, CRM

Introduction

The basic idea of Structural Equation Models (SEM) [1] [2] is that complexity inside a system can be explained taking into account a network of causal relationships, defined according to a theoretical model, linking latent complex concepts, called Latent Variables (LV), each measured by several observed indicators usually defined as Manifest Variables (MV). The several estimation techniques, well known in literature, can be grouped into two different approaches to Structural Equation Model estimation: the PLS (Partial Least Squares) approach to Structural Equation Models, also known as PLS Path Modeling (PLS-PM), belongs to the so called component-based estimation procedure different from the classical covariance-based LISREL-type approach. The main aim of component-based methods is to provide an estimation of the latent variables in the model in such a way that they are able to properly explain the causal relationships defined by the path diagram structure and, at the same time, the most representative of each corresponding block of manifest variables.

A PLS path model is described by two models: a measurement model relating each latent variable and the corresponding manifest variables, and a structural model that takes into account the relationships among the latent variables. The measurement model of PLS-PM use the idea that a latent variable (LV) is an unobservable variable (or construct) indirectly described by a block of corresponding manifest variables (MV). Different types of measurement model are available, depending on the direction of the relationships between the latent variables and the corresponding manifest variables [3]: the reflective model, the formative model and the MIMIC model (a mixture of the two previous models).

In a reflective model the block of manifest variables related to a latent variable is assumed to measure a unique underlying concept. Each manifest variable reflects its latent variable: it plays a role of endogenous variable in the block specific measurement model. As the reflective block reflects the (unique) latent construct, it should be homogeneous and unidimensional. In the reflective way, the block of MVs is unidimensional in the meaning of factor analysis. In PLS path modeling a priori knowledge is incorporated in the algorithm. At least at the theoretical level, The MVs of one block can always be built in a way that they are all positively correlated. On practical data this condition has to be checked. A block is essentially unidimensional if the first eigenvalue of the correlation matrix of the block MVs is larger than 1 and the second one smaller than 1, or at least very far from the first one.

The first principal component can be built in such a way that it is positively correlated with all (or at least a majority of) the MVs. There is a problem with MV negatively correlated with the first principal component and the literature suggest that these MVs are inadequate to measure the LV and consequently be removed from the measurement model.

Size effect in opinion surveys and PLS-PM

As is well known in literature and in practical applications, in opinion surveys, evaluation of intangible constructs is conditioned to the perception of the measurement scale suggested to respondents. Some studies deal with impact of scale perception in measurement of opinions or satisfaction on results of a segmentation process or on decomposition in principal components of the total variability.
In fact, the matrix of correlations among the attributes is always composed of coefficients of a positive sign, as
the respondents tend to provide their ratings by referring to a personal and latent mean vote. In applications to
the “thémascope” this type of data, as in all phenomena with a size effect or a latent trend, is to find a systematic
very important first eigenvalue. Conservation of this large “size effect” therefore generates a cluster of subjects
who wish all attributes and one that does not want anyone in particular, namely that seemingly indifferent to
any composition of the product or service under consideration. This result is not acceptable, since a market
segmentation should build groups which differ among them for the mix of attributes overall favorite. A possible
solution is to transform original data using a not linear row-standardization of the input matrix, eliminating the
size effect and different perception of the useful range of the scale [4].

Semiometric studies of size effect in opinion about intangible variables show that it is phenomenon linked to
the questionnaire composition. The first studies on factor analysis in the early twentieth century concerned notes
tests of intelligence or academic subjects. Consider this example: the first axis is often the "general factor of
ability" sometimes called, but it was a matter of debate, "intelligence" [5]. With few exceptions, all notes are
positively correlated with each other. There is, as a first approximation, good students and bad students. The
former have higher scores on average in almost all subjects, the latter notes lower. So much simpler if the length
is measured and diameter of chicken eggs, we find that these two measures are correlated because the main
source of variability is the size of the egg: the first principal axis will oppose large eggs in small eggs, hence the
name size factor. The second principal axis, sometimes called "Form factor" will instead show that in equal size,
egg are more or less spherical. It is possible to demonstrate that such a phenomenon is superimposed to the
notation of the effects in the case of Sémiométrie [6].

Level of involvement of the respondent in fact is the real cause of the first principal component, and therefore
the pls-pm assumption of unidimensionality is verified because the presence of a large size effect. The paper
discusses the operational and theoretical aspect of pls-pm in the presence of a size-effect generating a strong
first principal component in the opinion data.

The modern frame of Customer Relationship Management and the Micro-marketing approach to the use of
“analytics” in business increasingly will define a scenario in which the Satisfaction will be used as a predictor
of individual behaviors in a more analytical (gps, web behavior, predictive sentiment analysis, social network
behavior) measuring system of the customer track. The paper presents an innovative strategy of the size effect
treatment based on the idea that the innate optimism of individuals is a latent variable to include in the pls-pm
model and that its manifest variables should be measured preliminarily by the inclusion in the questionnaire of
psychographic and semiometric questions about the life-values of respondents.

Real applications in banking sector, in retail business (loyalty studies) and in citizen satisfaction analysis will
be presented.

References

[1] D. Kaplan, Structural equation modeling: foundations and extensions, Thousands Oaks, California, Sage,
2000.


[3] C. Fornell, F.L. Bookstein, Two structural equation models: LISREL and PLS applied to consumer exit-

[4] F. Camillo, La costruzione di tipologie su preferenze espresse con voti: problemi di misurazione e alcune
soluzioni di ricodifica dei dati, Micro e Macro Marketing / vol. VIII, n. 3, 423-440, December 1999

Psychology, 15, 201–292, 1904

A partial least squares algorithm handling ordinal variables

Gabriele Cantaluppi\textsuperscript{a}\textsuperscript{*} and Giuseppe Boari\textsuperscript{b}

\textsuperscript{a}Università Cattolica del Sacro Cuore, Milano, Italy
\textsuperscript{b}Università Cattolica del Sacro Cuore, Milano, Italy

Keywords: Ordinal Variables, Partial Least Squares, Polychoric Correlation Matrix

Introduction

The partial least squares (PLS) technique is largely used in socio-economic studies where path analysis is performed with reference to the so-called structural equation models with latent variables (SEM-LV).

It often happens that data are measured on ordinal scales; a typical example concerns customer satisfaction surveys, where responses given to a questionnaire are on Likert type scales assuming a unique common finite set of possible categories.

In several research and applied works, averages, linear transformations, covariances and Pearson correlation coefficients are conventionally computed also on the ordinal variables coming from surveys. A more traditional approach would require appropriate procedures to be adopted in order to handle manifest indicators of the ordinal type. Within the well-known covariance-based framework, several approaches are suggested in order to appropriately estimate a SEM-LV model; in particular, [1], [2] and [3] make the assumption that to each manifest indicator there corresponds an underlying continuous latent variable.


As observed by [7] in the procedure by [4] a value is assigned to the impact of each explanatory variable on each category of the response, while the researcher may be interested in the impact of each explanatory variable on the response as a whole. The same issue characterizes the techniques illustrated in the Chapter 5 by [8].

The present proposal goes in this direction: a reformulation of the PLS path modeling algorithm is introduced allowing us to deal with variables of the ordinal type in a manner analogous to the covariance based procedures.

In this way we recall the traditional psychometric approach, by applying a method for treating ordinal measures according to the well-known [9] scaling procedure, that is assuming the presence of a continuous underlying variable for each ordinal indicator. The polychoric correlation matrix can be defined. We show that by using this matrix one can obtain parameter estimates also within the PLS framework.

When the number of points of the scale is sufficiently high the value of the polychoric correlation between two variables is usually quite close to that of the Pearson correlation; in these situations there would be no need to have recourse to polychoric correlations and the traditional PLS algorithm may be applied. However, to make the response of interviewed people easier, it is common practice to administer questionnaires whose items are measured on at most 4 or 5 point alternatives: in these circumstances the proposed modification seems to work better than the traditional PLS algorithm.

1 A reformulation of the PLS algorithm

We resume the sequence of steps defining the reformulation of the PLS algorithm. See [10] for a more detailed presentation of the algorithm, and on how to estimate latent scores in presence of ordinal manifest variables. We

\textsuperscript{*}Corresponding author. E-mail: gabriele.cantaluppi@unicatt.it.
underline that the algorithm has the characteristic of avoiding the determination, at each step, of the composites scores \( \hat{y}_{js} \) and of the instrumental variables scores \( z_{js} \).

Compute \( \Sigma_{XX} \) (which in case of ordinal items is the polychoric correlation matrix)

Define the starting weights resumed in the diagonal block matrix \( \mathbf{W} = [w_1, \ldots, w_j, \ldots, w_n + m] \).

**Iterative phase**

Set \( \mathbf{W}_{TEMP} = \mathbf{W} \)

Compute:

\[
\Sigma_{\hat{Y}\hat{Y}} = \mathbf{W}^T \Sigma_{XX} \mathbf{W}
\]

\[
\mathbf{sW} = \mathbf{W} \left( \left| \mathbf{W}^T \Sigma_{XX} \mathbf{W} \right| \ast \mathbf{I} \right)^{-1} = \mathbf{W} \left[ \Sigma_{\hat{Y}\hat{Y}} \ast \mathbf{I} \right]^{-1}/2, \text{ with } \mathbf{sW} \text{ the updated weights}
\]

\[
\Sigma_{\hat{Y}\hat{Y}} = \mathbf{sW}^T \Sigma_{XX} \mathbf{sW}
\]

\[
\Upsilon = \left( \mathbf{T} + \mathbf{T}' \right) \ast \text{sign}(\Sigma_{\hat{Y}\hat{Y}}), \text{ where } \mathbf{T} \text{ is the inner path matrix}
\]

\[
\Sigma_{XZ} = \Sigma_{XX} \mathbf{sW}\Upsilon
\]

\[
\Sigma_{X\hat{Y}} = \Sigma_{XX} \mathbf{sW}
\]

\[
\mathbf{C} = \chi_{\mathbf{W}} \ast \Sigma_{XZ}, \text{ where } \chi_{\mathbf{W}} \text{ is an indicator matrix with ones if } w_{ij} > 0
\]

\[
\pm = \text{sign} \left\{ \mathbf{1}'_{p+q} \left[ \text{sign}(\mathbf{\chi}_{\mathbf{W}} \Sigma_{XF}) \right] \right\}
\]

Update the weights \( \mathbf{W} = \mathbf{C}[\text{diag}(\mathbf{1}'_{p+q} \mathbf{C})]^{-1} \text{diag}(\pm) \)

Obtain \( \mathbf{sW} \)

Check if \( ||\mathbf{W} - \mathbf{W}_{TEMP}|| < \varepsilon \).

**References**


Principal Covariates Clusterwise Regression

Eva Ceulemans \textsuperscript{a}*, and Eva Vande Gaer\textsuperscript{a} and Henk A. L. Kiers\textsuperscript{b} and Iven Van Mechelen\textsuperscript{a} and Tom F. Wilderjans\textsuperscript{a}

\textsuperscript{a}KU Leuven, Leuven, Belgium
\textsuperscript{b}University of Groningen, Groningen, the Netherlands

Keywords: Multicollinearity, dimension reduction, clusterwise regression, multilevel data

Introduction

In the behavioral sciences, many research questions pertain to a regression problem in that one wants to predict a criterion on the basis of a number of predictors. Although in many cases ordinary least squares regression will suffice, sometimes the prediction problem is more challenging, for three reasons: First, many predictors can be available, making it difficult to grasp their mutual relations as well as their relations to the criterion. In that case, it may be very useful to reduce the predictors to a few summary variables, on which one regresses the criterion and which at the same time yield insight into the predictor structure. Second, the population under study may consist of a few unknown subgroups that are characterized by different regression models. Third, the obtained data are often hierarchically structured, with for instance observations being nested into persons. Although some methods have been developed that partially meet these challenges (i.e., Principal Covariates Regression -PCovR-, clusterwise regression -CR-, and structural equation models), none of these methods adequately deals with all of them simultaneously. To fill this gap, we propose the PCCR method, which combines the key ideas behind PCovR \cite{Jong1992} and CR \cite{Spath1979}. The PCCR method is validated by means of a simulation study and by applying it to data gathered in daily life on eating disorders.

References


\textsuperscript{∗}Corresponding author. E-mail: Eva.Ceulemans@ppw.kuleuven.be.
PLS Path Modelling of the molecular origins of gene expression noise

Guilhem Chalancon*  
*MRC Laboratory of Molecular Biology, Cambridge, United Kingdom

Keywords: Gene Expression Noise, PLS-PM, Systems Biology, Regulatory Genomics

Introduction

In this study, we propose to introduce the use of Partial Least Squares Path Modelling (PLS-PM) in Regulatory Genomics and Systems Biology. In particular, we seek to investigate the molecular origins of a poorly understood phenomenon: gene expression noise. Recent years of single-cell research have unveiled that clonal cells maintained in fixed conditions express many of their genes in a surprisingly fluctuating manner, raising the question of what causes fluctuations (i.e. gene expression noise) to vary in a gene-specific manner [1-2]. To date, little is known about what in the architecture and the information content of genes specifically contributes to making their expression noisy. In addition, a quantitative understanding of how such features contribute to expression noise is lacking.

Here we integrated single-cell measurements of expression noise together with data sets that describe attributes of expression and of regulatory control in thousands of yeast genes, ranging from mRNA synthesis to protein degradation. Using PLS-PM, we show that determinants of transcript stability, translation efficiency and protein stability contribute to expression noise in comparable proportions to that of transcription initiation, thereby bringing new light on how expression noise might be regulated. Further, we show that these molecular features have an additive impact on expression noise, suggesting that their presence or absence in a single gene might yield predictive amounts of expression noise. These results have two implications: first, they demonstrate that PLS-PM can successfully be applied to the study of gene expression regulation. Second, they indicate that it might be possible to genetically engineer expression noise at any level of the central dogma of Biology.

1 Data Integration and Model Construction

Newman et al. measured expression noise for >2,000 of yeast genes, by computing the coefficient of variation in protein abundance of thousands of clonal cells expressing a gene’s protein tagged with a fluorescent reporter [2]. We integrated these data with >250 features that have high coverage in the yeast genome and that relate to the control of gene expression (Figure 1A-C). To study how such features influence variation in expression noise across hundreds of genes, we built a path model that accounts for the influence of molecular abundance and expression dynamics on expression noise [1]. The structural model was composed of formative blocks describing the four sequential stages that characterise gene expression regulation (mRNAs birth, mRNA death, protein birth, and protein death), and reflective blocks describing outcomes in expression (Figure 1D).

2 Global model and analysis

We applied a greedy feature selection algorithm in order to select molecular features that are predictive of abundance – in addition to being biologically relevant – and thereby defined our formative blocks. We then applied PLS-PM and validated outer weights, loadings and path coefficients by bootstrap, arriving at a model where 59% of the variance in noise is explained. Furthermore, we identified specific manifest variables whose weights and direct effects suggest that they impact on expression noise, and showed that individual genes carrying those features tend to have higher levels of noise, irrespective of their abundance and expression dynamics. Together, these results open the road for experimentally validating the identified predictors, which
could empower the genetic engineering of expression noise, leading to applications in both fundamental biology and biotechnology.

Figure 1. Modelling of gene expression noise in yeast. A Gene Expression Noise was quantified as the coefficient of variation (CV) in fluorescence of clonal cells expressing a protein with a fluorescent tag. B Genome-wide measurements of steady-state abundance and expression dynamics of both mRNAs and proteins were integrated to model four reflective blocks. Corresponding data sources are indicated, and the labels of each block are shown in D. C Gene-specific Molecular Features with a regulatory effect on expression were integrated from various sources, and assigned to four distinct formative blocks. D Path model obtained for a set of 26 manifest variables (out of which 13 are molecular features). Red and blue edges indicate negative and positive $\beta$, respectively. Bootstrapping was systematically applied on weights, loadings and path coefficients. All analyses were performed in R with in-house routines and the ‘plspm’ library.

References


Time-Variant Effect on PLS Regression Models of Complex Industrial Systems

Chan, R.\textsuperscript{a}, Pratley A. J.\textsuperscript{b}, & van Voorthuysen, E. J\textsuperscript{a}.

\textsuperscript{a} University of New South Wales, School of Mechanical and Manufacturing Engineering, Sydney, Australia
\textsuperscript{b} University of Sydney, Business School, Sydney, Australia

Keywords: PLS regression, Time-variant effect, Statistical process control

1 Introduction

With the evolution of technology and increase in complexity of machinery and equipment comes the evermore-important requirement to be able to characterise and understand complex behaviour [1], and to understand how this changes as a function of operational use and time. It is of particular interest to industry since, in the absence of appropriate decision support systems, critical machine operating and maintenance decisions are often clouded by human subjectiveness [2]. A complex system is a process that can include many different input and output measures as well as interactions, apparent as well as hidden. Good decision-making requires knowledge and understanding. This in turn requires information, analysis, data and transparency. Furthermore, continuous improvement as well as adaptive control and process management relies on information feedback and ongoing process characterisation and identification. PLS modelling has been used to characterise an industrial process, in terms of a set of key process performance and operational characteristics.

2 Time-Variant Effect on PLS Regression Models

The aim of this paper is to investigate the time-variant effect on important parameters in PLS regression models. These parameters include the loadings and weights for reflective indicators and the path coefficients (regression coefficient) between two components in an inter-battery (Mode A) PLS regression analysis. According to literature, a well-defined construct has to satisfy both theoretical and empirical criteria [3]. Theoretical considerations include the nature of the construct, the direction of causality, and the characteristics of the indicators. Empirical considerations include indicator error, inter-correlation and co-linearity. It is common, not only in PLS-based modelling, that, once a construct is defined and the path coefficients result in an acceptable fit, in this case an acceptable R-squared value, the model can be considered robust and ready for use in a predictive or explanatory sense. This paper argues that, for models in general and PLS in particular, the validity based on observation data collected in a particular time period may not necessarily hold over time. Furthermore, the way that a component is defined for a particular construct may not hold its validity and robustness over time. This research aims to test whether the quality of PLS models can be monitored and maintained using charting techniques combined with statistical process control.

This research is supported by a medical supply manufacturer that produces intravenous bag for patients with Continuous Ambulatory Peritoneal Dialysis (CAPD). The production line consists of twenty-two work stations designed as an automated transfer-line, where the bulk of materials handling, indexing and process operations are performed by pneumatic actuation controlled by programmable logic controllers (PLC) under varying degrees of sophistication in terms of open-loop as well as closed-loop control. The nature of the machine is considered to be complex due to the non-linear performance and interaction of more than six hundred
pneumatic cylinders. One major concern that is faced by the operator is a less than optimal level of overall equipment effectiveness (OEE) [4] in recent years.

To support the argument and hypothesis put forward, a Mode A regression analysis was conducted on an observation set based on process data collected between two key work stations on this production line. Performance data were collected over a period of five and a half years and then divided into eleven sub-groups for analysis. Results clearly demonstrate the time-variant nature of the behaviour of the process and the variant behaviour of the PLS model derived from this data. This leads to the development of a method to monitor the critical PLS parameters and to ensure the ongoing validity of a PLS-model representation of the production process.

References


Wynne W. Chin\textsuperscript{a*}, Jason B. Thatcher\textsuperscript{b}, Ryan T. Wright\textsuperscript{c} and Doug Steel\textsuperscript{d}

\textsuperscript{a} C. T. Bauer College of Business, University of Houston, Houston, TX, U.S.A., 
\textsuperscript{b} College of Business and Behavioral Science, Clemson University, Clemson, SC, U.S.A. 
\textsuperscript{c} School of Management, University of Massachusetts - Amherst, Amherst, MA, U.S.A., 
\textsuperscript{d} School of Business, University of Houston Clear Lake, Houston, TX, U.S.A.

Keywords: Abstract format, template, PLS 2014 Proceedings, Word, max 5 keywords

Introduction

Common Method Variance (CMV) continues to be an important issue for social scientists \cite{1, 2, 3, 4, 5}. To date, methodologists have yet to agree upon a best practice for detecting and controlling for CMV. In a recent paper \cite{6}, the unmeasured latent marker variable approach, a frequently employed technique, was shown to be incapable of detecting or controlling CMV in PLS analyses. Spurred by that paper, a new MLMV approach was introduced as a way for teasing out CMVs influence on structural paths\cite{7}. The process requires collecting multiple unrelated measures at the same time as collecting data related to the primary research model. This contrasts with the unmeasured latent marker variable approach \cite{6}, which uses indicators of manifest variables from the primary research model to estimate CMVs influence. According to the authors, a critical aspect of the MLMV approach is to select a set of measures that reflect underlying constructs that have no nomological relationship with the particular study in question while using the same survey format and scale to reflect the common method effects. These measures in turn are modeled as capturing an underlying CMV and labeled as an MLMV. In their paper, Chin et al. \cite{7} provided two approaches for minimizing CMV: a construct level correction (CLC) and an item level correction (ILC). They demonstrated the efficacy of these approaches using simulated data. To the best of our knowledge, empirical examples have not been conducted. Thus, this paper presents the results of a survey study using response scale formats commonly employed in social science research (i.e., Likert and Semantic). Our results demonstrate the viability of the MLMV approach in both providing an estimate of and controlling for the effects of CMV.

References


* Corresponding author. E-mail: wchin@uh.edu


PLS and Functional Neuroimaging: Bias and Detection Power across different Resampling Schemes

Nathan W. Churchill\textsuperscript{a,*}, Babak Afshin-Pour\textsuperscript{a}, Stephen C. Strother\textsuperscript{a,b}

\textsuperscript{a} Rotman Research Institute, Toronto, Canada
\textsuperscript{b} University of Toronto, Toronto, Canada

Keywords: neuroimaging, split-half, bootstrap, resampling, task-PLS

Introduction

Functional neuroimaging studies use PLS to estimate shared information between brain function and extrinsic variables (i.e. subject groups, task conditions). However, high-dimensional neuroimaging data are challenging to analyze, leading to ill-posed analyses. Moreover, significant noise confounds are often present in the data. It is therefore critical to assess the reliability of latent variables (LVs) extracted by PLS, commonly using empirical significance estimates obtained via resampling schemes. The most widely-used PLS model employs Bootstrap resampling (BR) \cite{1}; however, this model produces biased estimates. Two alternatives have recently been proposed, based on split-half subsampling methods: a model that estimates the stability of reconstructed LVs (SR) \cite{2}, and a model that estimates the reproducibility of independently-estimated LVs (SI) \cite{3}. We evaluate BR, SR and SI resampling schemes using simulated functional Magnetic Resonance Imaging (fMRI) data, and compare the models under a range of different simulated experimental parameters.

1 Task PLS and resampling

We compared resampling methods using task-PLS. For each of \(C\) different task conditions, we obtain \(N\) image volumes. The images are averaged per group and stacked in \((P\) voxels \(x\) \(C\)) matrix \(D\). We then perform Singular Value Decomposition (SVD) \(D = U \Lambda V^T\), where \(u_k, v_k\) form the \(k\)th pair of spatial and condition LVs, respectively. For BR, we resample subjects with replacement to compute \(D^*\) and estimate singular values \(\lambda_k^*\), averaged over 1000 resamples. For SR, we estimate \(U, V\) from full dataset \(D\), then randomly split data into halves \(D_1, D_2\). We reconstruct \(U_i = D_i V S^{-1}\) and \(V_i = D_i^T U S^{-1}\) \((i = 1,2)\) in each split, then measure correlations \(r_{\text{spat}}(u_{1k}, u_{2k})\) and \(r_{\text{cond}}(v_{1k}, v_{2k})\), averaged over 100 resamples. For SI, we randomly split data into halves \(D_1, D_2\), independently estimate \(D_i = U_i \Lambda_i V_i^T\) \((i = 1,2)\) per split, and measure average \(r_{\text{spat}}\) and \(r_{\text{cond}}\) for 100 resamples. For all models, reliability \(p\)-values are obtained by comparing test statistics \((\lambda_k^*\) or \(r_{\text{spat}}, r_{\text{cond}}\)) against a null distribution, obtained by randomly permuting subjects between groups for 1000 resamples. We compute \(p\)-values per model for 100 simulated datasets, and compare median \(p\)-values of BR, SR and SI.

2 Simulation modelling

We performed task-PLS on simulated fMRI data, using the framework developed in \cite{4} to model functional brain networks. For a fixed set of 3 task LVs, we varied four parameter settings: signal variance (defined as the proportion of noise variance) \(V = (0\text{ to }4)\); number of image volumes per condition \(N = (10\text{ to }80)\); and number of conditions \(C = (6\text{ to }20)\). We also simulated a data heterogeneity confound, by randomly permuting condition labels, which created a percentage of “null” images \(N_{\text{null}} = (0\%\text{ to }60\%)\).

3 Comparing resampling models

Figure 1 displays median \(p\)-values for spatial eigenimages \(u_k\) \((r_{\text{spat}})\) and Bootstrapped significance of \(\lambda_k^*\). The models all show comparable sensitivity as a function of signal variance \(V\). However, BR and SR models have
lower sensitivity compared to SI, for small samples $N$, small numbers of conditions $C$, and greater percentages of “null” data $N_{null}$. Figure 2 displays median $p$-values for condition eigenvectors $v_k$ ($r_{cond}$) and Bootstrapped significance of $\lambda_k^*$. For the BR and SR models, reliability $p$-values of $v_k$ are higher than for $u_k$, whereas they remain stable for SI. Moreover, SR estimates of $v_k$ are less sensitive than BR for parameters ($V, C, S_{null}$).

Figure 1. heat maps of median reliability $p$-values for a single “spatial” LV $u_k$, for BR (Bootstrapped reliability of $\lambda_k^*$), SR (Split-half, reconstructed LVs from $r_{spat}$) and SI (Split-half, independently estimated LVs from $r_{spat}$).

Figure 2. heat maps of median reliability $p$-values for a single “condition” LV $v_k$, for BR (Bootstrapped reliability of $\lambda_k^*$), SR (Split-half, reconstructed LVs from $r_{cond}$) and SI (Split-half, independently estimated LVs from $r_{cond}$).

4 Conclusions

We evaluated the sensitivity of a set of resampling approaches for simulated fMRI data. Model choice has a greater impact on detection for “condition” LVs $v_k$, compared to “spatial” LVs $u_k$. Moreover, independent split-half estimation provides the greatest sensitivity over a wide range of simulated experimental designs.

References


The impact of social media on brand value
- an empirical study.

Anatoli Colicev*, Peter O’Connor*, Vincenzo Esposito Vinzi*

*Essec Business School, Paris, France

**Keywords**: Social media, PLS Path Modelling, Facebook

**Introduction**

Facebook has the third largest population on Earth (after China and India), reaching one billion users in September 2012. Half a billion users collectively send more than 400 million tweets per day on Twitter and four billion minutes of videos are watched by over 800 million people on YouTube every month. As consumers increasingly turn away from traditional media [3], the question is raised on how this shift affects brand value, one of the critical success factors for many companies. The effect of social media on brand and brand value is potentially strong [4]. One of the key challenges is justifying the investment required to meaningfully engage in social media marketing [5]. This paper aims to address this research gap, with the objective of developing a framework to assess the impact of social media on brand value. It focuses on two research questions: (1) How to measure the effect of participating in social media channels on brand value? and (2) Which social media efforts have a positive effect on brand value. Data from both Interbrand and Brand Finance are used to measure brand value, with the study population defined as Interbrand’s top 100 brands. Social media actions metrics were selected using a portfolio of automated tools, with Partial Least Squares path modeling (PLS PM) used to validate the resulting model. We argue that our selected constructs have a positive influence on brand value, and that the more the brand invests in communicating and marketing to consumers over social media using these specific metrics, the higher will be its brand value.

**1. Design and Methodology**

The intended sample for this study was Interbrand’s 100 most valuable brands. Data on brand value was collected from Interbrand and Brand Finance brand value rankings. Data on social media activity of the target population was using a series of automated web based tools each of which had been previously validated by manually crosschecking its output against actual data from the brand’s social media presence. We analyzed only the official presence on each system, excluding regional pages from the analysis. When no official global brand page existed we used the US brand page. Data were collected for 31st of December 2012 and on the 31st December of 2013. The variables in 2012 and 2013 represent a cumulative value of a metric such as a total number of likes in 2012 and 2013 or a total number of retweets on the end of 2012 and 2013. Then each variable was operationalized as a difference between 2013 and 2012 value. The software package XL-Stat was used to apply PLS PM and evaluate the model. Overall, in our model we obtain a good internal consistency, convergent and discriminant validity and therefore we can proceed to evaluating the structural model. Moreover the significance of the loadings has been assessed using the bootstrap procedure. The results of 5000 samples indicate that all the loadings are significant at 0.05 level.

**2. Figures, tables and references**

In Figure 1 we present the conceptual model and in Table 1 the results of this mode. The hypotheses H2, H3, H4 and H6 are supported. We can see that the hypotheses that are not supported are H5 (p-value=0.425) and H7 (p-value=0.224). Moreover we find an opposite effect of what we expected for the first hypotheses. [7] suggests that the path values should not be lower than 0.20 and in our case this criterion is respected.

* Corresponding author. E-mail: b00410707@essec.edu, Essec Business School, Paris
Figure 1 Conceptual model

Table 1 Results of the structural model

<table>
<thead>
<tr>
<th>Construct</th>
<th>Path Coef.</th>
<th>T-stat</th>
<th>Hypothesis</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Actions on Facebook to Brand Value</td>
<td>-0.21</td>
<td>-2.39</td>
<td>H1</td>
<td>Not Supported</td>
</tr>
<tr>
<td>Brand Actions on Facebook to Brand Value</td>
<td>0.27</td>
<td>2.73</td>
<td>H2</td>
<td>Supported</td>
</tr>
<tr>
<td>Brand Actions on Facebook to User actions on Facebook</td>
<td>0.40</td>
<td>3.97</td>
<td>H3</td>
<td>Supported</td>
</tr>
<tr>
<td>User Actions on YouTube to Brand Value</td>
<td>0.69</td>
<td>8.66</td>
<td>H4</td>
<td>Supported</td>
</tr>
<tr>
<td>Brand Actions on Twitter to Brand Value</td>
<td>-0.28</td>
<td>-1.12</td>
<td>H5</td>
<td>Not supported</td>
</tr>
<tr>
<td>Brand Actions on Twitter to User actions on Twitter</td>
<td>0.21</td>
<td>1.98</td>
<td>H6</td>
<td>Supported</td>
</tr>
<tr>
<td>User Actions on Twitter to Brand Value</td>
<td>-0.09</td>
<td>-0.80</td>
<td>H7</td>
<td>Not supported</td>
</tr>
</tbody>
</table>

References


Using PLS for Variable Selection: A Simulation Study

Ian Cox\(^a\) and Marie Gaudard\(^b\)

\(a\)SAS Institute Inc., Manchester, UK
\(b\)North Haven Group, Orlando, USA

Keywords: VIP, Variable Selection, Simulation

Introduction

The difference between ‘explaining’ and ‘predicting’ has recently been articulated again as part of a more general framework discussing information quality (InfoQ, [1]). Explaining involves establishing cause and effect relationships and unbiased modeling, whereas predicting requires minimizing prediction errors and provides tools for managing operational systems. If the ultimate goal of an analysis is improvement to, or innovation of, the system under study, there is a strong argument that explaining is more important than predicting.

Explanation necessarily implies the partitioning of measured variables into two (or more) groups, which we designate \(Y\) (outcomes or outputs), and \(X\) (inputs). Motivated by the desire for parsimony, a key goal of explanation is the selection of only that subset of variables in \(X\) that are active, in the sense that they really do have an impact on \(Y\). Given its importance, the question of variable selection has received considerable attention in many branches of statistical modeling [2, 3]. For example, in the case of multiple linear regression (MLR), a standard approach is to use stepwise regression [4]. It should be noted, though, that even within this framework there is considerable complexity and sophistication, leading to a number of alternative prescriptions [5, 6]. Although discussed several times [7, 8], in the case of partial least squares (PLS) detailed studies of variable selection seem to be less common than in other methodological areas.

The current paper describes a simulation study in which data from random realizations of a known model were subject to PLS analysis, active variables being selected using several prescriptions found in the literature [9, 10]. This selection was then compared with the known model to determine false positive and false negative selections. To broaden the validity and utility of the results, various classes of known model were considered, along with an analysis that considered the variables in \(Y\) collectively (PLS2) and separately (PLS1).

The Structure of the Study

In terms of mimicking data likely to arise in real studies with a synthetic model, and denoting \(Xs\) (\(Ys\)) as the list of variables in \(X\) (in \(Y\)), there are several important things to consider:

1. The numbers of \(Xs\), \(Ys\), and observations.
2. The correlation structure of the \(Xs\).
3. The number of active \(Xs\), and how the active and inactive \(Xs\) are arranged within the correlation structure.
4. The functional form that links the active \(Xs\) to the \(Ys\).
5. The level of intrinsic noise.

Our simulation study attempts to deal with all these aspects in a limited but useful way. All simulations were done in JMP [11], but used JMP’s scripting language (JSL, [12]) directly rather than any of the inbuilt JMP platforms.

Corresponding author: ian.cox@jmp.com
The study considered 6 Ys \((nY = 6)\) and 20 Xs \((nX = 20)\), with 20 observations for each combination. Two types of X correlation structure were considered, denoted as cType = 1 and cType = 2: The first has only positive correlation values decaying away from the leading diagonal, the second has both positive and negative values in bands parallel to the leading diagonal. Details of the first (second) correlation structure are controlled by the parameters \(\rhoX\) and \(\text{divisorFac}(\text{peakSize} \text{ and } n\text{Peak})\), but note that \(\text{divisorFac}\) was held fixed at 0.25. An X matrix was generated by simulating values from a multivariate Gaussian with the chosen correlation structure.

The number of active Xs \(nX\text{active}\) was assigned, and a coefficient budget \(\text{coeffTot}\) was used to constrain the \(nX\text{active}\) terms in their relationship with the Ys. The locations of the \(nX\text{active}\) terms in the list of \(nX\) terms was controlled by another variable, \(\text{coeffType}\), described in the full paper.

Once coefficients were assigned to the \(nX\text{active}\) terms, true values for all the Ys were calculated and concatenated with the X matrix generated above to make a combined matrix. Random Gaussian noise was added to the Ys to make the matrix to be used for PLS modeling. PLS1 and PLS2 models with a different number of latent factors, \(a\), were fit via NIPALS [14] and SIMPLS [15] estimation. Finally Cutoff values of 0.8 and 1.0 were applied to the calculated VIP [16] and VIP* [17] values for each term to select those Xs considered as active. Using a full factorial combination of the simulation parameters mentioned above led to design matrix with 360,000 rows, each row generating 1,000 random realizations.

**Selected Simulation Study Results**

Figure 1 shows the overall error rate in the identification of active Xs for cType = 1 and NIPALS estimation. It clearly shows the superiority of choosing a VIP cut-off of 1.0 to segregate active and inactive terms. In contrast, an unscientific survey of the literature tends to advocate the value of 0.8.

Other figures in the full paper show the impact on the overall error rate of factors 1 to 5 above, along with the details of the PLS fitting procedure and the choice of VIP* rather than VIP as a summary metric relating to the PLS projection. More granular results for true positives and false positives are also presented.

![Figure 1. Overall error rates in isolating active terms using VIP and NIPALS for cType = 1.](image)

**References**

Estimating and correcting optimism bias in multivariate PLS regression: application to the study of the association between Single Nucleotide Polymorphisms (SNPs) and multivariate traits in Attention Deficit Hyperactivity Disorder (ADHD)

Erica Cunningham\textsuperscript{a}, Antonio Ciampi\textsuperscript{a,}\textsuperscript{*}, Ridha Joober\textsuperscript{b} and Aurelie Labbe\textsuperscript{a}

\textsuperscript{a}McGill University, Montreal, Quebec, Canada
\textsuperscript{b}Douglas Mental Health University Institute, Verdun, Quebec, Canada

Keywords: multivariate PLS, bias, correlation, genetic data, ADHD

Introduction

PLS regression has become a useful and increasingly popular tool in genomics, in particular in the study of the associations between genotypes and phenotypes. Given two sets of measurements \( Y \) and \( X \), multivariate PLS determines two sets of orthogonal linear combinations of variables, \( T = (t_1, t_2, \ldots) \) and \( U = (u_1, u_2, \ldots) \) such that the correlations \( \rho_k = \text{corr}(t_k, u_k) \) are maximized. In genomic studies, the \( \rho_k \)'s are used as measures of association between genome-level measurements, \( X \), and phenotype-level measurements, \( Y \). Since the correlations are the parameters of central scientific interest, it is important to study the statistical properties of their PLS estimates.

One can expect that they are vulnerable to optimism bias, i.e. they may overestimate their actual values due to potential overfitting. In this work, we evaluate the optimism bias by a limited simulation and assess the effectiveness of a bootstrap-based bias correction. The simulation is inspired by a real data analysis problem: the study of the association between SNP measurements (genotype) and a number of behavioural and cognitive measurements (phenotype).

1 Simulation to estimate bias

The model of Figure 1 was used as a basis for the simulations. It is a realization of a simple situation, in which one gene affects behaviour and another gene affects cognition without pleiotropy. A linear combination of 30 SNP variables defines a genetic score \( t_1 \), and a linear combination of 7 behavioural traits defines a behaviour phenotype score \( u_1 \); similarly, a linear combination of nine (distinct) SNP variables defines another genetic score \( t_2 \), and a linear combination of nine cognitive traits defines a cognitive phenotype score \( u_2 \). The coefficients of these linear relationships are known as loadings. The relationship between genotype and phenotype is embodied in \( \rho_1 \) and \( \rho_2 \), the correlations between \( t_1 \) and \( u_1 \), and \( t_2 \) and \( u_2 \), respectively. We have implemented several biologically motivated scenarios by varying the values of the correlations and of the loadings.

2 Results

For each scenario, we simulated the \( X \) and \( Y \) variables in samples of size 300 according to the corresponding model and added an error term, the variance of which was fixed at three different levels. We applied multivariate PLS regression to each sample, obtaining an estimate of \( \rho_1 \) and \( \rho_2 \), hence the difference between the estimates and the true values. The bias was calculated by taking the average of these differences over 500 replications.

As can be seen in Figure 2, bias is present. In general, the bias tends to increase when there is more noise in the SNP or trait information (when there are uninformative SNPs or traits). Bias tends to be larger when estimating a true correlation of zero compared to a non-zero correlation.

\textsuperscript{*}Corresponding author. E-mail: 1020 Pine Avenue West, Montreal, QC, Canada. antonio.ciampi@mcgill.ca.
Figure 1. Model for the simulation study.

Bootstrap methods [3] were incorporated into the simulations in order to correct for the estimated bias. Table 1 presents the results of the bootstrap correction.

Table 1. Bias in correlations between X scores \((t_1\) and \(t_2)\) and Y scores \((u_1\) and \(u_2)\) in simulations.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Simulated correlation</th>
<th>Estimated bias</th>
<th>Bootstrap-corrected bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\rho_1 = \text{corr}(t_1, u_1))</td>
<td>0.2</td>
<td>0.1436</td>
<td>0.0461</td>
</tr>
<tr>
<td>(\rho_2 = \text{corr}(t_2, u_2))</td>
<td>0.2</td>
<td>0.1522</td>
<td>0.0737</td>
</tr>
<tr>
<td>(\rho_1 = \text{corr}(t_1, u_1))</td>
<td>0</td>
<td>0.4125</td>
<td>0.3154</td>
</tr>
<tr>
<td>(\rho_2 = \text{corr}(t_2, u_2))</td>
<td>0</td>
<td>0.3911</td>
<td>0.3147</td>
</tr>
</tbody>
</table>

References


Quantile PLS Path Modeling

Cristina Davino\textsuperscript{a*} and Vincenzo Esposito Vinzi\textsuperscript{b}

\textsuperscript{a}University of Macerata, Italy
\textsuperscript{b}ESSEC Business School, France

\textbf{Keywords:} quantile regression, pls path modeling

\section*{Introduction}

The methodological framework of the paper is represented by one of the most widespread methods used to study complex concepts that cannot be directly measured: PLS (Partial Least Squares) path modeling (PLSPM) \cite{1} \cite{2} \cite{3}. It is a consolidated method able to provide a combination of a set of indicators taking into account both the main aspects of the complex phenomenon and their relative importance. Such a method is in essence based on the use of linear combinations of the indicators and of a unique system of weights. Moreover the PLS algorithm is an iterative process based on simple and multiple OLS regressions.

The paper aims to introduce quantile regression (QR) \cite{4} \cite{5} in the estimation process of a PLS path model. It is a matter of fact that QR allows to analyze the impact that the regressors exert on the whole distribution of the dependent variable. The proposed quantile PLS path model can enhance PLSPM potentialities because it is able to distinguish regressor effects on the different parts of the dependent variable distribution. For example, in the business/market framework, it can be interesting to evaluate if and how much the impact of consumer preferences is different among very, medium or low satisfied customers thus differentiating leverages to increase the satisfaction. Moreover, introducing QR in a PLSPM allows to handle heteroskedastic relationships among the dependent variables and the regressors. This event, frequently occurring when the dependent variable is highly skewed, is typical in subjective measurements.

\section{PLS Path Modeling and Quantile Regression}

PLS path modeling is a statistical approach for modeling complex multivariate relationships among observed (also known as manifest variables, MVs) and latent variables (LVs) that are actually defined as composites. Typically, such a modeling can be faced through Structural Equation Models that allow the estimation of a theoretical network of linear relationships linking latent complex concepts, each measured by means of a number of MVs. PLS path modeling represents a consolidated approach alternative to Structural Equation Models where linear predictive relationships are privileged. The idea behind the model is to measure the LVs through the MVs and to describe the dependence relationships among the LVs. A PLS path model is made of a measurement model relating each set of MVs to the corresponding LVs and a structural model connecting the LVs in accordance with a network of linear relationships.

Let $\xi_j$ be one of the $J$ LVs, it is measured by a set of $x_{jh}$ ($h = 1, \ldots, H_j$) MVs and impacted by a set of $\xi_{ji}$ ($i = 1, \ldots, I_j$) LVs. In the measurement model, each LV is a linear function (i.e. a composite) of its MVs:

\begin{equation}
\xi_j = \sum_{h=1}^{H_j} w_{jh}x_{jh}
\end{equation}

\textsuperscript{*}Corresponding author. E-mail: cristina.davino@unimc.it.
In the structural model, a set of linear equations allows to relate LVs:

$$\xi_j = \beta_{j0} + \sum_{i=1}^{l_j} \beta_{ij}\xi_i + \epsilon_j \quad (2)$$

The PLSPM algorithm is based on an iterative process. According to the original Wold’s PLS approach, starting from an arbitrary vector of weights, they are used for the external estimation of LVs that are then normalized to unitary variance. These LVs are then updated by considering the relationships with adjacent latent variables in the dependence network so as to yield internal estimates. Upon convergence of the iterative procedure, the next step carries on the estimation of the structural equations by individual ordinary least squares regressions or PLS regressions [6].

Quantile regression (QR), as introduced by Koenker and Basset in 1978, may be considered as an extension of classical least squares estimation of conditional mean models to the estimation of a set of conditional quantile functions. Quantile regression allows indeed estimating the conditional quantiles of a response variable \(y\) as a function of a set \(X\) of covariates. Although different functional forms can be used, the paper will refer to linear regression models. The QR model for a given conditional quantile \(\theta\) can be formulated as follows:

$$Q_{\theta}(y|X) = X\hat{\beta}(\theta) \quad (3)$$

where \(0 < \theta < 1\) and \(Q_{\theta}(\cdot|\cdot)\) denotes the conditional quantile function for the \(\theta^{th}\) quantile. The parameter estimates in QR linear models have the same interpretation as those of any other linear model: each \(\beta_i(\theta)\) coefficient can be interpreted as the rate of change of the \(\theta^{th}\) conditional quantile of the dependent variable distribution per unit change in the value of the \(i^{th}\) regressor.

## 2 Quantile PLS Path Model

QR is proposed as an estimation method both in the measurement and in the structural part of a PLSPM. The result is a quantile PLS path model where the regressors can play a different role if referred to units with different values of the dependent variable. It follows that, for each quantile of interest, a set of coefficients for both the measurement and the structural part is obtained. The evaluation of the statistical significance of the differences among the coefficients related to the different quantiles can be carried out exploiting the classical tools to test interquantile differences [7] as well as the bootstrap and permutation-based approach classically used in PLSPM.

A further strength of the proposed method is the lack of error distribution assumptions required by QR so that in the parameter-testing phase QR makes use of distribution-free statistical inferential procedures for estimating and testing parameters. Moreover, the proposal will allow detecting extreme values.

## References


A Markov-switching regression model with non Gaussian errors for investigating stock market behavior

L. De Angelis and C. Viroli*

Department of Statistical Sciences, University of Bologna, Italy

Keywords: Markov Switching, Normal Inverse Gaussian Distribution, Risk Measure

Introduction

In the recent financial crisis in 2007-2009 the strong interdependence among the different institutions of the financial system has been one of the aspects of greatest concern. Trying to understand the nature of relationships between financial institutions in a market and investigating the impact of one institution on the global market are two fundamental steps that allow to quantify the so-called systemic risk [1].

In this work we use Markov-Switching Regression models (MSR) to assess the underlying dependence structure of asset returns during the recent financial crisis. Markov-Switching Regression models are particularly suitable for describing the time series dynamics because they incorporate a latent discrete random variable whose states correspond to distinct dynamic patterns during different time periods [2].

Several authors have shown that MSR models play a relevant role in investigating regime switching of returns and volatilities in stock markets because of their ability to capture frequently observed stylised facts (see [3, 4, 5, 6] among the others).

In this work we analyze the dynamic behaviour of the returns of the main systemically important financial institutions with the aim of measuring the interconnectedness in the finance and insurance sectors.

Since it is well-known that the return distribution is skewed and leptokurtic, we consider the Normal-Inverse Gaussian distribution (NIG) [7] to model the conditional distribution of the observations, as an alternative to the Gaussian distribution and the Generalized Error Distribution (GED). The NIG distribution has four parameters and can take a variety of different shapes. In so doing, it can account for many extremal observations, skewness and heavy tails in the data.

The model

Consider a vector of observed returns, \( r_t \), with \( t = 1, \ldots , T \) and a set of \( q \) explanatory variables \( x_t \). Let \( s_t \) be a latent discrete variable taking the values in the domain \( \{ 1, \ldots , M \} \) where \( M \) is the number of latent states, interpretable as market regimes. We assume that the regressors can affect both the observed returns and the transition probabilities among the different states whose latent process is governed by a Markov chain, which can thus be either homogeneous or non-homogeneous.

Given the status of a latent discrete variable, \( s_t = i \), with \( i = 1, 2, \ldots , M \), a general MSR model specification is given by:

\[
    r_t | \{ s_t = i \} = \mu_i + \theta_i x_t + \varepsilon_t,
\]

\(*\)Corresponding author. E-mail: cinzia.viroli@unibo.it, Department of Statistical Sciences, via delle Belle Arti, 41, University of Bologna, Italy.
where $\mu_i$ are the switching intercepts and $\theta_i$ the vectors regression coefficients, that in this general formulation they are states-varying. Alternatively, in a restricted formulation they could be assumed to be constant across the regimes. The innovations $\varepsilon_t$ are assumed to be distributed according to different probabilistic models with regime specific parameters. We consider the Gaussian distribution, and the more flexible GED and NIG distributions. This latter probabilistic model has a specific parameter to capture skewness, besides very extreme tails. More specifically, the innovation term is modelled as a NIG distribution if it has density:

$$f(\varepsilon_t; \alpha_i, \beta_i, \xi_i, \delta_i) = \frac{\alpha_i}{\delta_i} \exp \left( \delta_i \sqrt{\alpha_i^2 - \beta_i^2} - \beta_i \xi_i \right) \phi(\varepsilon_t)^{-1/2} K_1 \left( \delta_i \alpha_i \phi(\varepsilon_t)^{1/2} \right) \exp(\beta_i \varepsilon_t),$$ (2)

where $\xi_i$ are location parameters assumed to be zero, $\delta_i$ are scaling parameters, $\beta_i$ are asymmetry parameters and the quantities $\alpha_i \pm \beta_i$ determine the heaviness of the tails. Moreover, $\phi(\varepsilon_t) = 1 + \left[(\varepsilon_t - \xi_i)/\delta_i \right]^2$ and $K_1(\nu)$ denotes the modified Bessel function of the third kind of order 1 evaluated at $\nu$. By taking all of some parameters constant across the regimes a family of possible and more parsimonious sub-models can be defined. The choice between the different distributions and their parameter specification is empirically tested on real data.

The approach is applied to the analysis of return series of several stocks of financial and insurance companies in the United States. We consider as regressors a set of $q$ macroeconomic and financial indicators. Model (1) and its multivariate extension are estimated through the EM algorithm. Once the maximum likelihood estimates have been obtained, the standard errors of the regression coefficients are computed by evaluating the observed information matrix in order to detect the significant covariates within each regime.

The proposed approach provides important insights in understanding the relationships among the main financial institutions and, in so doing, it eases the measurement of the systemic risk.

References


Identifying common and distinctive processes underlying multiset data

Katrijn Van Deun\textsuperscript{a}\textsuperscript,* and Age K. Smilde\textsuperscript{b} and Henk A.L. Kiers\textsuperscript{c} and Iven Van Mechelen\textsuperscript{a}  
\textsuperscript{a}KU Leuven, Leuven, Belgium  
\textsuperscript{b}University of Amsterdam, Amsterdam, The Netherlands  
\textsuperscript{c}University of Groningen, Groningen, The Netherlands

Keywords: multiblock, multiset, common and distinctive, data integration

Abstract

In many research domains it has become common practice to rely on multiple sources of data pertaining to the same set of entities. Examples include a systems biology approach to immunology with collection of both gene expression data and immunological readouts for the same set of subjects, and the use of several high-throughput techniques for the same set of fermentation batches. A major challenge is to find the processes underlying such multiset data and to disentangle therein the common processes from those that are distinctive for a specific source. Several integrative methods have been proposed to address this challenge including canonical correlation analysis \cite{1}, simultaneous component analysis \cite{2}, OnPLS \cite{3}, generalized singular value decomposition \cite{4, 5}, DISCO-SCA \cite{6, 5}, and ECO-POWER \cite{7}. To get a better understanding of the relations between these methods, this paper brings the methods together and compares them both on a theoretical level, as well as in terms of analyses of high-dimensional micro-array gene expression data obtained from subjects vaccinated against influenza.

References

\begin{itemize}
  \item \cite{1} A. Tenenhaus and M. Tenenhaus, “Regularized generalized canonical correlation analysis,” \textit{Psychometrika} \textbf{76}(2), pp. 257–284, 2011.
\end{itemize}

\textsuperscript{*}Corresponding author. E-mail: katrijn.vandeun@ppw.kuleuven.be
Component-based Redundancy Path Modelling

Pasquale Dolce\textsuperscript{a}, Vincenzo Esposito Vinzi\textsuperscript{b}, Carlo Lauro\textsuperscript{a}

\textsuperscript{a} University of Naples “Federico II”, Naples, Italy
\textsuperscript{b} ESSEC Business School, Paris, France

Keywords: PLS Path Modeling, Redundancy Analysis, Component-Based Method

1 Introduction

Partial Least Squares Path Modeling (PLSPM) is a method aimed to model a network of dependence relationships between blocks of variables [1] where each block is summarized by a construct, i.e. a linear composite of its own manifest variables.

In the literature, it is known that PLSPM presents some inconsistencies in terms of coherence with the direction of the relationships specified in the path diagram [2]. The directions of the links in the inner model do not play a role in the algorithm apart from the specific case of the so-called path weighting scheme for the inner estimation.

In order to show such inconsistencies of PLSPM, let us consider the case of three blocks of variables, consisting of two explanatory blocks $X_1$ and $X_2$, and a block $Y$ to be explained. Whether we establish path directions from the two blocks $X_1$ and $X_2$ to the block $Y$ or from the block $Y$ to the two blocks $X_1$ and $X_2$, the PLSPM algorithm produces the same results, in terms of weights and loadings.

Even though PLSPM analyzes networks of dependence relationships among constructs, the estimation process amplifies interdependence among them. PLSPM provides composite scores that are as much correlated as possible to each other while being somehow representative of each corresponding block of manifest variables. In the search for optimally correlated constructs, PLSPM misses to distinguish between dependent and explanatory blocks.

In order to overcome such incoherence of PLSPM, we propose a more suitable non-symmetric approach that aims at maximizing the explained variance of the dependent manifest variables in one block given the others, i.e. a new approach based on the optimization of a redundancy-related criterion in a multi-block framework [3]. Currently, the methodological core of our approach is based on redundancy analysis [4], but it could also exploit other multivariate explicative statistical methods (e.g., PLS2 regression, ridge regression, PCR and so forth so on), in order to inherit their prediction oriented objective [5, 6] as well as a direct estimation of the path coefficients.

In this perspective, we propose a new algorithm based on extracting and utilizing all the information in the blocks that is relevant to maximizing the explained variances of manifest variables in dependent blocks.

2 Extensions of Redundancy Analysis

Given an explanatory block $X$, and a block $Y$ to be explained, the redundancy analysis proposed by Wollenberg (1977) [4], derives successively orthogonal components of the predictors $X$ which optimally explain the variance of the $Y$-variables. Redundancy analysis as developed by Wollenberg shows how the optimal $X$-components should be chosen, but it does not provide $Y$-components simultaneously with the $X$-components.

Following this argument, Johansson (1981) [3] suggested two alternative transformations for the $Y$ set which are naturally associated with the transformation for the $X$ set. Given the weights $w$, defining the $i$-th $X$-component, the corresponding $Y$-component is defined via another vector of weights, $v$, which satisfy desirable orthogonality properties. In particular, the two solutions proposed by Johansson are directly based on the optimal $X$-components from redundancy analysis, for which the condition $w^\prime X^\prime X w = 0$ is fulfilled. The first solution, based on a least squares approach, satisfies the orthogonality condition between components of the same block so that
\[ v_i^\prime Yv_j = 0, \ i \neq j, \] but not the orthogonality condition between components of different order across blocks so that \( w_i^\prime X Yv_j \neq 0, \ i \neq j. \) The second solution, based on a restandardized procedure, fulfills opposite conditions, so that \( v_i^\prime Yv_j \neq 0, \ i \neq j, \) and \( w_i^\prime X Yv_j = 0, \ i \neq j. \)

We exploit the first solution proposed by Johansson because the specific orthogonality conditions are useful for interpretation purposes and then we generalize the analysis to more than two blocks.

## 3 The Method

Our approach is based on a multi-step algorithm. At first, we extract as many components as allowed by the ranks of blocks from each exogenous block, based on a redundancy analysis with respect to adjacent dependent blocks as defined by the inner model relationships. Then, the components of each dependent block are extracted by means of a redundancy analysis applied to the manifest variables of the dependent blocks with respect to the components of the exogenous adjacent blocks extracted at the previous step. This is then repeated for the subsequent dependent blocks where the sequence is defined by the prediction flow specified in the inner model.

At the end of the second step, the manifest variables of all the blocks are replaced by components. Then, the same steps are applied by replacing the original manifest variables with the newly extracted components so as to update these components. As some of the blocks play a role of both exogenous and endogenous block, we apply redundancy analysis considering these components firstly as dependent variables, then as exploratory variables.

In this new approach, we overcome the theoretical difference between formative and reflective schemes for the measurement model. We only make a distinction between explanatory blocks and dependent blocks, in the sense of redundancy analysis. Furthermore, we do not assume unidimensionality within each block.

We apply a backward selection procedure on the set of the extracted components, in order to remove noise, to simplify interpretability and to yield final components’ scores.

We can also integrate useful variable selection features, since we can identify the manifest variables that do not improve the capability of the model in terms of variance explained on the dependent variables and thus implement the method as a sparsifying tool.

In order to assess the quality and validity of results, we provide a new goodness-of-fit index based on redundancy criterion and prediction capability together with a classical bootstrap-based inferential approach.

Finally, we show the functioning of the proposed algorithm (implemented in a R code) through a simulation study and a real data application in the area of healthcare performance, and also compare it to PLSPM and other component-based methods such as Generalized Canonical Correlation Analysis [7] and Generalized Structured Component Analysis [8].

## References


Kronecker Products within NIPALS algorithms

Pasquale Dolce\textsuperscript{a}, Laura Trinchera\textsuperscript{b}, Vincenzo Esposito Vinzi\textsuperscript{c}, Carlo Lauro\textsuperscript{a} and Alfred O. Hero\textsuperscript{d}

\textsuperscript{a} University of Naples “Federico II”, Naples, Italy
\textsuperscript{b} NEOMA Business School, Campus Rouen, France
\textsuperscript{c} ESSEC Business School of Paris, Cergy-Pontoise, France
\textsuperscript{d} University of Michigan, Ann Arbor, USA

Keywords: Three-Way Factor Analysis, Kronecker Principal Component Analysis, NIPALS Algorithm

Introduction

Three-way data are characterized by observation of different variables under several conditions. Three-way data are indexed by three different indices. Thus, \( y_{ipq} \) may denote the score of observation \( i \), with respect to the variable \( p \), under the condition \( q \). Extension of the two-way factor model to the three-way factor model has been well studied in literature \cite{1}. In his thesis \cite{3}, and in some subsequent work \cite{2}, Lohmöller proposed an iterative algorithm for Kronecker matrix decomposition that he called the Kronecker Principal Component (KPC) algorithm. Here we discuss a method which use the KPC algorithm combined with the Non-linear Partial Least Squares (NIPALS) algorithm \cite{4}, as an alternative method for factor analysis.

1 The Kronecker Principal Component model

According to Lohmöller \cite{2, 3} a three-way array (\( A \)) can be decomposed as a Kronecker product of two matrices (\( B \) and \( C \)) following the model:

\[
A = B \otimes C + E.
\]

where \( B \) and \( C \) are called Kronecker principal component matrices of \( A \), if they are estimated under the least squares criterion:

\[
LS = \text{trace}(E' E) = \text{min}.
\]

The model in Equation (1) is called the Kronecker principal component (KPC) model.

Lohmöller and Wold (1980) provided an iterative algorithm for estimation of the model \ref{1} and proved the convergence.

2 KPC and NIPALS algorithm for Three-Way Factor Analysis

Let us assume that we observed \( P \) variables over \( Q \) conditions. Thus, we have a total of \( P \times Q \) observed variables. Let \( y_{(Q \times P) \times 1} \) denote a vector of manifest variables with covariance matrix \( S_{(Q \times P) \times (Q \times P)} \) and the variables indexed by a combined index:

\[
y_{(q,p)} \quad q = 1, \ldots, Q; \quad p = 1, \ldots, P.
\]

Let us denote with \( H_1 \) the number of components we use for reducing the space of the conditions (mode 3), and with \( H_2 \) the number of components we use for reducing the space of the variables (mode 2). \( \eta_{(H_1 \times H_2) \times 1} \) denote a vector of latent variables with covariance matrix \( \Phi_{(H_1 \times H_2) \times (H_1 \times H_2)} \) and the variables indexed by a combined index, \( \eta_{(h_1, h_2)} ; h_1 = 1, \ldots, H_1; h_2 = 1, \ldots, H_2. \)

\*Corresponding author. E-mail: pasquale.dolce@unina.it
The general form for a factor model can be written as:

\[ y = A\eta + e \]  \hspace{1cm} (4)

In the three-way case, we can write the factor model in Equation (4) as:

\[ y = (A_1 \otimes A_2)\eta + e \]  \hspace{1cm} (5)

where \( A_1 \) and \( A_2 \) are pseudo loading matrices, and \( e \) is the vector of measurement errors. Thus, we assume that the loading matrix \( A \) can be decomposed into two matrices:

\[ A = A_1 \otimes A_2 + A_0 \]  \hspace{1cm} (6)

where \( A_0 \) is the residual matrix which is assumed to be zero in equation (4).

Assuming that \( cov(\eta;e) = 0 \), and that the covariance among latent variables \( \Phi \) can be decomposed into two pseudo covariance matrices, \( \Phi = \Phi_1 \otimes \Phi_2 \), it follows that the covariance matrix \( S \) among the observed variables can be decomposed as:

\[ S = S_1 \otimes S_2 + C. \]  \hspace{1cm} (7)

where \( C = cov(e) \).

The estimation of model in Equation (5) can be done by an extended NIPALS algorithm which includes additional computational steps in order to decompose the loading matrix \( A \) into the two matrices \( A_1, A_2 \).

We will compare results from the three-way NIPALS algorithm to the classic three-way data analysis methods on real data.

References


Multigroup and multiblock PLS regression

Aida Eslami\textsuperscript{a}*, El Mostafa Qannari\textsuperscript{a} & Stéphanie Bougeard\textsuperscript{b}

\textsuperscript{a}LUNAM University, ONIRIS, Sensometrics and Chemometrics Laboratory, Nantes, F-44307, France

\textsuperscript{b}French agency for food, environmental and occupational health safety (Anses), Unit of Epidemiology, Technopole Saint Brieuc Armor, F-22440 Ploufragan, France

Keywords: Multigroup and multiblock data, PLS regression

Introduction

In various domains of application, data are often organized in two blocks of variables consisting of an independent dataset, X, and a dependent dataset, Y. Moreover, we consider the case where individuals are a priori divided into several known groups. We refer to this type of data as two-block and multigroup data. The objective of two-block and multigroup analysis is to investigate the relationships between X and Y taking into account the group structure among individuals.

In order to investigate the links between X and Y a first strategy of analysis is to ignore the group structure and perform a PLS regression on concatenated data Y and X. However, in this case the total variance mixes up both between- and within-group variances. A second strategy of analysis is to study the groups separately by means of several PLS regressions. This leads to too many parameters to be estimated and in addition it is difficult to sum up the results since there is no common structure among these separate analyses.

To counteract these problems, some methods have been presented in the literature such as Least Square PLS (LS-PLS) (Jørgensen et al. 2007), Sequential and Orthogonalized PLS (SO-PLS) (Næs et al. 2010) and Parallel Orthogonalized PLS (PO-PLS) (Måge et al. 2012).

We propose an extension of PLS regression to the case of multigroup data, so called Multigroup PLS (mgPLS) (Eslami et al. In Press). mgPLS is introduced on the basis of a maximization problem whose solution leads to an eigenanalysis. mgPLS can be considered as a compromise PLS regression among several groups of individuals in order to find out common variable structure among groups. The method of analysis is extended to the case of multigroup and multiblock data. Illustrations on the basis of real case studies are presented.

References


Tree-structured Partial Least Squares with an Application to Othodontics Data

Soo-Heang Eo\textsuperscript{a}, Shin-Jae Lee\textsuperscript{b}, Sungwan Bang\textsuperscript{c} \textsuperscript{*} and HyungJun Cho\textsuperscript{a} \textsuperscript{†‡}

\textsuperscript{a} Department of Statistics, Korea University, Seoul, South Korea
\textsuperscript{b} Department of Orthodontics, Seoul National University School of Dentistry, Seoul, South Korea
\textsuperscript{c} Department of Mathematics, Korea Military Academy, Seoul, South Korea

Keywords: PLS, Model-based recursive partitioning, GUIDE, Orthodontics data

Introduction

Partial least squares (PLS) regression is an alternative to classical regression for handling multicollinearity by combining the merits of principal component analysis and multiple linear regression. The aim of PLS is to predict a set of responses from a set of predictors by finding latent variables that capture the variability in both responses and predictors. Latent variables used in PLS are defined as the linear combinations of original variables. Among predictors, some variables can come from an extremely different nature or some can be indirectly related to responses. In that case, taking all predictors into account for extracting latent variables is not suitable from the model selection point of view. More latent variables can be selected for model fitting. If the variable from other natures has a different role rather then a predictor in PLS, it can be possible to improve the prediction performance by using the less number of latent variables. A disadvantage of PLS would be the deficiency in data visualization and model interpretation among others, which may be as important as building an optimal predictive model.

A recursive partitioning algorithm emerges as one of the solutions capable of achieving these purposes. We here propose new piecewise regression by combining the model-based recursive partitioning and PLS in order to improve prediction performance, and provide a visually interpretable model. Recursive partitioning, also known as tree-structured modeling, has been widely used because they allow us to give easy data visualization and interpretation. The objective of our proposed method is to select the most relevant predictors recursively, and provide more accurate prediction performance. Instead of fitting a global PLS to the whole data, one might fit a collection of local PLS to subsets of the data so that a better fit and higher predictive accuracy are obtained. We alleviate variable selection bias using the merits of the residual analysis approach of [1] and conditional inference of [2]. Our developed software program can be obtained from the authors upon request.

Proposed Method

Let \((Y, X, Z)\) be matrices of the response, predictor, and partitioning variables where \(Y\) and \(X\) are the \(q\)- and \(p\)-dimensional spaces, and \(Z = (z_1, z_2, \ldots, z_r)\) with \(r\)-dimensional space, respectively. As a basic model, we employ

\textsuperscript{*}Corresponding author. E-mail: wan1365@hanmail.net
\textsuperscript{†}Corresponding author. E-mail: hj4cho@korea.ac.kr.
\textsuperscript{‡}This research was supported by the National Research Foundation of Korea (NRF) funded by the Korean Government (2012-0007545) for S-J Lee, funded by the Ministry of Science, ICT and Future Planning (NRF-2013R1A1A1007536) for S. Bang, and funded by the Ministry of Education, Science and Technology (2010-0007936) for H. Cho.
the following PLS regression

\[
X = TP^T + E, \\
Y = UQ^T + R,
\]

where \(T\) and \(U\) are \(n \times K\) matrices that are projections of \(X\) and \(Y\), i.e., the score matrices of \(X\) and \(Y\). The \(P\) and \(Q\) are \(p \times K\) and \(q \times K\) orthogonal loading matrices, respectively. It is assumed that the error matrices \(E\) and \(R\) are independent and identically distributed random normal. Simply, the NIPALS algorithm is used to obtain parameters [3]. The basic model is fitted to have estimates that are different at each node so that terminal nodes might have different estimates. As an impurity function at node \(t\), \(R(t)\), mean squared errors are used.

For the partitioning, the partitioning variables \(Z\) are used to divide the data set into two subregions in which the basic model is fitted. Split rules are utilized to find the best split variable and point by evaluating the impurities between parent and child nodes over all possible splits. We here propose new split rule algorithms to provide the unbiased variable selection and visually interpretable model. The approach selects split rules by investigating the randomness of residuals after fitting the basic model to the data at each node, resulting in fast and ignorable-bias selection of split rules [1, 2]. This selection rule consists of univariate and bivariate algorithms. For the determination of tree size, the permutation test with multi-step stopping rule [4] is utilized.

Results and Conclusions

We applied the proposed method and the standard PLS regression to the craniofacial X-ray data set used in [5, 6]. The data sample consists of a lateral cephalogram, an X-ray image of craniofacial area on which the variables are gathered by the two-dimensional coordinates. The response variable \(Y\) is the coordinates on the post-surgery facial landmarks. The coordinates on the skeletal structure before and after surgery and pre-existing facial landmarks are used for predictors \(X\). Moreover, demographic variables, such as age, gender, interval time after surgery, operation type, and so on are used to partitioning variables \(Z\).

For comparison, the mean squared error of prediction (MSEP) is calculated by 10-fold cross validation. The overall MSEP of the proposed method is 3.70 with standard deviation 0.82, compared to those of the standard PLS are 4.51 (mean) and 1.02 (SD). When it comes to the errors of each response, our proposed method also outperforms the standard PLS. Our proposed method provides easy interpretation of the interrelation between terminal nodes by data visualization since the structure of the tree has a biplot per each terminal node. Throughout data analysis, we see that our proposed method attains better performance in terms of the prediction and the data visualization compared to the standard PLS.

References


A least squares approach to latent variables extraction in formative-reflective models

Marco Fattore\textsuperscript{a}, Matteo Pelagatti\textsuperscript{b,\textasteriskcentered}, Giorgio Vittadini\textsuperscript{a}

\textsuperscript{a}Department of Statistics and Quantitative Methods
Universit\`a degli Studi di Milano-Bicocca, Italy
\textsuperscript{b}Department of Economics, Management and Statistics
Universit\`a degli Studi di Milano-Bicocca, Italy

Keywords: path model, formative-reflective model, least squares, reduced rank regression, PLS-PM

Abstract

The use of formative latent constructs in data analysis is spreading. Formative models are becoming a standard tool in socio-economic research, particularly in the fields of causal modeling and multidimensional evaluation. Although at theoretical level the debate on the role and the use of formative models is still very fired, in practice these models are involved in many applied studies \cite{1}, making it urgent to develop new and effective statistical tools to deal with them. Despite the relevance of the topic, indeed, there are still unsolved methodological problems when dealing with structural equation models comprising formative constructs. This is particularly true for formative-reflective schemes, where manifest formative variables and manifest reflective variables are causally connected via a relational structure comprising both exogenous and endogenous latent variables. Such schemes, often used in causal modeling, are usually addressed using the Lisrel algorithm, which is affected by indeterminacy problems \cite{2}, or using the PLS-PM algorithm, which cannot handle formative relationships properly.

The limitations of the Lisrel \cite{3} and PLS-PM \cite{4} algorithms in this framework have been extensively discussed in \cite{5}, where an original algorithm, called RA-PM, has also been proposed as a valuable alternative. RA-PM overcomes many of the issues of Lisrel and PLS-PM, being consistent with the formative and the reflective relationships in the model and providing unique latent scores. However, RA-PM turns out to be only conditionally optimal and it does not take into account the formative side of the model adequately. For this reason, in this paper we propose a new procedure for the extraction of exogenous and endogenous latent variables in a formative-reflective scheme, which shares the same benefits of RA-PM, but satisfies a global optimum condition as it takes into account both the formative and the reflective sides of the model in a balanced way.

In a formative-reflective scheme, the exogenous latent variables play a double role. On the one hand, they should summarize their formative blocks; on the other hand, they should mediate, via the system of endogenous latent variables, the causal relationships linking the formative side to the reflective side. Realizing this, the proposed procedure extracts the exogenous latent variables balancing between these two aspects. An interesting feature of the tool is that it provides a way to check the correct specification of the model, that is to check whether the formative and the reflective sides of the model can indeed be consistently connected through the structural equation model. This feature is very important when one realizes that latent variables can be defined in a meaningful way only embedding them in larger relational structures, as discussed later in this paper. The extraction procedure has been designed for a formative-reflective scheme with unidimensional blocks of manifest variables, but it can be easily extended and adapted to the case where blocks are multidimensional. The algorithm

\textsuperscript{\textasteriskcentered}Corresponding author. E-mail: matteo.pelagatti@unimib.it.
can be effortlessly implemented in any programming language with matrix algebra and numerical optimisation capabilities. We implemented it in a user-friendly Ox object class [6], freely available from the second author.

References


Probabilistic Disjoint Principal Component Analysis

Carla Ferrara\textsuperscript{a}, Francesca Martella\textsuperscript{a} and Maurizio Vichi\textsuperscript{a}
\textsuperscript{a} Sapienza University of Rome, Department of Statistical Sciences, Rome, Italy

Keywords: Probabilistic model · Partition of variables · Maximum Likelihood estimation · Latent factors

Introduction

One of the most crucial and questionable topic related to Latent Variables Methods (LVMs), such as Principal Component Analysis (PCA) and Factor Analysis (FA), is the interpretation of latent variables (LVs) (also known as factors). Several authors have expressed reservations mainly for the difficulty to interpret univocally factors by using an objective approach (see for example Jolliffe, 1986). Difficulties of interpretation, according to Zou et al. (2006), are related to the fact that each factor typically is a linear combination of all manifest variables (all loadings are usually non-zero), even though only few manifest variables are relevant for the corresponding factor. Several methodologies have been proposed to specify subsets of variables that most explain factors, where the general idea is that the loading matrix should have as many zero coefficients as possible to show a simplest structure and presumably to allow the most meaningful interpretation. It is worth to note that in Confirmatory Factor Analysis (CFA), the general idea is to specify the simplest (most parsimonious) model that explains the correlations between variables via a reduced set of factors by specifying specifics manifest variables associated to factors frequently formed by disjoint classes of correlated variables.

In this paper, firstly we will consider the Disjoint Principal Component Analysis (DPCA) model, which has been proposed together with the clustering of objects by Vichi and Saporta (2009). This work has been proposed in a least square context and is characterized by factors formed by disjoint classes of manifest variables leading to a more easy interpretation. Secondly, we will extend the DPCA in a maximum likelihood context, highlighting advantages arising from this approach. In details, we will obtain a probabilistic formulation of the DPCA, named Probabilistic DPCA (PDPCA), from a Gaussian latent model by estimating the DPCA with an isotropic Gaussian error model in a maximum likelihood context. Finally the performance of the new methodology is evaluated on both simulated and real data sets.

1 Disjoint Principal Component Analysis (DPCA)

In order to simply the interpretation of factors, following the idea of CFA, Vichi and Saporta (2009) have introduced, together with the clustering of the objects, the DPCA model, which consists in a constrained PCA aiming at identifying factors associated to disjoint classes of variables. In particular, the DPCA model can be formally written as a PPCA model (Tipping and Bishop, 1999) without considering any distributional assumptions on factors and errors and imposing that each factor is a linear combination of one subset of variables explaining maximum variance only.

In details, the factor loading matrix, say \( A \), is reparametrized as the product of two matrices, \( A = BV \) where \( V = [v_{jk}] \) is a \((J \times K)\) binary and row stochastic matrix defining a partition of variables into \( K \) classes identifying \( K \) factors, with \( v_{jk} = 1 \), if the \( j \)th variable belong to \( k \)th class, \( v_{jk} = 0 \), otherwise; while, \( B \) is a \((J \times J)\) diagonal matrix weighting variables. Thus, the DPCA model can be specified as follows

\[
X = YBV + E
\]

where \( Y \) is a linear transform of the form \( Y = XBV \). The least squares estimators of the model (1) are the optimal solutions of the following quadratic problem with respect to unknown \( B \) and \( V \), given by

\[
\min_{B,V} \| X - XBVV'B \|^2
\]
with the constraints
\[ V = \left[ v_{jk} : \forall v_{jk} \in \{0,1\} \right] \] (3)
\[ V I_k = I_j \] (4)
\[ B = \text{diag}(b_1, \ldots, b_J) \] (5)
\[ V^T B B V = I_k \] (6)

For more details on the LS estimators see Vichi and Saporta (2009).

2 Probabilistic Disjoint Principal Component Analysis (PDPCA)

Let us now propose an isotropic error model that joins the PPCA characteristics with the DPCA. It will be named Probabilistic Disjoint PCA (PDPCA).

The PDPCA model can be specified as follows
\[ X = Y V^T B + E \] (7)

with constraints (3)-(5) and the additional restriction
\[ -1 \leq b_j \leq 1, \quad j = 1, \ldots, J \] (8)

used to specify correlation between variable \( j \) and the corresponding factor.

Matrices \( Y = [y_i : i = 1, \ldots, n] \), (where \( Y \) is standardized) and \( E = [e_i : i = 1, \ldots, n] \), have probability distributions:
(i) \( y_i \sim N_k(0, \Sigma_Y) \),
(ii) \( e_i \sim N_J(0, \Sigma_E) \) and \( \Sigma_E = \sigma^2 I_J \),
(iii) \( \text{Cov}(e_i, y_j) = \Sigma_{EY} = 0 \).

Note that only for PPCA and PCA principal components are uncorrelated and their loadings are orthogonal, while in DPCA loadings are orthogonal but the uncorrelated components condition are not explicitly imposed either. Furthermore for PDPCA the same strategy is adopted and therefore in PDPCA condition \( \Sigma_Y = I_J \) is sacrificed.

In order to estimate the PDPCA parameters, the following log-likelihood function has to be maximized:
\[ l(V, B, \sigma^2; X) = -n/2 \{ \ln \left| BV \Sigma_Y V^T B + \sigma^2 I_J \right| + \text{tr} \left( (BV \Sigma_Y V^T B + \sigma^2 I_J)^{-1} S \right) + J \ln(2\pi) \} \] (9)

subject to constraints (3)-(5) and (8).

3 Conclusions and further research

In this paper, a new latent variable model, PDPCA, which produces factors formed by disjoint classes of variables is proposed. The methodology is formulated in a maximum-likelihood framework and will be seen as a constrained version of the Probabilistic Principal Component Analysis.

The performance of the proposed approach has been discussed on both simulated and real data sets and generally exhibits accuracy of the estimates and good results.

In future researches, it would be interesting to consider uncorrelated component condition and, moreover, imposing other types of restrictions on the loading matrix \( A \).

References


A mixed method approach to help demonstrate saturation in qualitative research: applying Partial Least Square regression to qualitative data

Fatoumata Fofana\textsuperscript{a,b,*}, Angélique Bonnau-Boniguet\textsuperscript{b} and Antoine Regnault\textsuperscript{a,b}

\textsuperscript{a} Mapi, Lyon, France
\textsuperscript{b} UPRES EA 4275 (SPHERE) “Biostatistics, Pharmacoeconomics and Subjective Measures in Health Sciences”, Nantes, France

Keywords: PLS regression, saturation, qualitative research, mixed methods

Introduction

Mixed methods research is defined as the third research paradigm in research after the traditional qualitative and quantitative researches.[1] Mixed methods research involves the methodical mixture of qualitative and quantitative approaches in a research process.[2]

A key notion of qualitative research is saturation as it dictates the amount of data to be collected in the research. Saturation is said to be reached (i.e. enough data has been collected) when no new concepts/relevant information emerge with additional interviews.[3] In the classical qualitative approach, saturation is assessed based on the researcher's interpretation of the data collected during the interviews. This creates a risk of introducing a researcher bias. Quantitative methods could be helpful to address this issue.

The aim of our research was to explore how a mixed research method can be applied to the question of saturation. More specifically, we explored whether PLS regression can contribute to the demonstration of saturation.

1 Methods

The approach we propose is conducted using data from qualitative research studies involving exploratory or semi-structured interviews. In this setting, data collected during the interviews are coded by qualitative researchers. This coding is an analytic process to derive a list of concepts that have been elicited by the interviewees during their interview. After this step of coding, qualitative data are quantitized into a dataset: interviews are considered as variables and concepts as observations; each cell contains the number of times an interviewee mentioned the concept during the interview. In our approach, saturation is assessed based on these newly quantitized data.

By definition, saturation occurs if no new relevant concepts have been elicited in the last interviews. Our purpose is to be able to establish a link between the block of first interviews and the block of last interviews. In our approach, this is done by exploring whether the structure of the last interviews fits the structure of the first interviews.

Thus our dataset has to be split into two blocks: a block constituted of the first interviews and a block of last interviews. PLS regression is then applied to this dataset as follows:

\[(X_{j+1} \ldots X_n) = (X_1 \ldots X_j) B_{\text{PLS}} + E\] (1)

Where:

- \(X_j\) corresponds to the j-th interview, with \(n=\text{total number of interviews}\)
- \(B_{\text{PLS}}\) is the block of regression coefficients from the PLS
- \(E\) is the block of residuals

This model is based on the assumption that if concepts elicited in the block of first interviews (block of \(X_1\)–\(X_j\)) could predict concepts elicited in the block of last interviews (block \(X_{j+1}\)–\(X_n\)) then saturation would be demonstrated.

2 Illustrative example

2.1 Material
We applied our approach to data collected in an actual qualitative research study that aimed to investigate the impact of *Clostridium difficile* infection (CDI) on nurses’ work in the hospital. Semi-directive interviews were conducted with nurses who had already managed patients with CDI.

A total of 12 nurse interviews were carried out (6 in the US and 6 in France). All qualitative data from these interviews were coded by the researcher during the analysis process. As a result, 67 concepts were identified. The final dataset was built up, composed of the 12 nurse interviews as variables and 67 concepts spontaneously elicited by nurses as observations.

PLS was applied to this data to predict the concepts elicited in the last three interviews based on the concepts elicited in the first nine interviews, given by the following equation:

\[
(X_{10}, X_{11}, X_{12}) = (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9) B_{PLS} + E
\]

### 2.2 Results

The PLS factors summarizing all the concepts elicited in the first nine interviews explained 50.9% of the variability of all the concepts elicited in the last three nurse interviews. Besides, 3 groups of nurse interviews emerged from the correlation loading plot (Group 1: Nurse 8, Nurse 9 and Nurse 10; Group 2: Nurse 1, Nurse 4 and Nurse 11; Group 3: Nurse 6 and Nurse 12). Nurses of these groups shared common experience when caring for patients with CDI.

![Correlation loading plot of the 2 PLS factors](image)

### 3 Conclusion

We successfully applied PLS regression to data collected in a qualitative research setting, enabling the question of saturation to be quantified and addressed. PLS regression is a promising approach to study saturation and more generally to be used in a mixed research methods setting. Yet some practical questions still remain, in particular regarding the interpretation of PLS results in this setting. Specifically, a critical question is the characterization of the cumulative percentage of variation in the last interviews accounted for by the PLS factors that could be considered as warranting saturation. Further research is on-going to gain a better understanding on this question.

### References


Handling Unobserved Heterogeneity in PLS Path Modeling Using REBUS: Case of TAM Apply to Social Media Adoption

Samuel Fosso Wamba*a and Laura Trincheraa

“NEOMA Business School, Rouen, France

Keywords: SEM, Unobserved heterogeneity, TAM, Social media Adoption, REBUS

Introduction
In a recent study by [1], the authors identified that very few articles from top IS journals using structural equation models (SEM) over the last 20 years have “examined unobserved heterogeneity”. Therefore, they are calling for research on methods that investigate unobserved heterogeneity when using SEM, especially for mature theories (e.g. technology acceptance model (TAM)). Consequently, this research is an initial effort in bridging this knowledge gap. The objective is to apply the REBUS algorithm [2] to handle unobserved heterogeneity in the context of the application of TAM to social media adoption and use.

1 TAM and research model
The TAM was first developed and proposed by Davis in 1986 to assess an individual’s acceptance of an information technology (IT) artifact [3]. Afterwards, the model has undergone many extensions [4]. At the core of these models, we have that the behavioral intention of a potential adopter of an IT artifact is explained jointly by two interrelated beliefs, namely perceived usefulness, which is "the degree to which a person believes that using a particular system would enhance his or her job performance." (p.320) [3] and perceived ease of use or "the degree to which a person believes that using a particular system would be free of effort (p. 320) [3]. The model also theorizes that perceived usefulness is influenced by perceived ease of use (p. 320) [3]. In short, we have the following hypotheses (Figure 1):

![Figure 1: Research model](image)

Hypothesis H1.: Perceived ease of use has a positive effect on intention to adopt.
Hypothesis H2.: Perceived ease of use has a positive effect on perceived usefulness.
Hypothesis H3.: Perceived usefulness has a positive effect on intention to adopt.

The TAM has been used for theory testing in various fields of research leading to an important accumulation of strong empirical support for the model, and has been proven effective in predicting approximately 40% of an IT artifact’s adoption and use [5]. However, the TAM also holds some limits including: (i) lack of consideration of different user task environment and limits [6]; (ii) lack of assessment of the role of facilitating conditions; and (iii) assumption of data homogeneity in empirical studies, which may leading to potential invalid conclusions [1].

2 Methodology: Data collection, REBUS and data analysis
In this study, a web-based questionnaire was used to collect data from 2556 social media users from UK, US, Canada, India and Australia in January 2013. REBUS algorithm [2] allows unobserved heterogeneity detection by providing at the same time latent class composition and local model parameter estimations.

---

*a Samuel Fosso Wamba, samuel.fosso.wamba@neoma-bs.fr
3 Results and discussion

REBUS detects 3 groups of social media users, each characterized by different values for model parameters. From Table 1., we can see that all Cronbach’s alpha, composite reliability and AVE from all groups are respectively greater than the threshold of 0.7, 0.7 and 0.5 except α for ‘PU’ for G2; and AVE for ‘PU’ for G2 and G3. While all the loadings of items measuring our constructs exceed 0.707 for all constructs in GM and G1; those related to PU and IU3 in G2 and G3 are less than the threshold.

Table 1. Factor loadings, Cronbach’s alpha, composite reliability and AVE: GM: Global model, Gi=group i

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Items</th>
<th>Standardized loadings</th>
<th>Cronbach’s alpha</th>
<th>D.G. est (PCA)</th>
<th>AVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived ease of use</td>
<td>PEU1</td>
<td>0.933 0.933 0.914 0.920</td>
<td>0.881 0.905 0.866 0.885</td>
<td>0.933</td>
<td>0.940 0.918 0.929 0.820</td>
</tr>
<tr>
<td></td>
<td>PEU2</td>
<td>0.841 0.840 0.833 0.848</td>
<td>0.897 0.905 0.866 0.885</td>
<td>0.933</td>
<td>0.940 0.918 0.929 0.820</td>
</tr>
<tr>
<td></td>
<td>PEU3</td>
<td>0.940 0.936 0.916 0.934</td>
<td>0.999 0.922 0.557 0.758</td>
<td>0.955</td>
<td>0.942 0.738 0.838 0.808</td>
</tr>
<tr>
<td>Perceived usefulness</td>
<td>PU1</td>
<td>0.722 0.828 0.799 0.758</td>
<td>0.999 0.922 0.557 0.758</td>
<td>0.955</td>
<td>0.942 0.738 0.838 0.808</td>
</tr>
<tr>
<td></td>
<td>PU2</td>
<td>0.920 0.852 0.553 0.706</td>
<td>0.999 0.922 0.557 0.758</td>
<td>0.955</td>
<td>0.942 0.738 0.838 0.808</td>
</tr>
<tr>
<td></td>
<td>PU3</td>
<td>0.957 0.840 0.370 0.665</td>
<td>0.999 0.922 0.557 0.758</td>
<td>0.955</td>
<td>0.942 0.738 0.838 0.808</td>
</tr>
<tr>
<td></td>
<td>PU4</td>
<td>0.957 0.911 0.227 0.685</td>
<td>0.999 0.922 0.557 0.758</td>
<td>0.955</td>
<td>0.942 0.738 0.838 0.808</td>
</tr>
<tr>
<td></td>
<td>PU5</td>
<td>0.929 0.876 0.682 0.687</td>
<td>0.999 0.922 0.557 0.758</td>
<td>0.955</td>
<td>0.942 0.738 0.838 0.808</td>
</tr>
<tr>
<td>Intention to use Social Media</td>
<td>IU1</td>
<td>0.645 0.747 0.643 0.758</td>
<td>0.993 0.893 0.792 0.849</td>
<td>0.933</td>
<td>0.926 0.866 0.900 0.776</td>
</tr>
<tr>
<td></td>
<td>IU2</td>
<td>0.818 0.918 0.868 0.929</td>
<td>0.993 0.893 0.792 0.849</td>
<td>0.933</td>
<td>0.926 0.866 0.900 0.776</td>
</tr>
<tr>
<td></td>
<td>IU3</td>
<td>0.865 0.886 0.675 0.689</td>
<td>0.993 0.893 0.792 0.849</td>
<td>0.933</td>
<td>0.926 0.866 0.900 0.776</td>
</tr>
<tr>
<td></td>
<td>IU4</td>
<td>0.513 0.917 0.882 0.921</td>
<td>0.993 0.893 0.792 0.849</td>
<td>0.933</td>
<td>0.926 0.866 0.900 0.776</td>
</tr>
</tbody>
</table>

Table 2. shows that the standardized path coefficients for all models are significant at a level of 0.001: all our hypotheses are supported for all models. However, the strength of the relationship PU and PEU is higher for G1 and G2, than G3. For the relationship strength between IU and PEU, the highest value is with G2. The same group represents the weaker relationship strength between IU and PU.

Table 2. Structural model

Finally, a post-hoc analysis of each group shows that metropolitan geographic location; postgraduate education level, country (e.g., UK and India) and ages range from 18 to 24 and 25 to 34 are the places where we can find main differences that depict the three discovered social media users groups.

References

Sparse Partial Least Squares for Macroeconomic Forecasting

Julieta Fuentes \textsuperscript{a*}, Pilar Poncela \textsuperscript{b} and Julio Rodríguez \textsuperscript{c}

\textsuperscript{a}Universidad Carlos III de Madrid and Banco Central, San Salvador, El Salvador
\textsuperscript{b}Universidad Autónoma de Madrid, Madrid, España
\textsuperscript{c}Universidad Autónoma de Madrid, Madrid, España

Keywords: Factor Models, Forecasting, Large Datasets, Partial Least Squares, Sparsity.

Introduction

Factor models have been applied extensively for forecasting when high dimensional datasets are available. In this case, the number of variables can be very large. For instance, usual dynamic factor models in central banks handle over 100 variables. However, there is a growing body of the literature that indicates that more variables do not necessarily lead to estimated factors with lower uncertainty or better forecasting results. This paper investigates the usefulness of partial least squares techniques, that take into account the variable to be forecasted when reducing the dimension of the problem from a large number of variables to a smaller number of factors. We propose different approaches of dynamic sparse partial least squares as a means of improving forecast efficiency by simultaneously taking into account the variable forecasted while forming an informative subset of predictors, instead of using all the available ones to extract the factors. We use the well-known Stock and Watson database to check the forecasting performance of our approach. The proposed dynamic sparse models show a good performance in improving the efficiency compared to widely used factor methods in macroeconomic forecasting.

1 Methodology

PLS is a dimension reduction technique which constructs a scheme for extracting orthogonal unobserved components based on the covariance between the predictors (X) and the dependent (Y) or forecasting variable (X’Y). It has been implemented in a static way while its implementation taking into account the dynamic behavior of the target is scarce (see [1] and [2]).

Our goal is to forecast \( y_{t+h} \) given the available information of the target up to time \( t \), as well as from many other predictors, that we denote as \( X_t \), and their lags. Since \( X_t \) can incorporate a large number of predictors, we would like to extract the information that is valuable for forecasting \( y_{t+h} \) in a parsimonious way. If the common information in \( X_t \) coincides with the useful information for forecasting \( y_{t+h} \), we can use factor techniques to extract it. We use the term “factor” in a broad sense, meaning the unobserved component or signal that might be common to several variables (although not necessarily to many of them).

The forecasting model is specified and estimated as a linear projection of a \( h \)-step ahead transformed variable \( y_{t+h} \) onto \( t \) dated predictors. The predictors are the estimated factors, their lags and lags of the variable to be forecasted. That is,

\[
y_{t+h} = \mu + \phi(L)y_t + \beta'(L)\hat{F}_t + \eta_{t+h}
\]

where \( y_{t+h} \) is the variable to be forecasted at period \( t+h \) as a function of its own lags \( \phi(L) y_t \) and of the factors and their lags estimated in the previous step \( \beta'(L) \hat{F}_t \). The \( h \)-step ahead prediction error is denoted by \( \eta_{t+h} \).

The factor methods differ both in the way in which the factors are extracted and in the way in which the projection of the common component is made.

To apply the PLS method, we review the basic static application and examine two types of approximations that account for the dynamics of the time series: (i) the factors are based on the covariance between the
original set of predictors ($X$) and the target variable ($Y_{t+h}$). The lags of the target variable are included in the forecasting equation (1); (ii) the factors are based on applying PLS between an expanded set of predictors ($X_e$), enlarged with lags of the target variable ($Y_{t+h}$) and the target variable ($X_eY_{t+h}$). The forecasting equation (1) does not include lags of the target and (iii) the factors are based on applying PLS between the original set of predictors and the residuals of an AR (p) process fitted for the target variable. The lags of the target variable are included in the forecasting equation (1).

SPLS imposes sparsity within the context of PLS, is a technique that besides taking into account the response variable for the component estimation such as PLS, allows a variable selection process to be performed to construct a factor-forming subset. We apply the SPLS method to time series; we proceed as in the previous three approaches to PLS to take into account the dynamics in the time series.

2 Conclusions

The empirical comparisons are performed in terms of the Root Mean Square Error (RMSE) of the forecast errors. We compare PLS, SPLS and the most widely used methods: principal components, targeted predictors and LARS using well-known Stock and Watson database. We find that there is some room for refinement in the factor forecasting methodology that increases with the forecast horizon $h$. The dynamic SPLS methodology introduced in this paper shows a good prediction performance, improving the forecast efficiency of the alternative factor methods.

Among the different possibilities analyzed to apply PLS and SPLS to time series data, it seems that applying directly the PLS techniques between the target variable and the predictors yields the better forecasting results. Enlarging the data set of predictors, by including the lags of the target variable in it, does not seem to be a good alternative for PLS when applied to time series data, although this is not necessarily the case when the sparse version is applied. The PLS method gives weight to all the forecasting predictors, so the dependence between the target variable and its past can be obscured if there are too many predictors. On the contrary, including the lags of the target variable explicitly on the forecasting equation seems to be the best way of capturing the dynamic behavior of the target.

The forecast performance of SPLS improves with the forecast horizon. This might reflect the fact that when the dynamics of the own lags die out, the predictive content of the cross section emerges. This is observed in most of the approaches analyzed in contrast to the pure AR(4). Taking into account that in the very short run ($h=1$), the forecasting results given by all the methodologies are much closer; the dynamic SPLS approaches seem to perform quite well. When the dynamic relationship is integrated through the inclusion of the lags of the target as additional predictors in the original dataset, the selection process seems to weight the relevant information for forecasting purposes appropriately. In particular, the presences of variables that have a negligible effect on the response do not lessen the participation of $zt$ and its lags. For the updated dataset, the isolation of the AR(p) process effects, before PLS estimation, shows also a good performance at all forecasting samples.

References


Seemingly unrelated linear regression models based on multivariate mixtures

Giuliano Galimberti and Gabriele Soffritti

Department of Statistical Sciences, University of Bologna, Italy

Keywords: Finite mixture model, EM algorithm, model selection

1 Seemingly unrelated regression models

In seemingly unrelated regression models a set of equations is used to model the dependence of \( D \) variables (\( D \geq 1 \)) on one or more regressors. The regressors may be different from one equation to another. When all the equations have the same regressors, classical regression models are obtained as special cases. Furthermore, the error terms are allowed to be correlated and, thus, the equations should be jointly considered. Namely, a seemingly unrelated regression model can be defined as

\[
Y_i = \beta_0 + X_i'\beta + \epsilon_i, \quad i = 1, \ldots, I,
\]

where \( Y_i = (Y_{i1}, \ldots, Y_{iD})' \) is the vector of the \( D \) dependent variables for the \( i \)th observation; \( \beta_0 = (\beta_{01}, \ldots, \beta_{0D})' \) contains the intercepts of the \( D \) regression equations; \( \beta = (\beta_{1}', \ldots, \beta_{d}', \ldots, \beta_{D}')' \), with \( \beta_d \) denoting the \( P_d \)-dimensional vector whose elements are the regression coefficients of the \( P_d \) regressors on the \( d \)th dependent variable; \( \epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iD})' \) contains the \( D \) values of the error terms, and \( X_i \) is the following \( P \times D \) partitioned matrix:

\[
\begin{bmatrix}
    x_{i1} & 0_{P_1} & \cdots & 0_{P_1} \\
    0_{P_2} & x_{i2} & \cdots & 0_{P_2} \\
    \vdots & \vdots & \ddots & \vdots \\
    0_{P_D} & 0_{P_D} & \cdots & x_{iD}
\end{bmatrix},
\]

with \( 0_{P_d} \) denoting the \( P_d \)-dimensional null vector and \( P = \sum_{d=1}^{D} P_d \).

These models can be successfully employed to analyse cross-section data, time-series data and repeated measures (see, for example, [1]). Seemingly unrelated regression models have been studied through many approaches, such as feasible generalized least squares estimation methods, maximum likelihood methods, bootstrap methods, likelihood distributional analysis and Bayesian analysis. Most of these approaches have been developed under the assumption that the distribution of the error terms is Gaussian.

2 Using mixtures for modelling the error distribution

A drawback of the just described solutions is that the observed data may be drawn from a distribution that does not belong to the pre-specified parametric distribution family. That is, such methods rely on models that may be incorrectly specified. A convenient framework in which to model unknown distributional shapes is represented by finite mixture modelling. It is well known that, through an appropriate choice of its components, a finite mixture model is able to model quite complex distributions, and can handle situations where a single parametric family is unable to provide a satisfactory model for local variations in the observed data [2]. Thus, solutions to the described
problem can be obtained by modelling the unknown distribution of the error terms using a finite mixture of $D$-dimensional components. Finite mixtures of Gaussian and Student $t$ distributions have been recently employed in multiple and multivariate linear regression analysis [3, 4, 5, 6]. For example, in [4], the following assumption is introduced:

$$
\varepsilon_i \sim \sum_{k=1}^{K} \pi_k N_D(v_k, \Sigma_k), \quad i = 1, \ldots, I,
$$

where $\pi_k$’s are positive weights that sum to 1, the $v_k$’s are $D$-dimensional mean vectors that satisfy the constraint $\sum_{k=1}^{K} \pi_k v_k = 0_D$, the $\Sigma_k$’s are $D \times D$ positive definite symmetric matrices and $N_D(v_k, \Sigma_k)$ denotes the $D$-dimensional Gaussian distribution with parameters $v_k$ and $\Sigma_k$.

In this paper seemingly unrelated linear regression models are studied by assuming that the distribution of the errors is a finite mixture of Gaussian components. Identifiability conditions for the resulting models are provided. Parameter estimation is performed using the maximum likelihood method and an expectation-maximisation algorithm. This approach allows to exploit the flexibility of mixtures for dealing with non-Gaussian errors. In particular, the resulting models are able to handle asymmetric and heavy-tailed errors and to detect and capture the effect of relevant nominal regressors omitted from the model. Furthermore, by setting the number of components equal to one or by constraining all the equations to have the same regressors, some solutions already described in the statistical literature can be obtained as special cases.

This approach to seemingly unrelated regression modelling based on finite mixtures can be improved in several ways. For example, parsimonious models can be obtained by introducing some constraints on the component-covariance matrices through their spectral decomposition [7, 8]. The seemingly unrelated linear regression models obtained this way could provide a good fit for some datasets by using a lower number of parameters. Another improvement can be achieved by using a mixture model with skewed components; the resulting models will be able to describe asymmetric errors by a lower number of components.

When several parsimonious Gaussian mixture models are specified in conjunction with a seemingly unrelated linear regression problem, model selection has to be performed by comparing models that differ about the number of components, the parameterisation of the component-covariance matrices and the regressors included in the set of equations. Thus, whenever an exhaustive search for the best model is unfeasible, other strategies for dealing with the problem of model selection have to be adopted.

In this presentation we will describe the results obtained by developing some of these ideas and by evaluating their effectiveness and usefulness from a theoretical and applied point of view.

References


Tri-PLS for compositional data

Gallo M.\textsuperscript{a}\textsuperscript{*}, Trendafilov N.\textsuperscript{b}, Todorov V.\textsuperscript{c} and Martín-Fernández J.A.\textsuperscript{d}

\textsuperscript{a}Department of Human and Social Sciences, University of Naples-L’Orientale, P.zza S.Giovanni 30, Naples (It-80134), Italy
\textsuperscript{b}Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK
\textsuperscript{c}United Nations Industrial Development Organization, Wagramerstr 5 P.O. Box 300 A-1400, Vienna, Austria
\textsuperscript{d}Departament IMAE, Universitat de Girona, Campus Montilivi, Edif. P-IV, Girona (E-17071), Kingdom of Spain

\textbf{Keywords:} Compositions, Aitchison geometry, Log ratio, NPLS, trilinear models

\section*{Introduction}

Compositional data (CoDa, \cite{1} and \cite{2}) consist of vectors of positive values summing to a unit, or in general to some fixed constant. They can often be found in many disciplines and appear as proportions, percentages, concentrations, absolute and relative frequencies. Unfortunately, the constant-sum constraint that characterizes compositions is frequently disregarded or improperly incorporated into statistical modeling and a misleading interpretation of the results is given. Due to these specifications, several difficulties arise when dealing with CoDa. The first word of warning came already in 1897 from Karl Pearson who showed the dangers of underestimating spurious correlations.

There are several approaches to incorporate CoDa into statistical modeling when it is not realistic to assume a multinomial distribution of the data. Based on the log-ratio transformations, Aitchison \cite{1} proposed preprocessing the compositional data by means of log-ratio transformations, and successively analyzing them in a straightforward way by ‘traditional’ methods. Following Aitchison’s approach, the high dimensionality of CoDa in many scientific fields has encouraged the use of bilinear and trilinear decomposition models. Thus, in attempts to find adequate low-dimensional descriptions of compositional variability, CoDa are collected into two or three-way arrays (\cite{3}, \cite{4}, \cite{5}, \cite{6}, \cite{7}). On the other side, Hinkle and Rayens \cite{8} examined the problems that potentially occur when one performs a partial least squares (PLS) on compositional data.

The principal goal of this talk is to extend the PLS regression to three-way compositional data, following the approach proposed by Bro \cite{9} and Bro and al. \cite{10}. Both Candecomp/Paraafac (CP - \cite{11} \cite{12}) and Tucker3 \cite{13} models can be viewed as latent variables models extending principal component analysis to three-way data set. However, the most fundamental properties of PCA cannot be extended to these two models. PCA is an optimal representation of a two-way array with respect to the criteria of best low-rank approximation in least squares sense and the best approximation of the data within a joint low-dimensional subspace, while Tucker3 is only the best approximation of a three-way array within a joint low-dimensional subspace and CP is the best low-rank approximation in a least squares sense.

The proposed extension of PLS to three-way compositional data is illustrated on real data sets and a software implementation will be available in the R package rrcovHD.

\textsuperscript{*}Corresponding author. E-mail: mgallo@unior.it, nickolay.trendafilov@open.ac.uk, v.todorov@unido.org, josepantoni.martin@udg.edu.
References


Model Based Composite Indicators. Why to use the PLS Path Modeling approach

Grassia M.G.\textsuperscript{a}\dagger Lauro N.C.\textsuperscript{b} Marino M.\textsuperscript{b}

\textsuperscript{a} Department of Social Science, University "Federico II", Naples, Italy
\textsuperscript{b} Department of Economics and Statistics, University "Federico II", Naples, Italy

Keywords: Composite Indicators, PLS-PM, Non Metric data, Hierarchical data, Moderator variables

Introduction

Composite indicators (CIs) are increasingly used for measuring very complex phenomena like poverty, progress, well-being, etc. Using a model based approach in building CIs differently from the data driven approach, it is possible to take into account the role (formative and reflective) of the manifest variables and to model the relationships between the CIs. In this paper, first we present Structural Equation Modeling (SEM) as tool for building model based CIs. Among the several methods developed to estimate SEM, we show the Partial Least Squares Path Modeling (PLS-PM), because of the key role that estimation of the latent variables (i.e. the CIs) plays in the estimation process. Moreover, we demonstrate how some recent developments in PLS-PM, in order to the treatment of non metric data, hierarchical data, longitudinal data and multiblock data can strongly help to building composite indicators. It is easy to realize, for example, that as consequence of considering nominal and ordinal data the knowledge about a phenomenon synthesized by a composite indicator is considerably extended and improved especially for operational use. In order to highlight the potentiality of the proposed approach, the construction of some Composite Social indicators is discussed.

1 Model Based Composite Indicators

The goal of many research in social, economic and political fields, is often to obtain a whole description of the various facets of a complex phenomenon, through a suitable synthesis of elementary indicators associated. The growing interest in the scientific community to this area of research, is highlighted through the work of international organizations such as OECD, EU, IMF, UNDP, which developed various Composite Indicators, with the ultimate goal of ranking countries or compare the change of a country over the time. Their use can refer to both aggregate or individual data. According to Saisana \textit{et al.} \cite{Saisana} a Composite Indicator (CI) is defined as \textit{a mathematical combination of single indicators that represent different dimensions of a concept, the description of which is the objective of analysis}. Furthermore the construction of a CI implies the search for a suitable synthesis of a number of observed or manifest variables (Mvs) in order to achieve a simple representation of a multidimensional phenomenon. Accordingly a CI can be considered as a latent concept (Lv), not directly measurable, whose estimate can be obtained through the values of Mvs. For the construction of CIs three different approaches are available: theory based, data driven and model based. Structural Equation Models (SEMs) \cite{Hoppner} can be used to buildig model based composite indicators. SEMs include a number of statistical methodologies that allow us to estimate the causal relationships, dened according to a theoretical model, linking two or more latent complex concepts (i.e. the composite indicators), each measured through a number of observable indicators.

\dagger Corresponding author. Vico Monte della Pietà 1, 80138 Naples, Italy. E-mail: mgrassia@unina.it.
2 The PLS-PM approach for building Composite Indicators

Among the several methods that have been developed to estimate SEMs we focus on the PLS-PM [3], because of the key role that estimation of the latent variables (i.e. the composite indicators) plays in the estimation process. The PLS-PM use in building CIs presents numerous application opportunities: it is prediction oriented; it is not theory based; the parameters of each block are estimated separately, as in the Path Analysis and by simple/multiple regression; it treats both reflective and formative indicators; it has the better estimation of the measurement model, because optimizes the prediction of the Lvs and the relationship between the Mvs and Lvs; the estimates became consistent when the sample size gets larger; it can estimate the model also in presence of multi-collinearity and missing data; it is possible to estimate the model even when the number of observations is smaller than the number of the Mvs. Methodological advances provide researchers much more flexibility in modeling relationships and thus allow for a better building of CIs. Amongst others, these advances include treatment of non metric data [4] [5], hierarchical and multiblock data [6], moderator and mediator variables [7]. In order to highlight the potentiality of methodological advances in PLS-PM, the construction of two Social CIs are discussed: the Social Cohesion Composite Indicator with non metric data; the Poverty Social Indicators with multiblock data.

![Social Composite Indicators](image)

**Figure 1. Social Composite Indicators**

References


Robust Kronecker Product PCA for Spatio-Temporal Covariance Estimation

Kristjan Greenewald\textsuperscript{a} and Alfred O. Hero III\textsuperscript{a}

\textsuperscript{a}University of Michigan, Ann Arbor, USA

In this paper, we develop a method for the estimation of spatio-temporal covariances and apply it to multivariate time series modeling and parameter estimation. The covariance for spatio-temporal processes manifests itself as a multiframe covariance, i.e. the covariance not only between pixels or features in a single frame (time point), but also between pixels or features in a set of nearby frames. In streaming applications, at each time $t$ the covariance may be estimated over a sliding time window of $p_t$ frames. If each frame contains $q$ spatial components, e.g., pixels, then the covariance is described by a $p_t q$ by $p_t q$ matrix:

$$\Sigma_t = \text{Cov} \left[ \left\{ I_n \right\}_{n=t-p_t+1}^{t} \right]$$

(1)

where $I_n$ denotes the $p_t q$ pixels or other features of interest in the $n$th video frame. We make the standard piecewise stationarity assumption that $\Sigma_t$ can be approximated as unchanging over each consecutive set of $p_t$ frames.

As $p_t q$ can be very large, even for moderately large $p_t$ and $q$ the number of degrees of freedom ($p_t q (p_t q + 1)/2$) in the covariance matrix can greatly exceed the number $n$ of training samples available to estimate the covariance matrix. One way to handle this problem is to introduce structure and/or sparsity into the covariance matrix, thus reducing the number of parameters to be estimated.

A natural non-sparse option is to introduce structure by modeling the covariance matrix $\Sigma$ as the Kronecker product of two smaller matrices, i.e.

$$\Sigma \approx T \otimes S.$$  

(2)

When the measurements are Gaussian with covariance of this form they are said to follow a matrix-normal distribution [1]. This model lends itself to coordinate decompositions [2]. For spatio-temporal data, we consider the natural decomposition of space (variables) vs. time (frames) [2, 3]. In this setting, the $S$ matrix is the “spatial covariance” and $T$ is the “time covariance,” both determined up to a multiplicative constant.

An extension to the representation (2) discussed in [2] approximates the covariance matrix using a sum of Kronecker product factors

$$\Sigma \approx \sum_{i=1}^{r} T_i \otimes S_i$$

(3)

where $r$ is the separation rank. We call this the Kronecker PCA (KronPCA) covariance representation.

This model with $r \geq 1$ has been used in various applications, including video modeling [3], synthetic aperture radar, and MEG/EEG covariance modeling (see [2] for references). In [4] it was shown that any covariance matrix can be represented in this form.

This allows for more accurate approximation of the covariance when it is not in Kronecker product form but most of its energy is in the first few Kronecker components. An algorithm (Permuted Rank-penalized Least Squares (PRLS)) for fitting the model (3) to a measured sample covariance matrix was introduced in [2] and was shown to have strong high dimensional guarantees in MSE performance. This KronPCA model does not naturally accommodate additive noise since the diagonal elements (variances) must conform to the KronPCA structure.

In [3] we extended this KronPCA model, and the PRLS algorithm of [2], by adding a structured diagonal matrix to (3). This model is called Diagonally Loaded KronPCA (DL-KronPCA) and, although it has an additional $p_t$ parameters, it was shown that it performs significantly better for both inverse covariance estimation and inference [3].

The DL-KronPCA model [3] is the $r+1$-Kronecker model (where $U_1, U_2$ are diagonal)

$$\Sigma \approx \left( \sum_{i=1}^{r} T_i \otimes S_i \right) + U_1 \otimes U_2.$$  

(4)
Following rearrangement of $R$ to form $B$ as in [2, 5], this becomes the problem of finding a low rank approximation to a matrix $B$ where the intersections of a set of rows and columns are not included in the LS objective function [2]. A weighted least squares solution to this problem is given in [3].

In this work, we consider extending DL-KronPCA to the case of sparse loading. In other words, we model the covariance as the sum of a low separation rank matrix and a sparse matrix $\Sigma_{\text{sparse}}$:

$$\Sigma \approx \left( \sum_{i=1}^{r} T_i \otimes S_i \right) + \Sigma_{\text{sparse}}.$$

(5)

DL-KronPCA is trivially a special case of this model. The motivation behind this extension is that while the KronPCA model may work well for most variables in the problem, there are sometimes a few variables (or correlations) that cannot be well modeled using KronPCA, either because of their nature or their importance for the application. Thus, putting these variables and correlations in their own term allows the separation rank of the remainder to decrease, thus reducing the number of parameters if $\Sigma_{\text{sparse}}$ is sufficiently sparse relative to the possible reduction in separation rank. We call this model Robust Kronecker PCA (Robust KronPCA).

After rearrangement of the covariance as in [3], this model can be shown to be equivalent to a low rank plus sparse model. This model has been well studied under the name Robust PCA. In particular, the problem is of the form

$$B \approx B_{\text{sparse}} + \sum_{i=1}^{r} T_i S_i^T.$$

(6)

We solve this Robust PCA problem (low rank + sparse + noise) using optimization, with a quadratic penalty on the error of the approximation. We also optionally add a block Toeplitz constraint (i.e. the $T_i$ are made Toeplitz), which corresponds to a time stationarity constraint [3]. This imposes further structure on the covariance, and also results in a significant reduction in computational cost due to a reduction in the size of the SVD similar to [6].

As an application, we consider the yeast (S. cerevisiae) metabolic cell cycle dataset used in [7]. The dataset consists of 9335 gene probes sampled approximately every 24 minutes for a total of 36 time points, and about 3 complete cell cycles [7].

In [7], it was found using a variety of algorithms that many of the gene expression intensities exhibit periodic behavior in the dataset due to the periodic cell cycle. Our goal is to establish that periodicity can also be detected in the temporal component of the Kronecker spatio-temporal correlation model for the dataset. We learn the spatiotemporal covariance using block toeplitz robust KronPCA and then analyze the time factors ($T_i$) to discover periodicity. This allows us to consider the overall periodicity of the gene set, taking into account relationships between the genes, as opposed to the univariate analysis in [7]. The sparse correction to the covariance allows for the partial or complete removal of genes and correlations that are outliers in terms of their temporal behavior from the estimation of the overall temporal behavior.

References


Variable selection in high-dimensional dataset: comparison of sPLS with other approaches in an HIV vaccine trial

Boris P. Hejblum\textsuperscript{a, b, c*}, Robin Genuer\textsuperscript{a, b, c}, and Rodolphe Thiébaud\textsuperscript{a, b, c}

\textsuperscript{a}Univ. Bordeaux, ISPED; Inserm U897, Bordeaux, France
\textsuperscript{b}Inria SISTM team, Bordeaux, France
\textsuperscript{c}VRI, Bordeaux, France

Keywords: sparse Partial Least Squares, variable selection, random forest, VSURF, HIV vaccine

Introduction

High-dimensional data with a number of variables much higher than the number of observations leads to the challenging issue of variable selection. Several sparse methods are supposed to deal with this issue, including sparse Partial Least Squares (sPLS) and other supervised approaches. Here, we compare the application of sPLS and another recently proposed method called VSURF based on random forests \cite{1} to a phase 1 trial against HIV.

The DALIA trial is a therapeutic vaccine trial against HIV using ex-vivo generated Dendritic-cells loaded with HIV-lipoepitopes in 19 patients on antiretroviral therapy (ART). We have previously reported that this vaccine was well tolerated and immunogenic. Vaccine-elicited HIV-specific T cell responses were associated with improved control of viral replication following antiretroviral interruption (ATI) \cite{2}. The question is whether the gene expression is associated with the observed virological and immunological responses measured by several markers.

Gene expression was measured at 14 time points (in whole blood through Illumina Human HT-12 v4 microarrays). Multivariate immune response was measured at week 16, when the 4th shot of vaccine was given. Several cytokines of interest were measured. In addition, the vaccinees clustered into 2 distinct groups after treatment interruption: those with a moderate viral rebound and those with a high viral rebound.

Methods

Here we use sparse Partial Least Squares \cite{3} to relate gene expression after vaccination (week 16) with immune response (7 cytokines of interest measured also at week 16) and viral rebound.

We compare these results to VSURF \cite{4} (using univariate responses in turn, each time explained by gene expression data). VSURF performs variable selection with random forests. It relies on a stepwise ascending variable introduction strategy, based on the random forests variable importance score.

The two main differences between the methods are the following: i) sPLS uses all the immune responses at once in a multivariate responses analysis whereas random forests and therefore VSURF can only accommodate univariate responses; ii) sPLS is a linear method whereas random forests are nonlinear.

Post vaccination (at week 16) immune responses were evaluated using ICS, Multiplex cytokine secretion and Interferon-γ ELISPOT. Peak of viral load was the maximum observed plasma HIV RNA during ATI. Longitudinal analysis of gene expression data performed using a new approach (TcGSA) based on hierarchical models, allowing heterogeneity in predefined gene sets (Chaussabels functional modules \cite{5}), lead to identification of the 5,399 probes studied in this integrative analysis. We focused on this subset of 5,399 probes out of the 47,231 probes measured, belonging to relevant gene sets.

*Corresponding author. E-mail: boris.hejblum@isped.u-bordeaux2.fr
Both sPLS and VSURF fail to select several highly correlated predictors. This is due to the very nature of those methods which excludes redundant information from the predictor matrix. Therefore, it makes sense to consider selected genes as representative of the gene set they belong to, and to think at a gene set level when interpreting the results.

**Results**

The sPLS approach shows a clear inverse relationship between HIV-specific responses (production of IL-2, IL-13, IL-21, IFN-γ, CD4 polyfunctionality, i.e., production of at least two cytokines) and the peak of viral load during ATI. Those cellular immune responses were positively correlated to genes associated with T cell functional modules (M4.1, M4.15) at week 16 and negatively correlated to genes associated with inflammation (M7.1, M5.7, M3.2, M4.13, M4.2). Interestingly, the last three gene sets were also found to be associated with the response to pneumococcal vaccine [6].

The VSURF approach gave more clearcut results concerning which gene sets interact with which cytokines. Furthermore, it revealed an association with neutrophils.

**Conclusions**

Although the two methods differed quite substantially, biological conclusions were quite similar when interpretation was made at the gene set level. Thus, results from sPLS are mostly confirmed by the VSURF approach, and we are all the more confident in the biological findings of the analysis.

**References**


A PLS Path Model for Predicting Impact of Job Characteristics on Work-Related Stress

Mounia N. Hocine\textsuperscript{a*}, Giorgio Russolillo\textsuperscript{a}, Gilbert Saporta\textsuperscript{a}
Conservatoire National des Arts et Métiers, Paris, France.

Keywords: work-related stress, job stressors, PLS path modeling

Introduction

Work-related stress has become an increasingly major public health issue as it has negative effects on both physical and psychological health \[1\]. Although stress is an inevitable part of organizational life, effort can be made to reduce its negative effect on health. To reduce effectively work-related stress, job characteristics (stressors) impacting a high level of perceived stress should be well-documented and the impact of each stressor should necessarily be measured in a multidimensional way. However, if theoretical frameworks have been particularly successful in generating and collecting data on work-related stressors, there is a limited attention in the literature on quantitative assessment of stressors impact on work-related stress considering the multidimensional aspect of this type of data. In France, the most important system in evaluating and monitoring professional stress in companies (Stimulus) uses the Cooper index that is based on a bivariate statistical approach to identify stressors with a high level of stress \[2\].

The aim of this research is to investigate the impact of individual, organizational and environmental factors on a high level of perceived work-related stress. We propose to use PLS path modeling \[3\] to predict stressors requiring priority action from managers to reduce work-related stress of their company employees.

1 Study data : items, dimensions and scales

This study is carried out on a sample of 10 000 anonymous employees randomly drawn from the Stimulus database. Work-related stress is measured using a 25 items questionnaire (MSP25), and job stressors are measured using a 58 items questionnaire \[4\]. The 58 stressors are grouped into 5 dimensions as illustrated on Table 1.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|}
\hline
Dimensions of job stressors & No. of Items \\
\hline
Work context & 14 \\
Job control & 14 \\
Relationship & 12 \\
Tasks & 12 \\
Recognition & 6 \\
\hline
\end{tabular}
\caption{The Stimulus five dimensions}
\end{table}

* Corresponding author. Conservatoire National des Arts et Métiers, Paris, France E-mail: mounia.hocine@cnam.fr.
2 PLS modeling: usefulness and expected results for stress data

To predict job stressors impacting individual work-related stress, the use of PLS modeling is more appropriate than an estimating approach as it allows developing a system of weights through several indexes measuring each one a dimension of stressors (see Figure 1 for conceptual model). The PLS approach has been used recently for modeling work-related stress and employee health and wellbeing in a Chinese higher education institution [5].

This approach could be a useful tool for the policy of psychosocial risks management at workplace, as it allows simultaneously:
1. to build a stress scale using the 25 items from the MSP25 questionnaire,
2. to integrate this scale into a multidimensional model in which the five dimensions of job stressors are predictors of a high level of stress, and
3. to hierarchise the stressors (system of levers) following their impact on individual stress level.

![Figure 1. Conceptual model for work-related stress](image)

References


PLS-DA for compositional data with application to metabolomics

Karel Hron\textsuperscript{ab}, Alžběta Kalivodová\textsuperscript{ab}, Peter Filzmoser\textsuperscript{c}, Lukáš Najdekr\textsuperscript{d}

\textsuperscript{a}Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, Olomouc, Czech Republic
\textsuperscript{b}Department of Geoinformatics, Faculty of Science, Palacký University, Olomouc, Czech Republic
\textsuperscript{c}Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria
\textsuperscript{d}Laboratory of Metabolomics, Institute of Molecular and Translational Medicine, University Hospital Olomouc, Palacký University, Olomouc, Czech Republic

Keywords: compositional data, ilr coordinates, partial least squares, metabolomics, significance testing

Introduction

Compositional data (or compositions for short) are multivariate observations with positive components, and they can be represented without loss of information as data with a constant sum constraint like proportions or percentages [1, 2]. Consequently, the sum of the compounds (parts) is not important and the only relevant information is contained in the ratios between the parts. Hence, the constant sum is just a representation, not an inherent property of the data, describing quantitatively parts of a whole and following a relative scale. The concept of compositional data (in the following sense) frequently occurs in a wide range of applications involving geochemistry, chemometrics and related fields.

Metabolomics aims at studying metabolites, their dynamic changes, interactions and responses to stimuli. It is applied on metabolism of plants, bacteria, animals and humans. In humans all biological materials from biofluids (blood, urine) till tissues are analyzed. Although absolute values of biomarkers compared with reference ranges (data from healthy population) is the most frequently used approach, in diagnostic procedures that interpret data based on "profiling" (data on more variables in patient’s biofluid by common techniques, e.g. organic acids in urine by gas chromatography - mass spectrometry), relative changes are more relevant/informative than absolute values. Data from metabolomic resembles thus can be indeed rather considered as observations carrying relative information, i.e. compositional data.

Concerning the statistical analysis itself, the problem occurs because more metabolites (in hundreds) than biological materials (only tens) are present in these data sets. Therefore, suitable methods must be applied for these kind of observations. One of them is partial least squares regression (PLS regression), concretely its popular special case partial least squares discriminant analysis (PLS-DA) [3, 4]. It is well-known that PLS-DA is devoted to a particular regression problem, where the response is formed by categorical variables, whose values represent single groups that occur in the data set. Nevertheless, the standard PLS-DA method needs to be adapted to compositional data, because (due to specific geometrical properties of compositions) using raw observations could lead to useless results. The aim of this contribution is thus to introduce the logratio approach to PLS-DA of compositional (metabolomical) data. The methodology, that is introduced in a more detail in the next section, is

\textsuperscript{*}Corresponding author. 17. listopadu 12, 771 46 Olomouc, Czech Republic. E-mail: hronk@seznam.cz.
applied to two metabolomical data sets in order to demonstrate advances of the logratio approach compared to standard statistical techniques.

1 Methodological aspects of PLS-DA for compositional data

The natural geometry of compositions, called the Aitchison geometry, accounts for the features mentioned above (see, e.g. [2] for details). The Aitchison geometry has all the usual properties that are known from the Euclidean geometry, for which the standard statistical methods are designed [5]. However, the concepts of the Aitchison geometry are different from the Euclidean one. For this reason, usual multivariate statistical analysis including the family of partial least squares methods cannot be directly applied to compositional data, since otherwise the interpretations of the results and the conclusions would be misleading [6, 2].

Statistical data analysis is usually carried out in the Euclidean geometry and not in the Aitchison geometry. Thus, the central idea is to express \( D \)-part compositions from the simplex, their sample space with the Aitchison geometry, in the Euclidean real space and then to apply the standard multivariate methods. Due to lack of existence of a canonical basis on the simplex, we search for interpretable orthonormal coordinates with respect to the Aitchison geometry. As these coefficients are built up using logarithms of ratios of compositional parts (log-contrasts [1]), we refer to isometric logratio (ilr) coordinates [7]. The case of PLS-DA then requires to construct \( D \) coordinate systems, where always just one coordinate (closely related to the corresponding original compositional part) is of primary interest. This corresponds to the case of linear regression with compositional explanatory variables, where the mentioned approach was successfully applied [8].

2 Examples from metabolomical practice

The above methodology is applied to real-world data sets from metabolomics. The data are related with the diagnosis of inherited metabolic disorders. The first example analyzes the significance of the corresponding regression parameters (metabolites) using a small data set resulting from targeted metabolomics, where just a subset of potential markers are selected. The second example - the approach of untargeted metabolomics - was used for the analysis detecting almost five hundred metabolites. The significance of the metabolites is investigated by applying PLS-DA, accommodated according to the compositional approach. The significance of important metabolites (markers of diseases, i.e. regression parameters) is much more visible with the compositional method in both examples.

References

Three-way Generalized Structured Component Analysis

Heungsun Hwang
McGill University, Montreal, Canada

Keywords: Generalized structured component analysis, 3-way data, alternating least squares

Introduction

Three-way data consist of three different types of entities simultaneously (e.g., subjects, variables, and occasions), each of which is often called a mode. These data abound in various fields. A few examples include brain imaging data in neuroscience (e.g., [2]), multivariate longitudinal data in developmental psychology (e.g., [6]), and gene expression data in genetics (e.g., [3]).

Generalized structured component analysis (GSCA) ([4], [5]) is a component-based approach to structural equation modeling. It involves the specification of three sub-models to specify a structural equation model: measurement, structural and weighted relation models. The measurement model specifies the relationships between observed and latent variables. The structural model expresses the relationships among latent variables. The weighted relation model is used to explicitly define a latent variable as a weighted composite or component of observed variables. GSCA combines the three sub-models into a single equation. It estimates parameters by minimizing a single least squares criterion. An alternating least squares (ALS) algorithm is used to minimize the criterion. In the algorithm, an entire set of parameters is divided into two subsets, each of which is alternately updated while the other is fixed temporarily. GSCA provides overall model fit measures that can be used for assessing the variance of the data explained by a given model and for comparing different models.

GSCA is currently geared for the analysis of two-way data. In this paper, GSCA is extended to analyze three-way data. This proposed extension is called three-way GSCA.

1 Three-way GSCA

As in (two-way) GSCA, three-way GSCA involves the specification of three sub-models. Let $X_p$ denote the $p$th three-way dataset ($p = 1, \ldots, P$), arranged in a block of $I$ by $J_p$ by $K_p$, where $I$ is the number of entities in the first mode (e.g., subjects), which is assumed to remain the same across all $P$ datasets, $J_p$ is the number of entities in the second mode (e.g., variables), and $K_p$ is the number of entities in the third mode (e.g., occasions). Let $X_p$ denote a matricized version of $X_p$. Let $w_p$ denote a vector of component weights for the $p$th latent variable. Let $\gamma_p$ and $\gamma_{2p}$ denote vectors of loadings relating $\gamma_p$ to the second and third modes of the $p$th dataset, respectively. Let $\Gamma = [\gamma_1, \ldots, \gamma_P]$ denote a matrix consisting of all latent variables. Let $B$ denote a matrix of path coefficients relating each dependent latent variable to its explanatory latent variables.

In three-way GSCA, the weighted relation model can be given as follows:

$$\gamma_p = X_p w_p,$$

where $X_p$ is an $I$ by $J_pK_p$ matrix derived from aligning each $I$ by $J_p$ frontal matrix of $X_p$ $K_p$ times next to one another.

The measurement model specifies the relationship between $X_p$ and its latent variable, as follows:

$$X_p = \gamma_p (c_p^2 \otimes c_p^1) + E_{1p},$$

(2)
where $E_{lp}$ is the residual for $X_{lp}$ and $\otimes$ indicates the Kronecker product.

The structural model expresses hypothesized path-analytic relationships among $P$ latent variables. It can be generally written as follows:

$$\Gamma = \Gamma B + E_2$$  \hspace{1cm} (3)

where $E_2$ is the residual for $\Gamma$.

In three-way GSCA, weights, loadings and path coefficients will be estimated. To estimate these parameters, the following least squares criterion is minimized:

$$\phi = \sum_{p=1}^{P} SS \left( X_{lp} - \gamma_p (e_p^2 \otimes e_p^1)^\prime \right) + SS \left( \Gamma - \Gamma B \right),$$  \hspace{1cm} (4)

subject to $\gamma_p ' \gamma_p = 1$, $e_p^1 ' e_p^1 = 1$, and $e_p^2 ' e_p^2 = 1$, where $SS(M) = \text{tr}(M'M)$.

An ALS algorithm is developed to minimize (4). It alternates three steps until convergence. In the first step, all weights ($w_p$'s) are updated for fixed loadings and path coefficients. In the second, all loadings ($e_p^1$'s and $e_p^2$'s) are updated for fixed weights and path coefficients. In the third, path coefficients in $B$ are updated with weights and loadings fixed. The bootstrap method ([1]) can be used to estimate the standard errors or confidence intervals of parameter estimates.

2 Example

The proposed method will be applied to part of the National Longitudinal Survey of Youth 1979 (NLSY79) data. In this example, three three-way datasets will be considered that were measured on 72 individuals. The first dataset consists of six variables of problem behaviors, which were repeatedly measured at 5 time points. The second dataset consists of three variables of cognitive performance, which were measured at 5 time points. The last dataset includes two variables (cognitive stimulation and emotional support), which were measured at 5 time points.

References

A software for Generalized Structured Component Analysis

Emmanuel Jakobowicz\textsuperscript{ab*}

\textsuperscript{a}Addinsoft, Paris, France
\textsuperscript{b}CEDRIC-CNAM, Paris, France

Keywords: GSCA, XLSTAT, Path modeling, software

Introduction

The component-based structural equation model domain is expanding quickly with new users and new methods. In this paper, we propose to focus on Generalized Structured Component Analysis (GSCA). The world of structural equation models was dominated for more than two decades by the covariance-based structural equation models also called SEM, but these last years new approaches based on different principles have appeared, the so-called component-based SEM. SEM methods have been so popular for two reasons: their theoretical foundations and the availability of well-known commercial software (LISREL, AMOS…). The component-based structural equation model has suffered from two problems: the lack of global criterion and the lack of commercial software to apply these methods. These two problems are an old story thanks to new methods like Generalized Structured Component Analysis (GSCA) or Regularized Generalized Canonical Correlation Analysis (RGCCA) and the XLSTAT software that offers a full toolbox for component-based SEM.

1 GSCA: a flexible and complete method for path modeling

Component-based structural equation models are tightly related to the expansion of PLS path modeling. However, since PLS path modeling has become a mainstream approach, new methods have been introduced to tackle its weaknesses. PLSPM has shown great results but still lack of some important properties.

Ten years ago, Hwang and Takane (2004) [2] introduced the Generalized Structured Component Analysis (GSCA) as “an alternative method to partial least squares for path analysis with components”. Since its introduction, this method has been adapted to many specific cases and has gain in popularity among structural equation models users. However, since there was no implementation of this method in reliable commercial software, it still did not gain enough popularity.

XLSTAT now offers a full toolbox for component-based SEM including the classical PLS Path Modeling method with many advanced features including interaction, indirect effects, second order constructs, multigroup analysis and segmentation methods. It also includes new approaches like Generalized Structured Component Analysis (GSCA) or Regularized Generalized Canonical Correlation Analysis (RGCCA, Tenenhaus and Tenenhaus, 2011 [6]).

The GSCA method and its strengths

GSCA was introduced by Hwang and Takane in 2004 [2], it took the idea that there should be a method that has the flexibility of PLS Path Modeling and the theoretical foundations of classical Structural Equation Models.

The proposed method replaces factors by exact linear combinations of observed variables. It employs a well-defined least squares criterion to estimate model parameters. As a result, the proposed method avoids the principal limitation of partial least squares (i.e., the lack of a global optimization procedure) while fully
retaining all the advantages of partial least squares (e.g., less restricted distributional assumptions and no improper solutions) (Hwang and Takane, 2004 [2]).

GSCA proposes a global criterion to optimize the model but other indexes to measure global fit of the model including FIT, AFIT, SRMR, & GFI which are inspired from the SEM theory and local fit (FIT_{measurement}, FIT_{structural}, reliability indexes, AVE, etc.) are available.

One of the strength of GSCA compared to PLS Path Modeling is the ability to impose constraints on the parameters of the model which is very useful in the framework of Structural Equation Models.

Even if many papers have been published on that method, its use has not yet widespread because it lacked a software implementation. XLSTAT development teams together with Prof. H. Hwang developed new functions allowing GSCA to be available in a statistical software.

2 GSCA: A method for advanced component-based structural equation models

GSCA is able to answer very important questions for SEM users that cannot be easily addressed by PLS Path Modeling.

For example, it allows imposing across-group equality constraints on parameter estimates when doing multi-group comparisons. Multilevel GSCA allows using higher order construct [4]. As an alternative to REBUS-PLSPMS, GSCA Fuzzy Clustering [4] is also available. It allows identifying segments of observations and fitting a given model to each segment at the same time.

A major advance in the development of GSCA was also the introduction of regularized GSCA. It can be used to deal with multicollinearity among latent variables as well as among observed variables.

Conclusion

Generalized Structured Component Analysis is one of the most promising methods amongst the component-based SEM, its great flexibility and its high theoretical background makes it one of the best choice when the number of observations is small and the model is complex. The availability of that method together with all component-based method in a comprehensive software will allow component-based structural equation modeling to grow quickly and become a standard amongst SEM users.

References


Multilevel Dynamic Generalized Structured Component Analysis

Kwanghee Jung\textsuperscript{1} and Hyun Jung Park\textsuperscript{2}

\textsuperscript{1}University of Texas Health Science Center, Houston, USA
\textsuperscript{2}Baylor College of Medicine, Houston, USA

Keywords: Multilevel analysis, generalized structured component analysis, brain connectivity analysis

Introduction

As a novel approach to structural equation modeling (SEM), Dynamic GSCA (Generalized Structured Component Analysis) \cite{1} extended the original GSCA (\cite{[2]}, \cite{[3]}) by incorporating a multivariate autoregressive model to deal with longitudinal/time series data. Dynamic GSCA is equipped with both measurement and structural models, thereby latent variables (i.e., ROIs; Regions of Interest) are defined by simultaneously taking into account both the indicator variables (e.g., BOLD signals in fMRI) specific to ROIs and their relations to other ROIs. Furthermore, Dynamic GSCA incorporates new features such as time lagged and stimulus effects. Time lagged effects denote that the earlier state of the ROIs exert influence on the later state. Stimulus effects include direct effects of experimental stimuli on specific ROIs and modulating effects on connections between ROIs. These new features capture the dynamic nature of time series data and accommodate more elaborate research designs such as event-related designs in functional neuroimaging studies. For parameter estimation, Dynamic GSCA uses an alternating least squares algorithm, and thus it benefits computationally from a least squares parameter estimation method without rigid distributional assumptions and avoids model identification problems and improper solutions.

In this paper, we extend Dynamic GSCA to enhance its data-analytic capability in brain connectivity analysis of multi-subject functional neuroimaging data. Functional neuroimaging data are typically hierarchically structured, where time points are nested within participants who are in turn nested within an experimental group. The proposed approach, named Multilevel Dynamic GSCA, explicitly accommodates the nested structure in functional neuroimaging data.

1 The Two-Level Dynamic GSCA

We begin with the measurement model. Let $Z_j$ denote a $T \times V$ matrix of observed variables in the $j$-th subject ($j=1, \ldots, J$), where $T$ and $V$ indicate the number of time points and the number of observed variables, respectively. Let $\Gamma_j$ denote a row block matrix of latent variables for $j$-th subject with $\gamma_p$ as $p$-th column vector.

$$\Gamma_j = [\gamma_1, \gamma_2, \ldots, \gamma_P].$$

Also, define a block diagonal matrix $D_W$ with $w_p$ ($p=1, \ldots, P$) as $p$-th diagonal block,

$$D_W = \text{bdiag}([w_1, w_2, \ldots, w_P]),$$

where $w_p$ is a vector of component weights for latent variables ($\Gamma_j$), and a similar block diagonal matrix for $j$-th subject with $c_p$ as the $p$-th diagonal block,

$$D_{C_j} = \text{bdiag}([c_1, c_2, \ldots, c_P]),$$

where $c_p$ is the vector of weights to applied to $\gamma_p$ to best approximate $Z_j$ for the latent variable $p$ in the $j$-th subject. Then, the Level-1 measurement model can be written as
\[ Z_j = \Gamma_j D_{Cj}^i + E_{Mj} = Z_j D_W D_{Cj}^i + E_{Mj}, \]  
(4)

where \( \Gamma_j = Z_j D_W \). Next, let \( D_{Cj}^i \) denotes a matrix of Level-2 fixed loadings. Let \( D_{Aj}^i \) is a matrix of Level-2 random loadings, which are assumed to vary across subjects. The level-2 model is given by

\[ D_{Aj}^i = D_{Cj}^i + D_{Aj}^i. \]  
(5)

The Level-2 model indicate that Level-1 loadings (\( D_{Cj}^i \)) is conceived as varying over the population of a Level-2 unit, and the variance of each random effect represents the inter-subject variability of the corresponding loading. Thus, the measurement model for the Two-Level Dynamic GSCA is as follows:

\[ Z_j = \Gamma_j D_{Cj}^i + \Gamma_j D_{Aj}^i + E_{Mj} = Z_j D_W D_{Cj}^i + Z_j D_W D_{Aj}^i + E_{Mj}. \]  
(6)

Next, we specify the structural model for the Two-Level Dynamic GSCA. In Dynamic GSCA, shift matrices are introduced to capture time lagged effects \([1]\). The shift matrix with time lag 0 (\( S_0 = I_T \); the identity matrix of order T) represents contemporaneous effect. The matrix \( S_l (l = 1, \ldots, q) \) represents time lagged effects among latent variables. The subscript \( l \) indexes the order of lags. Let \( A_{ij}^l \) denote a matrix of Level-1 path coefficients of latent variables, which are conceived to be different across subjects. Then the Level-1 structural model is given by

\[ \Gamma_j = \sum_{l=0}^q S_l \Gamma_j A_{ij}^l + E_{Sj}. \]  
(7)

Next, let \( A_{ij}^l \) denote a matrix of Level-2 fixed path coefficients, and let \( \Theta_{ij}^l \) denote a matrix of Level-2 random effects of path coefficients, which may vary across subjects. The Level-2 model is given by

\[ A_{ij}^l = A_{ij}^l + \Theta_{ij}^l. \]  
(8)

Thus, the structural model for the Two-Level Dynamic GSCA is as follows:

\[ \Gamma_j = \sum_{l=0}^q (S_l \Gamma_j A_{ij}^l + S_l \Gamma_j \Theta_{ij}^l) + E_{Sj}. \]  
(9)

2 Example

Part of the data from a Sternberg working memory task collected on 30 participants is analyzed to demonstrate the effectiveness of the proposed approach. In the analysis, we specify a fully and bidirectionally connected structural model with time-lagged effects. Then, we apply the two-level Dynamic GSCA to fit the specified model to the data, while accounting for their nested structure, where BOLD signals (level 1) are nested within the participants (level 2). A user-friendly program for Multilevel Dynamic GSCA is also discussed, developed with interactive input/output features and a good graphical interface.

References


Brand Nostalgia and Consumers' Relationships to Luxury Brands: a Continuous and Categorical Moderated Mediation Approach

Kessous A\textsuperscript{a}, Magnoni F\textsuperscript{b} and Valette-Florence P\textsuperscript{b*}

\textsuperscript{a}INSEE, Paris, France
\textsuperscript{b}IAE & CERAG, Grenoble, France

Keywords: Continuous moderation, categorical moderation, mediation, moderated mediation, PLS

Introduction

This article presents a comparison between consumer relationships with two luxury car brands: nostalgic luxury car brand (Mini) vs. futuristic luxury car brand (Infiniti). The research questions are therefore the following: how does the relationship to brands vary according to the perceived nostalgic vs. futuristic character of the luxury car brands? In order to create a stronger attachment to luxury car brands is it relevant to use nostalgia? The relevance of this article is two-fold. On the one-hand, it contributes to a better explanation of the nature of the links that consumers have with these two types of luxury car brands. On the other hand, it extends Park et al.’s work (2010) by introducing brand nostalgia as an antecedent of brand attachment and by including moderating variables such as past temporal orientation and consumers’ need for uniqueness. It has been further validated in a different cultural environment and with new product categories.

1 Conceptual Background and Hypotheses.

Brand nostalgia, attachment and separation distress have been selected as the three main latent concepts. In addition, the need for uniqueness and the orientation toward the past have been chosen as continuous moderator variables, whereas both a nostalgic luxury car brand (Mini) and a futuristic luxury one (Infiniti) represent categorical moderators. Figure 1 presents the tested causal model.

![Tested causal model](image)

According to the aforementioned model, the following set of hypotheses is tested:

H1: Nostalgic luxury car brands (vs. futuristic luxury car brands) lead to higher: a) brand nostalgia; b) brand attachment; b) separation distress.
H2: Brand nostalgia has a direct positive effect on: a) brand attachment; b) separation distress.
H3: Brand attachment has a direct positive effect on separation distress.

* Corresponding author: IAE de Grenoble, BP 47, 38040 Grenoble Cedex 9, France. pvalette@upmf-grenoble.fr
H4: Past temporal orientation reinforces the relationship between: a) brand nostalgia and brand attachment; b) brand nostalgia and separation distress.

H5: Consumers’ need for uniqueness reinforces the relationship between brand attachment and separation distress.

2 Data analysis, hypotheses testing and main results.

All in all, the sample comprised 255 clients, who were quasi-equally distributed between the two brands. 132 Mini car owners and 123 Infiniti car owners responded to the survey in full. The sample is male (78% men) and relatively young (80% under the age of 54). As for the analyses, a PLS approach has been selected because of its minimal demands on sample size and suitability to handle higher order latent constructs and violation of multivariate normality. Moreover, the present study relies on rather small sample sizes and the model is complex involving several second order reflective latent variables. In this research, the estimation of the different PLS models follows a two steps procedure.

To test the hypotheses, we relied on a latent MANOVA and a step-down analysis. Since that from the outset the latent variables define a causal model, we decided to conduct a latent analysis of variance because it was necessary at that stage to delve deeper into the joint effects of the nostalgic vs. futuristic orientation of the luxury car brand (Mini vs. Infiniti) on the latent brand relationships variables encompassed by this research. Another advantage of analyzing variance at the latent level using a structural equations model is the ability to compare the strength of the effect between different dependent latent variables and to perform a step-down analysis at the latent level. When there is a causal relationships network among the dependent variables, step-down analyses provide useful information as to whether the mean difference in a dependent variable is due to the direct effect of the experimental manipulation or its dependence on other variables. A step-down analysis proceeds into two sequential steps. The first stage begins with a latent MANOVA performed on all dependent variables. If the path estimates point to a rejection of equal means, then the next step consists of testing the dependent variables in the hypothesized causal network while partialling out all remaining dependent variables as covariates. As a result, the researcher can then assess the relative impact of the experimental manipulation, while taking into account the causal order between all the dependent latent variables.

All in all, results show that brand nostalgia, brand attachment and separation distress latent scores are significantly greater for Mini than for Infiniti, hence supporting H1a, H1b and H1c. In addition, brand nostalgia has a direct positive impact on brand attachment (0,508) and separation distress (0,572), supporting H2a and H2b. Brand nostalgia has also an indirect effect on separation distress (0,181) and brand attachment influences directly separation distress (0,357). These results support H3. Hence, brand attachment is a partial mediator between brand nostalgia and separation distress. Moreover, when performing a latent step-down analysis, the direct influence of the nostalgic vs. futuristic orientation of the luxury car brand is no longer statistically significant at all. This means that all the effects are now due to the causal relationships between the dependent variables. This result seems both theoretically and managerially important. However, even if the nostalgic vs. futuristic orientation of the luxury car brand doesn’t have any direct impact, it still has an important indirect effect. Two points deserve attention. Firstly, all the indirect effects are now greater than when the nostalgic vs. futuristic orientation of the luxury car brand was solely taken into account (for brand attachment for example, 0,107 vs. 0,250). This means that the encompassed latent variables indirectly amplify the effect of the nostalgic vs. futuristic orientation of the luxury car brand. Secondly, there is once again an attenuation of the incidence of the brand orientation on the dependent variables. This indirect effect is indeed greater for separation distress (0,370) than brand attachment (0,250). Once again, this result puts the stress on the influence of brand nostalgia on separation distress, either directly or indirectly.

Finally, in order to study the joint effect of the two latent moderator variables, we relied on the product indicator approach, hence following recent recommendations made by Henseler and Chin (2010) in the case of complex moderation investigations (we recall that we jointly model 3 moderations through either first or second order latent variables). Far and foremost, the past orientation positively moderates attachment (β=0,204; p=0,001). In other words, the past orientation reinforces the impact of brand nostalgia on brand attachment, supporting H4a. A similar effect arises as for the double moderation of the past orientation and the need for uniqueness on separation distress. Once again, this is the past orientation that has the greatest moderating incidence (β=0,194; p=0,016), compared to need for uniqueness (β=0,077; p=0,046). The past orientation and the need for uniqueness both reinforce the impact of either brand nostalgia or brand attachment on separation distress, supporting H4b and H5. However, one can notice that the moderating influence of the past orientation on brand nostalgia is almost three times higher than the moderating influence of the need for uniqueness on band attachment. In other words, brand nostalgia and past orientation seems the more important to predict separation distress.

References

Treating Similar Aspects Differently, and Different Aspects Similarly

Henk A.L. Kiers*
University of Groningen, Groningen, The Netherlands

Keywords: Nonsymmetric correspondence analysis, Symbolic data analysis, PCA of interval data, PCovR

Introduction

Arguably Carlo Lauro’s main influence has been in two fields, which he actually also brought together: Nonsymmetric (multiset) Data Analysis and Symbolic Data Analysis. In addition to his own work on these two fields, he stimulated and fostered collaboration between many researchers (quite a few of them from his own Naples school) on these two broad topics.

1 Nonsymmetric Data Analysis

Lauro’s first contribution on Nonsymmetric Data Analysis was his work, jointly with D’Ambra, on the development of nonsymmetric correspondence analysis. Based on Rao’s paper [5], and independent invention by, for instance, Van den Wollenberg [6], Lauro and D’Ambra [3] designed a nonsymmetric version of correspondence analysis, in which one variable (or possibly a set of variables) is considered as reference or independent variable, and the other as dependent variable. Many variants of this have later been proposed, among which partial nonsymmetric correspondence analysis, and many elaborations and properties have been described in the literature (e.g., see [2] for an overview).

2 Symbolic Data Analysis

As far as Lauro’s contributions to Symbolic data analysis are concerned, his possibly most salient contribution to the literature was in the form of several papers (mainly with Francesco Palumbo) on Principal Component Analysis of interval data. Lauro and Palumbo [4] proposed a highly original approach of handling interval valued data by both analyzing the interval midpoints data matrix, and the ranges data matrix, separately by means of PCA’s, and by ingeniously combining these into results of the same form as the original data (i.e., data boxes, or more generally hyperboxes).

3 Contrasts and a compromise

In this presentation, the two above approaches will be reviewed, and it will be made clear why these approaches can both be seen as counterintuitive: NSCA is a class of methods where similar aspects of data are treated differently, and the PCA of interval data approach is a case where different aspects (midpoints and ranges) are treated similarly. And it will be underlined why these counterintuitive properties make these methods highly ingenious.

Finally, upon realizing that NSCA is a qualitative variables variant of redundancy analysis, an exploration will be made towards methods for qualitative variables based on alternatives to redundancy analysis, like PCovR [1], comprising redundancy analysis and PCA as special cases, thus comprising the symmetric and

* Corresponding author. E-mail: h.a.l.kiers@rug.nl
nonsymmetric approaches as special cases. In this way, a class of compromises can be developed between NSCA and Multiple Correspondence Analysis. In other words, a class of methods will be proposed where the degree of nonsymmetry can be tuned at wish.

References


Athanasios Kondylis\(^a\)* and Joe Whittaker \(^b\)

\(^a\) Philip Morris International, Applied Cigarette Research, Quai Jeanrenaud 3, 2000 Neuchâtel.

\(^b\) Lancaster University, Department of Mathematics and Statistics, LA1 4YF Lancaster, UK.

Keywords: Preconditioning, Krylov spaces, Partial Least Squares, Discrete wavelet transform

Introduction

In linear regression analysis the coefficient vector \( \beta \) is estimated from the \( n \) sample observations stacked in \( y \) and \( X = (x_1, \ldots, x_p) \) as the solution to the optimization problem:

\[
\hat{\beta}_{\text{ls}} = \arg\min_{\beta} \{ (y - \tilde{y})'(y - \tilde{y}) \}, \quad \text{where} \quad \tilde{y} = X\beta \quad \text{and} \quad \beta \in \mathbb{R}^p,
\]  

where \( p \) denotes the number of predictor variables. The solution requires solving the normal equations

\[
X'(y - X\beta) = 0.
\]

When predictors are many and highly correlated regularized regression methods are better used to trade bias for variance. The final regression estimate gets smaller mean square error.

Using the notation \( A = X'X \) and \( b = X'y \), the PLS regression coefficient \( \hat{\beta}_{\text{pls}}^m \) of order \( m \) is defined by:

\[
\hat{\beta}_{\text{pls}}^m = \arg\min_{\beta} \{ (y - \tilde{y})'(y - \tilde{y}) \}, \quad \tilde{y} = X\beta \quad \text{and} \quad \beta \in \mathcal{K}_m(A, b),
\]

where

\[
\mathcal{K}_m(A, b) = \text{span}(b, A b, \ldots, A^{m-1} b),
\]

introduce the Krylov sequence, see [1], that serves as a basis to approximate the regression solution. If \( A \) has rank \( p \) and \( m = p \) then \( \mathcal{K}_m = \mathbb{R}^p \) and the PLS solution coincides to the OLS solution at (1). When \( m < p \) the PLS regression truncates the least squares solution to the span of the first \( m \) components of the Krylov sequence. To achieve the PLS regression solution in expression (3) various algorithms have been proposed in the literature. We use the orthogonal loadings PLS regression algorithm.

Preconditioning Krylov spaces

Preconditioning techniques, see [2], consist of pre-multiplying the linear system (2) by a non-singular preconditioning matrix \( M' \in \mathbb{R}^{p \times p} \) to give the preconditioned normal equation:

\[
M'X'(y - X\beta) = 0.
\]

*Corresponding author. E-mail: athanasios.kondylis@pmi.com.
The preconditioned PLS solution seeks for:

$$\tilde{\beta}_m^s = \arg\min_{\beta} \left\{ (y - \tilde{y})'(y - \tilde{y}) \right\}, \quad \tilde{y} = \hat{X}\beta \text{ and } \beta \in \mathcal{K}_m(M' A, M' b).$$

(6)

Preconditioning techniques have mostly attracted the interest of numerical analysts, for accelerating the convergence of iterative methods. We use preconditioning in order to deal with data that are functional, and to select features in the functional domain. Solutions are then back-transformed, if needed, in the original coordinates system.

### PLS regression for functional data

Let \( \mathcal{W} \) be an orthogonal basis matrix appropriate for functional data. Following [3], we use the discrete wavelet transform matrix \( \mathcal{W} \), and we precondition the linear system in order to approximate the final Krylov solution in \( \mathcal{W} \mathcal{K}_m(A, b) \). Given that \( \mathcal{W} \) is orthogonal, the resulting space is preserved, that is: \( \mathcal{K}_m(\mathcal{W} A \mathcal{W}', \mathcal{W} b) = \mathcal{W} \mathcal{K}_m(A, b) \). The preconditioned linear system \( \mathcal{W} A \tilde{\beta} = \mathcal{W} b \), is solved in the transformed coordinates \( \mathcal{W} A \mathcal{W}' = \mathcal{W} b \), while the final solution may be transformed back in terms of the original coordinates system by applying the inverse discrete wavelet transform.

The functional PLS solution lives in a Krylov subspace, as does the ordinary PLS regression vector.

Consider solving (6) iteratively for a non-singular diagonal matrix \( M^{ts} \) (\( s \) denotes the iteration step), with entries reflecting the relevant features on the wavelet domain, given as:

$$\mu_j^s = \lambda \frac{\tilde{\beta}_j^{fpls, s}}{\sqrt{\sum_{j} (\tilde{\beta}_j^{fpls, s})^2}},$$

(8)

Equation (8) constructs importance based on relevance for prediction, not from the ability to reconstruct the original functions. The Importance factors in (8) determine a partition of the wavelet coefficients falsely allocated to \( A \), and its choice could be viewed as a multiple testing problem. Therefore the proposed method is to iterate until the relevant subset \( A^r \) no longer changes. The output set \( A^r \) indicates the relevant functional coefficients subset. These are scaled according to \( M^{r} \), the others are thresholded to zero and the final solution is

$$\tilde{\beta}_j^r = \begin{cases} \frac{\tilde{\beta}_j^{fpls, r}}{\beta_j} & \text{for } j \in A^r, \\ 0 & \text{otherwise.} \end{cases}$$

(9)

To recover the sparse PLS solution in terms of the original coordinates system, one needs to pre-multiply the solution \( \tilde{\beta}_j^r \) for \( j \in A^r \) by \( \mathcal{W}' \).

### References


E-Health Website Adoption: A Test of the Theory of Trying

Lesley Land\textsuperscript{a}, Natasha Kamalesan\textsuperscript{a}, Wynne W. Chin\textsuperscript{b} and R. Ryan Nelson\textsuperscript{c}

\textsuperscript{a}University of New South Wales, Australia
\textsuperscript{b}University of Houston, Texas, United States of America
\textsuperscript{c}University of Virginia, Virginia, United States of America

Keywords: E-health adoption, Perceived Characteristics of Innovation, Theory of Trying

Introduction

Whilst the potential for using health focused Information Technology to improve health care has been acknowledged, these technologies are still not always accepted by potential users for a variety of reasons including poor device usability, insufficient training on how to use the technology, lack of computer skills, and low self-efficacy [1, 2, 3, 4, 5]. This is a significant concern for consumers and health care organizations; patients who reject health focused IT innovations will not benefit from them, and rejection means a loss of return on investment for health care organizations [5]. Consequently, this creates a need to understand the factors that influence patients who reject health focused IT innovations (i.e. non users or non adopters).

There is extensive research examining the acceptance, use and adoption of new information systems (IS) and information technologies (IT) by both individuals and organisations. However, these various models have also presented two limitations: comparatively low explanatory power, and inconsistent influences of the factors across studies [6]. Researchers have called for greater theoretical richness in IT acceptance models and not just greater parsimony [7, 8]. As such, we present the results of applying Partial Least Squares analyses on a model that combines factors from the Perceived Characteristics of Innovation (PCI) and processes from the Theory of Trying (ToT) [9].

References


\textsuperscript{*} Corresponding author. E-mail: wchin@uh.edu


Analysis of Structured Data: All Roads Lead to Naples

Ludovic Lebart*
Telecom-ParisTech, Paris, France

Keywords: Nonsymmetrical Correspondence Analysis, Structured Data Analysis, Data Mining, Text Mining

Introduction

The upsurge of Exploratory Data Analysis in the late nineteen sixties has led to an extensive use of a series of basic techniques such as Principal Components Analysis, Simple and Multiple Correspondence Analysis, often in combination with Clustering Techniques. The next step was then to answer the following question: How to use what we know already about the data to improve their description? (i.e.: how to involve metadata in the context of exploration?). Carlo Lauro and the researchers from the Naples School of Data Analysis he has founded and stimulated have accomplished several pioneering works along that avenue of research.

1 Brief overview of some symmetrical / nonsymmetrical methods

Let us summarize (Table 1) a small selection of methods for the eigen-analysis of data tables \( R \) made up of two blocks \( R = (X, Z) \): we give only the formulas of the matrices to be diagonalized. In cases where \( X \) and \( Z \) are arrays of numerical variables, they are centered.

<table>
<thead>
<tr>
<th>Matrix to be diagonalized</th>
<th>And / Or alternative form</th>
<th>Method name</th>
</tr>
</thead>
<tbody>
<tr>
<td>((X'X)^{-1}XZ (Z'Z)^{-1}Z'X)</td>
<td>(X'Z (Z'Z)^{-1}XZ)</td>
<td>(Canonical Analysis) [CA]</td>
</tr>
<tr>
<td>(XZ (Z'Z)^{-1}Z'X)</td>
<td>And: (Z'(X'X)^{-1}X'Z)</td>
<td>(Projected Analysis) [NSCA]</td>
</tr>
<tr>
<td>(X'(I - Z (Z'Z)^{-1}Z')X)</td>
<td>And: (Z' (I - X (X'X)^{-1}X') Z)</td>
<td>(Partial Analysis)</td>
</tr>
<tr>
<td>(X'Z'X)</td>
<td>Or: (Z'X X'Z)</td>
<td>(Orthogonal Procrustean Analysis)</td>
</tr>
<tr>
<td>((Z'Z)^{-1}Z'X)</td>
<td>And: ((X'X)^{-1}X'Z)</td>
<td>(Procrustean Analysis)</td>
</tr>
</tbody>
</table>

Despite the attractiveness of the mathematical properties of Correspondence Analysis, most applied statisticians recognize that in real life, two partitions of a same data set rarely play symmetric roles. It is one of the merits of Lauro and d’Ambra [1], [2], [3] to have maintained, against the general trend during the nineteen eighties, that in some specific circumstances, the formulas of CA should be amended to take into account this lack of symmetry. These first pieces of research led to a series of noteworthy works and papers from students and former students of Carlo Lauro. Some of them are described and commented in [4] and [5].

2 Specific application fields, including textual data

Numerous theoretical extensions or applications papers in various fields have been published since the first mentioned pioneering works, such as [6], [7]. In this section, we will deal with the potential of the analyses of

* Corresponding author E-mail: ludovic@lebart.org
Structured qualitative data in the field of textual data analysis, a domain of research that has benefited from several contributions by a former student of Carlo Lauro: Simona Balbi, [8], [9].

2.1 Textual data: highly structured qualitative data.

The main feature of textual data sets is certainly the huge amount of available meta-data (syntactic structure, semantic networks). Moreover, a marked disymmetry between rows and columns of a lexical contingency table is often observed. Statistical units constituting the rows (observations, subjects, individuals) can be either documents (described by their titles, abstracts or full texts) in documentary databases, respondents (described by their responses to open questions) in surveys, or segments of texts (sentences, context units, paragraphs, chapters, novels) in literary applications. They may also be generated by an automated cutting up of the corpus into segments of text having about the same length (the Elementary Context Units of Max Reinert). In the other dimension (columns) of the lexical table, we find words (or lemmas [after parsing], or segments) generally selected using a frequency threshold. Rows and columns play quite different roles for the user.

2.2 Contrasting several approaches

To illustrate the advantages and drawbacks of the tools proposed by the Naples School of Data Analysis, we will contrast both symmetrical and nonsymmetrical analyses in the framework of Text Mining using several corpuses of scientific texts,… involving Carlo Lauro.

References


Discriminant Analysis for Multi-way Data

Lechuga G., Le Brusquet L., Perlberg V., Puybasset L., Galanaud G. and Tenenhaus A.

Supélec Sciences des Systèmes - EA4454 (E3S), Gif-sur-Yvette, France
Sorbonne Universités, UPMC Univ Paris 06 UMR S 1146, CNRS UMR 7371, Laboratoire d’Imagerie Biomédicale, F-75005, Paris, France
AP-HP, Pitié-Salpêtrière Hospital, Surgical Neuro-Intensive Care Unit, Paris, France
AP-HP, Pitié-Salpêtrière Hospital, Department of Neuroradiology, Paris, France

Keywords: Discriminant analysis, multi-way data, medical imaging, high dimension

Introduction

In standard multivariate data analysis, individuals \( \times \) variables data table is usually considered (two-way data table). However, from a practical viewpoint this simple data structure appears to be somehow limited. It is the case for instance when individuals are characterized by the temperature at different locations sampled over different times, leading to a three-way data structure. Such multi-way structure can be viewed as a stack of matrices \( X = \{ X_{ijk} \} \) from which the \( I \) horizontal slices describe the individuals \( i = 1, ..., I \), the \( J \) lateral slices describe the variables (temperature) \( j = 1, ..., J \) and the \( K \) frontal slices describe the different time points \( k = 1, ..., K \). Many two-way data analysis methods have been extended to the multi-way configuration. For instance, a multi-way formulation of Partial Least Squares Regression (N-PLS) has been proposed in [1]. N-PLS relies on the maximization of a covariance criterion but explicitly takes into account the multi-way structure of the input data. In this paper, we present a multi-way formulation of Fisher Discriminant Analysis (MFDA) in an attempt to improve the interpretability of the resulting model compared with the results obtained with unfolded methods. MFDA is illustrated on a real multi-modal Magnetic Resonance Brain Imaging (MRI) dataset.

1 Multi-way FDA analysis

FDA is defined by the optimization problem: \( w^* = \arg \max_w \frac{w^T S_B w}{w^T S_T w + \lambda w^T w} \), where \( S_B \) is the between covariance matrix, \( S_T \) is the total covariance matrix and \( \lambda w^T w \) is an additive regularization term required in the high dimensional setting. MFDA seeks to maintain the natural tensor structure of the input data by constraining \( w \) to be of the form \( w = w^K \otimes w^J \). \( w^K \) is a weight vector associated with the \( K \) modalities while \( w^J \) is the weight vector related to the \( J \) variables. From the following equalities an alternating algorithm is developed to maximize the FDA optimization problem subject to the structural constraint that \( w = w^K \otimes w^J \).

1Multi-way FDA analysis

\[ w^T S_B w = (w^K \otimes w^J)^T (X^T \otimes M_{B/T} X^T) (w^K \otimes w^J) \]
\[ = (w^J)^T (X^T (w^K \otimes I_J)) (X^T (w^K \otimes I_J))^T (w^K) \]
\[ = (w^K)^T (X^T (I_K \otimes w^J)) (X^T (I_K \otimes w^J))^T (w^K) \]

\[ w^T S_T w = (w^K \otimes w^J)^T (X^T \otimes M_{T} X^T) (w^K \otimes w^J) \]
\[ = (w^J)^T (X^T (w^K \otimes I_J)) (X^T (w^K \otimes I_J))^T (w^K) \]
\[ = (w^K)^T (X^T (I_K \otimes w^J)) (X^T (I_K \otimes w^J))^T (w^K) \]

\[ \lambda w^T w = \lambda (w^K \otimes w^J)^T (w^K \otimes w^J) \]

*Corresponding author. E-mail: gisela.lechuga@supelec.fr.

1This study was funded by a grant from the French Ministry of Health (Projet Hospitalier de Recherche Clinique registration #P051061 [2005]) and from departmental funds from the Assistance Publique-Hôpitaux de Paris. The research leading to these results has also received funding from the program “Investissements d’avenir” ANR-10-IAIHU-06.
$\mathbf{M}_{B/T}$ are positive semi-definite matrices that depend only of the label vector $\mathbf{y}$ and $\mathbf{X}^u$ is the unfolded matrix. The alternating algorithm is described in Algorithm 1, which starts by assigning random initial values for $\mathbf{w}^d$ or $\mathbf{w}^K$ and iterates a sequence of FDA problems. More specifically, each update boils down to perform FDA between $\mathbf{X}^d$ and either $\mathbf{X}^s$ or $\mathbf{X}^K$ where $\mathbf{X}^d = \mathbf{X}^u(\mathbf{w}^K \odot \mathbf{I}_J) = \sum_{j=1}^J (\mathbf{w}^K)_j \mathbf{X}_j$, and $\mathbf{X}^K = \mathbf{X}^u(\mathbf{I}_K \odot \mathbf{w}^d) = \sum_{i=1}^I (\mathbf{w}^d)_i \mathbf{X}_i$. We can note that $\mathbf{X}^d$ (resp. $\mathbf{X}^K$) is a $I \times J$ (resp. $I \times K$) matrix. Algorithm 1 allows to calculate $(\mathbf{w}^d_1, \mathbf{w}^K_1)$, which corresponds to the first discriminant axis. The following $C - 1$ discriminant axes, $\mathbf{w}^d_s, \mathbf{w}^K_s, s = 2, \ldots, C - 1$ are obtained subject to the additional orthogonality constraint between $\mathbf{w}^K_s$ and $\mathbf{w}^d_{s+1}$ [2].

2 Results

MFDA is applied on multi-modal diffusion images acquired on individuals divided into 3 classes: 39 controls, 65 coma patients with a positive outcome and 39 coma patients with a negative outcome ($I = 143$). 4 diffusion images namely fractional anisotropy (FA), mean diffusivity, axial diffusivity and radial diffusivity were acquired from the entire brain of the patients and controls ($K = 4$). Each image has a size of $91 \times 109 \times 91$ voxels, reshaped into a $1 \times 902629$ vector ($J = 902629$). We mention that due to the dimensionality of the dataset, a kernel version of FDA is used. The leave-one-out test error rate obtained with MFDA is equal to 71% whereas for the unfolded method the accuracy is equal to 76%. This slight loss in accuracy is compensated by an improvement in the interpretability of the obtained classifier as seen in Figures 1 and 2 which shows an axial cut at a central slice. Such improvement is partly due to the chosen structure for modeling $\mathbf{w}$: MFDA clearly separates the influence of spatial positions and the influence of the modalities. FDA applied to $\mathbf{X}^u$ results in 8 weight matrices (4 for each eigenvector), which complicate the interpretability, opposed to only 2 weight matrices obtained with MFDA which integrate all the modalities. Interestingly in our application, we have exhibited from MFDA that the discriminating voxels are located, as expected, within the main white matter bundles. Indeed, traumatic brain injury is characterized by the presence of diffuse axial injury mainly located within deep and axial white matter bundle.

![Figure 1. MFDA obtained weights.](image1)

![Figure 2. FDA FA obtained weights.](image2)

References


PLS-inspired multivariate linear modeling

Ying Li\textsuperscript{a} and Dietrich von Rosen\textsuperscript{a,}\textsuperscript{*}

\textsuperscript{a}Swedish University of Agricultural Sciences, Sweden

\textbf{Keywords:} PLS, two-step regression, Krylov space, growth curve model

\section*{Introduction}

Many real life applications produce “near-collinear” data which among others causes prediction problems. It is therefore often natural to apply some kind of regularized regression methods when analyzing them. For example, sub-set selection, ridge regression, lasso regression, principal components regression or partial least squares regression (PLS). Most of these methods are defined in an iterative algorithmic fashion. In each step a component (basis vector) is identified and thereafter residuals are formed which are taken as inputs in the next step. Cross validation is often used to choose the optimal number of components. The predictor of interest is in many applications a projection of the response variable on the space generated by the components. In the special case of PLS the space is Krylov structured.

Due to experimental conditions studies often follow special designs, including studies comprising different treatment groups. Usually in the above mentioned algorithms the design of an experiment is not taken into account. The advantage with our modeling approach is that mean and dispersion structures can be handled when performing prediction.

\section*{1 PLS and multivariate linear models}

In the proposed method we start with a fairly general model and apply a two step estimation approach. Let \( y \) be the dependent variable which should be predicted from \( x \) with \( \omega = \text{Cov}[x, y], \Sigma = D[x], E[x] = \mu_x, E[y] = \mu_y \) and then (in the subsequent suppose \( \omega \) to be known)

(i) Suppose \( x = \omega \delta + \epsilon \), where \( \epsilon \) is a normally distributed error and \( \delta \) an unknown parameter, i.e. \( x \) follows a linear model.

(ii) Predict via conditional expectation: \( y_0 = \omega \delta \Sigma^{-1}(x_0 - \mu_x) + \mu_y \) for given \( x_0 \) and estimated parameters.

Through (i) information from \( x \) is summarized, i.e. \( \mu_x \) is obtained by deriving the maximum likelihood estimator (MLE) of \( \delta \) and then in (ii) the predicted response \( y_0 \) is obtained from \( \mu_x \). However, there exists a fundamental problem of how to estimate \( \Sigma^{-1} \). If there are more variables then observations or variables are collinear then the usual sample covariance matrix will be singular or close to singular leading to a poor estimator of \( \Sigma^{-1} \) and a non-informative predictor. Note that a ridge approximation (Tikhonov regularization) is given by \( \Sigma^{-1} \approx (\lambda I + \Sigma)^{-1} \) for some \( \lambda \). Moreover, instead of \( \Sigma^{-1} \), the Moore-Penrose inverse can be used which is another common used shrinkage method, meaning that the eigenspace corresponding to small–eigenvalues of \( \Sigma^{-1} \) will be removed. Our research, however, is inspired by the Caley-Hamilton theorem which is implying that when (\( \Sigma \) is of size \( p \times p \))

\[
\Sigma^{-1} = \sum_{i=1}^{p} c_i \Sigma^{i-1}
\]

\textsuperscript{*}Corresponding author. E-mail: Department of Energy and Technology, Box 7032, SE-750 07 Uppsala, Sweden. dietrich.von.rosen@slu.se.
for some constants $c_i$ and therefore if $A = \Sigma G$, where $G = (\omega, \Sigma \omega, \ldots, \Sigma^{n-1} \omega)$
\[
x = \omega \delta + \varepsilon = \Sigma^{-1} \omega \delta \approx A \beta + \varepsilon.
\]

Instead of referring to the Caley-Hamilton theorem we may think of the appearance of the Krylov space in (1) as an effect of a polynomial approximation of $\Sigma^{-1}$. Anyhow, no matter how one is justifying the Krylov space, in both cases an alternative way of obtaining PLS-predictions has been obtained, i.e. the two-step approach mentioned above yields a predictor which is similar to the classical PLS-predictor.

The model in (1) is now extended. Let there be $n$ observations of the pair $(y, x)$ which are stored in $y$ and $X$. Then, instead of $x = A \beta + \varepsilon$ in (1),
\[
X = ABC + E,
\]
can be used, where $X: p \times n, A: p \times q$, is as in (1), $B: q \times k$, is an unknown parameter matrix, $C: k \times n$, is a known between individuals design matrix and $E \sim N_{p, n}(0, \Sigma, I)$. The matrix $C$ is the same design matrix as the one used to describe the models in a Gauss-Markov set up. The bilinear mean structure in (2) associates to the growth curve model (GMANOVA) which has been studied intensively for many years. Now, for given $X_0$ and $C_0$, prediction will take place according to
\[
y_0 = \omega' \Sigma^{-1} (X_0 - ABC_0) + \bar{y}.
\]

Remember that $A$ is a function of $\Sigma$ and therefore the mean structure and the dispersion structure are jointly parameterized, i.e. linked. However, under certain assumptions of the sample size, the interesting fact is that given $\omega$ explicit MLEs of $\Sigma$ and $B$ can be obtained (see [1]), which is far from obvious. Moreover, $\omega$ can always be estimated by exploiting the relation between $y$ and $X$. Hence, all parameters in the model can be estimated explicitly and it is possible to discuss details more thoroughly than if directly applying a PLS-algorithm.

Here are some consequences:

(i) A non-algorithmic prediction approach simplifying computations has been derived.

(ii) It offers greater flexibility than PLS and a better performance than at least ”standard PLS”. Moreover, it is now possible to combine several studies via the design matrix $C$.

(iii) Explicit MLEs of $\Sigma$ and $B$ are obtained where the estimator of $\Sigma$ differs from the sample covariance matrix.

(iv) Statistical model validation can take place.

References

Structured variable selection for generalized canonical correlation analysis

Tommy Löfstedt, Fouad Hadj-Selem, Vincent Guillemot, Cathy Philippe
Edouard Duchesney, Vincent Frouin and Arthur Tenenhaus

NeuroSpin, CEA, Saclay, France
Gustave Roussy, Villejuif, France
SUPELEC, Gif-sur-Yvette, France

Keywords: RGCCA, Generalised multiblock method, Variable selection, Structured penalty, Sparse penalty

Introduction

Regularized generalized canonical correlation analysis (RGCCA) [1], is a general framework for multiblock data analysis. Sparse GCCA [2] is a recent extension of RGCCA that includes $\ell_1$ penalties promoting sparsity for the weight vectors. In this paper, we propose to extend SGCCA so that it could exploit a pre-given structure between variables within blocks via structured and sparsity-inducing penalties [3]. Structured penalties were previously considered in a two-block setting with canonical correlation analysis [4]. However, combining such structured penalties with RGCCA poses new challenges for the optimisation techniques used. In this paper we propose a general multiblock algorithm that allows for structured and sparse penalties to be included in an RGCCA model.

1 Method

Let us consider $K$ blocks of variables, $X_1, \ldots, X_K$, of size $n \times p_k$ and an adjacency matrix $C = (c_{j,k})$, where $c_{j,k} = 1$ if $X_j$ and $X_k$ are connected, and $c_{j,k} = 0$ otherwise. The optimisation problem that is considered in this paper is defined by:

$$\min_{w_1, \ldots, w_K} \sum_{j=1}^K \sum_{k=1}^K c_{j,k} \left( \text{Cov}(X_j w_j, X_k w_k) \right)$$

subject to $P_k(w_k) \leq p_k$, $N_k(w_k) \leq n_k$, $\tau_k \|w_k\|^2_2 + (1 - \tau_k) \text{Var}(X_k w_k) \leq 1$, for all $k = 1, \ldots, K$,

where $w_1, \ldots, w_K$ are the weight vectors; $g$ is called the inner-weighting scheme (usually the identity, the absolute value or the square function); $P_k$ are convex functions that have a known projection operator; the regularisation parameters $\tau_k \in [0, 1]$ provide a way to control the trade-off between maximising correlation and maximising covariance. The $N_k$ are convex and possibly non-differentiable functions that can be expressed in the framework of Nesterov, as described in [3]. More precisely, $N_k$ can be expressed in the form

$$N_k(w_k) = \sum_{G=1}^{G_k} \| A_{k,G} w_k \|_q,$$

in which $G_k$ is the number of groups which are associated with $N_k$, $A_{k,G}$ is a linear operator for group $G$ associated with $N_k$, and $\| \cdot \|_q$ is the $q$-norm with dual norm $\| \cdot \|_{q'}$. The functions $N_k$ are smoothed using the technique of

*Corresponding author. E-mail: lofstedt.tommy@gmail.com.
Nesterov [3, 5] as follows:
\[
\hat{N}_k(\mu_k, w_k) = \langle A_k w_k, \alpha^*_k \rangle - \frac{\mu_k}{2} \|\alpha^*_k\|^2,
\]
for which \(\lim_{\mu_k \to 0} \hat{N}_k(\mu_k, w_k) = N_k(w_k)\). Then \(\hat{N}_k\) is convex and differentiable and is used instead of \(N_k\).

The solution to this problem is found by splitting Equation (1) in two parts: One part that can be expressed as a projection, and one part that is smooth and can be summed in a Lagrange formulation together with the objective function of Equation (1). The problem is reexpressed as

\[
\begin{align*}
\text{minimise} & \quad \hat{f}(w_1, \ldots, w_K) = \text{minimise} - \sum_{j=1}^{K} \sum_{k=1}^{K} c_{j,k} g(\text{Cov}(X_jw_j, X_kw_k)) + \sum_{k=1}^{K} \nu_k \hat{N}_k(\mu_k, w_k) \\
\text{subject to} & \quad w_k \in \mathcal{W}_k = \{w|w \in \mathbb{R}^{p_k} \land P_k(w) \leq p_k \land \tau_k \|w\|^2_2 + (1 - \tau_k) \text{Var}(X_kw) \leq 1\}, \forall k = 1, \ldots, K.
\end{align*}
\]

2 Algorithm

Equation (2) is a multiconvex function with a indicative constraint over a convex set. This optimisation problem is solved with the Algorithm outlined in Algorithm 1. This algorithm is related to alternating least squares, and the projected gradient method.

**Algorithm 1 A generalised RGCCA minimiser**

**Requires:** \(\hat{f}, \nabla \hat{f}, w_k^{(0)} \in \mathcal{W}_k, \varepsilon > 0\)

**Ensure:** \(w_k^{(s)} \in \mathcal{W}_k\) such that \(\varepsilon \in \partial \hat{f}(w_1^{(s)}, \ldots, w_K^{(s)}, v_1, \ldots, v_K)\)

1: repeat
2: \quad for \(k = 1\) to \(K\) do
3: \quad \quad repeat
4: \quad \quad \quad \quad \quad w_k^{(s+1)} = \text{proj}_{\mathcal{W}_k}(w_k^{(s)} - t_k \nabla w_k^{(s)} \hat{f}(w_1^{(s)}, \ldots, w_K^{(s)}))
5: \quad \quad \quad until \(\|w_k^{(s-1)} - w_k^{(s)}\|_2 \leq t_k \varepsilon\)
6: \quad end for
7: until \(\|w_k^{(s)} - w_k^{(s+1)}\|_2 \leq t_k \varepsilon\) for all \(k = 1, \ldots, K\)

3 Discussion and conclusions

This structured variable selection combined with RGCCA is exemplified using CGH and microarray data for the prediction of tumour location in glioma. We exemplified the method by using a total variation penalty on the CGH data and a group \(\ell_{1,2}\) penalty, with overlap, on the microarray data. The groups were obtained using the KEGG pathway database.

4 Acknowledgements

This work was supported by grants from the French National Research Agency: ANR IA BRAINOMICS (ANR-10-BINF-04), and a European Commission grant: MESCOG (FP6 ERA-NET NEURON 01 EW1207).

References


A study of the impact of Corporate Social Responsibility and price image on retailer personality and consumers’ reactions (satisfaction, trust and loyalty to the retailer)

Lombart C.\textsuperscript{a*} and Louis D.\textsuperscript{b}

\textsuperscript{a} Audencia Nantes. School of Management (Audencia – PRES LUNAM), Research Center in Marketing & Distribution In Situ, Nantes, France
\textsuperscript{b} IUT de Saint-Nazaire, LEMNA, Nantes, France

1 Conceptual framework and research hypotheses

The objective of this study is to examine the influence of the retailer’s CSR policy and price image on its personality. The influence of these two variables on the consequences of retailer personality – consumer satisfaction, trust and loyalty to the retailer (measured by their attitude and future behavioral intentions) – is also analyzed. The model proposed in this study (Figure 1) primarily considers the relations between two antecedents of retailer personality: retailers’ CSR policy and price image. Links between retailer personality and the consequences of this concept (consumer satisfaction, trust and loyalty to the retailer) are incorporated in the model. Lastly, links between the retailer’s CSR policy and price image and the consequences of retailer personality are integrated.

2 Methodology

2.1 Retailer studied

Système U is a French cooperative of food retail merchants. In early 2012, the retailer’s market share (in value) was 10.1\% (for hypermarket and supermarket), positioning it behind the retailers Carrefour, Leclerc and Intermarché. In terms of price image, Système U is situated behind Leclerc, which has highlighted its low price image in its institutional communication for several years. Système U also has a Corporate Social Responsibility (CSR) policy via its commitments to improve the environmental quality of its offering (organic products and ecological cleaning products), store design (efficient in terms of natural resources and energy), and reduction and recycling of waste in its stores (treatment and recovery of plastic).

2.2 Sample

The study was conducted on a convenience sample of 352 individuals, ages 20 to 25, within a store laboratory. This store is an exact replica of several food shelves (pasta and rice, fruit juice) and hygiene (shower products and shampoo) of a Système U supermarket. In addition, two shelves were dedicated to products (preserved vegetables, soups, pasta, biscuits, fruit juice, jams, etc.) that bore an organic agriculture label (French and/or
European), under the store brand. Another shelf was dedicated to ecological cleaning products (laundry products, housecleaning, dishwashing liquids, paper towels, etc.) that bore the European ecolabel, under the store brand.

3 Results

3.1 Test of measurement model

Before testing the model proposed in this study using structural equations modeling (SEM), confirmatory factor analysis was performed on the data collected using the partial least squares method (PLS) with a bootstrap procedure (200 iterations). In line with the literature, second-order factors were posited for Corporate Social Responsibility, retailer personality, trust and future behavioral intentions.

3.2 Test of structural model

The model proposed in Figure 1 was tested with PLS and a bootstrap procedure (200 iterations). Indices of fit for the external GoF (measuring the performance of the measurement model) and internal GoF (measuring the performance of the structural model) was 0.998 and 0.867 respectively. The measurement and structural models proposed are therefore satisfactory. Table 1 presents the results of the structural model.

<table>
<thead>
<tr>
<th>Path coefficients</th>
<th>t-value</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corporate social responsibility → Agreeableness</td>
<td>0.286</td>
<td>5.431**</td>
</tr>
<tr>
<td>Price image → Agreeableness</td>
<td>0.133</td>
<td>2.529*</td>
</tr>
<tr>
<td>Corporate social responsibility → Conscientiousness</td>
<td>0.249</td>
<td>4.798**</td>
</tr>
<tr>
<td>Price image → Conscientiousness</td>
<td>0.221</td>
<td>4.250**</td>
</tr>
<tr>
<td>Corporate social responsibility → Sophistication</td>
<td>0.162</td>
<td>3.065**</td>
</tr>
<tr>
<td>Corporate social responsibility → Disingenuousness</td>
<td>-0.187</td>
<td>-3.450**</td>
</tr>
<tr>
<td>Price image → Disingenuousness</td>
<td>-0.137</td>
<td>-2.530*</td>
</tr>
<tr>
<td>Agreeableness → Satisfaction</td>
<td>0.346</td>
<td>7.852**</td>
</tr>
<tr>
<td>Price image → Satisfaction</td>
<td>0.325</td>
<td>7.476**</td>
</tr>
<tr>
<td>Corporate social responsibility → Satisfaction</td>
<td>0.156</td>
<td>3.426**</td>
</tr>
<tr>
<td>Disingenuousness → Satisfaction</td>
<td>-0.197</td>
<td>-4.610**</td>
</tr>
<tr>
<td>Satisfaction → Trust</td>
<td>0.482</td>
<td>12.218**</td>
</tr>
<tr>
<td>Corporate social responsibility → Trust</td>
<td>0.360</td>
<td>9.261**</td>
</tr>
<tr>
<td>Sophistication → Trust</td>
<td>0.132</td>
<td>3.592**</td>
</tr>
<tr>
<td>Introversion → Trust</td>
<td>0.089</td>
<td>2.495*</td>
</tr>
<tr>
<td>Satisfaction → Attitude</td>
<td>0.765</td>
<td>23.521**</td>
</tr>
<tr>
<td>Sophistication → Attitude</td>
<td>0.088</td>
<td>3.436**</td>
</tr>
<tr>
<td>Corporate social responsibility → Attitude</td>
<td>0.101</td>
<td>3.357**</td>
</tr>
<tr>
<td>Trust → Attitude</td>
<td>0.072</td>
<td>1.969*</td>
</tr>
<tr>
<td>Attitude → Future behavioral intentions</td>
<td>0.449</td>
<td>6.795**</td>
</tr>
<tr>
<td>Satisfaction → Future behavioral intentions</td>
<td>0.390</td>
<td>5.939**</td>
</tr>
<tr>
<td>Sophistication → Future behavioral intentions</td>
<td>0.122</td>
<td>3.516**</td>
</tr>
<tr>
<td>Conscientiousness → Future behavioral intentions</td>
<td>0.116</td>
<td>3.248**</td>
</tr>
</tbody>
</table>

Note: * p < 0.05 and ** p < 0.01.

4 Conclusion

This research primarily contributes to better understanding the formation of retailer personality. It demonstrates the influence of perceived retailer’s CSR policy and price image on four (agreeableness, conscientiousness, sophistication, disingenuousness) and three (agreeableness, conscientiousness and disingenuousness) retailer personality traits respectively (out of five). Only the personality trait introversion was not influenced by the two variables (CSR and price image).
Formative versus reflective measurement: Fallacy or reality

George A. Marcoulides\textsuperscript{a*}, Andrew Hardin\textsuperscript{b}, and Jerry Chang\textsuperscript{b}

\textsuperscript{a} University of California, Riverside, USA
\textsuperscript{b} University of Nevada, Las Vegas, USA

**Keywords:** composite indicators, formative indicators, measurement models.

**Introduction**

There has been considerable debate in the extant literature surrounding the appropriateness of implementing composite and causal indicators in measurement models. In this panel we discuss recent literature concerning this much-debated area, including the necessary distinction between models comprised of these indicators, and the frequent lack of psychometric theory underlying their implementation in empirical research. Designing studies that avoid the inappropriate application of these respective indicators is advocated.
Multivariate regression modeling with clustered correlated predictors

F. Martella, D. Vicari\textsuperscript{1*}, M. Vichi
Dipartimento Scienze Statistiche, Sapienza Università di Roma, Rome, Italy

Keywords: Multivariate regression, partition of predictors, weakly-correlated factors

Summary

Applications where several dependent variables (responses) are predicted by a large set of explanatory variables have been considered in various disciplines (bioinformatics, brain imaging, data mining, genomics and economics). The standard use of the Ordinary Multivariate Regression Model to accommodate this issue is not always a good choice when the number of predictors is large. In such respect, in fact, not only we can encounter difficulties in interpretation of the (many) regression coefficients, but also the multivariate prediction can be affected by the presence of multicollinearity among predictors which poses problems because influences the stability and the prediction can be unreliable. In literature many proposals and strategies for dealing with such problems have been proposed such as standard variable selection methods, penalized (or shrinkage) techniques and dimensionality reduction methods (DRMs).

In particular, we focus on DRMs, which attempt to build small set of linear combinations of the predictors, used as input to the regression model, and differ in how the linear combinations are built. See among others, principal component regression (PCR) [1], canonical correlation regression (CCR) [2], partial least squares regression (PLSR) [3], [4] for the continuum regression (unified regression technique embracing OLS, PLSR and PCR), reduced rank regression (RRR) ([5], [6]), redundancy analysis (RA) [7]. Recently, [8] proposed a general formulation for dimensionality reduction and coefficient estimation in multivariate linear regression, which includes many existing DRMs as specific cases. Finally, [9] proposed a new formulation to the multiblock setting of latent root regression applied to epidemiological data and [10] investigated a continuum approach between MR and PLS. A drawback of DRMs is that they may generally suffer from a possible difficulty of interpretability of the resulting linear combinations which many researchers try to overcome through rotation methods.

In the DRMs framework, here we propose to build weakly correlated factors as linear combinations of disjoint classes of correlated predictors that best predict the responses in a pre-specified sense. In other words, we simultaneously perform clustering of the predictors and the estimation of the regression coefficients of the derived weakly correlated factors. This turns out to be a relevant gain in the interpretation of the regression analysis, since factors are formed by disjoint classes of predictors and, therefore easily interpretable and nicely displayed by a path diagram identifying the underlying relations between predictors, factors and responses. In particular, firstly we will consider the multivariate regression model based on the optimal partition of predictors (MRBOP) which has been proposed by [11]. MRBOP has been estimated in a least square context. Secondly, we will consider to derive MRBOP in a maximum likelihood framework, which is intuitively appealing for comparisons with other methodologies, for choosing the number of subsets leading factors and allowing inference on model parameters. The performance of the proposed model is illustrated by simulated and real data sets. The results are encouraging and would deserve further discussion.
References


Sparse nominal-level PLSR with interactions

From time series data to nonlinear differential equation model

Harald Martens\textsuperscript{a}\textsuperscript{*}

\textsuperscript{a} Prof. emerit., Inst. Electronics & Telecommunication, Norwegian U. of Science and Technology, Trondheim, Norway.
E-mail: martens@ntnu.no

Keywords: sparse, nominal, nonlinear interactions, ODE, PLSR

1 Summary
A new PLS-based regression method - sparse nominal-level PLSR with interactions – will here be introduced and applied to multivariate time series data. The purpose is to present a generic method for developing a compact, graphically interpretable ODE model of a complex system via nonlinear dynamic PLS subspace modelling. The method will here be used for illustrating how a complex, highly nonlinear mathematical model from e.g. computational biology may be simplified and assessed in terms of a compact differential equation model. Cross-model-validation and jack-knifed covariance estimation will be employed in order to optimize model rank and to assess the uncertainty of the resulting model structure. The present development is part of a PLS-based system for multivariate dynamic metamodelling to reveal how complex mathematical models actually behave under different circumstances, and to use this for faster computation.

2 Introduction
Mechanistic modelling allows the realization of existing knowledge and/or theories of causality, and it can give critical assessments and useful predictions. However, sometimes the existing “knowledge” may in fact be incomplete or wrong, and even if it were sufficiently well approximated, its mathematical realization in terms of coupled differential equations etc may be difficult to overview and slow to compute.

Explorative statistical modelling, on the other hand, may be very adaptive, and relatively easy to develop and validate, if enough good data are available. It can give good predictive ability under the condition observed. But alas, it may be misleading outside this realm. Moreover, it normally does not directly give causal representation of the system, even though the search for causality is at the heart of science.

The mechanistic modelling often consists of combining various well-defined mathematical equation elements. Here we focus on one class of mechanistic models, ordinary differential equations (ODEs). These models describe a system ‘from the inside’, in terms of how fast the system’s K state variables $x$=[x\textsubscript{1}, x\textsubscript{2},..., x\textsubscript{K}] (properties, e.g. concentration of chemical compounds) change under different conditions. The rate at which the K state variables change, $\frac{dx}{dt}$=[dx\textsubscript{1}/dt, dx\textsubscript{2}/dt,..., dx\textsubscript{K}/dt], are modelled as functions of the level of these state variables, $x$, as controlled by a set of model parameters, $\theta$:

$Rates = M(States, Parameters)$, \ i.e. \ $\frac{dx}{dt} = M(x, \theta)$

If the original model $M$ is complex, it is difficult to foresee how it will behave, theoretically. Therefore there is often a need for a reduced model, with fewer equations and a reduced set of state variables and parameters, $\frac{dx_{\text{Red}}}{dt} = M_{\text{Red}}(x_{\text{Red}}, \theta_{\text{Red}})$, that maintains the essential properties of the original model $M$ under the chosen...
conditions. The hope is that this can give simpler interpretation and faster computation. The purpose of the present paper is to generate such a reduced model based on jackknifed sparse nominal-level PLSR with interactions. Thus it represents an attempt at retaining the view ‘from the inside’, but in a simplified metamodelling framework.

3 Method

The linear ODE version of \( \frac{dx}{dt} = M(x, \theta) \) is \( \frac{dx}{dt} = x \cdot B \), where the Jacobian matrix \( B \) (K x K) defines the system’s dynamic behavior. When integrated over a time sequence at N time points, from an initial set of states \( x_{\text{init}=0} \) and with a given set of parameter values \( \theta \), e.g. by a numerical integration functionality such as Matlab’s function ode45, this results in K time series of data for the K state variables, \( X(N \times K) \). Assume first that \( B \) in this linear dynamic model is to be estimated from simulation state variable data \( X \), obtained by numerical integration at N time points for a given set of parameters \( \theta \), but from M different initial conditions \( x_{\theta(m),0} \), \( m=1,2,\ldots,M \). This results in a state matrix \( X(N \times M \times K) \). Taking the temporal derivatives \( \frac{dx}{dt} \) within each of the M time series results in a corresponding rate matrix \( Y(N \times M \times K) \). The Jacobian \( B \) may now be estimated from a (reduced-rank) linear regression model, as estimated by e.g. PLSR. Scores and loadings plots give overview. Cross-validation reveals the optimal model rank, A, and the linear dynamic model at rank A may be written \( Y = X \cdot B_A + F_A \) (ignoring offsets). Jackknifing gives the precision of the \( B_A \) estimate.

As showed at the 2009 PLS meeting in Beijing [1], this linear model is not good enough if the original model \( \frac{dx}{dt} = M(x, \theta) \) is highly nonlinear. Instead of expanding the PLSR model with polynomial terms or develop local PLSR models, as is usually the first remedies to handle nonlinearity with PLSR, a nominal-level PLSR was there presented: Each state variable was split into a set of e.g. 20 different indicator variables, and \( X \) (N x K) was thus replaced by the combined set of K-20 indicator variables. Again a cross-validated PLSR was employed to determine optimal rank, A. But the resulting regression coefficient matrix \( B_A \) was not of size (K x K) but (K-20 x K). To interpret the results graphically, each of the K x K rate/state relationships was displayed as a discontinuous curve with 20 points, each point with its jackknifed uncertainty estimate. From this, the owner of the data could see what type of nonlinearity was required, and formulate a continuous mathematical ODE model based on nonlinear main effects (i.e. no interactions between the state variables).

In the present paper, an extension of this main-effect nominal PLSR method presented at PLS’09 [1] and discussed at PLSR’12 [2], will be presented. The extension also searches for important interaction effects in the data from the original system \( \frac{dx}{dt} = M(x, \theta) \). This is done by including two- and three-variable interactions etc between all the nominal variables from all the K state variables. The PLSR structure model remains the same, \( Y = X \cdot B_A + F_A \), but a sparse version of PLSR is employed in order to avoid the over-fitting effect of having many near-irrelevant X-variables. Jackknife covariance estimation is presented and used for assessing the nullspace sloppiness due to neutral parameter sets, and hence the resulting model ambiguity.

4 References


SO-PLS as an exploratory tool for path modelling

Elena Menichelli\textsuperscript{a*}, Trygve Almøy \textsuperscript{b}, Oliver Tomic\textsuperscript{c}, Nina V. Olsen\textsuperscript{d}, Tormod Næs\textsuperscript{d,e}

\textsuperscript{a}Telenor Research, Telenor ASA, N-1331 Fornebu - Oslo, Norway
\textsuperscript{b}The Norwegian University of Life Sciences, N-1432 Ås, Norway
\textsuperscript{c}Norwegian Knowledge Centre for the Health Services, N-0130 Oslo, Norway
\textsuperscript{d}Nofima AS, N-1431 Ás, Norway
\textsuperscript{e}University of Copenhagen, 1958 Copenhagen, Denmark

Keywords: path modelling; SO-PLS; orthogonalisation; regression; multi-block.

Introduction

An important challenge in consumer science is to relate information from several sources to each other for the purpose of obtaining improved insight. Path modelling (PM) methods \cite{1} are important for approaching such problems, but there are a number of challenges to consider. For instance it is not always obvious how to split the variables into meaningful blocks and to set up the paths between the blocks. Another challenge is the uni-dimensionality assumption of the blocks which may be questionable in many practical situations \cite{2}.

In this paper we propose the use of a method from the multi-block tradition \cite{2} to solve problems in typical PM situations. The method is called SO-PLS (Sequential and Orthogonalised Partial Least Squares) regression \cite{2} and has many benefits. For instance, it is invariant to the relative scaling of the blocks and it can be used for determining the additional contribution of new blocks incorporated in the model. The method is based on an underlying idea about a structured relationship between the blocks, but it is more explorative and therefore more flexible. It allows for combining variables in one block, without any requirement of uni-dimensionality. The paths are merely proposals of important relations to consider rather than strict causal paths.

In this paper we discuss some of the properties of SO-PLS and how the method relates to standard PM. Focus will be on applications in consumer acceptance studies in which it has never been used before. Focus will also be on how the method handles multidimensionality of the blocks and how it can be used to combine variables.

1 The SO-PLS approach to PM

The SO-PLS approach to PM \cite{2} is based on the multi-block SO-PLS regression method. This is a method developed for estimating regression equations with $N$ blocks of independent variables, i.e.

$$ Y = X_1B_1 + X_2B_2 + \ldots + X_NB_N + E \quad (1) $$

where the $Y$ represents the matrix of dependent variables and the $X$'s are the different blocks of input variables. The estimation is based on sequential use of orthogonalization and PLS regression \cite{1}. Focus is thus on additional effects, measured by the additional explained variance of a block incorporated in the model. When used in PM, the SO-PLS is based on an underlying notion of a structural relationship between the blocks, but estimates an independent model for each endogenous block \cite{2}. Since the method handles any type of data blocks regardless of dimensionality, it also opens the possibility of combining closely related variables into larger blocks of data. In order to simplify interpretation, the PCP method \cite{3} is performed for each model, since it compresses all information in the SO-PLS model down to $X$- and $Y$-loadings plots and a scores plot.

2 Study case

2.1 Data set

In this study 12 iced-coffees were presented to 100 consumers. Products vary according to a fractional factorial design based on: calorie content (60 or 90 kcal per 100 ml), origin (Norway or Italy), price (17, 23 or 29 NOK) and type of coffee (“latte” or “espresso”). The consumers’ probability of buying was evaluated on a

* Corresponding author. E-mail: elena.menichelli@telenor.com
9-point scale. Consumers’ demographics, habits and attitudes were recorded. The data are structured according to the approach recommended in [4] for combining product information, consumers’ characteristics and acceptance. Data are thus split into data blocks according to their nature and related into a dependence diagram (Figure 1). Since there are four endogenous blocks, four independent models are estimated. In this abstract just the largest model is briefly described: the probability of buying is predicted from the consumers’ characteristics blocks, with neophobia orthogonalised to demographics, importance of attributes orthogonalised to demographics and neophobia, consumption habits orthogonalised to the previous blocks.

![Figure 1](image)

**Figure 1.** The dependence diagram shows the order of the consumer characteristics blocks that is considered in the SO-PLS approach to PM for explaining the probability of buying.

### 2.2 Results

From the root mean squared error of prediction (RMSEP) plot (not shown) the results with the best prediction ability are given by using 1 component in the demographics and neophobia blocks, 3 components in the block of the importance of the extrinsic attributes, no components in the consumption habits block. The latter is thus discarded from the model, since its additional contribution does not improve prediction. The first two PCP components explain together 81% of the variance of the predicted Y. The first axis in the Y-loadings plot (not shown) discriminates between espresso and latte iced-coffees. Country of production and price characterize the second component. The X-loadings plot (not shown) shows that both the neophobia and the importance of attributes variables are the ones spreading the space the most. The position of the demographic variables indicates that the female and young consumers give more importance to calorie content and price than to country of origin and coffee type. The consumers that stick to the usual food are opposite to the ones that like to try new food. The food-neophobic consumers will more likely buy the latte iced-coffees. The second component in the X-loadings plot suggests that consumers willing to experience new food and giving high importance to price will probably buy medium-low price Italian products.

### 3 Conclusions

For the SO-PLS methods there is no need for any prior consideration of uni-dimensionality, since PLS can be used directly for many and possibly collinear variables. This implies that one can also construct more general (or conceptual) blocks and in this way obtain information not only about how the variables in the block are related, but also about how they as a block influence other blocks. This may simplify analysis considerably. When one opens up the possibility of using several dimensions in each block, the same dimensions are not necessarily used for prediction and to be predicted, as in the present example (here not shown). Considering one model for each endogenous block may seem a bit complex, but each model can be simplified by the use of PCP. With too many blocks, one could also merge some of them, which is often impossible when using standard approaches. The approach is pragmatic and useful for exploratory purposes and, in consumer acceptance studies, is able to reveal important relations between consumers, products and acceptance data.

### References


Statistical properties of coefficient estimates in PLS-PM: reflective versus formative measurement model

Simona C. Minotti \(^a\) and Antonio Ciampi \(^b\) and Giuseppe Lamberti \(^c\) and Tomas Aluja-Banet \(^c\)

\(^a\) Università degli Studi Milano-Bicocca, Milan, Italy
\(^b\) McGill University, Montreal, Canada
\(^c\) Universitat Politecnica de Catalunya, Barcelona Tech, Spain

Keywords: PLS-PM, Reflective, Formative

Abstract

We investigate the statistical properties of PLS Path Modeling (PLS-PM) estimators in the case of misspecification of a measurement model. We also discuss possible criteria for choosing between reflective and formative measurement models. This is particularly useful for situations in which there is little theoretical basis to decide \textit{a priori} whether a construct is formative or reflective. The study is based on a Monte Carlo simulation analysis and some real applications.

1 Introduction

In the PLS-Path Modeling (PLS-PM) literature, the specification of a measurement model as reflective or formative is the object of a lively debate. The problem arises because in many cases there is no compelling underlying theory to guide the choice. However, recent research shows that even in correctly specified models, serious problems of interpretation may occur when a measurement model is specified as formative. Therefore, [1] suggest to temporarily suspend the use of formative measurement and use instead the alternative methods recently recommended in the literature. Yet, it appears highly unlikely that researchers will follow this drastic course of action.

In the present context there is a clear need for a systematic investigation of the consequences of misspecifying a measurement model through comprehensive simulations. While a few authors have carried out simulation studies to investigate the robustness of PLS estimates, their focus is the application of PLS-PM to the European Customer Satisfaction Index (ECSI) data [2, 3, 4].

In this work we investigate the statistical properties of PLS-PM estimators, focussing on the issue of measurement model misspecification in a general context. We outline a broad framework in which PLS-PM models of varying degrees of complexity can be specified and simulated, and we present some results. In addition, we discuss three applications to real data: the well known customer satisfaction data, student satisfaction data, and a new data set on mental health from a 3D study (Delirium, Dementia, Depression).

We conclude by discussing data based criteria for choosing between reflective and formative measurement models when indications based on theory are lacking or unclear.

\(^*\) Corresponding author. E-mail: Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy. simona.minotti@unimib.it
2 The results

As an example, we present preliminary results based on only three path models with two latent variables; these are shown in Figures 1, 2 and 3. Corresponding Tables 1, 2 and 3 show mean squared error, bias and variance of the path coefficient estimators when the model is well specified and under different kinds of misspecification.

![Figure 1 Reflective model.](image1)

![Figure 2 Formative model.](image2)

![Figure 3 Formative-reflective model.](image3)

<table>
<thead>
<tr>
<th>Table 1 Reflective model.</th>
<th>Table 2 Formative model.</th>
<th>Table 3 Formative-reflective model.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>BIAS</td>
<td>VAR</td>
</tr>
<tr>
<td>Reflective &amp; Reflective</td>
<td>0.1</td>
<td>0.00098</td>
</tr>
<tr>
<td>Reflective &amp; Formative</td>
<td>0.1</td>
<td>0.00024</td>
</tr>
<tr>
<td>Formative &amp; Reflective</td>
<td>0.1</td>
<td>0.00024</td>
</tr>
<tr>
<td>Formative &amp; Formative</td>
<td>0.1</td>
<td>0.00024</td>
</tr>
</tbody>
</table>

Although space limitations prevent us from presenting more complex models, it is clear that misspecification may have significant consequences on the estimates, and therefore on interpretation.

References


PLS and n-PLS for modelling the total antioxidant capacity of food extracts using excitation emission fluorescence spectroscopy

Joanna Orzel* and Michal Daszykowski
Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland

Keywords: EEM, ORAC, multivariate calibration, multi-way calibration

Introduction

Nowadays, antioxidants are one of the most popular food constituents that have ability to deactivate harmful for living organisms free radicals. Food is the main source of these vital compounds. The total antioxidant content of food is described by its total antioxidant capacity parameter (TAC). The TAC refers to the amount of active antioxidants contained in a sample and can be evaluated using various assays. Primarily, assays are based on a chemical reaction between artificially generated free radicals and compounds of interest. Reaction is monitored using spectrophotometry, fluorimetry or other instrumental technique [1]. All of proposed TAC assays require large amounts of reagents. Moreover, either incubation of samples at a particular temperature or measurements carried out over a long period of time are additional difficulty.

Many antioxidant compounds present in foods exhibit fluorescent properties. We focus on the development of a method facilitating evaluation of total antioxidant capacity of food samples on the basis of their fluorescence spectra and their chemometric modeling.

Excitation-emission fluorescence spectroscopy offers the possibility to observe simultaneously several fluorescent compounds contained in a complex sample. Excitation emission matrices (EEMs) are two-dimensional signals. Collected for a set of samples form a complex three-way data (third order tensor). To obtain information of TAC from EEMs a calibration model need to be constructed. In presented study models were constructed using designated for three way data – n-PLS [2] and compared with classic PLS models [3] constructed for unfolded signals.

The TAC of food samples was evaluated using the oxygen radical absorbance capacity assay (ORAC assay) [4]. EEMs were registered using quartz cuvette and a portable device – a fluorescent probe.

In presented study, coffee, which is known as reach in antioxidant food, was chosen to carry out the preliminary studies for a novel method for TAC evaluation.

1 Experimental

The ORAC assay is often recommended for examination of food TAC. It relays on a damage of a fluorescent probe (fluorescein) caused by free radicals, monitored as decrease of fluorescence intensity. Antioxidants present in a sample inhibit damage of the probe, thus the fluorescent signal is sustained. Simultaneously, fluorescence intensity of a blank probe (without antioxidants and with the same amount of free radical generator and fluorescence agent) is examined. Both reactions are driven to completion. TAC is expressed as the area contained between two curves defined for blank and tested samples. Water soluble vitamin E analog - Trolox served as standard for the ORAC assay. The TAC of samples were expressed as Trolox equivalents

* Corresponding author. E-mail: jorzel@us.edu.pl.
(TE) - μmol of Trolox per 5 g of a sample. In the presented study, fifteen various types of coffee were prepared in three laboratory replicates, resulting in 45 samples, were examined.

EEMs were collected for all coffee samples. Excitation range was set between 230 nm and 650 nm with 10 nm intervals and emission spectra were collected for each excitation wavelength in a range of 330 nm to 700 nm with 2 nm interval.

2 Modeling

Raw 2D spectra were preprocessed before construction of calibration models. Rayleigh scattering, which is typical artifact of EEMs, was handled using the Delaunay triangulation as described in [5]. Preprocessed data were randomly divided into training (30 samples) and test (15 samples) sets. The training set was modeled using n-PLS and PLS methods. Signals for n-PLS modeling were collected in a three-way tensor of size samples × emission wavelengths × excitation wavelengths, (30×186×43). For PLS modeling they were unfolded into matrix of size samples × (emission wavelengths × excitation wavelengths), (30×7998). See Fig. 1.

Performance of constructed models with optimal properties were evaluated using an independent test set. The results of multi-way and classic modeling were compared. Fit of models with optimal properties were equal 313 TE and 453 TE for PLS and n-PLS, respectively. However, the n-PLS model, which requires five factors, results in better prediction properties (465 TE) than PLS model (588 TE) constructed using one factor less.

3 Acknowledgement

JO appreciates the support of Polish National Science Center grant no. 2011/03/N/ST4/00713.

References


Partial Possibilistic Regression Path Modeling

Palumbo F.\textsuperscript{a} and Romano R.\textsuperscript{b}\textsuperscript{*}

\textsuperscript{a}Federico II University, Naples, Italy
\textsuperscript{b}University of Calabria, Cosenza, Italy

Keywords: Interval valued data, Possibilistic regression, Component-based SEM

Introduction

Structural Equation Models (SEM) \cite{1} consist of a network of causal relationships among latent variables (LVs) defined by blocks of manifest variables (MVs). Under the framework of component-based estimation methods, Partial Least Squares Path Modeling (PLS-PM) \cite{2} represents a statistical approach to SEM with an increasing popularity in several areas.

In PLSPM, similarly to classical least squares regression, the process of data analysis is represented by the simple equation: \textit{data} = \textit{model} + \textit{error}. Here the main source of uncertainty is randomness, i.e. random measurement errors in recording observations. However, there are other sources of uncertainty besides randomness. When human judgments are involved into the data analysis process the vagueness is the major source of uncertainty \cite{3}.

The present paper aims to propose a flexible methodology for analyzing phenomena characterized by a complex structure of relations among the variables and where the vagueness is the major source of uncertainty. This is achieved by combining the principles of PLS-PM \cite{2} in modeling the net of relations among variables, with the principles of Possibilistic Regression \cite{4}, which is oriented to model vagueness rather than randomness. A combination of these two approaches has already been proposed in previous papers (\cite{5}; \cite{6}), where the possibilistic regression was used to model relations in the structural model and the least squares regression was used to model relations in the measurement model. The novelty of the proposed approach consists in using the quantile regression (\cite{7}; \cite{8}) for modeling the measurement model. This choice allows to optimize the use of possibilistic regression, which is extremely sensitive to outliers.

1 Partial Possibilistic Regression Path Modeling

Partial Possibilistic Regression Path Modeling (PPR-PM) is a method to analyze phenomena whose description requires the analysis of a complex structure of relations among the variables inside the system, and where there is an additional source of complexity arising from the influential human beings involvement.

Let us assume \( P \) variables \((p = 1, \ldots, P)\) observed on \( N \) units \((n = 1, \ldots, N)\) and collected into a partitioned table \( X = [X_1, X_h, \ldots, X_H] \), where \( X_h \) is the generic block composed by \( P_h \) indicators.

In PPR-PM, an iterative procedure permits to estimate the latent variable scores and the outer weights, while path coefficients come afterward from possibilistic regressions between the estimated latent variables.

The algorithm computes the latent variables’ scores alternating the outer and inner estimation till convergence. The procedure starts on centered (or standardized) MV by choosing arbitrary weights \( w_{ph} \). In the external estimation, the latent variable is estimated as a linear combination of its own MV:

\[ v_h \propto \sum_{p=1}^{P_h} w_{ph} x_{ph} \quad X_h w_h \quad (1) \]

\textsuperscript{*}Corresponding author. E-mail: rosaria.romano@unical.it
where $v_h$ is the standardized outer estimation of the latent variable $\xi_h$ and the symbol $\propto$ means that the left side of the equation corresponds to the standardized right side. In the internal estimation, the latent variable is estimated by considering its links with the other adjacent $H'$ latent variables:

$$\hat{\theta}_h \propto \sum_{h'=1}^{H'} e_{hh'} v_{h'}$$

where $\hat{\theta}_h$ is the standardized inner estimation of the latent variable $\xi_h$ and the inner weights, according to the so-called centroid scheme, are equal to the sign of the correlation between the outer estimate $v_h$ of the $h$-th latent variable and the outer estimate of the $h'$ latent variable $v_{h'}$ connected with $v_h$.

These first two steps allow us to update the outer weights $w_{ph}$. The weight $w_{ph}$ is the regression coefficient in the quantile regression of the $p$-th manifest variable of the $h$-th block $x_{ph}$ on the inner estimate of the $h$-th latent variable $\theta_h$:

$$x_{ph} = w_{ph} \hat{\theta}_h + \varepsilon_{ph}$$

The quantile regression is an extension of the classical estimation of the conditional mean to the estimation of a set of conditional quantiles:

$$Q_{\theta}(x_{ph} | \hat{\theta}_h) = \hat{\theta}_h w_{ph}(\theta) + \varepsilon_{ph}$$

where $0 < \theta < 1$, $Q_{\theta}(\cdot | \cdot)$ denotes the conditional quantile function for the $\theta$-th quantile. In particular, PPR-PM considers only the case in which $\theta = 0.5$, i.e. the median is the single chosen quantile.

The algorithm iterates till convergence. After convergence, structural (or path) coefficients are estimated through possibilistic regression among the estimated LV:

$$\xi_j = \hat{\beta}_0 j + \sum_{h: \xi_h \to \xi_j} \hat{\beta}_{hj} \xi_h$$

where $\xi_j (j = 1, \ldots, J)$ is the generic endogenous (dependent) latent variable and $\hat{\beta}_{hj}$ is the generic interval path coefficient in terms of midpoint and spread $\hat{\beta}_{hj} = \{c_{hj}; a_{hj}\}$ interrelating the $h$-th exogenous (independent) variable to the $j$-th endogenous one. The higher the midpoint coefficient the higher the contribution to the prediction of the endogenous LV, while the higher the spread coefficient the higher the imprecision in the relation among the considered LV.

References

Robust penalized logistic regression modeling and tuning parameter selection

Park H.\textsuperscript{a*} and Konishi S.\textsuperscript{b}

\textsuperscript{a}Laboratory of DNA Information Analysis, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan.
\textsuperscript{b}Department of Mathematics, Faculty of Science and Engineering, Chuo University, Tokyo, Japan.

Keywords: Logistic regression, L\textsubscript{1}-type regularization, Robust modeling, Tuning parameter selection

1 Introduction

The logistic regression has been widely used to classify samples based on the predictor variables. The maximum likelihood method, however, often yields unstable estimation results in the high dimensional situation (i.e., \( p \) is greater than sample size \( n \)). To settle on the issue, penalized log-likelihood methods were proposed by using L\textsubscript{1}-type penalties. Although the penalized logistic regression has shown an exceptional performance for classification, it suffers from outliers, since the log-likelihood function is very sensitive to outliers.

We propose a robust penalized logistic regression based on a weighted log-likelihood function. In order to control outliers in high dimensional dataset, we use principal component spaces. Furthermore, it was shown that outliers are more clearly revealed in the principal components space than in ordinary data space, since the principal component describes the direction maximizing variance along each component and outliers increase variance of data set (Filzmoser et al., 2008, Park and Konishi, 2013). Thus, we use the weight based on the principal component space (Park and Konishi, 2013) for robust logistic regression modeling.

In robust L\textsubscript{1}-type regularization method, choosing the tuning parameters is a crucial issue, since the variable selection and estimation procedures heavily rely on the regularization parameters and outlier detection is based on a tuning constant. The cross-validation (CV) and traditional information criteria (i.e., AIC and BIC), however, cannot perform well for sparse logistic regression modeling, since the CV suffers from overfitting effect (Wang et al., 2007). Furthermore, the traditional information criteria are not suitable for the robust sparse logistic regression modeling, since they were derived under the assumption that the model is estimated by the maximum likelihood. To settle on the issue, we derive an information criterion for choosing the tuning parameters of robust penalized logistic regression in line with generalized information criteria (Konishi and Kitagawa, 2008).

2 Robust L\textsubscript{1}-type regularized logistic regression

Suppose we have \( n \) independent observations \( \{ (y_i, x_i); i = 1, \ldots, n \} \), where \( y_i \) are random response variables coded as either 0 or 1 and \( x_i = (x_{i1}, \ldots, x_{ip})^T \) is a vector of predictor variables. The logistic regression model assumes that: \( \Pr(Y_i = 1|x_i) = \pi(x_i) \) and \( \Pr(Y_i = 0|x_i) = 1 - \pi(x_i) \), where \( Y_i \) is a random variable distributed according to the Bernoulli distribution. The logistic regression model represents the class conditional probability as follow:

\[
\pi(x_i) = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \quad \text{and} \quad \log \frac{\pi(x_i)}{1 - \pi(x_i)} = x_i^T \beta, \quad \text{where} \quad x_i^T \beta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.
\]

We propose a robust penalized logistic regression based on a weighted log-likelihood function. In order to reduce influence of outliers, we use a weight based on the \( p^* \)-dimensional principal component space (Park

\textsuperscript{a}Corresponding author. E-mail: hwpark@ims.u-tokyo.ac.jp
and Konishi, 2013). The outliers increasing a variance of data set are more obviously revealed in the principal component space than in the original data space, since the principal components describe the direction that maximizes the variance (Filzmoser et al., 2008; Park and Konishi, 2013). It implies that we can effectively detect outliers in principal component space, and thus reduce effect of outliers in the sparse logistic regression modeling. The weight based on \( p^* \)-dimensional principal component is given as (Park and Konishi, 2013): 

\[
R_{p^*}^0 = \min\left(\sqrt{k/R.D_{p^*}\ell^2}, 1\right)/\sum_{i=1}^n \min\left(\sqrt{k/R.D_{p^*}\ell^2}, 1\right),
\]

where \( k = \chi^2(\text{df} = p^*) \) as a 95% quantile of the \( \chi^2(\text{df} = p^*) \) distribution (Khan et al., 2007), and \( R.D_{p^*}\ell^2 \) is a robust Mahalanobis distance based on the minimum volume ellipse (MVE): 

\[
R.D_{p^*}\ell^2 = \sqrt{(\mathbf{z}_i - \mathbf{T}^\text{pc})^T(C^\text{pc})^{-1}(\mathbf{z}_i - \mathbf{T}^\text{pc})}
\]

where mean \( \mathbf{T}^\text{pc} \) and covariance matrix \( C^\text{pc} \) is estimated by MVE with proper tuning constant based on \( p^* \)-dimensional principal components \( \mathbf{Z} \). By using the weight, we propose a robust penalized logistic regression based on a weighted log-likelihood,

\[
\ell(p^*_n)(\beta) = \sum_{i=1}^n R_{p^*_n}[y_i x_i^T \beta - \log\{1 - \exp(\alpha x_i^T \beta)\}] + \sum_{j=1}^p \left\{ \frac{1}{2} \alpha \beta_j^2 + (1 - \alpha) |\beta_j| \right\},
\]

where \( 0 \leq \alpha \leq 1 \) and \( \lambda > 0 \). By using the weight \( R_{p^*_n}^0 \), we control the effect of outliers on the sparse logistic regression modeling, and thus we can effectively perform classification of samples even in the presence of outliers.

### 3 Tuning parameter selection

The proposed robust logistic regression modeling procedure significantly relies on the choosing the tuning parameters. Although the regularization parameters have been often selected by the cross-validation, it is well known that the cross-validation suffers from the overfitting effect in the sparse regression modeling (Wang et al., 2007). Furthermore, the traditional model selection criteria, such as AIC and BIC, are not suitable for choosing the tuning parameters in the sparse regression modeling, since they were derived under the assumptions that the model is estimated by the maximum likelihood method, and they carried out in a parametric family of distributions including the true model (Konishi and Kitagawa, 2008). To overcome the limitations of the traditional model selection criteria, Konishi and Kitagawa (1996, 2008) proposed a generalized information criteria, which can be used to evaluate models estimated by various statistical methodologies not only the maximum likelihood method, based on a functional estimator.

We derive an information criterion for choosing the tuning parameters of the robust penalized logistic regression modeling in line with the generalized information criterion. To derive the generalized information criterion, the second differentiable functional estimator \( \hat{\beta} = \mathbf{G}(\hat{\mathbf{G}}) \) is required in calculation of an influence function, and thus we refer to the local quadratic approximation for deriving inderdifferentiability of \( L_1 \)-norm penalty in derivative of the information criterion (Fan and Li, 2001; Park et al., 2012).

### References

Eco-innovation and performance: What is the story in Luxembourg?

Anne Dubrocard and Leila Peltier - Ben Aoun

STATEC, Luxembourg, Luxembourg

Keywords: PLS, eco-innovation, performance, environment, Luxembourg, CIS, NAMEA

1 Abstract

Eco-innovation is evaluated through its “reduced” environmental impacts of greenhouse gas emissions, air and water pollution, energy consumption and soil contamination. In this context, eco-innovation should be seen as a subset of products or processes innovations and of organizational changes, characterized by a double externality and for which, therefore, regulation plays an essential role (regulatory push-pull effect) alongside the traditional factors of supply and demand (technology push and demand push). Indeed, Belin et al. (2011) point out that eco-innovation generates positive externalities both during the dissemination of knowledge phase and during the dissemination of eco-innovation phase. As a consequence of the underinvestment resulting mechanically from this double “market inefficiency”, the regulatory environment is a crucial determinant of corporate eco-innovative behaviour. Identifying the corporate behaviour in terms of eco-innovation and measuring its impact on growth has become an issue of public policy. In recent years many papers pay attention to the determinants but very few considers the performance output (to our knowledge only Mazzanti (2009)). A first insight into this behaviour is provided by the answers to a set of questions contained in the Community Innovation Survey (CIS) conducted in 2010 in Luxembourg. National accounts provide measurements of greenhouse gas emissions which are detailed in the classification of economic activities adapted to environmental approaches (NAMEA). These data sources are mobilized for this study which aims at exploring the causality between latent variables or components that could show the variation of the exogenous variables and of the dependent variables. Our exploratory search model of causality between latent variables will test the following hypothesis: objectives of innovation, innovation, eco-innovation and performance explained by sets of explanatory criteria. The objectives of the innovation should have a significant impact on the type of innovation deployed by innovative companies; in turn the type of innovation and the regulatory environment contribute in a differentiated way to eco-innovation, which we hope will partly determine the economic and environmental performances. Many empirical papers used standard econometric models (from OLS, probit to panel estimation) to study the determinants of eco-innovation, but those methods cannot test the causality. To go further in this direction we deploy the partial least squares (PLS) method which is a method of “soft modelling” based on a variance analysis which requires no assumption of variable multinormality. The application of PLS regression to the treatment of simultaneous equation model was carried out with the software SmartPLS (Ringle et al., 2006). It is therefore a matter of capturing the multidimensional and complex process that leads companies to innovate and then to eco-innovate and linking these decisions to their expected outcomes in terms of economic and environmental impacts. This technique is used to build predictive models when there are many and correlated explanatory variables compared to the number of observations. Moreover, since it can be deployed on observations of very small samples, it is particularly suited to testing hypotheses and concepts of eco-innovation based on the database built for Luxembourg. In the first part of this paper we will focus on determinants of eco-innovation. Many recent studies ([1], Rehfeld et al. (2007), Wagner (2007), [2], Nguyen Groff (2012)) focus on identifying the determinants of eco-innovation. The main determinants are related

*Corresponding author. E-mail: leila.ben-aoun@statec.etat.lu
to environmental regulations, to the information sources mobilized in the innovation process and to the types of innovation already deployed. Moreover, the different types of eco-innovation are closely linked to the types of innovation developed within the company Belin et al. (2011). In an empirical paper, these authors measure the impact of all possible combinations between innovation in products, processes, organizational change and marketing on the development of clean technologies and "end-of-pipe" eco-innovations. They get mixed results: the impact is significant for France but not for Germany. Zoboli Mazzanti (2006) and Wagner (2007) get a strong correlation between organizational innovation and the eco-innovation of process and product. On a theoretical level, environmental innovations should have impacts through improved performance of two types: environmental performance and economic performance. Economic performance can be measured through costs, turnover or changes in the number of employees [3]. Indeed, following the Porter hypothesis, costs should decrease over the long term thanks to a better use of inputs. Improved revenue can come from the creation of new markets or from the sale of products for which a larger (profit) margin is acceptable to consumers whose willingness to pay is higher for "organic" products for instance (Kammerer, 2009). However, the time lag and the difficulty of acquiring additional data make the measurement of the performance delicate. From our results we found that the deployment of an innovation process by the company seeks first to improve its competitive position regardless of the type of innovation considered (technological or organizational) and to develop new products or to improve its current products or the quality of its services. This second objective increases the likelihood of developing technological innovations. For instance, the opportunity to export and thus enter new markets is a determinant of innovation as suggested by Mazzanti et al. (2010) and, more generally, market conditions are major determinants of innovation, which confirms the results of Belin et al. (2011). The model was intended to clarify the determinants of eco-innovation. The importance of the regulatory environment is corroborated as a result of numerous empirical studies: Belin et al. (2011), et al. Ambec. (2008), and Jeff Palmer (1997), Frondel et al. (2007), Del Rio and Gonzalez (2005), Kesidou Demirel (2012) who argue that the regulatory context is the determining factor of corporate "green innovator" behaviour. Thus, Porter hypothesis, is confirmed and the regulatory framework strongly influences the eco-innovative behaviour of companies, whether in the production process or when the environmental benefit is intended for the final consumer as for France and Germany (Belin et al., 2011) and as suggested by the econometric estimates conducted on data from Luxembourg by Groff and Nguyen (2012). However, this result contradicts Kammerer (2009), for whom the impact of regulation varies depending on the purpose of eco-innovation. And the model does not retain any source of information as playing a convincing role in the eco-innovative behaviour of companies. Finally, some types of accumulated innovation experiences make companies more likely to introduce clean technologies. According to the results of Zoboli Mazzanti (2006), organizational innovation is strongly correlated with technological eco-innovation. However, technological innovation has no impact on the practice of eco-innovation regardless of the type considered. This reinforces the complementary role attributed to organizational innovation which ensures the reaping of the full benefits of the implementation of any other type of innovation including eco-innovation. In addition, Mazzanti and Zoboli (2009) and Cainelli et al. (2010) show that the more decentralized a company is, the easier the process of eco-innovation is. The last part of the model aims to highlight the direct impact of eco-innovation on the sector’s environmental or economic performance. The model does not support this hypothesis. Several reasons can explain this relative lack of success. On the one hand, as shown by Horbach et al. (2011), in most cases, companies which have eco-innovated do not observe changes in terms of cost reduction, turnover or employment, these are not the main goal assigned to this process. On the other hand, this results confirm Ambec et al. (2008) findings which have shown that regulation enforce eco-innovation and then results in a firm performance decreasing.

References


Multi-block Logistic Regression for High Dimensional Data

Cathy Philippe\textsuperscript{a}* , Arthur Tenenhaus\textsuperscript{b} , Vincent Frouin\textsuperscript{c} and Vincent Guillemot\textsuperscript{c}

\textsuperscript{a}UMR 8203 Gustave Roussy, Villejuif, France
\textsuperscript{b}Supelec, Gif-sur-Yvette, France
\textsuperscript{c}Neurospin, I2BM, CEA Saclay, Gif-sur-Yvette, France

Keywords: Multi-block Analysis, Logistic Regression, Ridge, High Dimensional Data

Introduction

The more our knowledge on the biology of cancer increases, the more the general landscape seems to be complex and necessitates to integrate amounts of data acquired at different functional levels of the tumors. Gene Expression or Copy Number data \cite{1} are respectively no longer self-sufficient and there is an unquestionable need for integrating these two complementary types of data. In this work, we propose to extend the general framework of RGCCA (Regularized Generalized Canonical Correlation Analysis) \cite{2} to logistic regression and thus provide a tool to predict a binary outcome from multi-block data.

1 RGCCA Framework and Multi-block Logistic Regression

RGCCA is a framework for studying associations between $J$ blocks of variables, $X_1, \ldots, X_J$, where each block $X_j$ (with $j = 1, \ldots, J$) represents a set of $p_j$ variables observed on the same set of $n$ individuals. RGCCA aims at extracting the information shared by the $J$ blocks of variables while taking into account a network of connections between these blocks. A design matrix $C = (c_{jk})$ defines the network of connections between blocks: $c_{jk} = 1$ if blocks $X_j$ and $X_k$ are connected, and $= 0$ otherwise. In the present case, a link between two blocks is modelled by the covariance and a link between a block and the outcome, $y$, is modelled by the logit link.

In RGCCA, the goal is to maximize the sum of the covariances (or their absolute values or their square, depending on the so-called “scheme”) between connected blocks under a constraint performing the compromise between the norm of the weights and the variance of the components.

In the same spirit, we propose to maximize globally the sum of all the likelihoods derived from a logistic link between a block and a binary variable $y$ and all the covariances between connected blocks:

$$\max_{a_1, \ldots, a_J} \left( \sum_{j=1}^{J} L(X_j, y, a_j) + \sum_{j,k=1; j \neq k}^{J} c_{jk}\text{cov}(X_j a_j, X_k a_k) \right)$$

(1)

To insure the convergence of the algorithm, we perform an Alternating Least Squares algorithm \cite{3}. It consists in computing, block per block, the outer weights $a_j$, as defined in RGCCA, via a Newton-Raphson algorithm. Once the $a_j$ are computed for a block $j$, they are injected in the calculation of the other outer weights. Analysing “-omics” data leads to dealing with high dimensional and highly correlated data, which can cause unstable parameter estimates. To avoid this pitfall, we refer to \cite{4} and propose to add, for each block, a Ridge penalty. For each block $j$, the optimization problem becomes:

$$\max_{a_j} f(a_j) = \max_{a_j} \left( \sum_{j=1}^{J} L(X_j, y, a_j) + \frac{1}{n} \sum_{j,k=1; j \neq k}^{J} c_{jk} a_j^T X_j^T X_k a_k - \frac{\lambda_j}{2} \|a_j\|_2^2 \right)$$

(2)

*Corresponding author. E-mail: cathy.philippe@gustaveroussy.fr.
We can compute, for each block, the gradient $\nabla f$ and the hessian $Hf$ of this new function $f$:

$$\nabla f(a_j) = X_j^T \left[ y - \pi_j + \frac{1}{n} \sum_{k=1, k \neq j}^J c_{jk} X_k a_k \right] - \lambda_j a_j$$

and

$$Hf(a_j) = -X_j^T W_j X_j - \lambda_j I,$$  

(3)

where $\pi_j = 1/(1 + \exp(-X_j a_j))$ and $W_j = \text{diag}(\pi(a_j)(1 - \pi(a_j)))$.

Maximization-Minorization (MM) algorithms are useful tools for maximization of concave or almost concave functions. In MM algorithms, the considered function is minorized at each iteration by a concave function in order to locally find an auxiliary maximum more easily computable than by using the original function. Since for all $j$, each component of the vector $\pi_j(1 - \pi_j)$ is inferior to $\frac{1}{4}$, we propose to refer to [5], and majorate $W_j$ by $\frac{1}{4}I$, thus allowing to generate the auxiliary function.

We obtain the following current iteration of the MM-Newton-Raphson algorithm for block $j$:

$$a_j^{(s+1)} = a_j^{(s)} - \left[ Hf \left( a_j^{(s)} \right) \right]^{-1} \nabla f \left( a_j^{(s)} \right) = 4X_j^T (X_j X_j^T + 4\lambda_j I)^{-1} \left( \frac{1}{4} X_j a_j^{(s)} + y - \pi_j + \frac{1}{n} \sum_{k=1, k \neq j}^J c_{jk} X_k a_k \right)$$

(4)

We note that $X_j X_j^T$ is an $n \times n$ matrix and can be computed once and for all Newton-Raphson iterations, as it does not depend on the current iteration. Moreover, we do not need to invert $W_j$ anymore, which allows us to tackle the problem of (quasi)-complete separation [6].

2 Results and Conclusion

We show promising results on simulated data [7].

We proposed in our work an adaptation of RGCCA to the prediction of a binary outcome by replacing covariance by a logistic link between each block of data and $y$, with very promising results. Works in progress include the application of multi-block logistic regression to a real multi-block dataset [1] and an extension to the prediction of a survival outcome.

References


The Use of PLS-SEM in Marketing Research: Advantages, Limits and Recommendations for Rigorous Applications

Karine Picot-Coupey and Julien Troiville

CREM UMR CNRS 6211, IGR-IAE, Université de Rennes1, Rennes, France

Keywords: Structural Equation Modeling (SEM), PLS, basis, pedagogy

If Partial Least Squares Structural Equation Modeling (PLS-SEM) enjoys a sustained success in several disciplines such as chemistry, medicine or biology [1], the interest it arouses in business research in general and in marketing in particular is much more mixed. The dominant Covariance-Based Structural Equation Modeling (CB-SEM) paradigm still remain the classic method when marketing researchers have to deal with multivariate analysis [2] [3]. Recent special issues (e.g. MIS Quarterly in 2009, Journal of Marketing Theory and Practice in 2011 or Longe Range Planning in 2012 and 2013) and methodological articles devoted to PLS (e.g. [4]) have outlined the founding principles but have also stimulated debates on its relevance as a scientific research method and on the robustness of its application [5] [2].

PLS-SEM in marketing research is often a default choice, especially because CB-SEM cannot be applied properly, because of “small samples”, “non-normal data” or “formative measures” [6] [7]. In this case, both the selection and the application of the approach are frequently not well justified which do not serve the quality of the research. However, PLS-SEM can be very relevant for many marketing researches, in consumer research as well as in strategic marketing (e.g. [8] [9]) to deal with latent variables and analyze cause-effect relationships. Moreover, issues related to the operationalization and the measurement model specification (e.g. [10] [11]) raise questions about data analysis method and should promote the use of PLS-SEM [12] [2] [3].

PLS-SEM technique often stands in the “shadow of its slightly elder sibling” [3], especially because some scholars view the approach as less rigorous due to a misconception about its implementation and an “easy-of-use” appearance. Unfamiliarity with the method is common that reflects both in the difficulty to find reviewers for research papers in which the method is employed [13] and in the few dedicated handbooks [1] [14]. There is an essential need to educate about PLS-SEM.

Considering the mitigated openness to this method, this article has an educational and methodological purpose. It presents in brief PLS-SEM basis, discusses its methodological foundations insisting on the advantages and limits to its selection and its use. Finally, it provides a clear and precise “user instructions” table to help marketing researchers in the implementation of the method.

This article is organized in three sections, as follows:

(1) Presentation of PLS-SEM basis and comparison with CB-SEM

CB-SEM methods are often used quite automatically and without a clear and precise link with the research objectives of the work and data specificity. In fact, CB-SEM can constrain the conceptual approach [15] [16] and suffers from limits, among which is the lack of managerial implications [2]. This, the PLS-SEM approach is not only the “plan B” or the spare wheel of CB-SEM such as LISREL [12]. Its basis makes it more suitable for many configurations that can be clearly identified.

| Table 1: PLS-SEM and CB-SEM founding principles |
| Summary of hard and soft modelling approaches |
| Estimation process, objectives, method and algorithm, conditions to fulfill |

(2) Discussion of the issues and potential limits addressed to PLS-SEM
On the basis of an extensive literature review conducted in the best marketing journals (JM, JMR, MKS, JCR, IJRM, JAMS, JBR, and RAM), benefits, limits as well as threats due to misuse are discussed.
Table 2: SEM-PLS: issues and recommendations related to:

| 1. Model specification: discussion about reflective versus formative measures, impact on model quality, recommendations on the conceptual frame |
| 2. Model identification |
| 3. The operationalization of measurement instruments, data collection and preparation |
| 4. Model estimation: reflective versus formative measures, inner and outer models’ evaluation, resampling procedures |
| 5. The reporting of the results |

(3) Implementation and recommendations

Hair et al. [4] insist on the pitfalls in PLS-SEM use and on improper reporting of results in marketing research using SEM-PLS. This 3rd section provides a PLS-SEM “user instructions” table in order to facilitate the work of marketing researchers when applying the method and reporting results. It develops guidelines for each key step and suggests strategies depending on the situations.

In doing so, the contribution of this article lies in an educational presentation of SEM-PLS in order to allow marketing researchers who are unfamiliar with it to:

(i) understand its methodological basis and decide if it is appropriate considering their objectives and data;
(ii) know how to rigorously use it and to overcome difficulties (PLS-SEM algorithm settings, software, procedures to evaluate results,…);
(iii) interpret, report and comment results in ways consistent with the most recent methodological requirements [2] [14] [17], thus contributing to improve the quality of marketing research using SEM-PLS.

References

Applications of Mode C analysis to understanding team performance

Pratley A. J.\textsuperscript{a*}, van Voorthuysen, E. J.\textsuperscript{b} & Chan R. T.

\textsuperscript{a}University of Sydney – Business School, Sydney, Australia
\textsuperscript{b}University of New South Wales – School of Mechanical and Manufacturing Engineering, Sydney, Australia

Keywords: Mode C, Team performance, Industry applications

1 Introduction

A convergence of factors in the past 50 years has resulted in an unprecedented proportion of the population either working in, or interacting with a team on a daily basis. In the 1960’s two significant factors interacted simultaneously with each other. The United States was losing control as the dominate manufacturer to Japan which resulted in a number of studies focused on understanding the ‘Japanese way’. Secondly the start of the ‘technological revolution’ began as we know it today, with computers beginning to become utilized within universities and business organizations.

The combination of a more egalitarian society with the proliferation of electronics and specialization of skills has resulted in more complex projects requiring a greater range of (specialist) skill sets (particularly in engineering). As a profession, engineering has both been a contributing factor and greatly influenced by the increase of workplace teams through the development of low cost electronics. As a case in point, consider the change in the average refrigerator over the past 50 years. It has gone from being an ice box with a motor and compressor, to a multi compartment, internet enabled device that is linked into the global supply chain.

Psychology was at the forefront of the use of structural equation modelling (SEM) across a variety of disciplines from the 1970’s onwards [1]. The interest in SEM originated from factor analysis even thought it had first been published at the start of the 20th century by Pearson [2]. Despite the ongoing development of both of these statistical methods for application to complex data sets used in psychology, there has been no significant use of PLS within psychology with the exception of a handful of articles [3], [4]. Both of these manuscripts have used PSL-PM, models with more than two components. There is currently no research using Partial Least Squares Two Block Analysis (PLS TBA) i.e. Mode A, B or C. Those researchers that have moved to using PLS PM are likely trying to find a one-to-one substitute of CB SEM. There is also the belief that complex models are required to understand and further develop current theories within team research and more broadly, industrial-organization psychology.

2 Application to industry

Over the period of four years three teams from two organizations were tracked through questionnaires. The data presented in the paper comes from one of the teams at a manufacturing facility in Australia of a global pharmaceutical company. The purpose of the project was two-fold. Firstly, the capacity of the production line was to be increased through engineering changes. The second aspect of the project was to improve the line efficiency through standardized methods, training of operators and engagement of a consulting firm on lean principles. Only the second part of the project was included due to the availability and interest of staff.
To gain an understanding of what was happening at the pharmaceutical company a number of open ended interviews were conducted. A member of staff from the Operational Excellence Group selected ten staff for interviews to present a broad cross section of the organization. The interviews generally ran for about an hour and were conducted one-on-one in visible locations (on site). The opening question was “What is unique about working at <your organization>? Workplaces have many similarities and these are of little interest. Of course professional staff send emails and attend meetings, the real question being asked is “What is different about these common activities?” The actual process of the interviews lies somewhere between an unstructured interview [5] and a semi-structured interview technique [6] tending more towards the unstructured end of the continuum.

3 Results and analysis

The eight organizational categories were expanded to twenty organizational variables in a 113 item questionnaire containing 74 formative questions and 39 reflective questions. The questionnaire was administered to the team on three occasions during the project with 13, 13 & 11 respondents respectively. The average r² across each of the 20 models for each time period was 0.847, 0.897 & 0.926 respectively. Of the 60 models only two had an r² below 0.7 and more than half had an r² of greater than 0.9. Bootstrapping was conducted on all models and confidence intervals reported on the r² values.

The results were repeated with a second team at the same pharmaceutical company the following year and again in a medical device company two years later using the methodology previously described. It could be argued the statistical results were even more impressive (based on r² values of the models). The results of individual models were reported to management with great interest. Never before had the management team understood the strength of a relationship and the amount of weight each indicator had.

Current thinking of teams suggests that large complex models are the solution to understanding teams in the 21st century. Certainly PLS PM can play a significant result in this area of analysis. A more important and often overlooked question is: “What small, simple and easy to implement model can provide direct and accessible insight to the non-statistically included reader?” This industry project demonstrates not only the viability, but the robustness and usability of PLS TBA specifically using Mode C analysis.

References


PLS regression for multivariate functional data

Cristian Preda\textsuperscript{a}\textsuperscript{*} and Gilbert Saporta\textsuperscript{b}

\textsuperscript{a}Université Lille 1, Villeneuve d’Ascq, France
\textsuperscript{b}CNAM, Paris, France

\textbf{Keywords:} multivariate functional data, PLS, regression

\textbf{Introduction}

Functional data or data represented by curves, is generally considered as sample paths of a real-valued stochastic process with continuous time, $X = \{X_t\}_{t\in[0,T]}$. Most of the approaches dealing with functional data consider the univariate case, i.e. $X(t) \in \mathbb{R}$, $\forall t \in [0,T]$, a path of $X$ being represented by a single curve. Despite its evident interest, the multivariate case, $X(t) = (X_1(t), \ldots, X_p(t)) \in \mathbb{R}^p$, $p \geq 2$
is, curiously, rarely considered in literature. In this case a path of $X$ is represented by a set of $p$ curves. The dependency between the $p$ measures provides the structure of $X$. One finds in [1] a brief example of bi-dimensional functional data, $X(t) = (X_1(t), X_2(t)) \in \mathbb{R}^2$, as a model for gait data (knee and hip measures) used in the context of functional principal analysis as an extension of the univariate case. For a more theoretical framework, we must go back to the pioneer works of [2] on random variables with values into a general Hilbert space. In [3]) the author provides a complete analysis of multivariate functional data from the point of view of factorial methods (principal components and canonical analysis). Recently, [4] considered model-based clustering for multivariate functional data and [5] introduced linear tools, similar to principal component analysis, for analysing such data.

In this paper we consider the linear regression model with multivariate functional random variable predictor and vectorial response,

\[ E(Y|X = x) = \int_0^T \sum_{i=1}^p \beta_i(t)x_i(t)dt, \quad Y \in \mathbb{R}^q, \beta_i \in (L_2([0,T]))^q, \forall i = 1, \ldots, p. \quad (1) \]

As an extension of the PLS approach for the functional linear regression model proposed in [6] for univariate functional data, we develop the PLS estimation in the case of multivariate functional predictor. The Tucker criterion provides the PLS components as eigen-vectors of the product of the Escoufier’s operators associated to the response and the predictor.

We present the PLS estimation when the predictor is approximated in a finite dimensional space of functions. A simulation study illustrates our methodology.

\textbf{References}


\textsuperscript{*}Corresponding author. E-mail: cristian.preda@polytech-lille.fr, Université de Lille 1, Villeneuve d’Ascq, France.
Fault detection in the Tennessee Eastman process using dynamic principal components analysis with decorrelated residuals

Tiago J. Rato and Marco S. Reis*
CIEPQPF, Department of Chemical Engineering, University of Coimbra, Coimbra, Portugal

Keywords: Multivariate statistical process control, Principal component analysis, Dynamic Principal component analysis, Missing data imputation, Tennessee Eastman benchmark process

Introduction

With the current data acquisition systems it is possible to easily collect large amounts of data, usually presenting correlation and autocorrelation [1]. These situations raise important challenges to traditional SPC schemes, which are based on univariate quantities and rely on the usual iid assumptions, motivating the development of new multivariate SPC charts. The multivariate statistics commonly used in the monitoring of industrial process are usually based on the Hotelling’s $T^2$ control chart applied to the variables, or to the first latent variables of a principal component analysis (PCA) model, usually complemented with the $Q$ statistic (also known as square predicted error, SPE), in order to monitor the variability not captured by the PCA model (MSPC-PCA). However, the MSPC-PCA method, assumes that the variables are uncorrelated in time, which is often not true in practice, especially with the high sampling rates currently achieved by modern instrumentation. To address this issue, Ku et al. [1] proposed an SPC procedure based on dynamic principal component analysis (DPCA), which is an extended version of PCA that includes time-lagged variables, in order to accommodate and tacitly model the dynamic behavior of variables within the same PCA model. Unfortunately, one can easily verify that the direct implementation of such method still leads to autocorrelated statistics, which raises implementation problems.

To handle these issues we propose a combination of DPCA and Missing Data methods, defined as DPCA with Decorrelated Residuals (DPCA-DR), that allow for the reduction of the autocorrelation effect present in the monitoring statistics. This method also led to an improvement of the monitoring performance on the studied systems.

1 Comparison study based on the Tennessee Eastman process

To assess the monitoring performance of the proposed method, the Tennessee Eastman process was used as a benchmark. This model was developed by Downs and Vogel [2] and has been widely used by the process monitoring community for comparison purposes. The simulation model has 41 measurements (XMEAS) and 12 manipulated (XMV) variables and include 21 process upsets, as described in [2, 3]. In this study we have used the same data sets as Russell et al. [3], which are composed by 960 observations with a sample interval of 3 min. Faults were introduced 8 hours after the onset of simulations. All the manipulated and measurement variables, except the agitation speed of the reactor’s stirrer (which is always constant), were collected, in a total of 52 variables.

To obtain comparable results, the UCL for the various methods were set to a false alarm rate of 1% under normal operation conditions and then the fault detection rate for each fault were determined, as presented in Table 1. From this analysis, we conclude that the DPCA-DR statistics are, for this case study, superior to PCA and DPCA, and were able to detect 19 of the 21 faults, failing only in the detection of faults number 3 and 9, where all methods also present problems, including CVA proposed by Russell et al. [3]. Furthermore, DPCA-DR

* Corresponding author. E-mail: marco@eq.uc.pt
DR statistics also presented lower autocorrelation and were able to maintain the out-of-control signals, while PCA and DPCA statistics returned to their in control range, creating the conditions for a false sense of safety.

Table 1. Fault detection rates for the various methods under study, regarding each faulty scenario. The top scores are signaled in boldface format.

<table>
<thead>
<tr>
<th>Fault</th>
<th>PCA</th>
<th>DPCA</th>
<th>DPCA-DR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T^2$</td>
<td>$Q$</td>
<td>$T^2$</td>
</tr>
<tr>
<td>1</td>
<td>0.991</td>
<td>0.995</td>
<td>0.980</td>
</tr>
<tr>
<td>2</td>
<td>0.985</td>
<td>0.984</td>
<td>0.035</td>
</tr>
<tr>
<td>3</td>
<td>0.036</td>
<td>0.165</td>
<td>0.999</td>
</tr>
<tr>
<td>4</td>
<td>0.257</td>
<td>0.980</td>
<td>0.989</td>
</tr>
<tr>
<td>5</td>
<td>0.989</td>
<td>0.999</td>
<td>0.986</td>
</tr>
<tr>
<td>6</td>
<td>0.974</td>
<td>0.968</td>
<td>0.973</td>
</tr>
<tr>
<td>7</td>
<td>0.034</td>
<td>0.010</td>
<td>0.030</td>
</tr>
<tr>
<td>8</td>
<td>0.367</td>
<td>0.154</td>
<td>0.439</td>
</tr>
<tr>
<td>9</td>
<td>0.414</td>
<td>0.638</td>
<td>0.340</td>
</tr>
<tr>
<td>10</td>
<td>0.985</td>
<td>0.925</td>
<td>0.990</td>
</tr>
<tr>
<td>11</td>
<td>0.943</td>
<td>0.950</td>
<td>0.943</td>
</tr>
<tr>
<td>12</td>
<td>0.988</td>
<td>0.999</td>
<td>0.990</td>
</tr>
<tr>
<td>13</td>
<td>0.035</td>
<td>0.007</td>
<td>0.059</td>
</tr>
<tr>
<td>14</td>
<td>0.174</td>
<td>0.137</td>
<td>0.217</td>
</tr>
<tr>
<td>15</td>
<td>0.787</td>
<td>0.905</td>
<td>0.790</td>
</tr>
<tr>
<td>16</td>
<td>0.893</td>
<td>0.901</td>
<td>0.890</td>
</tr>
<tr>
<td>17</td>
<td>0.115</td>
<td>0.059</td>
<td>0.046</td>
</tr>
<tr>
<td>18</td>
<td>0.340</td>
<td>0.423</td>
<td>0.408</td>
</tr>
<tr>
<td>19</td>
<td>0.362</td>
<td>0.414</td>
<td>0.429</td>
</tr>
</tbody>
</table>

2 Conclusions

In this paper we presented a methodology for conducting large-scale process monitoring of dynamical systems, called DPCA-DR, and compared its performance against PCA and DPCA. The overall superiority of DPCA-DR was also verified by paired t-tests between all the statistics, from where it can be concluded that, with a 5% significance level, the DPCA-DR statistics are indeed significantly better than all the PCA and DPCA monitoring statistics. Therefore, given is higher detection rates and lower autocorrelation, the DPCA-DR statistics are regarded as more effective, reliable and consistent than their counterparts tested in this study, features that make them a viable alternative to current monitoring statistics.

References


Extending the Non-Metric PLS approach to inwards directed path models

Giorgio Russolillo\textsuperscript{a}\textsuperscript{*}

\textsuperscript{a}Conservatoire National des Arts et Métiers, Paris, France

Keywords: non-linearity, non-metric data, path modeling, mode B

Introduction

Partial Least Squares Path Modeling (PLS-PM) was introduced by H. Wold \cite{1} for the analysis of path models with latent variables. The main idea of such models is that $Q$ latent variables (LVs) $\xi_1, \ldots, \xi_q, \ldots, \xi_Q$, linked in a network of linear relationships, can be used to study the relations between $Q$ blocks of observed variables (MVs) $X_1, \ldots, X_q, \ldots, X_Q$. Each model has two levels of relationships. The first level concerns relationships between LVs (structural model); these relations can be stored in a symmetric binary matrix $C$ of order $Q$, whose generic element $c_{qq'}$ is one if $\xi_q$ is connected to $\xi_{q'}$ and zero otherwise. The other level of relations concerns the links between each LV and its own block of MVs (measurement model).

PLS-PM assumes that LV of the $q$-th block can be estimated by a linear composite of the manifest variables $x_{1q}, \ldots, x_{pq}, \ldots, x_{pq}$ of the block. The weight system to be applied to the linear combination is obtained through an iterative two-step algorithm. In the first step an outer proxy $t_q$ of each latent construct is obtained as a linear composite of the observed variables; in the second step an inner proxy $z_q$ of each latent construct is computed as a linear combination of its adjacent latent constructs estimated using the outer proxies computed at the previous step. The inner proxies are then used to update the system of weights (i.e. the outer weights) to be applied to the observed variables. These two steps are iterated till numerical convergence on the outer weights. Refer to \cite{2} for a detailed presentation of the algorithm.

The PLS-PM algorithm is extremely flexible. Several options can be used to compute inner and outer proxies of the LVs, also called schemes and modes respectively. In this work we focus on the centroid and the factorial schemes to calculate the inner weights for computing the inner proxies $z_q$, and on the so-called new mode A \cite{3} and the Mode B to compute the outer weights $w_{pq}$ used in computing the outer proxies $t_q$. Any combination of these options leads the algorithm to converge to a function (depending on the scheme) of a sum of correlations (if mode B is used) or covariances (if new mode A is used) between composites of linked blocks \cite{4}.

PLS-PM only applies to quantitative (metric) data, as it implies two assumptions: i) each observed variable is measured on a interval (or ratio) scale ii) relationships between variables are linear and, consequently, monotonic. However, in many real applications data are observed on non-metric measurement scales.

The Non-Metric Partial Least Squares (NM-PLS) approach has been recently proposed by Russolillo \cite{5} to extend covariance-based PLS criteria to the treatment of non-metric variables and non-linearity. This approach is based on the concept of Optimal Scaling (OS) \cite{6, 7}. The OS principle sees observations as categorical, and represents each observation category by a scaling parameter. This parameter is subject to constraints deriving from the measurement characteristics of the variables. In this process each variable $x$ is transformed as $\hat{x} \propto X\phi$, where $\phi$ is a vector of numeric values (the scaling parameters) associated to the different values (or categories) of the variable $x$, and the matrix $X$ defines a space in which constraints imposed by the scaling level are respected. The symbol $\propto$ means that the left side of the equation corresponds to the right side normalized to unitary variance.

\textsuperscript{*}Corresponding author. E-mail: giorgio.russolillo@cnam.fr.
In order to apply the OS principle to the PLS Path Modeling, when New Mode A is used, the following criterion must be optimized:

\[
\begin{align*}
\arg \max_{w, \phi} & \quad \sum_q c_q q \gamma_q \text{ (1)} \\
\text{s.t.} & \quad \|w_q\| = \sqrt{n}, \|\hat{x}_{pq}\| = \sqrt{n}
\end{align*}
\]

where \(\hat{x}_{pq} = X_{pq}\phi_{pq}\) is the \(p\)-th variable of the \(q\)-th block, quantified by means of the scaling parameter \(\phi_{pq}\).

The criterion (1) depends on two set of parameters: the model parameters \((w_q)\) and the scaling parameters \((\phi_{pq})\). A NM-PLS algorithm has been proposed for solving Eq. (1) [5]. This algorithm alternately optimizes criterion (1) with respect to each subset, keeping the other fixed. In particular, keeping fixed PLS parameters \(w_q\), the optimal solution for \(\phi_{pq}\) is given by the quantification function \(Q(X_{pq}, z_q)\) which orthogonally projects \(z_q\) on the space spanned by \(\hat{X}_{pq}\) [5].

1 A new Non-Metric PLS Path Modeling algorithm for Mode B

A new algorithm is proposed, which extends the NM-PLS approach to the Mode B scheme and by consequence to whatsoever combination of the modes and schemes cited in previous section.

To extend the Mode B PLS-PM to OS, the following optimization problem must be solved:

\[
\begin{align*}
\arg \max_{w, \phi} & \quad \sum_q c_q q \gamma_q \text{ (2)} \\
\text{s.t.} & \quad \|w_q\| = \sqrt{n}, \|\hat{x}_{pq}\| = \sqrt{n}
\end{align*}
\]

It is possible to show that this criterion can be rewritten as:

\[
\begin{align*}
\arg \max_{w, \phi} & \quad \sum_q \text{cor}(\hat{X}_q w_q, z_q) \text{ (3)}
\end{align*}
\]

A NM-PLS algorithm can be used to maximize this criterion. For fixed scaling parameters, the optimization problem in (3) is solved with respect to \(w_q\) by using the usual PLS-PM iteration steps. In order to optimize the problem with respect to \(\phi_{pq}\) while keeping \(w_q\) fixed, we propose the following backfitting procedure (see [6, 8], among others). Following the back fitting principle, the optimal solution for \(\hat{x}_{pq}\) is obtained by applying the quantification function \(Q(X_{pq}, x_p)\) on the back-fitted variable \(x_{pq}^*\), obtained as:

\[
x_{pq}^* = (1/b_{pq})(z_q - \sum_{j \neq p} b_{jq} \hat{x}_{jq})
\]

where \(b_{pq}\) is the \(p\)-th element of the regression coefficient vector \(b_q = (\hat{X}_q^T \hat{X}_q)^{-1} \hat{X}_q^T z_q\).

References


Embedding PLS analysis into the theory building process – the case of consumer trust in mobile payments

Ahmed Shuhaiber\textsuperscript{a} and Hans Lehmann\textsuperscript{b}

\textsuperscript{a,b} Victoria University of Wellington

\textbf{Keywords:} Theory building, single-path PLS, multi-path PLS, mixed-method research, mobile payments

Little is known in recent and current research about the factors that contribute (or otherwise) to consumer trust in the specific case of mobile payment transactions – a rapidly growing subsection of mobile internet finance applications. A research project was undertaken to establish such factors - in a mixed research approach – firstly from literature with subsequent qualitative and quantitative empirical validation.

Following the classic theory building process, there were seven phases in the research:

**Step One:** Analysis of the current research landscape: this yielded little research that was specifically aimed at consumer trust in mobile payments (CTMP). Therefore a process of fusing literatures was employed:

a. Research into factors for building trust in general and in electronic commerce specifically; and
b. Research into factors that facilitate consumer acceptance of mobile payment transactions.

This resulted in an unweighted list of 18 (combined) factors that could now be postulated into hypotheses of what influences trust in mobile payments.

**Step Two:** First in the empirical validation, addition or rejection of CTMP factors a substantive geographical area was selected for preliminary empirical validation of the factor list: the United Arab Emirates, where mobile payments had been introduced recently and had experienced mixed reaction from consumers.

**Step Three:** The consolidation of the literature hypotheses with qualitative confirmation took three steps:

a. targeted qualitative semi-structured interviews, followed up by a theoretically sampled set of focus group interviews; their analysis enabled
b. establishing qualitative weightings for the influence strength of validated factors; it also lead to the
c. elimination of hypotheses that could not be substantiated empirically.

The results from this first validation yielded a qualitative model with indicative vector strengths for factor correlates to CTMP. This qualitative model is now ready for quantitative evaluation/validation.

**Step Four:** Using measurement models and instruments from the literature a survey instrument for quantitative analysis of the qualitative model was constructed and extensive psychometric proofing of the questionnaire was carried out.

\* Corresponding author. E-Mail: Hans.Lehmann@vuw.ac.nz

\begin{center}
\textbf{The qualitatively weighted Factorial CTMP Model Based on the Findings from the Qualitative Validation}
\end{center}
Step Five: Next, the implied factor structure in the instrument was validated in an exploratory Factor Analysis of the survey data. As a result, a number of posited factors with insignificant/immaterial factor loadings were eliminated from the model.

Now the first PLS was run and established the single-path model. The model largely confirmed the qualitative construct but, despite 80% explained variance overall (an $R^2$ of 0.795 for Trust Beliefs), most factors showed low path coefficients overall. The only two exceptions were (unsurprisingly) a positive influence of Providers’ Reputations, tempered by Perceived Security Risks.

Step Six: In order to identify the reasons for the weak factor strengths a second analysis of the cross correlates derived from the instrument building and exploratory Factor Analysis was carried out showed that there were significant cross-correlations between factors, which may have obstructed vector strengths in the single-path model. This needed to be further investigated.

Step Seven: A multi-path PLS was set up, following the pointers from the cross-correlation analysis. After further refinement (e.g. re-booting and eliminating immaterial paths), a final multi-path model of Consumer Trust in Mobile Payments (in the UAE) was established.

The dominant influential focal construct on trust in m-payments was found to be the Provider Characteristics (with provider’s reputation as the most influential), which is also related to Consumer Characteristics (such as awareness, propensity to trust, and experience), albeit only explaining around 22% of its variance. On the other hand, Provider Characteristics explain more than 46% of the model’s variance. In turn, they are impacted by Environmental Influences (mainly media and word-of-mouth) on the one hand and, negatively, by Perceived Risks (financial, security and technical), which account for over 50% of the variance, on the other. The perceived risks were themselves materially shaped by Environmental Influences and exclusively affected by Mobile-Device Characteristics (such as functionality and design). This interaction, with clear correlative direction and quantified strength, is essential for understanding how CTMP is built – which is new knowledge for the academic community and provides valuable pointers for practitioner’s marketing and consumer service activities.

In conclusion, we believe that this research contributes a clear and instructive demonstration of where SEM/PLS methods have their place in the classic theory building process and how (and when) they may be applied to construct a factorial model that is fully grounded in its empirical and theoretical environment. Furthermore, this process extends seamlessly into a clear direction for future research. An international research collaboration project has been initiated to replicate the study (from Step Two, albeit with a literature update for each new region/language) and eventually synthesize the individual results into a global model of trust factors for mobile payment services.
Partial Least Squares for Dependent Data

Marco Singer\textsuperscript{a*}, Tatyana Krivobokova\textsuperscript{a}, Axel Munk\textsuperscript{a} and Bert de Groot\textsuperscript{b}

\textsuperscript{a}Institute for Mathematical Stochastics, Georg August Universitäät, Göttingen, Germany

\textsuperscript{b}Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

Keywords: dependent data, spatio-temporal separability, partial least squares, protein dynamics.

Introduction

Dependency between observations can have a large effect on the predictive capabilities of regression techniques. We study an alteration of the PLS algorithm via a linear transformation of the data and proof theoretical results like the convergence in probability of the PLS estimator. As an application we consider protein dynamics.

1 PLS under dependence

We consider a linear regression model

\[ Y_t = \langle X_t, \beta \rangle + \varepsilon_t \]

where \( t \in [0, 1] \) and \( X := (X_t)_{t \in [0,1]} \) is a centered, \( \mathbb{R}^k \)-valued random process, \((\varepsilon_t)_{t \in [0,1]}\) is a centered process independent of \( X \) and \( \beta \in \mathbb{R}^k \). We assume that the covariance matrices \( K_{t,s} := \text{Cov}(X_t, X_s) \) are spatio-temporal separable, i.e., there exists a function \( \tau : [0,1]^2 \to \mathbb{R} \) and a positive definite, symmetric matrix \( \Sigma \in \mathbb{R}^{k \times k} \) such that

\[ K_{t,s} = \tau(t,s)\Sigma. \]

Since \( X \) can be nonstationary we define the weight vectors in the (population) PLS algorithm via

\[ w_i = \arg \max_{w \in \mathbb{R}^k} \int_0^1 \text{Cov}(Y_t, \langle X_t, \beta \rangle) \, dt \quad \text{s.t.} \quad \|w_i\| = 1, \quad \langle w_j, w_i \rangle = 0, \quad j < i. \]

It turns out that under spatio-temporal-separability the \( w_1, \ldots, w_i \) span the Krylov space \( \mathcal{K}(\Sigma\beta, \Sigma) \).

In the finite sample case we use a method inspired by the generalised least squares approach to estimate the population model, i.e., given data \( X_i := (X_{t_i,1}, \ldots, X_{t_i,n})^T \) and \( Y_i, i = 1, \ldots, n \), with \( 0 < t_1 < t_2 < \cdots < t_n < 1 \) assembled in a matrix \( X := (X_1, \ldots, X_n)^T \) and a vector \( Y := (Y_1, \ldots, Y_n)^T \) we assume that knowledge of the temporal covariance function \( \tau \) is available. Defining \( T := (\tau(t_i,t_j))_{i,j=1,\ldots,n} \) we take \( n^{-1/2}X^T T^{-1}X \) and \( n^{-1/2}X^T T^{-1}Y \) as estimators for \( \Sigma \) and \( \Sigma\beta \), respectively.

Blanchard and Krämer [1] studied the convergence in probability of the Kernel Conjugate Gradient algorithm when early stopping via the discrepancy principle is considered. In the case when the transformed variables \( T^{-1/2}X \) are rowwise independent and almost surely bounded we achieve similar results to their theorem 2.1 for PLS in the linear regression setting (1), when the temporal covariance in \( X \) and \( \varepsilon \) is the same. Thus when early stopping is considered convergence in probability to the true regression coefficients with rate \( n^{-1/2} \) can be attained for the PLS estimator.

\textsuperscript{*}Corresponding author. E-mail: msinger@gwdg.de, Göttingen, Germany.
2 Application to protein dynamics

As an application we consider the protein dynamics which were studied in [2]. The data matrix X consists of the coordinates of atoms in a protein over a timespan and some functional property of the protein as \( Y \). The trajectory of each (coordinate of an) atom is modelled as an ARIMA\((p,1,q)\)-process similar to [3]. We fit a model to the data by assuming spatio-temporal separability, i.e. all the columns of X follow the same ARIMA\((p,1,q)\)-process. We call the corresponding covariance matrix of the time series \( T \).

To evaluate the prediction performance we split the dataset into a training and test set. We build the PLS model on the training set once with the original data \( X \) and \( Y \) and once with the linearly transformed data \( T^{-1/2}X \) and \( T^{-1/2}Y \). Then the predicted values are compared to the real ones on the test set via the Pearson correlation coefficient and residual sum of squares (RSS). This is shown for the protein T4 Lyzoyme (T4L) in Figure 1 where half the data was used for training and half for testing. This protein is a catalyst for the hydrolisis of certain substances in the cell walls of the bacteriophage T4. The functional value is chosen to be the distance between the active side residues and the substrate, which is important for the catalysis.

As can be seen the prediction with only a few components is better when dealing with the correlation between observations first. If a high enough number of components is chosen then the predictive capabilities become similar.

![Figure 1](image)

Figure 1. Correlation (left) and RSS (right) of the prediction of the functional property and the real data on the test set for the protein T4L when using ordinary PLS and the modified version.

References


Functional Generalized Structured Component Analysis

Hye Won Suk\textsuperscript{a,*} and Heungsun Hwang\textsuperscript{b}
\textsuperscript{a}Arizona State University, Tempe, AZ, USA
\textsuperscript{b}McGill University, Montreal, QC, Canada

Keywords: Generalized structured component analysis, Functional data analysis, Basis function expansion, Penalized least squares

Introduction

Generalized Structured Component Analysis (GSCA) \cite{1} represents component-based structural equation modeling, which enables to examine directional relationships among multiple sets of responses by combining data reduction with path modeling. In GSCA, a component or a weighted composite is obtained from each set of responses in such a way that, given hypothesized relationships among the sets, it captures the variation in both observed responses and components as much as possible. In this paper, we propose an extension of GSCA, called functional GSCA, to deal with so-called functional data.

1 Issues in analyzing functional data by GSCA

Functional data refer to the data that are considered to arise from an underlying smooth function varying over a continuum such as time or space. Familiar examples are motion capture data, neuroimaging data, and polygraph data. The emergence of sophisticated measurement tools, e.g., motion capture devices, handheld computers, Bluetooth devices, eye-trackers, and brain scanners, has facilitated the collection of functional data. Although the data recorded by these tools are discrete and their trajectories are often rough, there is a scientific reason to believe that the true trajectory underlying the data is a smooth curve perturbed by error.

One might think that functional data are not different from multivariate data except that a set of responses consists of scores measured at multiple time points instead of on multiple variables. However, analyzing functional data by GSCA is not as straightforward as it may appear for several reasons. First, a smooth curve underlying functional data is infinite-dimensional whereas GSCA has been developed and extended for finite-dimensional multivariate data only. Second, if we analyze functional responses by GSCA as if there were multivariate, we ignore the fact that functional responses at adjacent time points tend to be connected and are not too different from each other, which may lead to highly fluctuating parameter estimates that may be difficult to interpret. Third, in order to deal with functional responses in GSCA, one may be tempted to transform functional data into multivariate data by calculating a few summary measures such as mean and variance over time, peak amplitude, and area under the curve. However, these summary measures might not be optimal; functional responses may not be summarized satisfactorily well by such summary measures, and more importantly, information on temporal variations in the data will be lost completely.

2 Functional GSCA

Functional generalized structured component analysis (functional GSCA) is proposed to address these issues. Technically, it integrates GSCA with basis function expansions into a unified framework and estimates parameters by minimizing a penalized least squares criterion. By using basis function expansions, we can

\* Corresponding author. E-mail: 950 S. McAllister, P. O. BOX 871104, Tempe, AZ 85287-1104, United States.
Hyewon.Suk@asu.edu.
represent infinite-dimensional curves onto a finite dimensional space spanned by well-known basis functions, which in turn enables us to deal with infinite-dimensional curves within the framework of multivariate analysis without losing important temporal variations in the data. To examine directional relationships among multiple sets of functional responses, functional GSCA extracts a component from each set of functional responses in such a way that it explains the variance of the set of responses and accounts for the directional relationships among the components as much as possible simultaneously. Thus, the extracted components will be optimal for summarizing the observed responses as well as for capturing the hypothesized directional relationships among the components. At the same time, functional GSCA estimates functional parameters such as loading functions and weight functions by minimizing a penalized least squares criterion that can take into account the smooth nature of the functions.

3 An empirical example

In this example, there are four sets of functional responses obtained from 152 movies: weekly box office revenue measured for 10 consecutive weeks from the release and advertising spending on the three different media, television, newspapers, and radios, measured for 15 consecutive weeks from 5 weeks before the release. The purpose of the analysis is to examine how the advertising spending on three different media predicts the movie revenue. Table 1 displays the estimated path coefficients and their 95% bootstrap confidence intervals. The advertising spending on each of the three media turns out to be significant in predicting the box office revenue. We can also see that the advertising spending on television and newspapers is more important to predict the box office revenue than the advertising spending on radios.

<table>
<thead>
<tr>
<th>Path</th>
<th>Estimate</th>
<th>95% Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
<td>To</td>
<td>Lower limit</td>
</tr>
<tr>
<td>Televisions</td>
<td>Box office revenue</td>
<td>0.3282</td>
</tr>
<tr>
<td>Newspapers</td>
<td>Box office revenue</td>
<td>0.3717</td>
</tr>
<tr>
<td>Radios</td>
<td>Box office revenue</td>
<td>0.1442</td>
</tr>
</tbody>
</table>

Figures 1(a) and 1(b) show the estimated loading functions and weight functions with their 95% bootstrap confidence intervals. The loading functions show the overall patterns of change of the four functional responses over time. The weight functions indicate which time periods are crucial for examining the effect of advertising spending on the three media in predicting the movie revenue.

References

Multigroup or multiblock data analysis with RGCCA

Arthur Tenenhaus\textsuperscript{a}\textsuperscript{*} and Michel Tenenhaus\textsuperscript{b}

\textsuperscript{a}SUPELEC, Gif-sur-Yvette, France
\textsuperscript{b}HEC, Jouy-en-Josas, France

Keywords: Regularized generalized canonical correlation analysis, multiblock data analysis, multigroup data analysis

Introduction

In the multi-block framework, the column partition $X = [X_1, \ldots, X_j, \ldots, X_J]$ is considered. In this case, each block $X_j$ is an $n \times p_j$ data matrix and represents a set of $p_j$ variables observed on a set of $n$ individuals. The number and the nature of the variables differ from one block to another but the individuals must be the same across blocks. The main objective of multiblock analysis is to understand relationships between blocks. In the multi-group framework, the row partition $X = [X_1', \ldots, X_i', \ldots, X_I']$ is considered. In this case, the same set of variables is observed on different groups of observations. Each $n_i \times p$ data matrix $X_i$ is called a group and represents a set of $p$ variables observed on a set of $n_i$ individuals. The number of observations of each group could differ from one group to another. The main aim is to investigate the relationships among variables within the various groups.

Regularized Generalized Canonical Correlation Analysis (RGCCA) \cite{1}, which is a unifying approach for multiblock data analysis, is extended to be also a unifying tool for multigroup data analysis \cite{2}.

1 A unified optimization problem for multiblock and multigroup data analysis

Multiblock and multigroup data analyses can be tackled with a single optimization problem that we present in this section. We consider $I$ matrices $Q_1, \ldots, Q_I$. Each matrix $Q_i$ is of dimension $m \times p_i$. We also associate to each matrix $Q_i$ a symmetric definite positive matrix $M_i$ of dimensions $p_i \times p_i$. Moreover, a design matrix $C = (c_{ij})$ is defined with $c_{ij} = 1$ if matrices $Q_i$ and $Q_j$ are connected, and $= 0$ otherwise. The core optimization problem considered in this paper is defined as follows:

$$\max_{a_1, \ldots, a_I} \sum_{i,j=1; i \neq j}^I c_{ij} g((Q_i a_i, Q_j a_j)) \quad \text{subject to} \quad a_i' M_i a_i = 1, \quad i = 1, \ldots, I$$

(1)

where $g$ stands for the identity, the absolute value or the square function. By setting $b = M_i^{1/2} b_i$ and $P_i = Q_i M_i^{1/2}$ optimization problem (1) becomes

$$\max_{b_1, \ldots, b_I} \sum_{i,j=1; i \neq j}^I c_{ij} g((P_i b_i, P_j b_j)) \quad \text{subject to} \quad b_i' b_i = 1, \quad i = 1, \ldots, I$$

(2)

A monotone convergent algorithm similar to the RGCCA algorithm can be developed for optimization problem (2). This algorithm is given in Algorithm 1. We stress that only first dimension components are built in Algorithm 1. Next dimensions can be easily obtained by following the same procedures on deflated blocks or groups with respect to the previous components.

\textsuperscript{*}Corresponding author. E-mail: arthur.tenenhaus@supelec.fr
1.1 Multi-block data analysis (for column partition)

In the multi-block framework, a column partition $X = [X_1, \ldots, X_j, \ldots, X_J]$ is defined. In this framework, the same set of variables is observed on different groups of observations. Each $n_i \times p_j$ data matrix $X_j$ is called a block and represents a set of $p_j$ standardized variables observed on a set of $n_i$ individuals. A $p_j \times p_j$ positive definite symmetric matrix $M_j$ is associated to $X_j$. Moreover, a design matrix $C = (c_{jk})$ is defined with $c_{jk} = 1$ if blocks $j$ and $k$ are related, and 0 otherwise. The following optimization problem is defined:

$$
\text{maximize } \sum_{j,k=1; j \neq k}^J c_{jk} \cdot g(\text{cov}(X_ja_j, X_ka_k)) \quad \text{subject to } a_j^T M_j a_j = 1, j = 1, \ldots, J
$$

(3)

where $g$ stands for the identity, the absolute value or the square function. Optimization problem (3) is a special case of optimization problem (2) with $Q_j = X_j$ and $m$ equals to the number $n_i$ of individuals. We note that RGCCA initially described in [1] is a special case of optimization problem (3) with $M_j = \tau_j I_{p_j} + (1 - \tau_j) \frac{1}{n} X_j^T X_j$, the shrinkage constant $\tau_j$ varying between 0 and 1.

1.2 Multi-group data analysis (for row partition)

In the multigroup framework a row partition $X = [X_1, \ldots, X_j, \ldots, X_J]^T$ is considered. In this framework, the same set of variables is observed on different groups of observations. Each $n_i \times p$ data matrix $X_i$ is called a group and represents a set of $p_i$ centered and normalized (unit norm) variables observed on a set of $n_i$ individuals. As previously, a $p \times p$ positive definite symmetric matrix $M_i$ is associated to $X_i$. A design matrix $C = (c_{il})$ and a function $g$ are defined as previously in problem (3). The following optimization problem is considered:

$$
\text{maximize } \sum_{i,l=1; l \neq i}^J c_{il} \cdot g(\langle X_i^T X_i a_i, X_i^T X_i a_l \rangle) \quad \text{subject to } a_i^T M_i a_i = 1, i = 1, \ldots, J
$$

(4)

Optimization problem (4) is also special case of optimization problem (2) with $Q_i = X_i^T X_i$ and $m$ and $p_i$ both equal to $p$. In this paper, $M_i = \tau_i I_p + (1 - \tau_i) X_i^T X_i$ is considered with $\tau_i$ equal to 0 or 1.

References


Changes in Post-Adoption Use of Information Systems: A Generalized Darwinism Framework

Vanesa Tennant\textsuperscript{a}, Annette M. Mills\textsuperscript{a,}\textsuperscript{*} and Wynne W. Chin\textsuperscript{b}

\textsuperscript{a} University of Canterbury, Christchurch, New Zealand
\textsuperscript{b} University of Houston, Texas, United States of America

Keywords: Post-adoption, Generalized Darwinism, change, evolution, use

Introduction

The ability to manage how people use information systems and technologies is critical for business practice to avoid the problem of technology underutilisation and to extract the most benefit from their investments. To achieve this requires some degree of change on the users’ part as they modify how the Information Systems (IS) is incorporated into their work practices [1,2]. The concept of change is therefore central to addressing the problem of underutilization, yet few studies have empirically examined change in IS use [3,2]. This research seeks to bridge a gap in the literature and respond to calls for further investigation of change and for alternative theoretical perspectives to understand changes in post-adoptive behaviors. To address this gap, we draw on evolutionary theory, that is, Generalized Darwinism and its key principles of variation, selection and retention, to explain how individuals’ use of an IS changes over time [4]. Furthermore, there were no prior studies found that have systematically developed and empirically tested measures for variation and retention either within or outside the IS literature. Thus, beyond our theoretical contributions on change in IS use, this study also provides a new set of adaptable measures for assessing the concepts of variation and retention.

Using a multi-method research design, this study includes an exploratory phase (qualitative) followed by a confirmatory phase (quantitative). For the qualitative phase, case studies were used to explore change in IS use; a cross-section of 39 users (i.e. basic, intermediate and advanced) of large-scale IS from three organizations were interviewed. The findings from the qualitative phase along with Generalized Darwinism principles of variation, selection and retention were used to develop a conceptual model that frames changes in post-adoption use for further analysis. The measurement and structural models were then tested using data collected from a field survey of 86 users of a Learning Management System and analyzed using the Partial Least Squares (PLS) approach to structural equation modeling. All measures were self-developed except for infusion, which was adapted from Jones et al [5].

1 References


* Corresponding author. Email: annette.mills@canterbury.ac.nz
On Estimation of Covariance Matrices Modeled as a Sum of Kronecker Products

Arun Venkitaramana and Magnus Janssonb

a, bDepartment of Signal Processing, KTH Stockholm, Sweden

Keywords: Covariance estimation, Kronecker products, ML estimation, Flip-Flop Algorithm, Biconvex function

Introduction

We address the problem of estimation of covariance matrices expressible as a sum of Kronecker products (KPs). Our goal is to arrive at estimates of the KP component matrices within a maximum-likelihood (ML) framework. Since the exact solution of the ML cost function is non-tractable, we propose a covariance-matching (CM) approach, noting that the estimates obtained from covariance-matching coincide asymptotically with those obtained from ML-estimation [1]. The minimization of the CM cost function is more tractable and can be solved efficiently as it is 'biconvex' in the components as we describe in Section 1. We look into two variants of the CM approach; in the first case, we consider the unconstrained solution to the CM cost function whereas in the second case, we restrict the estimated KP components be positive-definite (PD).

1 Covariance Matching for Sum of KP Model

Let \( R_0 \) denote the true covariance matrix that we wish to estimate such that \( R_0 = \sum_{\gamma}^K A_\gamma \otimes B_\gamma \), where \( A_\gamma \)s and \( B_\gamma \)s denote the component matrices, and \( K \) denotes the number of KP components assumed to be known. Let \( \hat{R}_n \) denote the sample covariance matrix obtained using \( n \) samples. We define our estimates \( A_\gamma \)s and \( B_\gamma \)s as the minimizers of the following cost function:

\[
C = \| \hat{R}_n^{-1}(\hat{R}_n - \sum_{\gamma}^K A_\gamma \otimes B_\gamma) \|^2_2, \tag{1}
\]

where \( \|A\|_2 \) denotes the Frobenius norm of matrix \( A \). The cost function is non-convex and has no closed-form solution. Notice, however, that if we fix all \( A_\gamma \)s, \( C \) is convex in \( B_\gamma \)s, and vice versa. In other words, \( C \) is a biconvex function over the set \( (A_1, A_2, \cdots) \times (B_1, B_2, \cdots) \) which motivates us to arrive at a flip-flop algorithm for estimation of KP components by alternately minimizing \( C \) with respect to \( A_\gamma \)s and \( B_\gamma \)s:

1. Initialize \( A = A_\gamma^{(0)} \) \( \forall \gamma \)
2. Set \( i = 0 \): Compute \( B_\gamma^{(0)} \) by minimizing \( C \) with respect to \( B_\gamma \)s after setting \( A_\gamma = A_\gamma^{(0)} \)
3. Set \( i = i + 1 \): Compute \( A_\gamma^{(i+1)} \) by minimizing \( C \) with respect to \( A_\gamma \)s after setting \( B_\gamma = B_\gamma^{(i)} \)
4. Set \( i = i + 1 \): Compute \( B_\gamma^{(i+1)} \) by minimizing \( C \) with respect to \( B_\gamma \)s after setting \( A_\gamma = A_\gamma^{(i+1)} \)

*Corresponding author. E-mail: magnus.jansson@ee.kth.se.
In our work, we choose the minimizers of the unweighted Frobenius norm for initialization, that is, $A^{(0)}_\gamma = A_S^\gamma$, where $\{A_S^\gamma\}_{\gamma=1}^{K}$ are obtained by minimizing $\|\hat{R}_n - \sum_{\gamma=1}^{K} A_\gamma \otimes B_\gamma\|_2^2$. It is a known result that $\{A_S^\gamma\}_{\gamma=1}^{K}$ are obtained by reshaping the column vectors obtained from the singular value decomposition (SVD) expansion of the reshuffled sample covariance matrix [2, 3]. In the absence of further conditions, it is not guaranteed that the resulting matrices are PD. In such a case, the estimated KP components $\{A_\gamma \otimes B_\gamma\}_{\gamma=1}^{K}$ cannot be interpreted as being individual covariance matrices.

1.1 Positive definiteness of the KP components

We also consider the estimation of KP components under the constraint that the estimated component matrices are PD, that is, $A_\gamma$ and $B_\gamma$ are PD for all values of $\gamma$. This is motivated by the observation that in many applications, for example in MEG/EEG signals, the underlying process is modeled as the resultant of multiple statistically independent source signals, in which case $A_\gamma$ and $B_\gamma$ are themselves covariance matrices [4]. As in the unconstrained case, we obtain the PD estimates by minimizing the cost function in (1) with the additional positive-definiteness constraint imposed on the components. The estimates are obtained using the same flip-flop algorithm as in the unconstrained case with the additional PD constraint. We use convex programming solvers to implement our algorithm. Since the feasible set is now restricted to the space of PD matrices, we use identity matrices of appropriate dimensions for initialization.

In Figure 1(a), we show the estimation error expressed as a function of the number of iterations for the unconstrained and constrained CM approaches for a synthesized example. We observe that in both cases, the algorithm converges after a sufficient number iterations. As expected, imposing the PD constraint results in a higher MSE when compared with the unconstrained case, and this manifests as a gap as shown in Figure 1. The gap reduces as the number of samples is increased. Figure 1(b) corresponds to the case when the actual covariance matrix is used in the estimation, that is, $\hat{R}_n = R_0$. We observe that after three to four iterations, the performance of the constrained and unconstrained cases coincide and the gap in between the corresponding errors is almost negligible.

References


Functional Linear Regression Analysis With Partial Least Squares And Its Application

Huiwen Wang and Lele Huang *
School of Economics and Management, Beihang University, Beijing, China

Keywords: Functional regression, Partial least squares, Group variable selection, Prediction

Introduction

With the rapid development of modern technology, especially the computer-related technique, data can be recorded densely over time (can be seen in the continuum), such as stock and future prices. Such type of data is termed functional or curve data. There are broad application prospects for functional data in a wide range of areas, which promotes the development of functional data analysis (FDA). There are vast literature on FDA, such as [1, 2, 3].

In the functional linear regression model, functional coefficients of functional predictors are to be estimated. With functional predictors in consideration, we have to transform them based on basis expansion. During the procedure, it’s crucial to choose basis function sequences and truncating parameter, see [4, 5].

Let $Y$ be a real-valued random variable defined on a probability space $(\Omega, \mathcal{B}, P)$, $\{X_j(t) : t \in F\}$ be a zero mean, second-order stochastic process defined on $(\Omega, \mathcal{B}, P)$ and $EX_j^2(t) < \infty$ for all $t \in F, 1 \leq j \leq p$. The sample paths of $\{X_j(t) : t \in F\}$ are in $L^2(F)$, the set of all square integrable functions on $F$. To express different kinds of functional data, such as curves, images and arrays, $F$ can be subsets of $R, R^p$ or other spaces. It’s assumed that the scalar response $Y$ is linearly related to the functional predictor $X_j(t)$ through the relationship

$$Y = \alpha + \sum_{j=1}^{p} \int F \beta_j(t)X_j(t)dt + \varepsilon$$

(1)

where the intercept $\alpha$ and $\varepsilon$ are scalars, $\varepsilon$ is a random error variable. The functional coefficient $\beta_j(t), 1 \leq j \leq p$ are to be estimated. Model (1) is applicable in a wide range of settings, especially where data are collected only through new developments in technology. Furthermore, we assume that among the $p$ functional predictors, only $p_0$ predictors are really correlated with $Y$ and we need to identify these predictors and estimate these slope functions $\beta_j(t)$.

1 Methodology

For selection and estimation, we first expand the predictors and their slope functions on the basis functions, which are obtained by PLS method, and then grouped variable selection method is utilized to select and estimate simultaneously.

1.1 PLS method

In the functional setting, if there is a single functional predictor, then the standard PLS basis is defined iteratively by choosing $\psi_p$ in a sequential manner, to maximize the covariance functional

$$\gamma_p(\psi_p) = cov \left(Y - g_{p-1}(X), \int X\psi_p \right)$$

(2)

*Corresponding author. E-mail: nanhuabiren@163.com.
subject to

\[ ||\psi_p|| = 1, \int \int \psi_j(s,t)K(s,t)\psi_p(t)dsdt = 0, \quad 1 \leq j \leq p - 1, \]

where \( g_{p-1}(x) = E(Y) + \sum_{j=1}^{p-1} f_j \cdot \psi_j \). According to [6], if \( E X^2 < \infty \), then the function \( \psi_p \) that maximizes \( \gamma_p \) in (2), expressed by \( \psi_1, \cdots, \psi_{p-1} \), is determined by

\[ \psi_p = c_0[K(b - \sum_{j=1}^{p-1} \psi_j \int b_j \psi_j) + \sum_{j=1}^{p-1} c_j \psi_j], \quad (3) \]

where, for \( 1 \leq j \leq p - 1 \), the constants \( c_j \) are derived by solving the linear system of \( p - 1 \) equations

\[ \int \int \psi_j K \psi_p = 0, \quad j = 1, \cdots, p - 1, \]

and where \( c_0 \) is determined by \( ||\psi_p|| = 1 \), \( K \) is the covariance function defined by \( K(s,t) = \text{cov}(X(s),X(t)) = E[X(s)X(t)] \).

1.2 Estimation and selection in multiple functional predictors regression model

Assume that we have obtained PLS basis functions for each \( X_j(t) \), then by an appropriate tuning parameter \( K_j \),

\[ X_j(t) \approx \xi_j \psi_j(t) = \sum_{k=1}^{K_j} \xi_{jk} \psi_{jk}(t), \quad \beta_j(t) \approx b_j \psi_j(t) = \sum_{k=1}^{K_j} b_{jk} \psi_{jk}(t). \]

Then we can select predictors and estimate the slope functions simultaneously by minimizing the objective function

\[ Q(B) = \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} \xi_j b_j^T)^2 + \sum_{j=1}^{p} P_\lambda(||b_j||), \quad (4) \]

where \( B = [b_1, \cdots, b_p]^T \) and \( ||b_j|| \) is the \( L_2 \) norm of \( b_j \).

Let \( \hat{B} = \arg\min Q(B) \), then \( \hat{B}_j(t) = \hat{b}_j \hat{\psi}_j(t), 1 \leq j \leq p \).

2 Simulation and real data

2.1 Simulation

2.2 Real data

Nowadays, in China people concern more and more about air quality index (AQI), which is an integrated index of concentration of \( PM2.5, PM10, CO, NO_2, SO_2 \). We collect these five concentration series of main cities over a period, which are treated as five functional predictors. The search frequency of respiratory system related words is the response. We try to find the significant predictors and estimate the slope functions.

References

Complex indicators for earthworm (*Lumbricus terrestris*) population viability – a PLS-PM path model approach

Sebastian Wolfrum\textsuperscript{a*}, Norman Siebrecht\textsuperscript{a}, Susanne Papaja-Hülsbergen\textsuperscript{a}, Gisela Lüscher\textsuperscript{b}, Philippe Jeanneret\textsuperscript{b}, Michaela Arndorfer\textsuperscript{c}, Jürgen K. Friedel\textsuperscript{c}, Maurizio G. Paoletti\textsuperscript{d}, Maximilian Kainz\textsuperscript{a}, Kurt-Jürgen Hülsbergen\textsuperscript{a} and Felix Herzog\textsuperscript{b}

\textsuperscript{a} Technische Universität München, 85354 Freising, Germany
\textsuperscript{b} Agroscope Reckenholz-Tänikon Research Station ART, Zurich, Switzerland
\textsuperscript{c} University of Natural Resources and Life Sciences, 1180 Vienna, Austria
\textsuperscript{d} Department of Biology, Padova University, 35121 Padova, Italy

Keywords: agriculture, ecology, organic farming, sustainability, soil fertility

Introduction

Information on biodiversity and species population parameters are of key importance for assessing related ecosystem services and impacts of agricultural management. However information on this parameters is often not directly measurable. Indicators are a convenient way to gain information on these variables. Here we present a Partial Least Squares Path Modeling (PLS-PM) approach to develop and evaluate complex indicators for biodiversity and related parameters [1]. As an example we use data on the common earthworm *Lumbricus terrestris* known as an “ecological engineer” and responsible for soil fertility and structure.

![Figure 1. Conceptual structure of PLS path model - LV = latent variable, M = outer model, S = inner model and hypothized relations (arrows)](image)

1 Material and methods

Eight organic and eight conventional randomly selected dairy cow farms in the South of Germany were selected. After a standardized habitat mapping [2] in each agricultural used habitat category a sample plot was placed randomly in one element. In total earthworms were sampled on 129 plots with a combined method [3]. For the analysis presented in this article abundance and biomass of *Lumbricus terrestris*, management...
parameters (tillage, fertilization, residues) and soil characteristics (N, C, Ph, texture, moisture) from 76 plots was used to create complex indicators for population viability, management intensity and site potential (Fig.1). SmartPLS [4] was used for calculating the PLS-PM.

2 Results

The model showed a significant negative relation of management intensity ($\beta=-0.222$) and a significant positive relation of site potential ($\beta=0.247$) with *Lumbricus terrestris* population viability ($R^2=0.162$). Arable plots had higher management intensity than grassland plots leading to lower viability scores on arable land. However grassland scores are highly variable. Averaged results for farm level (Tab. 1) allow a ranking of farms, but do not show a significant difference between management systems.

<table>
<thead>
<tr>
<th>Farm</th>
<th>Management system</th>
<th>Site potential (mean/farm)</th>
<th>Management intensity (mean/farm)</th>
<th>Population viability (mean/farm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>OF</td>
<td>0.231</td>
<td>0.610</td>
<td>-0.751</td>
</tr>
<tr>
<td>7</td>
<td>CF</td>
<td>0.101</td>
<td>-0.076</td>
<td>-0.630</td>
</tr>
<tr>
<td>12</td>
<td>CF</td>
<td>-1.792</td>
<td>0.907</td>
<td>-0.623</td>
</tr>
<tr>
<td>11</td>
<td>CF</td>
<td>-0.194</td>
<td>0.026</td>
<td>-0.615</td>
</tr>
<tr>
<td>3</td>
<td>CF</td>
<td>0.699</td>
<td>0.665</td>
<td>-0.282</td>
</tr>
<tr>
<td>5</td>
<td>CF</td>
<td>-0.963</td>
<td>0.294</td>
<td>-0.146</td>
</tr>
<tr>
<td>8</td>
<td>OF</td>
<td>-0.118</td>
<td>-0.126</td>
<td>-0.137</td>
</tr>
<tr>
<td>6</td>
<td>CF</td>
<td>0.449</td>
<td>-0.294</td>
<td>-0.029</td>
</tr>
<tr>
<td>15</td>
<td>OF</td>
<td>-0.611</td>
<td>0.393</td>
<td>-0.011</td>
</tr>
<tr>
<td>14</td>
<td>OF</td>
<td>0.517</td>
<td>-0.393</td>
<td>0.117</td>
</tr>
<tr>
<td>13</td>
<td>CF</td>
<td>-0.077</td>
<td>0.013</td>
<td>0.234</td>
</tr>
<tr>
<td>16</td>
<td>OF</td>
<td>0.555</td>
<td>-0.918</td>
<td>0.568</td>
</tr>
<tr>
<td>9</td>
<td>OF</td>
<td>0.202</td>
<td>-0.062</td>
<td>0.570</td>
</tr>
<tr>
<td>10</td>
<td>OF</td>
<td>0.910</td>
<td>-0.517</td>
<td>0.585</td>
</tr>
<tr>
<td>4</td>
<td>CF</td>
<td>0.333</td>
<td>-0.006</td>
<td>0.647</td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td>0.202</td>
<td>-0.006</td>
<td>-0.029</td>
</tr>
</tbody>
</table>

3 Conclusions

PLS-PM provides a flexible solutions to develop and evaluate complex biodiversity indicators. The models can be extended to other organisms or species assemblages. However, the main advantage is the possibility to assess the quality of models and the calculation of latent variable scores that can be used for further analysis.

References


Use of PLS path modeling to study customer satisfaction of two Lebanese mobile operators

Genane Youness a*, and Ahmad Ramadan a

a ISSAE -Cnam, Beirut, Lebanon.

Keywords: PLS path modeling, ECSI, Latent variable, satisfaction, bootstrap.

Introduction

The purpose of this study is to measure customer satisfaction of two mobile operators in Lebanon, Alfa and Mctouch, from a sample of the telephone customers for these two operators. 500 phone active numbers randomly selected from their system database, are drawn. After telephone contacts, we were able to record and complete 300 questionnaires referred to the EPSI (European Performance Satisfaction Index) questionnaire of 23 questions and that for each operators; the data collection was carried out between May 2012 and June 2012. The goal is to construct the structural equations modeling and to estimate the measure of customer satisfaction by PLS approach and to build a satisfaction index for each operator. The ECSI model (European Customer Satisfaction Index) is used [1], to solve the main difficulties in measuring customer satisfaction. It provides best reliable results and do better understand and monitor the quality of customer relationship business. The hypothesis are tested and discussed.

1 Empirical study.

It should be noted that all manifest variables to the corresponding latent variables have a reflective measurement model. After the convergence of the PLS algorithm, the structural path significance testing of both inner and outer model is verified, based on bootstrap procedure with 300 subsamples. The results showed that the factor Customer Complaints has to be removed from both models. For Alfa model, the relations of Perceived Quality, and Image to Satisfaction, are not significant and have to be removed. For Mctouch model, the relations of Perceived Quality, Image, and customer expectation, to Satisfaction have to be removed.

After applying these modifications and using the algorithm PLSPM by SmartPLS software [2], we obtained the final models for Alfa an Mctouch. For Alfa model, the coefficient of determination R² for Satisfaction is 0.55. This means that the variance in customer Satisfaction is explained 55% by Expectation and Perceived Value. For Loyalty, the important factor in mobile phone industry, R² is 0.482. This means that the variance of Loyalty is explained 42% by Satisfaction and Image. The inner model suggests that Expectation has the strongest effect on Satisfaction (0.501), followed by Perceived Value and Image. For Mctouch model, the variance of Satisfaction is explained 54.6% by Perceived Value. The variance of Loyalty is explained 57.6% by Satisfaction and Image. The inner model suggests that the only important impact on Satisfaction is Perceived Value (0.739). Satisfaction has the strongest effect on Loyalty (0.45) then Image (0.39). Image has no direct impact on Satisfaction.

For Alfa: Satisfaction = 0.5014Expectation + 0.3085PerceivedValue. Loyalty=0.427 Satisfaction+0.34 Image.
For Mctouch: Satisfaction = 0.7392PerceivedValue. Loyalty=0.45 Satisfaction+0.39 Image.

The analysis reveals that Perceived Value turned out to be such an important factor for the two models; it is the only strong predictor of Satisfaction for Mctouch operator. For Alfa, the most important impact is Expectations.

* Corresponding author. E-mail: genane.youness@cnam.fr, Beirut, Lebanon, BP 11- 4661 Beirut, Lebanon.
2 Comparison of the two models

For both mobile operators, the obtained quality indices considered by the communality, the redundancy, and the Goodness Of Fit, have good predictive and discriminant validity. The Results are shown in the following table 1:

Table 1. Comparison of results found for Alfa and Mctouch models.

<table>
<thead>
<tr>
<th></th>
<th>Alfa</th>
<th>Mctouch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial Model</td>
<td>Final Model</td>
</tr>
<tr>
<td>Reliability</td>
<td>0.8630</td>
<td>0.8630</td>
</tr>
<tr>
<td></td>
<td>0.8515</td>
<td>0.8515</td>
</tr>
<tr>
<td>A.V.E</td>
<td>0.6780</td>
<td>0.6778</td>
</tr>
<tr>
<td></td>
<td>0.6570</td>
<td>0.6539</td>
</tr>
<tr>
<td>R²</td>
<td>0.5644</td>
<td>0.5502</td>
</tr>
<tr>
<td></td>
<td>0.5920</td>
<td>0.5464</td>
</tr>
<tr>
<td>Redundancy</td>
<td>0.2738</td>
<td>0.3059</td>
</tr>
<tr>
<td></td>
<td>0.1065</td>
<td>0.3203</td>
</tr>
<tr>
<td>GOF</td>
<td>0.5283</td>
<td>0.5562</td>
</tr>
<tr>
<td></td>
<td>0.5208</td>
<td>0.5454</td>
</tr>
<tr>
<td>Satisfaction Index</td>
<td>7.7259</td>
<td>7.7256</td>
</tr>
<tr>
<td></td>
<td>7.7499</td>
<td>7.7480</td>
</tr>
</tbody>
</table>

The study led to the development of an overall composite index of customer satisfaction for Alfa and Mctouch. The average score in terms of overall satisfaction obtained for all Alfa or Mctouch customers is around 7.75. The Alfa and Mctouch models are not significantly different in term of satisfaction indices (for H0: dif=0 versus dif≠0, tvalue=0.2077 and Pvalue=0.8356).

Conclusion

In this paper customer satisfaction for both Alfa and Mctouch mobile operators were compared by PLS approach using ECSI format. Perceived value was found to be the only driver of customer satisfaction for Mctouch while expectation was found the most important one for Alfa;

References


Comparability Regarding Cross-Cultural Research: Establishing Data Equivalence Using Amos Multi-Group Analysis

Zheng L.-L.\textsuperscript{a}*, Plaisent M.\textsuperscript{b}, Bernard, P.\textsuperscript{b}

\textsuperscript{a}La Rochelle Business School, La Rochelle, France
\textsuperscript{b}University of Quebec in Montreal, Montreal, Canada

Keywords: measurement equivalence, Amos Multi-Graphic Analysis, Cross-cultural research

Introduction

Measurement invariance refers to whether items and factorial structure of a survey instrument are equivalent across different groups in terms of a variety of categories, such as, gender, culture, age, or occupation. In cross-group research, measurement equivalence issues (i.e., whether measurement accuracy, reliability and validity are achieved across samples) are critical (Byrne, 2004). When a measurement instrument is not equivalent in a cross-group study, items may have different meaning for different population groups, and thus the factorial structure of the measurement instrument might not hold across groups. The validity of research findings are then problematic and need further investigation. Multiple group SEM analysis has been suggested to be employed to test measurement equivalence. Despite the importance of measurement equivalence, “this is a road less traveled” in cross-group studies, particularly in cross-cultural comparative research (Byrne, 2004, p.272). In this paper, we intend to highlight the need of measurement invariance testing in the case of cross-cultural studies and use of multi-group SEM analysis using AMOS.

1. Establishing data equivalence in cross-cultural comparative research

Cross-cultural research is not an easy task (Cavusgil and Das, 1997). Compared to domestic research, in cross-national research the researcher must overcome many additional factors not inherent in typical research tasks. Cross-cultural Marketing research presents enhanced and specific methodological problems, stemming mostly from cultural construct equivalent issues (Craig and Douglas, 2000). To ensure equivalence in the conceptual model adopted, attention needs to be paid to construct equivalence to ensure that the constructs being studies are equivalent in all contexts and cultural settings, as well as measure equivalence, that operational measures are equivalent in all research settings.

1.1. Construct equivalence

Examination of construct equivalence requires examination of three distinct aspects. If the nature of a construct varies across cultures, then it is questionable as to whether a measure of that construct developed in one culture is a construct valid measure of the same construct in a different culture (Craig and Douglas, 2005). Thus, cross-cultural researchers must pay attention to conceptual and functional construct equivalency when measuring constructs in a cross-cultural context (Scroggins et al. 2010). According to Craig and Douglas (2000), the researcher must firstly assess whether a given concept serves the same function from country to country (i.e., functional equivalence). Second, the researcher must determine whether the same concepts occur in different countries and whether the way in which they are expressed is similar (i.e., conceptual equivalence). Finally, where the concept belongs to a specific class of objects or activities, the researcher must examine whether the same classification scheme of objects can be used across countries (i.e., category equivalence).
1.2. Measurement Equivalence

Measurement equivalence refers to whether the measurement operations yield measures of the same attribute in different countries (Steenkamp and Baumgartner, 1998). It has to be determined in terms of the “units of measurement used, as well as in the translation of verbal and nonverbal instruments to ensure comprehension by respondents in different socio-cultural environments” (Craig and Douglas, 2000, p. 20). To make valid cross-cultural comparisons, it is necessary to establish calibration, translation and metric equivalence (Craig and Douglas, 2000).

2. Testing measurement invariance using Amos multi-group analysis

Multiple group SEM analysis has been suggested as a reliable method for determining measurement equivalence (Steenkamp and Baumgartner, 1998). Generally, a measurement invariance testing requires the following five main steps:

**Step 1: The baseline model** As a prerequisite to testing for factorial invariance, it is customary to consider a baseline model, which is estimated for each cultural group separately.

**Step 2: Testing for the validity of the Hypothesized Model across cultural groups** As a preliminary step in testing for invariance across groups, we test for the validity of the hypothesized model. Parameters are estimated for two groups simultaneously (Byrne, 2004).

**Step 3: Testing for Invariance of the Fully Constrained Model across groups** Prior to testing for the equality of sets of parameters, it is worthwhile testing for the possibility that a fully constrained model is invariant across groups (Byrne, 2004). All factor loadings are constrained equal across cultural groups. In testing for the invariance of this constrained model, we compare its chi-square value with that for the initial.

**Step 4: Partial Measurement Invariance** Given findings of non-invariance at the level of all factor loadings, we then proceed to test for the invariance of all factor loadings in each subscale (i.e., all loadings related to the one particular factor). Given evidence of non-invariance at the subscale level, we then test for the invariance of each factor loading (related to the factor in question) separately.

**Step 5: Testing for Invariance of Structure Model across groups** First we test for invariance of all structural paths. We constrained all structural paths. If one or more regression weights of the paths in the structure model are different between groups, which means inequality exists on all corresponding structural paths across groups (p<0.05), we then need to test for the invariance of each structural path separately to pinpoint the non-invariance regression paths. To pinpoint these different regression paths, the orderly process of testing for the invariance of regression weight parameters is continued until all hypothesized parameters are found to be equivalent across groups, their specified equality constraints are retained (cumulatively) throughout the remainder of the invariance-testing process (Byrne, 2004).

References


Partial least square components representing frailty improve survival analysis with gene expression profiles

Yi Zhou\textsuperscript{a} and Siu-wai Leung\textsuperscript{a,b,*}

\textsuperscript{a} State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao, China
\textsuperscript{b} School of Informatics, University of Edinburgh, United Kingdom

Keywords: Partial least square components, frailty, survival analysis, gene expression data

Introduction

Gene expression profiles have been used in survival analysis for cancer prognosis to predict the patients’ survival outcomes [1]. Because the large number of genomic variables may cause overfitting and collinearity, it would be necessary to implement the dimension reduction methods. Partial least square (PLS) method has been used in dimension reduction to predict the survival probabilities in Cox regression [2]. However, the classic multivariate Cox regression assumes the follow-up time of patients to be independent with each other [3]. Considering the gene expression may become the frailty of patients who are frail to die earlier [4], we build a prognostic model adding PLS components to represent frailty, named PLS-frailty model, to improve the survival prediction.

1 Methods

We adopted partial least square (PLS) regression to extract $k$ PLS components with genomic variables as predictors and patient survival rates as responder. Additionally, the frailty of censoring was added into PLS regression. The most significant $q$ ($q < k$) PLS components selected by the univariate Cox regression were regarded as covariates to fit the frailty model for survival analysis. The $q^{\text{th}}$ component was calculated to follow Gaussian distribution and was added into the model to represent frailty. Let $y_i$ ($i = 1, 2, \ldots, n$) denote the patient survival time with or without censoring. $\lambda$ is the proportional hazard function, and $\lambda_0$ is the baseline hazard function. $T_j$ is the selected genomic PLS components, $\mathit{frailty}(T_q)$ is the gene frailty, and $\alpha$ is the coefficient vector. The PLS-frailty model is shown in Equation (1).

$$
\lambda(y) = \lambda_0(y) \mathit{frailty}(T_q) \exp \left( \sum_{j=1}^{q} T_j \alpha \right)
$$

We applied two breast cancer gene expression datasets. One is collected at Netherlands Cancer Institute (NKI) [2] and the other is collected in Stockholm [5]. Ten percent of the genomic variables were selected at random for analysis. The linear part of the survival model, named prognostic index (PI), was used to evaluate the discrimination and prediction performances in ten-fold cross-validation. In discrimination assessment, we predicted PIs in testing data to differentiate patients into low risk group (PI ≤ 0) and high risk group (PI > 0), and then assess the difference between the survival curves of the groups. In prediction evaluation, we used PIs to predict the overall patient survival probabilities and then evaluate the prediction errors (PEs) and relative prediction error reductions (RPERs) [7]. We compared the results in PLS-frailty model with that in PLS-Cox model [6]. PLS-Cox model is a combination of PLS regression and Cox regression for survival analysis without the addition of the frailties.
2 Results

PLS-frailty model had better performance in discriminating the patients into the high risk and low risk groups than traditional PLS-Cox model without frailty in the breast cancer datasets (Figure 1). The prediction performance of PLS-frailty model did not differ much from PLS-Cox model.

3 Conclusion

Our proposed PLS-frailty model added the frailties as random effects to improve the survival prediction of breast cancer from gene expression data. The model reduced the dimensions of genomic variables and outperformed PLS-Cox model in differentiating low risk and high risk patients.

References

Challenges in applying PLS to un-targeted serum NMR metabolomic data from a nested case-control study in the EPIC study

N. Assi, V. Viallon, A. Fages, B. Elena-Herrmann, M. Jenab and P. Ferrari

Abstract

Advances in technology developments in the field of analytical chemistry, mainly in nuclear magnetic resonance (NMR) spectroscopy and in mass spectrometry (MS), made it possible to generate massive sets of data for metabolomic analyses. The main objective of metabolomic studies is to identify relevant biomarkers associated with a higher risk of patho-physiological conditions and/or disease onset. While quantitative statistical methods are available for processing information on metabolomic profiles variability [1], the analytical strategy to simultaneously relate large sets of data is far from being an obvious one. In metabolomics, such methods are needed to fully characterize metabolomic profiles by integrating information on individuals’ environmental factors. Partial Least Squares (PLS) regression [2,3] is a very suitable multivariate technique for this purpose, as it generalizes features of principal component analysis (PCA) and multiple linear regression (MLR), by maximizing the covariance between two sets of variables, $X$ and $Y$, respectively. PLS iteratively extracts components that explain co-variability in $X$ and $Y$ [4].

Metabolomic profiles were assessed in a hepatocellular carcinoma (HCC) nested case-control study within the European Prospective Investigation into Cancer and Nutrition [5], a large prospective investigation with over 520,000 participants in 10 countries, whose design, data collection methods and rationale are detailed and described elsewhere [5,6]. HCC cases were identified during the study follow-up among study participants, through examination of relevant histology [7]. Pre-diagnostic serum samples from 114 HCC cases and 222 matched controls were processed using standard $^1$H NMR metabolic profiling protocols. As part of implementing a “meeting-in-the-middle” approach, an analytical strategy was developed to identify the link between lifestyle factors, including dietary habits, physical activity, anthropometry (the $X$ factors) and disease outcome, through metabolomic profiles (the $Y$ factors), possibly highlighting relevant markers of exposure on one hand, and predictive markers of disease outcome, on the other approach [8].

After removal of the structured noise, metabolomic spectra signals were first processed via an intelligent bucketing procedure, the Statistical Recoupling of Variables (SRV) [9,10]. This permitted to reduce the number of NMR variables to 285 cluster variables corresponding to reconstructed peak entities. The food and lifestyle matrix included 16 dietary variables, alcohol at recruitment, body mass index, height, measure of physical activity, smoking status, fasting status, diabetes, hepatitis, level of education, and a score for liver

Keywords: PLS, metabolomics, scaling, meeting-in-the-middle, hepatocellular carcinoma

Corresponding authors: 150, Cours Albert Thomas, 69008 Lyon, France. assin@students.iarc.fr, ferrarip@iarc.fr
damage. In this study, PLS was used to derive components by relating the Y matrix of NMR variables to the X matrix of lifestyle variables. Each factor was then related to HCC risk in conditional logistic regression models.

In this application, while the NMR signals were normalized to total intensity, normalized for batch effect and centered and Pareto scaled, the predictor matrix was comprised of continuous displaying different levels of variability, along with a few categorical variables. We therefore faced challenges related to i) the scaling of variables and ii) the choice of the optimal number of factors to retain using cross-validation procedures. Overall, X-variables were characterized by level of variability 10³- to 10¹⁰-fold larger than Y-signals, with a moderate-to-large impact on the findings. Different approaches were used to tackle this, including Pareto scaling, log-transformation and classical standardization. The leave-one-out cross-validation approach was primarily used. The number of selected PLS components was the one minimizing the predicted residual sum of squares, using van der Voet test [11].

References


Deviance residuals based sparse PLS and sparse kernel PLS regression for censored data

P. Bastien\textsuperscript{a,*}, F. Bertrand\textsuperscript{b}, N. Meyer\textsuperscript{bc}, M. Maumy-Bertrand\textsuperscript{b}

\textsuperscript{a}L’Oréal R&I, Aulnay, France  
\textsuperscript{b}IRMA, CNRS Université de Strasbourg, Strasbourg, France  
\textsuperscript{c}EA3430, INSERM Faculté de Médecine, Strasbourg, France

Keywords: PLS, Cox, sPLS, Deviance, Kernel

There has been a vast literature in the last decade devoted to relating gene expression profiles to subject survival or to time to cancer recurrence. Biomarker discovery from high-dimensional data, such as transcriptomic or SNP profiles, is a major challenge in order to allow more precise diagnosis. The proportional hazard regression model suggested by Cox, 1972, to study the relationship between the time to event and a set of covariates in the presence of censoring is the model most commonly used for the analysis of survival data.

However, like multivariate regression, it supposes that there are more observations than variables, complete data, and variables not strongly correlated between them. In practice when dealing with high-dimensional data, these constraints are crippling. Collinearity gives rise to issues of over-fitting and model mis-identification. Variable selection can improve the estimation accuracy by effectively identifying the subset of relevant predictors and enhance the model interpretability with parsimonious representation.

In order to deal with both collinearity and variable selection issues, many methods based on Lasso penalized Cox proportional hazard have been proposed since the seminal paper of Tibshirani, 1997: Fan and Li (2002), Gui and Li (2005), Segal (2006), Park and Hastie (2007), Zhang & Lu (2007), Zou (2008), Sohn (2009), Goemann (2009), Tibshirani (2009), Fan et al (2010), Simon et al (2011).

Regularization could also be performed using dimension reduction as is the case with PLS regression. PLS regressions has already been extended to Cox regression (Bastien (2001), Nguyen and Rocke (2002), Li and Gui (2004)). Recently, Chun & Keles (2010) provides both empirical and theoretical results that the performance of PLS regression is ultimately affected by the large number of predictors. In particular, existence of higher number of irrelevant variables leads to inconsistency of coefficient estimates in linear regression setting. Chun & Keles proposed sparse PLS regression which promotes variables selection within the course of PLS dimension reduction. Moreover, they demonstrated that for univariate PLS, the first direction vector of they sparse PLS algorithm is easily obtained by soft threholding of the original PLS direction vector.
We propose two original algorithms named sPLSDR and its non linear kernel counterpart DKsPLSDR, by using sparse PLS regression (sPLS) based on deviance residuals. We compared their predicting performance with state of the art algorithms based on reference benchmark datasets and simulations.

**Results:** As sPLSDR and DKsPLSDR compare favorably with other methods in their computational time, prediction and selectivity, as indicated from results based on benchmarks and simulated datasets, we view them as a useful addition to the toolbox of estimation and prediction methods for the widely used Cox’s model in the high-dimensional and low sample size settings.

**Availability and Implementation:** The R-package plsRcox is available on the CRAN and is maintained by Frédéric Bertrand.


* Corresponding author. E-mail: pbastien@rd.loreal.com.
Local classification by Locally Weighted Partial Least Squares-Discriminant Analysis (LW-PLSDA)

Marta Bevilacqua\textsuperscript{a} and Federico Marini\textsuperscript{a}\textsuperscript{*}

\textsuperscript{a}Department of Chemistry, University of Rome “La Sapienza”, Rome, Italy

Keywords: Local modeling, Locally Weighted - Partial Least Squares Discriminant Analysis (LW-PLSDA), classification

1. Introduction

Classification problems occur ubiquitously in chemometrics, from traceability to process monitoring, passing through metabolomics or image analysis. However, the development and widespread availability of more and more efficient instrumental platforms resulted in a corresponding increase in the complexity of the problems and tasks chemists are called to tackle and solve. In the framework of pattern recognition, this consideration translates to the possibility that many sources of variation, other than the class belonging, affect the instrumental signal, resulting in a non-linear separability of the categories in the feature space. Whenever this occurs, the use of non-linear classification methods constitutes an obvious answer to this issue; however, involving an higher number of adjustable parameters, these kind of methods usually have strict requirements in terms of the samples to variables ratio, and are more prone to overfitting. One way of overcoming these limitations, which has already been extensively used also for calibration problems, is to implement the non-linearity through the training of locally linear classifiers [1].

In this study, the possibility of extending the Locally Weighted-PLS approach originally proposed by Centner and Massart [2], to deal with non-linear classification problem was thoroughly investigated. In this context, the effect of applying different weighting schemes (uniform weighting or distance-based weighting) was tested both on spectral-like simulated data sets (Figure 1) and on real data, obtaining in all the cases significantly better results with respect to the linear global model.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Graphical representation of two of the three simulated data sets: non-linearly separable problems were simulated in the PC space as concentric spheres (left) and modified exclusive-OR (right) and then transformed to spectral-like data using suitable loadings.}
\end{figure}

\footnotetext[1]{Corresponding author. Email: P.le Aldo Moro 5, I-00185 Rome, Italy. \texttt{federico.marini@uniroma1.it}}
2. Results
Results obtained by using LW-PLSDA on all the five analyzed data set (three simulated and two real ones) were in all cases better than those obtained by global PLS-DA both in terms of model parsimony and in terms of classification accuracy. Moreover, the outcomes of the proposed methods were also compared with those of a simple local classifier (kNN) and of a globally nonlinear implementation of PLS-DA (kernel-PLSDA): while on the simulated data sets, similar results were obtained by using LW-PLSDA, kNN and kernel-PLSDA, on the two real data sets, LW-PLSDA significantly outperformed the other two methods in terms of classification accuracy. Moreover, with respect to kernel-PLSDA, the estimation of model parameters and the selection of metaparameters in LW-PLSDA are more straightforward and in general easier.

References

A new statistical insight on PLS through orthogonal polynomials

Mélanie B.∗, Fabrice G.† and Jean-Michel L.‡

∗ Institut de mathématiques-Université Paul Sabatier, Toulouse, France

Keywords: PLS, multicollinearity, constrained least square, orthogonal polynomials, prediction error.

Introduction

The PLS method [1], [2] was first proposed by Wold in 1985 to remove the problem of multicollinearity in a regression model. This method has been mostly investigated by practical experiments but some statistical properties still remain to be studied. This work aims at providing some controls on the rate of consistency of this method. Here we consider PLS as a constrained least square problem over polynomial subspaces. In particular we use the theory of orthogonal polynomial to provide a new formulation of the residuals in each eigenvectors directions. Then we use this formula to study PLS in the context of a high-dimensional multiple regression model. We provide bounds for the empirical risk and for the mean square error of prediction under the assumption of a low variance of the noise.

1 Link between PLS residuals and orthogonal polynomials

1.1 The statistical framework within which we study PLS

We consider the regression problem $Y = X\beta^* + \varepsilon$ with $\text{Var}(\varepsilon) = \sigma^2$. Here we are concerned with finding a good approximation $\hat{\beta}$ of the true parameter $\beta^*$ when only $n$ noisy observations $Y_i = x_i^T\beta^* + \varepsilon_i$ are available ($x_i \in \mathbb{R}^p$). We consider a high dimensional framework in the sense that $p \gg n$. We assume that the design matrix is fixed and of full rank $n$. We adopt an inverse problem point of view by considering PLS as a constrained least square problem over Krylov subspaces i.e $\hat{\beta}_k \in \arg\min_{\beta} \|Y - X\beta\|_2$ where $K_k = \{X^T Y, (X^T X)X^T Y, \ldots, (X^T X)^{k-1}X^TY\}$ and $k \leq n$.

1.2 A new expression for the PLS residuals

For $k < n$, we have $\hat{\beta}_k = \hat{P}_k(X^T X)X^TY$ where $\hat{P}_k$ is a polynomial of degree $k - 1$. Let $\hat{Q}_k(t) = 1 - t\hat{P}_k(t)$ the polynomial of degree less than $k$ and of constant term equals. We also have $\|Y - X\hat{\beta}_k\|_2^2 = \|\hat{Q}_k(X^T X)Y\|_2^2 = \min_{Q \in \mathcal{Q}_{k,1}} \|Q(X^T X)Y\|_2^2$ where $\mathcal{Q}_{k,1}$ is the set of all the polynomials of degree less than $k$ whose constant term equals one. Therefore $\hat{Q}_k$ is the solution of a minimization problem over polynomial subspaces. From this last result we deduce

Proposition: $(\hat{Q}_k)_{1 \leq k < n}$ is a sequence of orthogonal polynomials with respect to the measure $d\mu(\lambda) = \sum_{j=1}^n \lambda_j (u_j^T Y)^2 \delta_{\lambda_j}$ where $(\lambda_j, u_j)_{1 \leq j \leq n}$ are the eigenelements of $XX^T$ and $\lambda_1 \geq \ldots \geq \lambda_n > 0$.

∗Mélanie B. E-mail: melanie.blazere@math.univ-toulouse.fr.
†Fabrice G. E-mail: fabrice.gamboa@math.univ-toulouse.fr.
‡Jean-Michel L. E-mail: jean-michel.loubes@math.univ-toulouse.fr.
Then using the theory orthogonal polynomial we provide a new explicit formula for the residuals.

**Theorem:** Let $k < n$. We have for all $i = 1, \ldots, n$

\[
\hat{Q}_i(\lambda_i) = \sum_{(j_1, \ldots, j_k) \in I_k^*} \left[ \frac{\hat{p}_{j_1}^2 \lambda_{j_1}^2 \cdots \lambda_{j_k}^2 V(\lambda_{j_1}, \ldots, \lambda_{j_k})^2}{\sum_{(j_1, \ldots, j_k) \in I_k^*} \hat{p}_{j_1}^2 \lambda_{j_1}^2 \cdots \lambda_{j_k}^2 V(\lambda_{j_1}, \ldots, \lambda_{j_k})^2} \right] \prod_{t=1}^k \left(1 - \frac{\lambda_t}{\lambda_{j_t}}\right),
\]

where $I_k^* = \{n \geq j_1 > \ldots > j_k \geq 1\}$, $\hat{p}_{j_t} = u_{jt}^T Y$ and $V(\lambda_{j_1}, \ldots, \lambda_{j_k})$ is the Vandermonde determinant of $(\lambda_{j_1}, \ldots, \lambda_{j_k})$. The interest of such a formula and some of the properties of the residuals will be discussed in further details. We will show that we recover from this formula some of the well-established shrinkage properties of PLS [3], [4].

## 2 Upper bound for the empirical risk and prediction error

We further explore the statistical properties of PLS by investigating the quality and the accuracy of the PLS estimator through the study of the empirical risk and prediction error.

### 2.1 Mean Square Error

The Mean Square Error quantifies the fit of the model to the data set used.

**Proposition:** Let $C(X^T X) = \frac{\lambda_i}{\lambda_n}$. We have for $k < n$,

\[
\mathbb{E} \left[ \frac{1}{n} \| Y - X\hat{\beta}_k \|^2 \right] \leq \left( \frac{\sqrt{C(X^T X) - 1}}{\sqrt{C(X^T X) + 1}} \right)^{2k} \left[ \frac{1}{n} \| X\hat{\beta}^* \|^2 + \sigma^2 \right].
\]

### 2.2 Prediction error

Hereafter the $\epsilon_i$ are supposed to be i.i.d $\mathcal{N}(0, \sigma_i^2)$. We further assume (H.1): $\sigma_n^2 = O(\frac{1}{n})$ and (H.2): there exists a constant $L > 0$ such that $\min_{1 \leq i \leq n} \{p_i^2\} \geq L$, where $p_i = (X\hat{\beta}^*)^T u_i$.

**Theorem**

Let $k < n$ and assume that (H.1) and (H.2) holds. Then, with probability at least $1 - n^{1-C}$ where $C > 1$, we have

\[
\frac{1}{n} \| X\hat{\beta}^* - X\hat{\beta}_k \|^2 \leq \frac{1}{n} \left[ 2 \left( \frac{\sqrt{C(X^T X) - 1}}{\sqrt{C(X^T X) + 1}} \right)^{2k} + 4 \frac{\log(n)}{nL} \left( 1 + \left( \frac{\sqrt{C(X^T X) - 1}}{\sqrt{C(X^T X) + 1}} \right)^{2k} \right) \right] \| X\hat{\beta}^* \|^2
\]

\[
+ \frac{1}{n} \left[ 2kC \frac{\log n}{nL} \left( 1 + C \frac{\log n}{nL} \right)^2 \max_{k \leq j \leq n} \left( \prod_{t=1}^k \left| \frac{\lambda_t}{\lambda_{j_t}} - 1 \right| \right) \right] \| X\hat{\beta}^* \|^2,
\]

where $C(X^T X) = \frac{\lambda_i}{\lambda_n}$.

We will explain how and why the signal to noise ratio along the eigenvectors directions and the distribution of the eigenvalues play an important role in the accuracy of the estimation of the true parameter. We will further explore the impact of the eigenvalue distribution on the prediction error and specify this error for some particular distributions. At last we will provide some recommendations on the conditions where the PLS method is or is not suitable.

### References


A Comparison of variable selection methods in high-dimensional regression

S. Faria\textsuperscript{a}\textsuperscript{*}, S. Shahriari\textsuperscript{a}, and A. Manuela Gonçalves\textsuperscript{a}

\textsuperscript{a}DMA-Department of Mathematics and Applications; CMAT-Centre of Mathematics, University of Minho, Portugal

Keywords: High-dimensional data; Partial least squares regression; Principle component regression

Abstract

The problem of high-dimensional variable selection has captured the attention of many statisticians worldwide in the last decade. In this study, we describe a bootstrap based technique for selecting predictors in partial least squares regression (PLSR) and principle component regression (PCR) in high-dimensional data. A simulation study is carried out to compare the bootstrap approach with several variable selection approaches on PCR and PLSR.

1 Introduction

Motivated by interesting applications, there has been a dramatic growth in the development of statistical methodology in the analysis of high-dimensional data, particularly related to regression (model selection, estimation and prediction). Principle component regression (PCR) and partial least squares regression (PLSR) are well-known techniques in dealing with high-dimensional data whose dimension of unknown number of predictors parameters $p$ is much larger than sample size $n$ \cite{1}, \cite{2}.

Although PCR \cite{3} and PLSR \cite{4} are two popular techniques that can handle multicollinearity among predictors even when there are more predictors than observations, it is still necessary to eliminate unimportant predictors in high-dimensional data. Since PCR and PLSR use the linear combinations of all the predictors, the number of variables used is not always reduced and it is still necessary to eliminate unimportant predictors.

In this work, we conduct a simulation study in order to investigate the performance of several variable selection methods in high-dimensional modeling. We apply these methods that are based on sparse formulation, jack-knife (JK) and bootstrap (Boot) approaches on PCR and PLSR.

2 Simulation Study

A simulation study was performed in order to investigate the performance of variable selection algorithms in PCR and PLRS. The simulation is designed to evaluate the algorithms performance in terms of the predictive ability, dimension reduction and the accuracy of variable selection. We used the freeware R to develop the simulation program (R Development Core Team, 2011).

In our simulations we generated three models considering different scenarios (several highly correlated variables, uncorrelated variables and variables with partly large coefficients).
In order to examine the performance of the variable selection methods in terms of the prediction error ability, dimension reduction and the accuracy of variable selection, we report the root mean square error (RMSE), the average value of reduced predictors (RP), the average value of true positive rate (TPR), the average value of false positive rate (FPR).

3 Discussion

The analysis of the methods on the simulation data sets for all the models showed that using the variable selection methods on PCR and PLSR considerably reduced the RMSE value.

The variable selection methods on PLSR gave more reduction in the RMSE value in all the models. For the dimension reduction, the variable selection methods on PLSR in all the models preformed a considerable reduction in the number of variables. The variable selection methods on PCR in models with highly correlated variables reduced more predictors.

In all the models, Bootstrap approach on both PCR and PLSR could perform favorably as well as other considered approached in terms of the predictive ability.

For all the models, Bootstrap approach on both PCR and PLSR by performing a proper average reduction in the number of variables, could show a moderate average rate of truly selected variables while having the smallest (best) average rate of falsely selected variables among other considered variable selection methods.

Acknowledgements

This research was financed by FEDER Funds through "Programa Operacional Fac-tores de Competitividade - COMPETE" and by Portuguese Funds through FCT -"Fundação para a Ciência e a Tecnologia", within the Projects PEst-OE/MAT/UI0013/2014. The author Shirin Shahriari has a Ph.D. scholarship by FCT SFRH/BD/51164/2010.

References


PLS-Path model for Social Network Data

Giuseppe Giordano\textsuperscript{a,*}, Stefania Spina\textsuperscript{b} and Laura Trinchera\textsuperscript{c}

\textsuperscript{a}Dept. of Economics and Statistics, University of Salerno, Italy.
\textsuperscript{b}Dept. of Economics and Statistics, University of Napoli “Federico II”, Italy.
\textsuperscript{c}Dept. of Information Systems, Supply Chain and Decisions, NEOMA Business Schools, France.

Keywords: Network Autocorrelation Model, SEM, Social Influence, Social Network Analysis

Abstract

Network data can describe social contexts and a social context can influence the emerging of relationships \cite{7}. The measurement of attitudes, behaviors, human features and structural characteristic of a context are retained in classical statistical variables (attributes data). A number of statistical models have been developed for network structure analyses. The most part of them arise from geographical connections model developed by \cite{2} and assume that for a given actor in a network the value of an outcome is directly influenced by the values of the actors’ neighbors outcomes. The Network Effects Model (NEM) by \cite{5} introduces this idea in the framework of auto-regressive models. This class of models allows accounting for dependence among actors in a classical regression models, i.e. by assuming that all the variables are directly observed. However, many phenomena in social research cannot be directly measured and call for latent constructs.

This call for an extensive approach that will exploit the capability of the NEM to model the dependence structure among the units and the possibility to manage observed variables as well as latent constructs.

In Structural Equation Modeling (SEM) framework real complex phenomena can be studied taking into account causal relationships among a number of latent concepts (i.e. the Latent Variables - LV) each measured by several observed indicators defined as Manifest Variables - MV.

The inclusion of the relational data structure (adjacency matrix) in a SEM will offer the opportunity to model classical issues of Social Sciences, Economics and Marketing such as Social Influence and Homophily, for example. The advantage can be envisaged in the different specification of the path diagram. This model specification offers higher flexibility, it allows to consider i) separate or joint effect of intrinsic opinions of the social actors, ii) the extent to which they are influenced by their alters, and iii) how people with similar characteristics are more likely to form ties.

Two complimentary methods emerged in the field of SEM: the so-called covariance-based SEM (also referred as LISREL models) and the more recent component-based SEM.

Among the component-based approaches the PLS Path Modeling is the most widely used. The PLS Path Modeling (PLS-PM) has the advantage that it involves no assumptions about the population and the scale of measurement \cite{8}, so it works without distributional assumptions. This kind of modeling is known as soft-modelling \cite{10} in contrast with hard-modeling (i.e. maximum-likelihood estimation procedures) typical of covariance based SEM.

In order to analyse social networks dependencies, the Network Autocorrelation Models \cite{4} have been developed from the geographical connections models developed by \cite{2}. There are two types of network autocorrelation models: the network effects model and the network disturbances model. The network effects model assumes that the dependent variable, for a given individual, is a function of exogenous variables and the values of the same dependent variable observed on other individuals. The outcomes for actors are not statistically independent as

\textsuperscript{*}Corresponding author: E-mail: Dept. of Economics and Statistics, University of Salerno, Italy ggiordan@unisa.it
assumed by many regression models, leading to a complex correlation structure. These models use adjacency
data to model this correlation structure. Network effect enters through the construction of explanatory variables
in a regression context, assuming that it is given a $n \times n$ network matrix of interdependencies, that is an adjacency
matrix. Furthermore, attributes data may induce similarities among units and can be used as explanatory variables
in the regression model as well. In this model the OLS regression includes the network data as the adjacency
matrix $A$:

$$Y = \rho AY + X\beta + \varepsilon \sim N(0, \sigma^2 I)$$

(1)

where $\varepsilon$ denotes a vector of stochastic errors, under usual assumptions, $\rho$ is a scalar that measures the
magnitude of the network effect, and $\beta$ is a vector of regression coefficients ([11]). In particular, [6] defines
the network effects model where “social network dependence” is incorporated through the addition of a “lagged”
dependent variable $y$ on the right-hand side of the regression equation. In this type of model the actor outcome
can depend directly on the outcomes of its own alters. The vector $A_y$ contains, for each focal actor, the value of
the outcome sum for all its alters, as such, alters’ outcome contributes to $y$ in proportion to the influence on ego. Thus, $A_y$ is a network-lagged outcome.

Furthermore, social network dependence is analyzed in the structural equations modeling framework. This
alternative approach introduced by [3] let consider the presence of latent variables. This approach illustrates
social network dependence through the network lagged variables as latent variables in the structural model
while relationships between latent network lagged variables and observed attribute data are represented in the
measurement model ([9]).

We use the PLS-PM approach to define a component-based estimation procedure alternative to the classical
covariance-based approach. The derived Network Effect Structural Model is presented in terms of traditional
SEM specification and along with a feasible substantive interpretation in the scope of Social Science. Numerical
simulation seems to show significant results with more robust results of the proposed approach with respect to
traditional Network Effect Model.

References

In Environment and Planning A, pp. 1023–1030.


Geographical Analysis, 40 (2), pp. 152–166.


Networks, 34, pp. 32–46.


application of Partial Least Squares. In J. Kmenta and J. Ramsey (eds), Model Evaluation in Econometrics,
Raman Spectroscopy and Multivariate Calibration as a New Technique for Fragrances Flash Points Determination

Robson B. Godinho\textsuperscript{a,b,*}, Mauricio C. Santos\textsuperscript{b} and Ronei J. Poppi\textsuperscript{a}

\textsuperscript{a}Institute of Chemistry, University of Campinas, Campinas, Brazil
\textsuperscript{b}Givaudan do Brasil Ltda, São Paulo, Brazil

Keywords: Flash Point, Fragrance, Raman Spectroscopy, Multivariate Calibration

Introduction

Currently, the huge growth and availability of perfumes and perfumed products and the different flammability levels of fragrances has made the determination of the flash point indispensable after creation of new fragrance. This importance is based on the fact that flash point defines all of the care that must be taken in the appropriate handling, transport and storage of the product. As fragrances are homogeneous mixtures of different organic substances with varied chemical structures and with various levels of flammability because of the multiple molecular interactions [1], those products must conform to the current safety legislation for chemicals.

The flash point is defined as the lowest temperature at which the mixture of vapors from a liquid spontaneously becomes inflammable upon contact with air [2]. The flash points of pure chemical substances are widely reported in the literature; however, limited data is available for mixtures, because the experimental measurements are relatively expensive and time-consuming. Therefore, predictive theoretical methods have been studied to estimate the flash point of both pure compounds and mixtures [3].

This study aims to find an analytical method capable of efficiently correlating the Raman spectra with the flash point of fragrances using multivariate calibration based on partial least squares (PLS); this method could provide faster results, be inexpensive and nondestructive and could open the possibility of analysis during the creation of new fragrances.

1 Materials and Methods

The samples were obtained from small homogeneous portions of the finished products collected from 32 different manufacturing processes or the laboratory equivalents of 81 batches of mixtures produced by Givaudan Brazil. Of the 32 fragrances, 20 have more than one batch; i.e., there are 2 to 5 batches with the same composition; only one batch was produced for each of the remaining 11 samples.

The flash points of the oils were determined according to the ASTM D7094-04, Standard Test Method for Flash Point by Modified Continuously Closed Cup (MCCCFP) Tester, using Miniflash-FLA, Grabner Instruments. The Raman spectra of the samples were acquired on a RamanStation 400F spectrometer, Perkin Elmer, with a laser excitation source at 785 nm. All spectra were measured in the range of 200 - 3278 cm\(^{-1}\) with a spectral resolution of 2 cm\(^{-1}\).

The Raman spectra were acquired and processed with Perkin-Elmer Spectrum 6.3.5 software. Data processing and the development of the models were performed with Pirouette\textsuperscript{\textregistered} 3.11 software, Infometrix Inc.

\* Corresponding author. E-mail: Robson B. Godinho, Institute of Chemistry, University of Campinas, P.O.B. 6154, Campinas, SP, Brazil. robson.godinho@iqm.unicamp.br
2 Results and discussion

Fragrances are composed of a wide range of ingredients with large differences in chemical characteristics and always aim to delivering a pleasant olfactory sensation in the product. Because of safety restrictions and transportation, one of the main attributes evaluated for all fragrances created is the flash point.

Once the spectrum is associated with the molecular composition the Figure 1 illustrates how the fragrances have spectral differences. These differences are more significant in the second half of the spectrum, called the fingerprint region.

![Figure 1.](image)

As fragrances have different characteristics, a model was developed based on principal component analysis (PCA) to evaluate the variability of the data in the studied group. After use of the PCA model to investigate the homogeneity of the data and finding the two samples FrFlVa01 and FlWo01 with major differences from the set, a model based on partial least squares regression (PLS) was proposed. To build this model, 50 batches of 20 different fragrances and 713 variables were used. The variables were mean-centered and they were processed with the first derivative by the Savitzky-Golay algorithm. The model was built through the cross-validation procedure of "leave-one-out".

For more consistent assessment of all of the results, it was adopted the relative standard error of prediction (RSEP) and the relative standard error of cross validation (RSECV). The RSEP calculated for the 20 samples included in the model was 0.71 %. With the addition of the 9 fragrance samples not initially included in the model, there is an increase in the prediction error. Although the samples contain some substances in common, the ratios are quite different, and the samples may have other additional ingredients. This fact supports the observed results and impedes a good correlation of the data. Even so, in a general assessment of the samples, the RSEP for this new data set is 1.05 % and indicates that this method can be used successfully for an initial evaluation of the flash point of a fragrance.

3 Conclusions

Raman spectroscopy based on multivariate analysis provides a new method to determine the flash point of fragrances. The determination method is rapid, easy to operate, inexpensive and nondestructive. The high accuracy, especially for the samples considered by the model, indicates that an alternative method to check flash points during the development of fragrances may be achievable.

References


Development of Combinatorial Regression Method for Avoiding Chance Correlations.

Tomoko Hatta\textsuperscript{a}, Tamaki Takaya\textsuperscript{a}, Tatsuya Takagi\textsuperscript{b,c}, Norihito Kawashita\textsuperscript{b,c}, Kousuke Okamoto\textsuperscript{b}
\textsuperscript{a} Department of Pharmaceutical Sciences, Osaka University, 1-6 Yamadaoka, Suita, Osaka 565-0871 JAPAN
\textsuperscript{b} Graduate School of Pharmaceutical Sciences, Osaka University, 1-6 Yamadaoka, Suita, Osaka 565-0871 JAPAN
\textsuperscript{c} Genome Information Research Center, Institute for Microbial Diseases, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871 JAPAN

Keywords: chance correlation, principal component analysis, varimax rotation

Introduction

As is well known, PLS has been being applied to Quantitative Structure-Activity Relationships (QSAR) for Computer Aided Drug Design (CADD). However, the number of the descriptors, which can be available for QSAR, are so many that the problem called “Chance correlation” easily occur even if PLS is applied. In addition, it is not easy to select effective descriptors from many available ones. Thus we tried to develop a combinatorial PLS like regression method, Partial Combinatorial Least Squares (PCLS) method using Principal Component Analysis (PCA) and Cluster Analysis in order to exclude as many descriptors which show chance correlation as possible. In many cases, there must be some descriptors which cannot be excluded even if the descriptors are significant by chance. However, this novel method could exclude such descriptors and built better regression models in 8 out of 10 times.

1 Method

The algorithm of the novel method, PCLS, is as follows:
\begin{enumerate}
  \item PCA is carried out with all explanatory variables.
  \item Obtained principal components are transformed by varimax rotation method.
  \item PLS is carried out with rotated components.
  \item A Euclidean distance matrix is obtained from the PLS loadings.
  \item Cluster analysis is carried out using the distance matrix.
  \item Using the variables in not all the clusters but several ones obtained, Multiple Regression Analysis (MLR) is carried out.
\end{enumerate}

In this study, Ward method was adopted for cluster analyses. Ten artificial data sets, of which structure is shown in Table 1, were generated and the training data sets were analyzed along the abovementioned process. Since the varimax rotation method is not always the best one, other rotation method, such as promax rotation might give better results.

\* Corresponding author. E-mail: ttakagi@phs.osaka-u.ac.jp
Table 1  Structure of artificial data sets used in this study

<table>
<thead>
<tr>
<th>Training set (n=35)</th>
<th>$y = \sum_{p=1}^{7} \beta_{p}x_{pi} + \varepsilon_{i}$</th>
<th>response variable</th>
<th>used for generating response variable</th>
<th>random numbers</th>
<th>“chance correlation” variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test set (n=15)</td>
<td>$y = \sum_{p=1}^{7} \beta_{p}x_{pi} + \varepsilon_{i}$</td>
<td>response variable</td>
<td>should be significant for the response variable</td>
<td>random numbers</td>
<td>high correlation with $y$</td>
</tr>
<tr>
<td></td>
<td>$\sigma_{e} = (1/5)\sigma_{y}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2  Results

All the 10 data sets were analyzed using both ordinal PLS and the PCLS methods. PCLS showed better results than PLS in eight cases out of ten (Table 2). Considering that the PCLS $Q^{2}_{\text{ext}}$‘s are not so lower than PLS ones even in the other two cases, PCLS could exclude the variables’ effects which showed high correlations by chance.

Table 2  PLS and PCLS results of external validations using ten data sets

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
<th>VIII</th>
<th>IX</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS</td>
<td>0.629</td>
<td>0.596</td>
<td>0.737</td>
<td>0.382</td>
<td>0.213</td>
<td>0.461</td>
<td>0.306</td>
<td>0.292</td>
<td>0.401</td>
<td>0.680</td>
</tr>
<tr>
<td>PCLS</td>
<td>0.701</td>
<td>0.605</td>
<td>0.777</td>
<td>0.440</td>
<td>0.738</td>
<td>0.832</td>
<td>0.277</td>
<td>0.395</td>
<td>0.789</td>
<td>0.605</td>
</tr>
</tbody>
</table>

Italic numbers indicate the larger $Q^{2}_{\text{ext}}$ values.

We tried to know why the PCLS results of the two data sets, VII and X, gave lower $Q^{2}_{\text{ext}}$’s than the ordinal PLS using Multi-Dimensional Scaling (MDS) method. The MDS two dimensional plots of all the 35 variables of both training and test data sets showed that the coordinates of $X_{31}$-$X_{33}$ of the test data sets are scattered in the case of VII as well as X though the ones in the case of other eight data sets are close with each other. This indicates that it is impossible to exclude the variables which show high significances by chance in such cases when PCLS method is applied. It also might be impossible to exclude such variables even if any kinds of methods are applied to such data.

3  Conclusions

Although development of the PCLS method is still in progress, minimum requirements of the novel method is considered to be met. However, all the “chance correlation” variables cannot be excluded. For example, in the case of some data sets such as VII and X, it is considered to be impossible to distinguish between truly significant variables and “chance correlated” ones. Hereafter, we are going to apply this method to some real data sets such as QSAR. Now, this study is in progress.
A better method of predicting face changes after cosmetic surgery: the partial least squares regression

Tae-Joo Kang\textsuperscript{a}, Byoung-Su Kang\textsuperscript{b}, Shin-Jae Lee\textsuperscript{a}\textsuperscript{*} and Tae-Il Kim\textsuperscript{a}\textsuperscript{†}

\textsuperscript{a}Seoul National University School of Dentistry and Dental Research Institute, Seoul, Korea
\textsuperscript{b}Woori Plastic Surgery, Seoul, Korea

Keywords: Multivariate PLS regression; Cosmetic Surgery; 2-dimensional data application; Prediction accuracy.

Introduction

Accurate prediction of post-surgery face changes is an essential step in the treatment planning of cosmetic surgery. During cosmetic treatment in facial deformity patients, the face does not directly follow the surgical skeletal movements in the underlying bony structures. The algorithms of currently available commercial programs are all based upon the 1-to-1 correspondence ratio and/or the ordinary least squares methods, which is far from being accurate.

Predicting face changes after surgery requires a number of variables to consider. Human face landmarks in face photos and x-rays have a 2-dimensional entity. A landmark (or variable) has 2 measurements in the x-axis and the y-axis. A certain degree of vertical repositioning surgery induces horizontal relocation also and vice versa. Furthermore, the face response at a specific face landmark is highly dependent on its adjacent response, and its neighboring points are also dependent on each other. Applying the PLS method would be a solution for prediction and interpretation of this highly correlated 2-dimensional situation.

Therefore, the aim of the present study is to propose a better statistical method of predicting face changes after cosmetic surgery by applying the partial least squares regression.

In addition, we will also discuss a method to report error analyses for 2-dimensional data. Previously published error reports of 2-dimensional data sets have inappropriately applied 1-dimensional approaches, such as differences in distance or angular measurements. Our visualization method can be applied to 2-dimensional data sets. This method shows errors in both the x-axis and the y-axis simultaneously, which can also identify any between-group differences.

\textsuperscript{*}Corresponding author. E-mail: nonext@snu.ac.kr
\textsuperscript{†}Corresponding author. E-mail: periopf@snu.ac.kr

This research was supported by the Korean Government (NRF-2012-0007545) for SJ Lee and supported by SNUDH (04-2010-0041) for TI Kim.
Partial Least Squares for detection of diesel oil adulteration

B. Krakowska\textsuperscript{a}, J. Orzel\textsuperscript{a}, M. Daszykowski\textsuperscript{a}, I. Grabowski\textsuperscript{b}, G. Zaleszczyk\textsuperscript{b}, M. Sznajder\textsuperscript{b}

\textsuperscript{a}Institute of Chemistry, The University of Silesia, Katowice, Poland
\textsuperscript{b}Customs Chamber of Customs Laboratory in Biala Podlaska, Biala Podlaska, Poland

Keywords: Diesel oil, Adulteration, PLS-DA, UVE-PLS

Introduction

Huge consumption of petrol oil generates significant profits for petroleum industry. Petrol oil used for agricultural or heating purposes usually has a different market price than oil used for a regular transport. The differences in petrol prices are a result of imposed excise tax. The diesel oil designated to the agricultural and heating purposes has a rebated excise tax, whereas oil used for a regular transport has a full rate of the excise tax. In each EU country some chemical components are deliberately introduced into a rebated tax diesel oil in order to distinguish it from the more expensive one. According to the Polish law, to differentiate diesel oil, two substances are introduced into fuel: a fiscal marker (Solvent Yellow 124, SY124) and a red dye (Solvent Red 19, SR19 or Solvent Red 164, SR164) \cite{1}.

Considerable differences in excise tax encourage illegal process of removal fiscal components from fuel and its redistributing at a higher price. Removal of fiscal components is possible and can be carried out outside of a chemical laboratory. This illegal procedure leads to a substantial loss in government’s budget. Therefore, it is necessary to develop a fast and cost-efficient analytical method that will be useful in the detection of the discoloration \cite{2,3}. Due to a complex chemical composition of oil samples it is necessary to use technique able to characterize well its chemical content. Gas chromatography (GC) seems to be a straightforward choice since GC chromatogram contains a great amount of chemical information. With this respect it can be regarded as a chemical fingerprint (a pattern reflecting its chemical characteristics) \cite{4}. Analysis of such complex chromatographic data requires advanced chemometric techniques \cite{5}. To distinguish between original and illegally treated samples discriminant variant of partial least squares, PLS-DA was used \cite{6}. In addition, uninformative variable elimination partial least squares approach, UVE-PLS, has been applied to remove irrelevant variables \cite{7}. Both discriminant models were constructed for 2×31 GC-FID fingerprints of diesel oil samples registered before and after the discoloration procedure.

1. Chemometric analysis

Prior to modelling chromatographic fingerprints were preprocessed to remove baseline (penalized asymmetric least squares method (P-ALS) \cite{8} and correct peaks shifts (Correlation Optimized Warping (COW) \cite{9}). Then, GC-FID chromatograms were divided into model and test sets by means of the Kennard and Stone algorithm \cite{10}. Model set, containing 42 samples, and remaining 20 samples (the test set) were used to construct discriminant models and to assess their predictive abilities, respectively.

1.1 Partial Least Squares Discriminant Analysis

The aim of partial least squares discriminant analysis, PLS–DA, is to construct a discriminant model that distinguish samples from different groups. For a two-class discrimination problem, sample origin is indicated usually using a bipolar response variable, $y$ \cite{11}. In order to construct the PLS–DA model, the optimal number
of latent variables is selected to maximize the covariance between a set of latent variables and response variable y. In this work, the PLS-DA model was constructed in order to distinguish between original and discolored diesel oil samples. Constructed model, helped to distinguish correctly all samples from the model set and 95% of samples from the test set.

1.2 Uninformative Variable Elimination - Partial Least Squares

Only a part of the original variables contains relevant information explaining the differences between groups of samples. The UVE-PLS method is used in order to identify uninformative variables [8]. The experimental data matrix X is augmented with matrix N containing normally distributed variables of a low amplitude that do not influence the PLS model. Detection of uninformative variables is based on the stability of the regression coefficients defined as the ratio of the mean value of regression coefficients and their standard deviation. Uninformative variables (with lower absolute values of stability than the stability of artificial variables) are eliminated from the data and then the remaining (informative) variables are used for the model construction. With the UVE-PLS, usually a smaller number of latent factors is required to construct a model with comparable prediction properties to classic D-PLS. In our study, about 6% of variables were recognized as informative and relevant for the model construction. The discriminant model based on a smaller number of variables distinguished correctly all samples from model and test sets resulting in a better performance than the PLS-DA model constructed for the full set of original variables.

The satisfactory results obtained in this work encourage the practical use of discriminant analysis of chromatographic fingerprint data. The proposed methodology showed a high sensitivity and specificity. However, a larger number of samples should be further analyzed in order to introduce the methodology for an everyday use.

References

Exploring the impact of scaling techniques on classification accuracy of NMR metabolomic data using PLS-DA

J. Labuschagne
UNISA, Roodepoort, South-Africa

Keywords: PLS-DA, Scaling, 1H-NMR, Data pre-processing.

Introduction

Analysis of NMR data is highly sensitive to the method of scaling used. The classification accuracy of models as well as model significance is highly influenced by the chosen data pre-processing method [1]. There is a lack of consistency in articles that report on results in using different scaling techniques and vague guidelines as to which techniques are to be used for different analysis objectives. Prior studies have compared newly introduced scaling techniques to commonly used techniques such as Pareto and unit variance (auto-scaling) scaling, but cross comparisons with respect to implications on PLS-DA and OPLS-DA classification models are still being researched. The process flow for the analysis of 1H-NMR data is shown in Figure 1. This presentation will demonstrate the impact of these various scaling techniques on PLS-DA. This is an ongoing study which will include the tendency of over fitting, influence of sample size and comparisons between open source (R) and commercial software (Simca). Model validation tools used will include cross-validation, jack-knifing model parameters, permutation tests and bootstrapping.

The scaling methods used in this presentation can be seen in Table 1. Each description is cited different sources and illustrates the small amount of information available that can be used when deciding on the correct scaling technique.

![Figure 1: Process flow of 1H NMR Data Analysis](image)

### Table 1. Scaling methods used in the data-preprocessing step of analyzing NMR data.

<table>
<thead>
<tr>
<th>Scaling Method</th>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable stability scaling</td>
<td>$\bar{x}<em>{ij} = \frac{(x</em>{ij} - \bar{x}_i)}{s_i}$</td>
<td>&quot;Vast scaling can be considered as the sequential application of mean-centering and U.V scaling (auto scaling)&quot; [2]</td>
</tr>
<tr>
<td>Pareto Scaling</td>
<td>$\bar{x}<em>{ij} = \frac{x</em>{ij} - \bar{x}_i}{\sqrt{s_i}}$</td>
<td>&quot;Pareto scaling falls in between no scaling at all and auto-scaling and gives the variable a variance equal to its standard deviation instead of unit variance&quot; [3]</td>
</tr>
<tr>
<td>Auto Scaling</td>
<td>$\bar{x}<em>{ij} = \frac{x</em>{ij} - \bar{x}_i}{s_i}$</td>
<td>&quot;Inflation of the measurement errors&quot; [1]</td>
</tr>
<tr>
<td>Mean-Centering</td>
<td>$\bar{x}<em>{ij} = x</em>{ij} - \bar{x}_i$</td>
<td>&quot;When data is heteroscedastic, the effect of this pretreatment method is not always sufficient&quot; [1]</td>
</tr>
</tbody>
</table>
1 Methods

Experimental 1H-NMR data sourced from the University of South-Africa is used for the analysis. Three levels of spacing for cultivars are used for comparison. Analysis is done using the numa package (Metabolomics Univariate and Multivariate Analysis) in R, the different scaling methods (Auto, Pareto, Vast, Range; Mean-Centering) are applied. Principal component analysis followed by OPLS-DA & PLS-DA is then performed. For the purpose of this presentation, results will be shown with score plots of the PLS-DA models.

2 Results

A comparison of the score plots generated from the PLS-DA models using various scaling techniques is shown in Figure 2. The different grouping patterns clearly show the effect that scaling has on data analysis.

![Figure 2: Comparison of scaling methods illustrated by the score plots of PLS-DA models on experimental data (effect of three levels of spacing on identical cultivars).](image)

References


Hot PLS – a framework for Hierarchically Ordered Taxonomic classification by Partial Least Squares

Kristian H. Liland*,a,b, Achim Kohlerb,a, Volha Shapaval

a Nofima AS, Osloveien 1, N-1430 Ås, Norway
b Norwegian University of Life Sciences, Postboks 5003, N-1432 Ås, Norway

Keywords: Partial least squares, taxonomy, fixed hierarchy, local modelling, replicate measurements

Introduction

A novel framework for classification by partial least squares in a fixed hierarchy is presented. The hierarchical approach ensures flexible local modelling with varying complexity. It results in an intuitive classification path from the highest taxonomic levels down to species and beyond. Results are presented as phylogenetic trees with local diagnostic information to gain maximum information about the classification and help the researcher to focus on interesting phenomena. Information on sample replicates is included in the classification to increase performance and avoid misclassifications due to low quality measurements.

Training

Classification

Vote

Detection of samples coming from previously unobserved classes is enabled by estimating cut-off distances from the calibration data classes. To further increase flexibility and improve customization the canonical

* Corresponding author. Email: kristian.liland@nofima.no, Address: Nofima AS, Osloveien 1, N-1430 Ås, Norway
powered partial least squares algorithm [1] is used for modelling and classification together with linear
discriminant analysis. This opens up for additional sample response information and forced sharpening of
focus on important variables.

1 Results

1.1 Experimental data
A set of molds obtained from the mycological strain collection of the Norwegian Veterinary Institute (Oslo,
Norway) were previously measured and analyzed by Shapaval et al. [2]. The molds were grown in twelve
biological replicates, where six biological replicates type I (performed on a separate day) and two biological
replicates type II (performed for each biological replicate I), with two technical replicates each. The data were
previously analyzed using variable selection and artificial neural networks where the last biological replicate
of type I was used as a validation data set [2].

1.2 Classification results

![Phylogenetic tree with classification success and local confusion matrices indicated.](image)

Table 1. Success in classification on each level of taxonomy. The first line shows the overall error across all biological
replicates, while the last line corresponds roughly to the test set in Shapaval et al.

<table>
<thead>
<tr>
<th>Replicate</th>
<th>Division</th>
<th>Class</th>
<th>Genus</th>
<th>Sub-genus</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>All combined</td>
<td>100%</td>
<td>97.7%</td>
<td>95.8%</td>
<td>93.8%</td>
<td>86.4%</td>
</tr>
<tr>
<td>1st</td>
<td>100%</td>
<td>98.3%</td>
<td>96.5%</td>
<td>93.0%</td>
<td>82.5%</td>
</tr>
<tr>
<td>2nd</td>
<td>100%</td>
<td>98.3%</td>
<td>93.5%</td>
<td>91.7%</td>
<td>81.7%</td>
</tr>
<tr>
<td>3rd</td>
<td>100%</td>
<td>98.3%</td>
<td>96.6%</td>
<td>96.6%</td>
<td>93.2%</td>
</tr>
<tr>
<td>4th</td>
<td>100%</td>
<td>91.5%</td>
<td>89.8%</td>
<td>83.0%</td>
<td>76.2%</td>
</tr>
<tr>
<td>5th</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>98.3%</td>
</tr>
<tr>
<td>6th</td>
<td>100%</td>
<td>100%</td>
<td>98.3%</td>
<td>98.3%</td>
<td>86.4%</td>
</tr>
</tbody>
</table>

References

classification and regression problems” J. Chemom. 23, pp. 495-504, 2009.

A new bootstrap based stopping criterion in PLS components construction.

Frédéric Bertrand\textsuperscript{a,b}, Jérémie Magnanensi\textsuperscript{a,b,}\textsuperscript{*}, Myriam Maumy-Bertrand\textsuperscript{b} and Nicolas Meyer\textsuperscript{a}

\textsuperscript{a}Faculté de Médecine - Laboratoire de Biostatistique - EA3430 \\
\textsuperscript{b}Institut de Recherche Mathematique Avancée - LabEx IRMIA - UMR7501 \\
Université de Strasbourg, CNRS, Strasbourg, France

\textbf{Keywords:} Bootstrap, PLS, PLS-GLM, component

\subsection*{Introduction}

The extraction of the optimal number of PLS components is a real challenge. Indeed, this step serves to find out the real dimension of the link between the response $y$, which will be a $\mathbb{R}^{n \times 1}$ vector in our case, and the predictors matrix $X$. Considering $k$ as the optimal number of components, concluding in $k_1 < k$ components leads to a loss of information so that links between some predictors and $y$ will not be correctly modelled. On the other hand, concluding in $k_2 > k$ components means that some “noise” or useless information in $X$ will be used to explain $y$, so that it leads to an over-complex model which could fit the data well but will have a poor predictive ability. This complexity or sensibility of the model could be measured with the development of the degrees of freedom (DoF), which was done by N.Krämer and M.Sugiyama\cite{1}. They also adapted the AIC and BIC criterion with these DoF and applied them to the selection of the optimal number of PLS components.

An extension to GLM of the PLS regression, noted PLS-GLM, has been developed by Bastien \textit{et al}\cite{2}. In this case, fewer criteria can be used since DoF are not established yet and some criteria, like $Q^2$, are not suited. Our aim was to find out an “universal” criterion, in that it could work either in PLS and PLS-GLM cases. Furthermore, we wanted to obtain some criterion which could be used as a classical test in that an error risk level $\alpha$ could be fixed. So, since bootstrap techniques are used to empirically model some statistics distributions, it also became possible to create confidence intervals in order to test regression coefficients\cite{3}. In our case, we adapted the so-called bootstrapping pairs technique in order to test the significance of the successive PLS components and compare this criterion with the most used ones: the $Q^2$ obtained by leave-one-out cross-validation (CV), the $Q^2$ obtained by 5-fold CV and the adapted AIC and BIC criterion. This choice of 5-fold CV is due to results on the better $q$ to use in $q$-fold CV\cite{4}. We also compare, in a logistic case, our bootstrap approach to two others criteria, the 5-fold CV miss-classed value and a criterion developed by Bastien \textit{et al}\cite{2}, noted $p_{val}$, which define the non-significance of a new component $t_k$ as not any significant predictors within it.

To compare all these criteria, whether in PLS or PLS-logistic case, we used some important data simulations for different levels of noise added in $X$ and $y$, noted respectively sigma4 and sigma5 on Fig.1, either when $n > p$ and $n < p$, with $n$ the number of subjects and $p$ the number of predictors. The simulation algorithm is available on: \url{http://www-irma.u-strasbg.fr/~magnanensi/Algo_simul.pdf}. Our study show a better stability of our criterion and a globally better predictive accuracy compared to the others criteria, either in PLS or in PLS-logistic.

\section{The adapted bootstrap method.}

This method relies on the assumption that, once the components are build, they are considered as being independent of the response. Indeed, once this base is fixed, we want to test the significance of a new component $t_k$, not by

\textsuperscript{*}Corresponding author. E-mail: 7, Rue Rene Descartes 67084 Strasbourg Cedex, France. magnanensi@math.unistra.fr
simulating the real distribution of the coefficient linked to this component, which would be a positive one, but rather by approaching the conditional distribution of these coefficients given $T = (t_1, \ldots, t_k)$. First, let $j = 1$.

1. Compute the $j$ firsts components $t_j$.
2. Bootstrapping pair $(y, T)$, returning $R$ bootstrap samples noted $(y, T)^{b_1}, \ldots, (y, T)^{b_R}$.
3. For each couple $(y, T)^{b_i}, i = 1, \ldots, R$, do the (generalized) linear regression $y = \sum_{h=1}^{j} (c_{bh} t_h) + \hat{e}_j^{b_i}$.
4. Since $c_j > 0$, construct an unilateral $IC = [IC_1^j, +\infty)$ of level $\alpha = 0.95$ for $c_j$ with BCa technique.
5. While $IC_1^j > 0$, do $j = j+1$, and return to step 1. Else, the final extracted number of component is $k = j - 1$.

2 Results
Criteria predictive accuracies were compared. For that, results obtained in the $n < p$ case, where $n = 20$, were used, considering it as a training sample and 80 supplementary data were simulated as a test sample. Then, we measured their predictive performance by computing the normalized mean squares test error for each of the criterion and also by computing the number of predictive miss-classed values in the PLS-logistic case. Finally, these means were compared with Student tests (see Fig.1).

![Figure 1](image-url)

**Figure 1.** Upper: PLS case; Left: $BIC_{dof}$ better than BootYT (red), no significant difference (green), $BootYT$ better than $BIC_{dof}$ (blue). Right: $Q^2$ better than BootYT (red), no significant difference (green), BootYT better than $Q^2$ (blue). Under: PLS-logistic case; BootYT better (red), no significant difference (green).

References
Abstract

In text-book Analysis of Variance with fixed effects the cell sizes are usually fixed before the experiment from power or other considerations. There are situations however, in which the cell sizes are random variables and are observed after the experiment is done. This is common in survey-type studies in which the overall sample size is fixed but the units are classified after the sample according to sample characteristics. In a recent paper [1] considered the consequences of randomness in the two-way model with fixed effects. In general, while for some tests at least asymptotically the randomness makes no difference, in other situations in multi-way classification models the randomness cannot be ignored. One such situation occurs in testing main effects in the two-way model with possible interactions. Let \( N_{ij} \) be the number of observations in the \((i,j)\) cell, \( \pi_{ij} \) be the probability that a particular observation is in that cell and \( \mu_{ij} \) be the expected value of an observation in that cell. We assume that the \( \{N_{ij}\} \) have a joint multinomial distribution with parameters \( n \) and \( \{\pi_{ij}\} \). Then \( \bar{\mu}_i = \sum_j \pi_{ij} \mu_{ij} / \sum_j \pi_{ij} \) is the expected value of a randomly chosen observation in the \( i \)th row. There is no obvious test here for testing the hypothesis that the \( \bar{\mu}_i \) are equal. With the \( \{\pi_{ij}\} \) unknown, there is no obvious sum of squares and F-ratio computed by the widely available statistical packages for testing this hypothesis. In this article we consider new tests for this problem and examine the effect of randomness in ANOVA models.

Keywords: main effects, analysis of variance, unbalanced data, multinomial.

1 Problem Description

The Two-Way Model with fixed effects for unbalanced data has received considerable attention in the past 30 years. There has been much discussion of the definition of the effects, calculation of sums of squares and definitions of hypotheses. The model is the subject of many texts, prominent among those [2], [3], [4], [5], [6]. Several decompositions of sums of squares are presented in ANOVA tables, and testing the parameters of the model is routinely accomplished by subtracting sums of squares to isolate the contribution of a certain parameter. Searle’s [2] R-notation, for example \( R(\alpha/\mu/\beta) \) is usually involved in testing the contribution of including \( \alpha \) after \( \mu \) and \( \beta \) in the model. The R-notation implies orthogonalizations of the design matrix that lead to sums of squares that test a host of hypotheses that depend on the cell sizes. Many hypotheses have interpretation problems and several articles have been written on the hypotheses implied by using Searle’s R(. .) notation, e.g. see [7], [8]; for an extensive list see [9].

This paper considers the same two-way model under the assumption that the overall sample size \( n \) is fixed but the cell sizes are themselves random variables following the multinomial distribution with unknown parameters \( \pi_{ij} \). The setting has been introduced in [10], see also [11]. Here, we are concerned especially about the two-way model and in particular we focus on two issues: 1) the definition of (row) main effects; 2) Testing main effects in the presence of interaction, but for the case where the interaction is non-significant.

*Panagis Moschopoulos E-mail: pmoschopoulos@utep.edu
Applications of the new setting abide in the literature. For example, in survey studies the size-n sample from the overall population is divided into sub-populations (cells) according to sample characteristics, e.g. a random sample from a population is divided after the fact according to religion and socio-economic status, while the interest lies on some other variable, e.g. family size, the y-variable, for more examples see [12], [13]. Cochran [14] refers to this sampling scheme as poststratification. As it turns out, in the case of the two-way model the multinomial assumption results in a natural definition of main effects and leads to a sensible test for main effects in the presence of interaction. Such a test is used for testing the hypothesis of equality of row means defined in terms of parameters and this hypothesis is the same as the hypothesis of zero main effects. In this paper we develop the test statistic for this hypothesis and also use it to construct simultaneous confidence intervals for differences among the row means.

References


Linking plants metabolite profile and antitrypanosoma activity data with PLS analysis: An approach to discover new natural active compounds

Mauro S. Nogueira\textsuperscript{a*}, Danniela P. V. Faleiro\textsuperscript{b}, Fernando B. da Costa\textsuperscript{b}, Reto Brun\textsuperscript{c}, Marcel Kaiser\textsuperscript{c} and Thomas J. Schmidt\textsuperscript{a}

\textsuperscript{a}Institute of Pharmaceutical Biology and Phytochemistry, University of Muenster, Germany
\textsuperscript{b}School of Pharmaceutical Sciences of Ribeirão Preto, Ribeirão Preto, Brasil
\textsuperscript{c}Swiss Tropical and Public Health Institute, Basel, Switzerland

Keywords: plant extracts, metabolite profile, antitrypanosomal activity, PLS1, targeted isolation

Introduction

The isolation of natural products using classical phytochemistry followed by the evaluation of their biological activities has been for many years the method of choice to obtain bioactive compounds. However, these procedures are time consuming, laborious and often lead to redundant results. The advance of “hyphenated” techniques has mitigated redundancy in methods like bioguided isolation, since they are now aided by dereplication [1]. However, to access the bioactive compound(s) within complex mixtures like crude extracts without previous fractionation, especially from unknown plants, it still a great challenge. Recently, approaches in the natural products field have excelled the discovery of bioactive compounds, by combining metabolic profiling studies of plants, \textit{in vitro} pharmacological assays, and computer-aided techniques. In this work we propose an approach to rapidly access compounds with antitrypanosomal activity by correlating LC-MS metabolite profiles of several plant extracts and their \textit{in vitro} biological activity with PLS analysis, followed by targeted isolation.

1 Material and Methods

Dried leaves, flowers and/or aerial parts (250-500 mg) of 13 Asteraceae from the Brazilian “Cerrado” were extracted with solvents of different polarity. Sixty crude extracts of these Asteraceae were analyzed by UHPLC/ESI-qQTOF MS/MS, in positive ionization mode, to obtain detailed metabolite profiles with retention data and exact masses of the constituents present in each extract. Extracts and isolated compounds were tested \textit{in vitro} against \textit{Trypanosoma brucei rhodesiense} (etiologic agent of Human African Trypanosomiasis – HAT). Calibrated chromatographic data were processed prior to statistical multivariate analysis and aligned in both m/z and retention time directions. A data matrix with 58 samples and 2889 variables was generated in the software \textit{Profile Analysis} (Bruker Daltonik) and the dependent variables were mean-normalized. Correlative modeling for the dependent Y-variables (antitrypanosomal activity in percentage of growth inhibition – g.i.) and the set of independent X-variables (metabolite data: m/z and rt) was performed with PSL (orthogonalized PLSR algorithm for one Y-variable) in \textit{The Unscrambler} software v. 9.2 (CAMO). Calculation was based on the equations (1) and (2). The model was validated by means of full cross-validation, which leaves out one sample at a time, meanwhile the values for the remaining samples are predicted.

\begin{equation}
X = T \times P^T + E \tag{1}
\end{equation}

*Corresponding author: nogueira@uni-muenster.de. Institute of Pharmaceutical Biology and Phytochemistry, Muenster, Germany.
where $X(n, p)$ represents the chromatographic data matrix, vector $Y(n, 1)$ the values of percentage of g.i. for the biological activity, $T(n, n)$ is the score matrix, and $P^T(n, p)$ denotes the transposed loading matrix. $B(n, 1)$ is a loading vector, $E(n, p)$ and $F(n, 1)$ are the residuals. Compounds (loadings) exhibiting positive correlation with the activity were chosen as targets for isolation by conventional chromatographic techniques.

### 2 Results and Discussion

The PLS1 model described 83% of the total variance in the $Y$ data by four significant PCs. On the scores plot, highly active samples ($>80\%$ g.i. at 2 mg/mL) were distinguished from low and non-active samples (Fig. 1). Metabolite profiles of most of the highly active samples led to positive scores on both PC1 and PC2 directions. The presence of peaks assigned to variables positively correlated with antitrypanosomal activity was checked by re-inspecting the extract chromatograms and their mass spectra. Although full dereplication was not possible as various plants had not been previously studied chemically, many variables could be assigned to sesquiterpene lactones (STLs) of various skeletal types. Five of these STLs were isolated, two were identified as new substances, and showed high activity against *T. b. rhodesiense* ($IC_{50}$ values: $1=0.078 \mu g/mL$, $2=0.254 \mu g/mL$, $3=0.486 \mu g/mL$, $4=1.12 \mu g/mL$, and $5=0.605 \mu g/mL$). Our initial assumption that compounds (variables associated) relevant to the differentiation of active samples and simultaneously positively correlated with the activity on the PLS model represent constituents responsible or highly relevant for the activity of their extracts was indeed confirmed by the antitrypanosomal activity of the isolated compounds. PLSR was thus very suitable to expose the existence of causal relationship in our experimental data. This approach can help speeding up the discovery process of natural bioactive compounds and may be applied to find new active molecules against diseases lacking new medicines, e.g. HAT, whose treatments rely on old and toxic drugs.

![Figure 1. Scores and loadings plots for PC1 vs. PC2 ($R^2 = 0.89$ and $Q^2 = 0.54$). Samples (extracts) are colored according to their level of activity. Circles highlight variables (compound associated), and the plant extracts positively correlated with activity. *Variables assigned to unknown compounds.*](image)

### 3 Conclusion

Correlation of metabolite profile data of a large number of plant extracts and their *in vitro* biological activity with PLS analysis, combined with compound dereplication, allowed rapid access of the active compounds within complex mixtures directed at a targeted isolation. This approach might be a powerful tool to enhance the discovery of new bioactive compounds, especially for diseases lacking new medicines.

### References

A Review on Strategies to Factor Matrices and Parallelize Algorithms

Alba Martinez-Ruiza* and Cristina Montañola-Salesb

aIndustrial Engineering Department
Universidad Católica de la Santísima Concepción, Concepción, Chile
bStatistics and Operations Research Department
Universitat Politècnica de Catalunya - BarcelonaTech
Barcelona Supercomputing Center, Barcelona, Spain

Keywords: Factorization, parallelization, multi-block analysis, component-based methods, PLS methods

Analysis of large data sets is nowadays a challenging task for several scientific disciplines. Computer science scientific community is working in solving computational complications for allowing the analysis of large data sets such as data storing or parallel and distributed computing and processing. Two main approaches are largely developed. First, to distribute the data storage and analysis among parallel processors to then choose a best alternative, and second to combine results obtained from storing and analyzing data in a set of parallel processors. Both approaches are often mixed. A large data set is usually stored in a distributed file system such as Cloud, Grid or Cluster of computers. Pieces of data—for instance, sets of rows—are distributed in several computers. In very simple terms, some of the most important complications lie in creating algorithms—sometimes codes—to make possible to apply traditional statistical methods or procedures of data analysis to these distributed data. Thus, when we need to parallelize a method/operation, we are really needing an algorithm to apply this method/operation to data stored in several computers. This includes, for instance, to apply the operation on a part of a matrix, and then combine this result with the results obtained in other computers (where the operation has been applied in other portions of the same matrix). The amount of data in every computer will often depend on its storage capacity.

From the statistical point of view, we also face the challenge of investigating several unresolved issues. We can mention: (a) suitable factorization of matrices considering both running times in parallel computations and the effectiveness of matrix decomposition procedures (see for instance [1] and [2]); (b) investigating the performance of statistical methods when analyzing large volumes of data. Millions of rows and columns challenge the underlying assumptions of statistical methods of analysis (i.e. validation of linearity); (c) error/noise management; and (d) visualization of data for inspection and exploration.

Currently, there are already some tools for the algebra of matrices and statistical analysis on distributed data. See for instance the R-packages snow, parallel and Rmpi. But, so much work remains to be done. Multi-block data analysis, consisting of a set of well-established methods, is well positioned to meet all the challenges that the analysis of large data sets imposes, as the methods are based on the analysis of components.

*Corresponding author. I would like to thanks Universidad Católica de la Santísima Concepción for funding this research. E-mail: Industrial Engineering Department, Universidad Católica de la Santísima Concepción, Concepción, Chile. amartine@ucsc.cl.
In this communication, we will give a literature review on the strategies that are used to parallelize algorithms and how researchers deal with computational complications in the processing of large data sets. We also mention the advantages and disadvantages of different applied strategies of factorization. These strategies may be useful to extend the use of the multi-block and PLS methods in the frame of large data sets.

References


Correlation of gas chromatography-olfactometry and sensory descriptive analysis of oregano samples by PLSR.

Sirli Seisonen\textsuperscript{ab*}, Keili Kajava\textsuperscript{ab}, R. Kuldjärv\textsuperscript{ab}, K. Vene\textsuperscript{ab}

\textsuperscript{a} Competence Center of Food and Fermentation Technologies, Tallinn, Estonia
\textsuperscript{b} Tallinn University of Technology, Tallinn, Estonia

Keywords: Oregano, GC-O, sensory analysis, PLSR

Introduction

\textit{Origanum Vulgare} is an aromatic perennial of the \textit{Lamiaceae} family, which is native to Mediterranean and western Asia \cite{1}. It is widely used in culinary, especially for pasta sauces, pizza, roasted vegetables, salads, baked fish etc. The basic taste of oregano is slightly sharp, bitterish with a note of camphor and often lemony \cite{1}. Oregano is also used as a food preservative, because of its antioxidative and antimicrobial characteristics \cite{2}. The aim of the present study was to compare commercially available oregano samples with different origin and processing methods using instrumental analysis (GC-Olfactometry combined with GC/MS) and quantitative descriptive sensory analysis, to correlate key-odor compounds with sensory parameters and to see predictive performance of model based on key-odor compounds.

1 Materials and methods

Samples were acquired from different regions and processed by the producer. Descriptive sensory analysis was carried out in CCFFT sensory lab (ISO 8589:2007) with a trained panel of ten assessors. Aroma analysis was performed using gas chromatography-mass spectrometry (GC-MS) and gas chromatography-olfactometry (GC-O) equipped with DB-5ms (30m×0.25cm×1μm). Volatile compounds were extracted using automated headspace SPME (Car/PDMS/DVB fiber). Instrumental aroma analysis was carried out using posterior intensity method with five trained GC-O assessors. Modified frequency was used for GC-O data and all the data analysis were done with XLSTAT 2014.

2 Results

Overall 25 sensory attributes were used for sensory analysis and 40 odor-active compounds were detected with GC-O. Number of components to retain was determined by cross-validation. After fitting suitable number of components, sensory attributes having Cumulative $Q^2$ index less than 0.4 and GC-O variables with variable importance in the projection (VIP) less than 0.8, and though making a poor contribution to the model, were removed. Number of removed variables was 9 and 11 respectively. After eliminations model was refit with 5 components and $Q^2$ index for different attributes varied from 0.59 to 0.94. Fits of individual Y attributes (GC-O attributes) were mainly over 0.95 with just a few exceptions near 0.9.
Freshness odor and taste and cucumber odor have a strong positive correlation with hexanal (grassy), 3-hexanol (grassy), p-cymene (citrus) and methional (cooked potato). Menthol and jasmine odors have strong positive correlation with R-carvone (minty) and phenyacetaldehyde (honey). Hay-like odor and taste are positively correlated with α-terpineol (oily, anise), methyl eugenol (cinnamon, clove). Spiciness taste and odor and astringent taste have strong positive correlation with 1,3,8-p-menthatriene (sulphury, terpene). Herbal intensity odor have strong positive correlation with (E,Z)-alloocimene (cucumber).

Figure 1. Correlation map on the first 2 components, X and Y variables on the PLS dimensions.

The predictive performance of gained model for oregano sensory parameters is good. Model should be fitted to new oregano samples to see the predictive performance with totally new data to make more comprehensive conclusions.

References


PLS calibration model based on UV-visible signals to quantify and confirm the presence of formaldehyde after the migration testing from melamine kitchenware

Veronica Tobar\textsuperscript{a}, Luis Sarabia\textsuperscript{b*} and María Cruz Ortiz \textsuperscript{a}

\textsuperscript{a}Department of Chemistry, Faculty of Sciences, University of Burgos, Plaza Misael Bañuelos s/n 09001 Burgos, Spain
\textsuperscript{b}Department of Mathematics and Computation, Faculty of Sciences, University of Burgos, Plaza Misael Bañuelos s/n 09001 Burgos, Spain

**Keywords**: PLS, formaldehyde, melamine, migration test, false non-compliant

**Introduction**

This work is aimed to develop a fast and unexpensive screening method for the determination of the content of formaldehyde in samples of migration testing from melamine tableware. In the European legislation the specific migration limit (SML) for this analyte is set at 15 mg/kg. The implemented method is a variant of that proposed in the technical guidelines (EUR 24815 EN 2011) \textsuperscript{[1]} to determine formaldehyde by UV-visible spectroscopy. The developed method uses a soft calibration based on partial least squares (PLS) using the full spectrum in the range from 350 nm to 650 nm instead of a single wavelength to build the calibration model. The advantage of using a multivariate calibration as PLS is that, through the Q and T\textsuperscript{2} indices, it is possible to detect the presence of interferents whose absorbance is taking place in the same spectral range that of the formaldehyde thus; the applicability of the method is guaranteed. The method enables to detect a possible lack of specificity.

1 Results and discussion

The migration test was conducted with simulant B 3\% (w/v) acetic acid, as it has been demonstrated that this simulant represents the worst case for the migration of formaldehyde from melamine kitchenware, the solvent for the calibration standard was also 3\% w/v acetic acid in water. A multivariate calibration model with a PLS regression using the spectra matrix (11 x 151) was performed. By means of a cross-validation step it was concluded that two latent variables need, which explain 99.8\% of the variance of Y-block. The first two columns of Table 1 show the details of this PLS model.

**Table 1.** PLS model with two latent variables. Parameters of accuracy line: calculated concentration versus true concentration; Decision limit, (CC\textsubscript{α}) and capability of detection (CC\textsubscript{β}) for x=0 and x= 15 mg/kg (probabilities of false positive (or false non compliance) and false negative (or false compliance) fixed at 0.05).

<table>
<thead>
<tr>
<th>PLS calibration model</th>
<th>Accuracy line</th>
<th>CC\textsubscript{α} mg/kg</th>
<th>CC\textsubscript{β} mg/kg</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Matrix size</strong></td>
<td><strong>Explained Variance (%)</strong></td>
<td><strong>Range in mg/kg (# samples)</strong></td>
<td><strong>P</strong></td>
</tr>
<tr>
<td>2-PLS (11 x 151)</td>
<td>93.4 of X 99.8 of Y</td>
<td>1-12 (n=8)</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-30 (n=11)</td>
<td>0.999</td>
</tr>
</tbody>
</table>

\(\alpha\) : correlation coefficient; (*) Joint confidence interval contains 0 and 1 respectively (95\% of confidence)

* Corresponding author. E-mail: lsarabia@ubu.es
The accuracy (trueness and precision) of the method was ensured and its precision is given as the standard deviation of the regression of the the accuracy line (PLS calculated concentration versus true concentration) for the two calibration ranges considered (see Table 1), being equal to 0.43 and 26 mg/kg respectively. The compliance of the specific migration limit for the quantity of formaldehyde in a migration test is mandatory. So assessing the efficiency of the UV-visible-PLS procedure in terms of the decision limit (CCα) and detection capability (CCβ) evaluating the probabilities of false non-compliance (α) and false compliance (β) is needed. For the specific migration limit, set for this migrant in 15 mg/kg [2], it was possible to determine CCα and CCβ as 15.8 and 16.6 mg/kg respectively. For these estimations the probabilities α and β were fixed at 0.05. Figure 1 shows the two indices for every standard and sample: the residual sum of squares (Q) and the Hotelling distance (T²). All calibration standards and migration samples had values for Q and T² indices that were lower than the threshold limits (see Figure 1). Therefore, the model allowed the quantification of the content of formaldehyde in melamine tableware. The quantities found after migration testing from three glasses and three cups of melamine, were in the range from 1.13 to 3.16 mg/kg. All of them were above the decision limit of the method (0.595 mg/kg) but did not exceed the specific migration limit.

Figure 1. Q and T² indices for calibration standards (black circles) and migration samples from melamine tableware (in red triangles) for the PLS model built with two latent variables. Dashed lines are the threshold values for both indices.

2 Acknowledgments

The authors thank the financial support provided by projects of the Ministerio de Economía y Competitividad (CTQ2011-26022) and Junta de Castilla y León (BU108A11-2).

3 References

[1] EUR 24815 EN 2011, “Technical guidelines on testing the migration of primary aromatic amines from polyamide kitchenware and of formaldehyde from melamine kitchenware”

Fast and inexpensive method for the determination of bisphenol A in a migration test from polycarbonate tableware using fluorescence spectroscopy with multiway techniques

Miguel Villalba\textsuperscript{a}, Luis Sarabia\textsuperscript{b} and María Cruz Ortiz\textsuperscript{a,*}

\textsuperscript{a}Department of Chemistry, Faculty of Sciences, University of Burgos, Plaza Misael Bañuelos s/n 09001 Burgos, Spain
\textsuperscript{b}Department of Mathematics and Computation, Faculty of Sciences, University of Burgos, Plaza Misael Bañuelos s/n 09001 Burgos, Spain

Keywords: PLS, Bisphenol A, excitation-emission fluorescence, migration test, false non-compliance.

Introduction

Bisphenol A (BPA) is a chemical widely used in manufacturing polycarbonate and epoxy resin, materials that are commonly employed in the production of various types of food and beverage containers, including the lining of metal cans. BPA may be transferred to food and beverages by migration processes from these containers. Recent studies have found that human exposure to BPA results in adverse health effects because BPA is an endocrine disrupter. Several regulations ban the use of BPA in baby bottles \cite{1} and a specific migration limit (SML) has been established for other containers. In the European legislation this limit is set at 600 \(\mu\)g kg\(^{-1}\). In this work a fast and inexpensive procedure for determination of BPA is carried out. To implement it BPA is extracted into acetic acid in water 3\% w/v (Simulant B). The migration testing from a polycarbonate tableware and the standards are analyzed by molecular fluorescence spectroscopy joint to Partial Least Squares (2-PLS or 3-PLS) calibration models that are the most commonly used for multivariate signals. Nevertheless, the chemicals that appear in the migration process from polycarbonate tableware make necessary the use of Parallel Factor Analysis (PARAFAC) for the analysis. Besides, unequivocal identification of BPA is mandatory according to legal regulations so a three-way technique with the second order property, like PARAFAC, is required. Additionally, the capability detection is evaluated (probabilities \(\alpha=\beta=0.05\)).

1 Experimental

The excitation-emission fluorescence measurements were performed on a PerkinElmer LS 50B Luminescence Spectrometer. The emission spectra were recorded between 295 nm and 350 nm at excitation wavelengths between 253 nm and 289 nm. Solutions of BPA were prepared in 3\% w/v acetic acid in water. The migration test was carried out with the simulant B (3\% w/v acetic acid with water) suitable for hydrophilic food with pH < 4.5. After filling each container with the food simulant it was heating up to 85°C for 10 hours and was stored at room temperature until its analysis. Three kind of samples were used: i) Calibration standards in the range from 0 to 640 \(\mu\)g L\(^{-1}\) (first case) and from 0 to 1120 \(\mu\)g L\(^{-1}\) (second case); ii) Prediction samples (160, 500, 700 \(\mu\)g L\(^{-1}\)); iii) Six migration test samples (one of them without preconcentration, and the remaining five at different levels of preconcentration). PLS and PARAFAC models were performed with the PLS_Toolbox 6.0.1 (Eigenvector Research Inc.), the linear regressions with STATGRAPHICS (Statpoint Technologies, Inc.) and CCC\(\alpha\) and CCC\(\beta\) evaluating probabilities of false positive and false negative according to ref \cite{2}.

2 Results and discussion

Firstly a multivariate calibration model with a PLS regression using 24 emission spectra recorded at excitation \(\lambda_e = 281\) nm was performed. By means a cross-validation step (leave-one-out) it was concluded that three latent

* Corresponding author. E-mail: mcortiz@ubu.es
variables were need, which explained 99.93% of the that variance of Y-block (concentration). The model is unbiased because the slope and intercept (of the accuracy line) are significantly equal to 1 and 0 respectively (at 5% significance level). The first row of Table 1 shows the details of this calibration model. The results of the 3-PLS model built with excitation-emission matrices can be seen in the second row. The performance of both models was very similar. For the prediction samples, the mean of the absolute values of relative errors, MAE, were 5.28% and 4.27% respectively. However, migration samples exhibit Q and T² exceeding the threshold values for both models and, therefore, the models did not allow the quantification of BPA. This fact, together with the need of unequivocal identification of the BPA in samples of the migration test made necessary the use of a method with the second-order property. The PARAFAC model built with the tensor containing only samples prepared in solvent, gave a negative CORCONDIA index, i.e. the tensor is not trilinear. This lack of trilinearity is caused by the fluorescent signal of the solvent which is constant for all samples. Once the excitation-emission matrix of the solvent was subtracted from all matrices of calibration samples, a trilinear tensor was obtained, which has a single factor in the PARAFAC decomposition of the tensor containing only the calibration standards in solvent. However, when the samples from the migration testing were included in the tensor, two factors were required. In both cases, the CORCONDIA indices were 98%. The excitation and emission profiles obtained in the PARAFAC decomposition match with experimental spectra, being the cosines of the angle equal to 0.981 and 0.991 for excitation and emission respectively. Consequently, the bisphenol A was unequivocally identified in the migration test samples. With the loadings of the sample profile of the factor associated with BPA, two calibration models were built. The results are shown in the last two rows of Table 1. The prediction errors (MAE) obtained were 4.98% and 3.39% respectively. For samples after migration testing, and taking into account the preconcentration level of each one, the average concentration obtained was 40.86 μg L⁻¹ (standard deviation for n= 5 equal to 3.03 μg L⁻¹).

Table 1. Details of several calibration models and parameters of the accuracy line (calculated vs. true concentration). Decision limit (CCα) and capability of detection (CCβ) for probabilities of false positive and false negative fixed at 0.05.

<table>
<thead>
<tr>
<th>Range and number of calibration standards (μg L⁻¹)</th>
<th>Calibration model</th>
<th>Type / Data size</th>
<th>Explained Variance (%)</th>
<th>#</th>
<th>ρ</th>
<th>Residual standard deviation</th>
<th>intercept</th>
<th>slope</th>
<th>CCα/μg L⁻¹ r=1</th>
<th>CCβ/μg L⁻¹ r=1</th>
<th>CCβ/μg L⁻¹ r=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 1120</td>
<td>2-PLS</td>
<td>24 x 56</td>
<td>99.9 of X</td>
<td>3</td>
<td>0.996</td>
<td>10.66</td>
<td>0.331</td>
<td>0.999</td>
<td>18.91/12.77</td>
<td>37.85/23.66</td>
<td></td>
</tr>
<tr>
<td>0 – 1120</td>
<td>3-PLS</td>
<td>24x56 x10</td>
<td>97.3 of X</td>
<td>3</td>
<td>0.998</td>
<td>19.89</td>
<td>1.012</td>
<td>0.998</td>
<td>33.71/21.07</td>
<td>67.01/42.16</td>
<td></td>
</tr>
<tr>
<td>0 – 640</td>
<td>PARAFAC</td>
<td>13x51x10</td>
<td>99.2 of X</td>
<td>1</td>
<td>0.982</td>
<td>43.92</td>
<td>0.765</td>
<td>0.982</td>
<td>83.9/54.8</td>
<td>168.4/110.0</td>
<td></td>
</tr>
<tr>
<td>100-1120</td>
<td>PARAFAC</td>
<td>18x51x10</td>
<td>87.9 of X</td>
<td>2</td>
<td>0.993</td>
<td>44.60</td>
<td>0.009</td>
<td>1.001</td>
<td>83.7/55.6</td>
<td>167.7/111.4</td>
<td></td>
</tr>
</tbody>
</table>

#: number of latent variables in PLS or factors in PARAFAC; ρ:correlation coefficient. r: number of replicates

3 Acknowledgments

The authors thank the financial support provided by projects of the Ministerio de Economía y Competitividad (CTQ2011-26022) and Junta de Castilla y León (BU108A11-2).

4 References


An empirical comparison between PLS, Lasso, Elasticnet and other models for highly correlated data.

GenaneYouness\textsuperscript{a}\textsuperscript{*} and WassimNassar\textsuperscript{a}

\textsuperscript{a} ISSAE- CNAM, Beirut, Lebanon

Keywords: Multicollinearity, Ridge regression, PLS, Lasso, and Elasticnet.

Introduction

Seven regression techniques to solve the question of multicollinearity in multiple regression, such as: PCR, PLS regression, with Ridge, Lasso, Lars, Adaptive Lasso and Elasticnet, are compared. An application based on real data providing from the Lebanese National Center for Scientific Research (CNRS) on the area of fire in Lebanon forests in 2005, is made. The comparison of these techniques is discussed along with some of their advantages and disadvantages.

1 Methods of multicollinearity

When we have correlation between predictor variables, some methods have to be used for dealing with this problem: dimension reduction methods such as PCR and PLS regression [1], and shrinkage penalized methods such as Ridge, Lasso [2], LARS, Adaptive Lasso and Elasticnet[3]. Elasticnet combines the penalty terms of Lasso and Ridge.

2 Empirical data

The seven techniques are applied to 160 burned areas with five factors affecting the behavior of forest fires such as elevation, perimeter, mean slope gradient, mean vegetation density (NDVI), and mean evaporation. These parameters were calibrated using the map burned areas extracted from the visual interpretation of satellite images. The multicollinearity is shown by the high values of the variance inflation of mean slope (13.56) and mean evaporation (20.89).

To validate our models, the data are divided into two sets randomly selected. The first one is the training set, based on the two thirds part of the data, and used to estimate the model, and the second one is a test set, based on the third part of the data, used to test the good behavior of the model by calculating $R^2$, the coefficient of determination, and the mean square error $\sigma^2$.

2.1 Empirical Study

In order to compare the different methods, we kept all predictors; they are summarized in the following graph and table:

Figure 1. Repartition of the coefficient of determination and the mean square error for all methods

\textsuperscript{*} Corresponding author. E-mail: genane.youness@cnam.fr, Beirut, Lebanon, BP 11- 4661 Beirut, Lebanon.
From these results, it should be noted that the Elasticnet method is the best one according to the highest coefficient of determination $R^2$ (0.891) and the minimum of the means square error (0.121). Moreover, the second best $R^2$ (0.88) is obtained by the PLS regression.

Table 1. Summary table of estimated parameters of the different methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>PCR</th>
<th>PLS</th>
<th>Ridge</th>
<th>LARS</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>Elasticnet</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERIMETER</td>
<td>0.88</td>
<td>0.88098</td>
<td>0.86825</td>
<td>0.8809797</td>
<td>0.88097972</td>
<td>0.881156354</td>
<td>0.89253770</td>
</tr>
<tr>
<td>MEAN_ELEV</td>
<td>0.33</td>
<td>0.334051</td>
<td>0.27028</td>
<td>0.3340512</td>
<td>0.33405121</td>
<td>0.332155233</td>
<td>0.22963326</td>
</tr>
<tr>
<td>MEAN_SLOPE</td>
<td>0.048</td>
<td>0.047948</td>
<td>0.08716</td>
<td>0.0479485</td>
<td>0.04794850</td>
<td>0.047307248</td>
<td>0.04549356</td>
</tr>
<tr>
<td>MEAN_NDVI</td>
<td>-0.007</td>
<td>-0.00703</td>
<td>-0.03329</td>
<td>-0.0070262</td>
<td>-0.00702625</td>
<td>-0.00599746</td>
<td>-0.04415432</td>
</tr>
<tr>
<td>MEAN_EVAPO</td>
<td>0.23</td>
<td>0.228305</td>
<td>0.20763</td>
<td>0.2283054</td>
<td>0.2283053</td>
<td>0.226201495</td>
<td>0.12130083</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.37505</td>
<td>0.150727</td>
<td>0.3396437</td>
<td>0.1388883</td>
<td>0.142111</td>
<td>0.12383</td>
<td>0.1209646</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.8786</td>
<td>0.887</td>
<td>0.8786</td>
<td>0.854</td>
<td>0.8666</td>
<td>0.8785</td>
<td>0.8913236</td>
</tr>
</tbody>
</table>

Table 1 shows that the estimated mean square error across methods has small values except for Ridge and PCR methods. According to the coefficient of determination, it can be noted that they are no large differences between the methods, and the most efficient method seems to be Elasticnet.

2.2 Using 10-fold cross validation

In addition and to complete the validation process, we used the cross validation method "10-fold" on the data. The test error for each group and its averaging CV(f), which is the estimator of the test error by cross validation, are calculated. This method uses the smallest root mean square error RMSE to find the optimal model of the 10 rounds. We obtained the following results shown in the table 2:

Table 2. Summary table of by cross validation 10-fold.

<table>
<thead>
<tr>
<th>Method</th>
<th>PCR</th>
<th>PLS</th>
<th>Ridge</th>
<th>LARS</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>Elasticnet</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.519</td>
<td>0.394</td>
<td>0.388</td>
<td>0.392</td>
<td>0.383</td>
<td>0.384</td>
<td>0.377</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.764</td>
<td>0.824</td>
<td>0.839</td>
<td>0.852</td>
<td>0.827</td>
<td>0.831</td>
<td>0.833</td>
</tr>
<tr>
<td>CV(f)</td>
<td>0.344</td>
<td>0.399</td>
<td>0.151</td>
<td>0.307</td>
<td>0.292</td>
<td>0.152</td>
<td>0.143</td>
</tr>
</tbody>
</table>

Due to the highest $R^2$ and the lowest CV(f) values, the Elasticnet method can be considered as the best one.

References


Estimation of NBA Rookies Performances Using Partial Least Squares Regression Model

Andi Zheng and Zhihong Zou *
School of Economics and Management, Beihang University, Beijing, China

Keywords: Multiple regression models, Partial least square, NBA players

Introduction

For NBA managers, it is crucial to select promising players from NCAA league because it affects the team’s long-term development. To criteria to distinguish a potentially good player from others is based on dozens of statistical indexes which record a player’s performance, such as height, weight, age, assists, rebounds, blocks, steals, personal fouls, etc. And these indexes can measure the player’s overall ability, like how well he can cooperate with his fellow players, or whether a player has good defensive awareness. Analyzing these data can be really helpful for managers to select new members from NCAA league.

However, these indexes are various and high-dimensional, so it’s hard to evaluate how well they will play in future quickly. If we could find some relationship between the performances of players over his whole career, and their statistical indexes before or shortly after entering the league according to history data, it will be helpful to search for the future NBA stars.

Specifically, first we can utilize the career records of retired or nearly retiring players to build the model and analyze the relationship, also we can find the most correlated components to the performances of players by partial least squares (PLS). Considering the fact that there are some top-star NBA players with not-satisfying statistical indexes in NCAA, we will include the data of good players who ranked relatively low as a rookie. Then we will use several player’s records to verify whether these components will evaluate the player’s potential more appropriately. Finally, we are going to apply the model to estimate NBA rookies’ future career. We believe it will prove to be useful helping managers make decisions when selecting team members.

About PLS, it is a projection method that models the relationship between the response $Y$ and the predictors $X$, see [1, 2, 3]. Over the years it has found useful applications in many fields, such as chemistry, behavioral sciences, and environment, see [4, 5].

The remainder of the paper is organized as follows. Section 2 introduces the PLS regression method and the specific steps to calculate each component. The details of the case study are presented in Section 3.

1 Methodology

PLS regression model is of special interest because, unlike other multiple linear regression models, it can analyze data with strongly correlated predictor variables, and model response variables simultaneously. In other words, it is a supervised projection model with uncorrelated latent variables.

Let $(y_1, y_2, ..., y_n)$ be the observations of a single response variable $Y$, and let $(x_1, x_2, ..., x_p)$ be the predictor variables. After $n$ independent observations, we have the following dataset:

*Corresponding author. E-mail: zouzhihong@buaa.edu.cn.
Also let $F_0$ be the standardized variable of $Y$, and $E_0$ be the standardized matrix of $X$. The first step is to extract the first PLS component $t_1 = E_0 w_1$ through solving the following equation:

$$\max \langle E_0 w_1, F_0 \rangle \quad \text{s.t.} \quad w^T w_1 = 1$$

thus we have:

$$t_1 = E_0 w_1 = \frac{1}{\sqrt{\sum r^2(x_i, y)}} [r(x_1, y)E_{01} + \cdots + r(x_p, y)E_{0p}].$$

Next apply regression on $t_1$: $E_0 = t_1 p^T_1 + E_1, F_0 = t_1 r^T_1 + F_1$, where $p_1, r_1$ are the estimated regression parameters: $p_1 = \frac{E_0^T t_1}{|t_1|^2}, r_1 = \frac{E_0^T t_1}{|t_1|^2}$, and $E_1, F_1$ are the residuals. After replacing $E_0$ with $E_1$, and $F_0$ with $F_1$, we can repeat the first step to obtain the second component $t_2 = E_1 w_2$, and the regression model: $E_1 = t_2 p^T_2 + E_2, F_1 = t_2 r^T_2 + F_2$.

Repeat the process until we have the most significant $m$ PLS components. After that we can apply multivariate regression model using the retained independent PLS components to interpret the response variable:

$$F_0 = r_1 t_1 + r_2 t_2 + \cdots + r_m t_m + F_m$$

## 2 Case study

In this section, we are going to apply the PLS regression model into the NBA playground. We all know that the ultimate evaluation of a player is his average points per game, but there are dozens of statistical indicators recording his overall performance, such as height, weight, age, assists, rebounds, blocks, steals, personal fouls, etc. And for further speaking, the performance of the whole team, even the whole league, other than the personal statistics above, would play an important role to the player’s final points. All these data can be decisive when a manager is picking new players for the team.

However, these high-dimensional variables are not necessarily independent. To say specifically, if a player contributes more assistance to his fellow teammates, it would certainly affect his attempts and his field goals among many other affected indicators; also the pace of the whole team and the opponent team would significantly influence the player’s final score. Therefore it can be hard and confusing to accurately evaluate a player’s potential using all the statistical indexes. Based on this consideration, we decide to build a PLS regression model using data from players who already have long-term professional careers or are already tired, and use the model to analyze the performance of NBA rookies, resulting in an estimated future average points. This should prove to be helpful especially when selecting new members.

## References


