Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

Report ISSN 0946-8838

Big data clustering: Data preprocessing, variable selection, and dimension reduction

Hans-Joachim Mucha (Ed.)

submitted: January 31, 2017

Weierstrass Institute
Mohrenstr. 39
10 117 Berlin
E-Mail: hans-joachim.mucha@wias-berlin.de

No. 29
Berlin 2017

2010 *Mathematics Subject Classification.* 62-07, 62H30, 62H25, 62P10, 90-08.

Key words and phrases. Cluster analysis, classification, big data, variable selection, dimension reduction, data preprocessing.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Foreword and Acknowledgments

I am pleased to present the report of the talks given at the 38th annual meeting of the working group "Data Analysis and Numerical Classification" (AG DANK) of the German Classification Society at WIAS in autumn 2016. This book, also published online at the web site http://www.wias-berlin.de/publications/wias-publ/, is dedicated to Prof. Dr. Hans-Hermann Bock on the occasion of his 75th birthday. He is the founder and famous ambassador of the GfKl and, here, especially of the AG DANK - see the festschrift of Prof. Gunter Ritter in this volume. Bock was the first and long-term chairman of the AG DANK from 1979 until 2001.

The autumn meeting took place at the Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Berlin, from Friday, Nov. 18 till Saturday, Nov. 19, 2016. Already 20 and 8 years ago, WIAS had hosted the traditional annual meeting with special focus on clustering, classification, and multivariate graphics (Mucha and Bock, 1996, Mucha and Ritter, 2009). The present workshop continued the topic of the previous meeting held at KIT in Karlsruhe: *Recent Developments of Big Data Analysis and Data Science*. Concretely, in Berlin, the focus was mainly directed on statistical problems of (necessary) data preprocessing such as transformations, variable selection, and dimension reduction in clustering and classification. Here, the aim was bringing together leading statisticians and scientists working in the life sciences for discussing applications of classification and clustering to neural sciences, market research, genetics, archaeometry and the like. The program started with four invited lectures of distinguished scientists, namely Dr. Karsten Tabelow (WIAS Berlin), Prof. Willi Sauerbrei (University of Freiburg), Prof. Thorsten Dickhaus (University of Bremen), and Dr. Markus Weber (ZIB Berlin). Altogether 16 talks were presented. Among them, four discussion papers dealt with statistical analyses of the special data set issued in advance. The present publication does not cover all talks presented at the autumn meeting, but only a selection of contributions. However, all presentations of the workshop are available online as pdf files at the website http://www.wias-berlin.de/workshops/dank2016/. The meeting was attended by altogether 24 participants who contributed interesting discussions.

The working group AG DANK of the German Classification Society ("Gesellschaft für Klassifikation e.V. (GfKl) Data Science Society") deals with all statistical, mathematical, and computational aspects of data analysis and classification problems (clustering, discriminant analysis, supervised/unsupervised classification, pattern recognition, data mining) and with their

DOI 10.20347/WIAS.REPORT.29 Berlin 2017

applications to the life sciences, economy, engineering, archaeometry, and administration. Founded in 1977, the Gesellschaft für Klassifikation Data Science Society is a transdisciplinary scientific society that aims at promoting methods of classification and data analysis in theory and application. It celebrated its 40th Annual Conference at the Fourth Joint Statistical Meeting of the Deutsche Arbeitsgemeinschaft Statistik "Statistics under one Umbrella" in Göttingen, Germany, in March 2016.

There was also a "lite" anniversary at the autumn meeting: Dr. Christian Hennig's twenty years of active work in the AG DANK. I am proud to say that, to my knowledge, he started his scientific career in GfKl at the autumn meeting at WIAS in 1996 with his talk "Analyse des ausgesendeten Datensatzes" and his related paper (Hennig 1996). Twenty years later, he delivered a very difficult dataset for our competition. Thank you, Christian. Now he is a distinguished scientist and, in 2014, he became the secretary of the IFCS, the umbrella society of the national classification societies. His career should encourage young scientists in the field of mathematical statistics to join the AG DANK.

The editor would like to thank all who have contributed to this report. I'm especially grateful to Prof. Ritter for his additional contribution "Happy 75th birthday, Prof. Bock." The corresponding lecture was already presented during the previous autumn meeting at KIT in Karlsruhe. My special thanks go to the board members of WIAS for their active support; they sponsored the three book prizes for the competition and the printout of this volume. Special thanks go to Christine Schneider for her thorough preparation of the workshop (catering during the event, websites, and the like).

Hans-Joachim Mucha
Research Group Stochastic Algorithms and Nonparametric Statistics
Chair of AG-DANK

References

HENNIG, C. (1996): Analyse des ausgesendeten Datensatzes. In: Mucha, H.-J. and Bock, H.-H. (Eds.): Classification and multivariate graphics: Models, software and applications. Report no. 10, WIAS, Berlin, 93–96.

MUCHA, H.-J. and BOCK, H.-H. (Eds.) (1996): Classification and multivariate graphics: Models, software and applications. Report no. 10, WIAS, Berlin.

MUCHA, H.-J. and RITTER, G. (Eds.) (2009): Classification and clustering: Models, software and applications. Report no. 26, WIAS, Berlin, http://www.wias-berlin.de/report/26/wias_reports_26.pdf.

Contents

DOI 10.20347/WIAS.REPORT.29 Berlin 2017

Figure 1: Prof. Hans-Hermann Bock talking to Prof. Tadashi Imaizumi

Part I

Festschrift on the Occasion of Prof. Bock's 75th Birthday

1 Happy 75th Birthday, Prof. Bock[1]

Gunter Ritter
Faculty of Informatics and Mathematics
University of Passau, Germany
ritter@fim.uni-passau.de

Abstract

In 2015, Hans-Hermann Bock celebrated his 75th birthday. Two months later, the annual meeting of the working group AG DANK took place at the KIT at Karlsruhe/Germany. At this meeting, I contributed a talk on the occasion of Bock's birthday. The present paper is the written record of this talk.

[1] Talk given on the occasion of the 75th birthday of Prof. Dr. Dr. h.c. H.-H. Bock at the Autumn Meeting 2015 of the AG DANK at the KIT, Karlsruhe/Germany

DOI 10.20347/WIAS.REPORT.29 Berlin 2017

Introduction

Bock is a man of remarkable foresight. At an early age, he recognized what would be important 40 years later – a gift that only few people share with him. The choice of his doctoral thesis “Statistische Modelle für die einfache und doppelte Klassifikation von normalverteilten Beobachtungen” (Statistical models for simple and double classification of normally distributed observations) presages the increasing importance that data analysis and classification gain today. An article in the newspaper “Frankfurter Allgemeine” states that Big Data drives the fourth industrial revolution; see Figure 3. Indeed, data analysis helps to earn billions of dollars in today’s economy.

Bock is a highly motivated scientist as well as a gifted organizer. He founded the German Classification Society (GfKl). He has served as its first chairman. A later chairman once said “Bock is not a member of the GfKl, he is the GfKl.” He has played a significant role in the foundation of the International Federation of Classification Societies (IFCS), the umbrella organization of a large number of national classification societies. He served also as its first president and organized its first conference at the RWTH Aachen University, Germany. He founded the working group AG DANK and has served as its chairman for 22 years. It is not exaggerated to say that not all of these institutions would exist without his initiative.

A few years after his thesis, Bock [3] wrote his book “Automatische Klassifikation” which contains everything that was then known on cluster analysis. He has later been a driving force in the foundation of the Springer scientific series “Studies in Classification, Data Analysis, and Knowledge Organization”, where he is still a managing editor. Moreover, he is a founding editor of the renowned statistical journal “Advances in Data Analysis and Classification.” He was its first editor-in-chief and still serves as a managing editor. Moreover, he is on the editorial board of various scientific journals.

He has received numerous honors. He was awarded the title of Doctor honoris causa by the Cracow University of Economics. He is the first recipient of the IFCS Research Medal and a recipient of the “DAGStat Medaille für besondere Verdienste um die Statistik in Deutschland,” conferred by the Deutsche Arbeitsgemeinschaft Statistik. He is an Honorary Member and Honorary President of the German GfKl and an Honorary Member of the Belarussian Statistical Association.

Bock has dealt with many statistical fields, his knowledge is comprehensive. Astonishingly, Bock was able to gain insight into statistical contexts without doing much programming work, as it is usual today. His theoretical insight was sufficient to obtain valid results. His fields of work are

(a) k-means and similar algorithms;

(b) significance tests;

(c) symbolic data;

(d) two-mode clustering;

Figure 2: Prof. Hans-Hermann Bock discussing with colleagues. On his side, his wife.

(e) neural networks and SOMs;

(f) clustering of time series;

(g) explorative data analysis;

(h) dissimilarity matrices;

(i) multivariate scaling;

(j) fuzzy clustering.

Bock has cooperated with well-known scientists, for instance, F.A.T. de Carvalho, P. Brito, I. Van Mechelen, P. De Boeck, W.H.E. Day, E. Kubicka, G. Kubicka, F.R. McMorris, V. Schmitz are his coauthors. He maintains close links to international colleagues and he has spent numerous sabbaticals in France, Japan, Poland, and in the USA. Discussion with him is always a gain for everybody, whether it is about statistics or about everyday questions.

1.1 Dr. Bock and the first years of the AG DANK

In 1979, Dr. Bock founded within the German Classification Society (GfKl) a section dedicated to scientists mainly interested in the mathematical and probabilistic foundations of statistics, in particular, classification, and clustering. The list of attendants and the program of the first meeting on April 4, 1979 are shown in Fig. 4. The section was first called SIG-NK, renamed SEK-DANK in 1985, and received its current name AG DANK in 1990. Bock was

Big Data trifft Industrie 4.0

Big Data treibt die vierte industrielle Revolution unaufhaltsam voran. Doch die Verschmelzung dieser beiden Faktoren erfordert ein neues Denken. Ein Pilotprojekt.

VON HANNES SCHWADERER

Produzieren ist teuer. Unternehmen weltweit versuchen deshalb ihre Produktionskosten zu senken. Für viele Industrien hängt die Absicherung ihrer Wettbewerbsfähigkeit davon ab – ganz besonders in Mitteleuropa. Produktionsmanager überlegen hier ganz genau, wie sie ihre Produktion so effizient wie möglich gestalten können. Was, wenn es wirklich einen Weg gibt, mehrere Millionen Dollar bei der Produktion zu sparen? Was wie futuristisches Wunschdenken klingt, könnte schon sehr bald Realität werden.

Noch vor wenigen Jahren löste der Begriff „Industrie 4.0" eher fragende Gesichter aus. Mittlerweile ist sie aber aus den Diskussionen nicht mehr wegzudenken. „Big Data" und das „Internet der Dinge" sind es, die die vierte industrielle Revolution antreiben. Erste Unternehmen sind sich bereits sicher: Industrie 4.0 kann den Produktionsprozess massiv optimieren. Je mehr Daten man aufnehmen und analysieren kann, desto genauer wird auch die Vorstellung davon, wie die Faktoren innerhalb eines Produktionsprozesses miteinander in Beziehung stehen. Dadurch entstehen intelligentere und besser vernetzte Produktionsschritte. Industrie 4.0 liefert in diesem Fall genau die Daten, die für eine ausführliche Analyse notwendig sind. Anschließend beleuchten Big Data Analytics die Zusammenhänge der Daten in einem bisher nicht denkbaren Maßstab.

Predictive Maintenance spart Zeit

Der amerikanische Halbleiterhersteller Intel hat ein Pilotprojekt gestartet und genau diese zwei Komponenten in der eigenen Produktion vereint. Neben Datenanalysen des eigenen Produktionsnetzwerks hat das Unternehmen die Leistung der internen Ausstattung und Sensoren näher beleuchtet. Können Industrie 4.0 und Big Data Analytics wirklich dabei helfen, den Produktionsprozess zu optimieren? Die Ergebnisse sprechen eine deutliche Sprache: Sie können.

Die Analysen helfen, Wartungszeiten optimal zu planen und Ausfälle sicher vorherzusagen. Ingenieure können sich wesentlich besser auf mögliche Wartungen und Reparaturen vorbereiten und diese im besten Fall sogar verhindern. Die sogenannte „Predictive Maintenance" spart Zeit und Geld. Ein Beispiel dieser Technik ist der Einsatz von Bildanalysen im Produktionsprozess. Die Analysen scannen und kategorisieren Einheiten. Intakte Einheiten werden weiterverarbeitet, während unvollständige Einheiten zur Inspektion weitergeleitet werden. Der manuelle Prozess ist zeitaufwendig und dauert etwa acht Stunden. Bildanalysen können den Vorgang um das Zehnfache beschleunigen.

Herkömmliche Sicherheitssysteme reichen nicht aus

Das ist erst der Anfang: Auch Lieferzeiten werden durch intelligente Steuerung verkürzt und der Einsatz von Ressourcen optimiert. Insgesamt wird die Industrie 4.0 anders als die vorangegangenen Industrierevolutionen nicht nur die gesamte Produktivität, sondern auch die Flexibilität erhöhen. Was genau bedeutet ein flexibler Produktionsprozess? Das Pilotprojekt bei Intel zeigt auf, dass durch smarte Datenanalysen eine individuelle Massenproduktion möglich wird, die Unternehmen weltweit mehr Flexibilität bieten wird.

Die Verschmelzung von Industrie 4.0 und Big Data Analytics erfordert aber auch ein neues Denken. Beachtet man die Menge und Relevanz der Daten, so muss Sicherheit an erster Stelle stehen. Schon mit Beginn der vierten industriellen Revolution sollten Unternehmen erste Maßnahmen zum Schutz ergreifen. In Zukunft werden herkömmliche Sicherheitssysteme nicht mehr ausreichen, Unternehmen vor IT-Angriffen zu schützen. Viele Unternehmen befürchten, dass ihre Daten nicht mehr sicher sein werden. Doch auch hier gibt es Lösungen, die ein Maß an Sicherheit bieten, so dass sich auch große Unternehmen damit wohl fühlen. Wichtig ist hier, dass es einen Ausgangspunkt gibt, der sicher und unveränderbar ist. Diese „Root of Trust" ist idealerweise direkt in der Hardware verankert, denn Silizium kann im Nachhinein nicht mehr geändert werden.

Das Pilotprojekt hat vor allem eines gezeigt: Durch die Optimierung von Produktionsprozessen können wertvolle Kostenvorsprünge erreicht werden. In einer zunehmend globalisierten und digital vernetzten Welt kann die Zusammenarbeit von Industrie 4.0 und Big Data ein entscheidender Faktor sein, die Wettbewerbsfähigkeit speziell von europäischen Unternehmen zu sichern.

Hannes Schwaderer ist Geschäftsführer Intel GmbH & EMEA Energy Sector Director.

Figure 3: Article from the Frankfurter Allgemeine (October 2016)

elected the first chairman in 1979; he was reelected five times and served until 2001 when he thought it was time for a change. I succeeded him as chairman in 2001.

A list of some events in the time period 1979 – 2001 is shown in Table 1. The co-chairs are also given. The members of AG DANK meet regularly at different places all over Germany once a year. Cities and times until 2001 are displayed in Fig. 5.

Table 1: Some events in the first 23 years of AG DANK.

Date of election	Chair; Event	Co-chair	Name
4. 4. 1979	H.-H. Bock, Aachen		SIG-NK
16. 4. 1980		M. Schader, Karlsruhe	SIG-NK
24. 3. 1983	executive board: SIG-NK becomes SEK DA-NK		
7.12. 1984	H.-H. Bock, Aachen	M. Schader, Karlsruhe	SEK DA-NK
9. 2. 1985	executive board: SEK DA-NK becomes member of IFCS		
11. 4. 1989	H.-H. Bock, Aachen	J. Hansohm, Essen	
12. 3. 1990	general meeting: SEK DA-NK becomes AG DANK		
1. 4. 1992	H.-H. Bock, Aachen	K. Ambrosi, Hildesheim	AG DANK
9. 3. 1995	H.-H. Bock, Aachen	R. Ostermann, Siegen	AG DANK
4. 3. 1998	H.-H. Bock, Aachen	G. Ritter, Passau	AG DANK
15. 3. 2001	G. Ritter, Passau	Chr. Hennig, Hamburg	AG DANK

I would now like to say some words on the scientific achievements of Prof. Bock. I choose the fields of probabilistic cluster analysis, Subsections 1.2 and 1.3, and symbolic data, Subsection 1.4.

1.2 Decision-theoretic foundation of clustering algorithms

With the exception of a singular paper by Pearson [13] in 1894, the state of art in clustering before 1986 was hierarchic and partitional. Representatives of the hierarchical view of clustering were T. Sørensen [17], 1948, K. Florek et al. [8], 1951, and Joe H. Ward, Jr. [21], 1963. Ward [21] had detected on heuristic grounds his well-known sum-of-squares criterion

$$\sum_{j=1}^{m} \sum_{i \in C_j} \|x_i - \bar{x}_{C_j}\|^2. \tag{1}$$

The partitional point of view was studied by Robert L. Thorndike [20], 1953, H. Steinhaus [18], 1956, S.P. Lloyd [11], 1957, A.W.F. Edwards and L.L. Cavalli–Sforza [7], 1965,

GESELLSCHAFT FÜR KLASSIFIKATION
Spezielle Interessengruppe (SIG)
"NUMERISCHE KLASSIFIKATION"

Königstein, den 4.4.1979

Programm zur Sitzung der
SIG Numerische Klassifikation
Königstein, 4.4.1979, 14.00 - 17.30

Ⓐ Vorträge:

1) P. Bollmann, TU Berlin
Effizienzuntersuchungen für ein Single-Pass-Verfahren

2) P. Ihm, Universität Marburg
Distanzgeometrien in metrischen Räumen

3) M. Schader, Universität Augsburg
Ein Austauschverfahren zur Klassifikation qualitativer Daten

15 Minuten Pause

4) U. Schulze, Ges. f. Math. Datenverarbeitung, St. Augustin
Erfahrungen mit Bewertungen von Klassifikationen

5) H.-H. Bock, RWTH Aachen
Unscharfe Klassifikation

Ⓑ Meinungsbildung bezüglich weiterer Vorgehensweise

Figure 4: List of attendants and program of the first SIG-NK meeting in 1979.

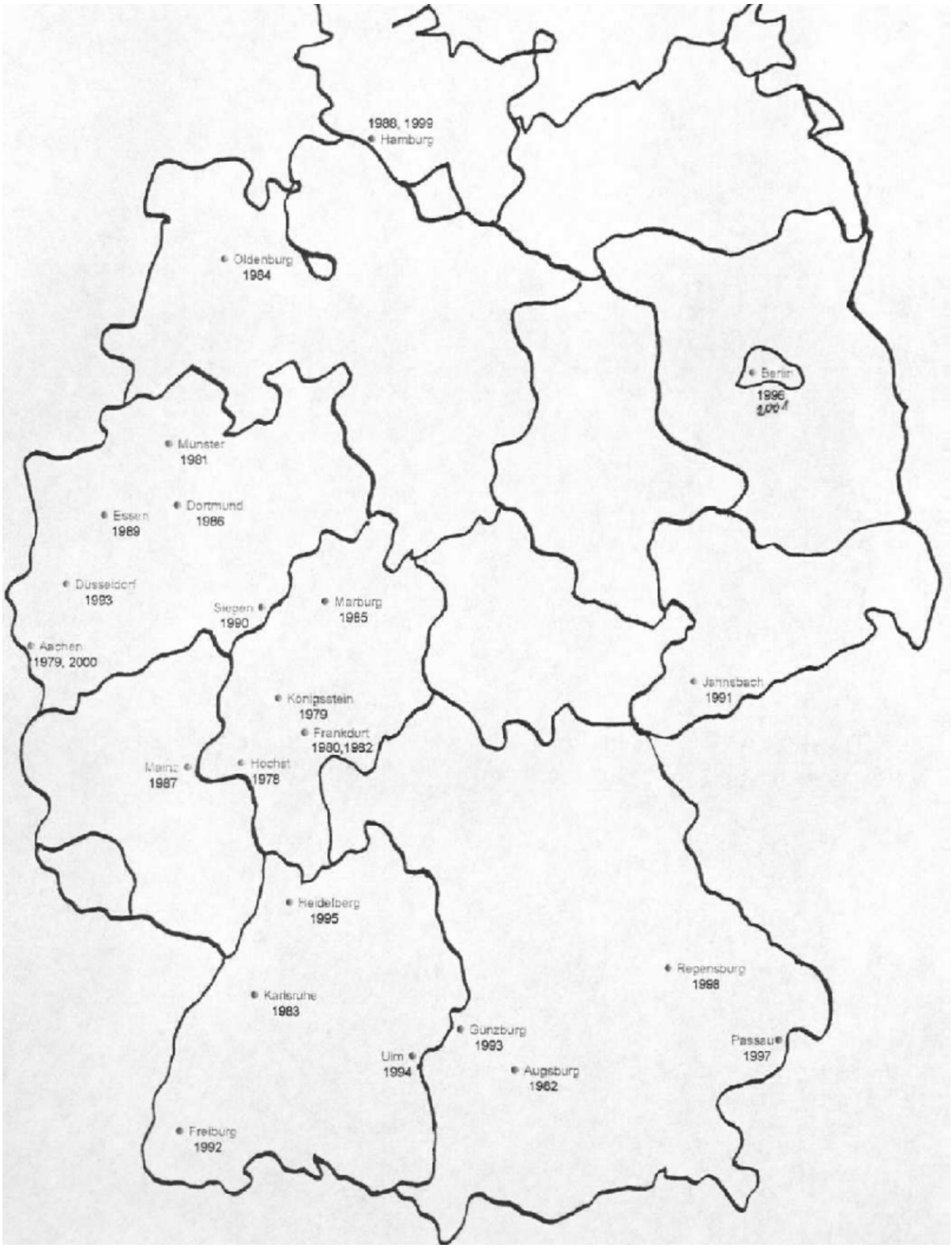


Figure 5: Meeting venues during the first 23 years.

John C. Gower [9], 1967, and J. MacQueen [12], 1967. These authors substantiated clustering algorithmically and geometrically. None of them had presented a probabilistic foundation of clustering.

As a mathematical statistician, Bock was not satisfied with this state of affairs. Recognizing clustering as a statistical estimation problem, he applied in his doctoral thesis [1] at the Mathematisch-Naturwissenschaftliche Fakultät of the University of Freiburg the methods of statistical decision theory. Excerpts of his thesis were later published in Metrika. The leading page of this paper is presented as Fig. 6.

At this point, I would like to go in more detail. Let us follow Bock in considering n p-dimensional observations $\mathbf{x} = x_1, ..., x_n$ from m unknown classes. Bock regards them as observations drawn from n independent, observable, normal random vectors $X_1, ..., X_n$ in $\mathbb{R}^p$ that are subdivided in $m < n$ (unknown) groups $A_1, ..., A_m$ with (unknown) mean values $a_j,\ j = 1, \ldots, m$ and a common spherical covariance matrix $\sigma^2 I_p$. This establishes a probabilistic model of the data. He seeks the quantities $\mathscr{A} = \{A_1, ..., A_m\}$, σ and $\mathbf{a} = (a_1, ..., a_m)$.

Statistical decision theory provides us with the following general solution: Let $L(\mathscr{A}, \mathbf{a}; \mathscr{B})$ be the loss incurred if $(\mathscr{A}, \mathbf{a})$ is true and $\mathscr{B}$ is estimated. An example is $L(\mathscr{A}, \mathbf{a}; \mathscr{B}) = 1 - \delta_{\mathscr{A};\mathscr{B}}$ Let λ be an a priori measure on the set of partitions $\mathscr{A}$ and their vectors of mean values $\mathbf{a} = (a_1, ..., a_m)$. An example is given by the product of the polynomial and the normal distribution, $\lambda(\mathscr{A}, \mathbf{a}) = q_{\mathscr{A}} f(\mathbf{a} \mid \mathscr{A}),\ f(a_j \mid \mathscr{A}) = N_{\delta_j(\mathscr{A}), \lambda_j(\mathscr{A}) \cdot \sigma^2 \cdot I_p},\ j = 1, \ldots, m$. The expected loss to be minimized is generally

$$\mathrm{E} \int L(\mathscr{A}, \mathbf{a}; \mathscr{B}(X))\, \mathrm{d}\lambda(\mathscr{A}, \mathbf{a})$$

In the special case above, the solution is provided by maximizing the a posteriori density

$$\begin{aligned} P(\mathscr{A} \mid \mathbf{x}) &\sim q_{\mathscr{A}} f(\mathbf{x} \mid \mathscr{A}) = q_{\mathscr{A}} \cdot \int_{\mathbf{a}} f(\mathbf{x} \mid \mathscr{A}, \mathbf{a})\, \mathrm{d}f(\mathbf{a} \mid \mathscr{A}) \\ &= \ldots \\ &= \frac{1}{(2\pi\sigma^2)^{np/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \|x_i\|^2 \right\} \\ &\quad \cdot \frac{q_{\mathscr{A}}}{\prod_{j=1}^{m} (\lambda_j n_j + 1)^{p/2}} \exp\left\{ \frac{1}{2\sigma^2} \sum_{j=1}^{m} \frac{n_j \lambda_j \cdot \|\bar{x}_{A_j}\|^2 + 2 \cdot \bar{x}_{A_j} \delta_j - \|\delta_j\|^2}{\lambda_j + 1/n_j} \right\}. \end{aligned}$$

Here, $f(\mathbf{x} \mid \mathscr{A}, \mathbf{a})$ is the probability density function of $\mathbf{x}$ under $\mathscr{A}$ and $\mathbf{a}$. This is a cluster criterion for general loss functions and prior probabilities. Bock discusses subsequently a number of special cases for the quantities L and λ.

Some of them make up the table in Fig 7. Note that Bock retrieves in row III Ward's sum-of-squares criterion. Thus, it emerges from probabilistic considerations, a fact that was unknown before Bock's thesis. His theory reappears in his book [3] with Vandenhoeck & Ruprecht in 1974; see Fig. 8. Apart from the classification model, his book contains everything that was known on clustering at this time: Separation and homogeneity measures, ultrametrics, hierarchical clustering, scaling.

Statistische Modelle und Bayessche Verfahren zur Bestimmung einer unbekannten Klassifikation normalverteilter zufälliger Vektoren [1])

Von H. H. Bock, Freiburg [2])

Zusammenfassung: Für N beobachtbare, unabhängige, normalverteilte Vektoren des R^p existiere eine unbekannte Einteilung in m disjunkte Klassen, innerhalb deren die Erwartungswerte gleich, aber unbekannt seien. Es werden Bayessche Verfahren zur Bestimmung dieser unbekannten Klassifikation entwickelt, wobei verschiedene Verlustfunktionen und a-priori-Verteilungen zugrunde gelegt werden. – Die Überlegungen können auf den Fall unbekannter Klassenanzahl m und auf den Fall der Klassifikation nach mehreren Gesichtspunkten erweitert werden.

Summary: Suppose that for N independent normally distributed vectors in R^p there exists an unknown classification into m disjoint classes such that in each class all vectors have the same (unknown) expectation. For several a-priori-distributions and loss functions Bayesian procedures are developped for the problem of identifying the unknown classification. – The methods can be extended to the case of an unknown number m of groups and to the case of many-way classifications.

1. Problemstellung

Es seien N p-dimensionale Beobachtungen $Y_1,\ldots,Y_N$ vorgegeben, welche in eine Anzahl m disjunkter, nichtleerer, möglichst homogener Gruppen (Klassen) einzuteilen sind. Zur statistischen Behandlung dieses Problems kann man die Beobachtungen als Realisierung zufälliger Vektoren ansehen [3]) und eine eventuelle Gruppierung durch gewisse Verteilungsannahmen charakterisieren. Beschränkt man sich hierbei auf Normalverteilungen und wählt als Gruppierungs-(Klassifikations-)merkmal den Erwartungswert, dann wird man bei bekannter Gruppenanzahl m das folgende Modell ansetzen:

a) $Y_1,\ldots,Y_N$ seien unabhängige, normalverteilte, beobachtbare zufällige Vektoren aus dem R^p mit gemeinsamer Kovarianzmatrix $\sigma^2 \cdot V$, die bis auf einen Faktor σ^2 bekannt sei; o.B.d.A. sei $V = I_p$ (p-reihige Einheitsmatrix).

---

[1]) Gekürzte Fassung einer von der Mathematisch-Naturwissenschaftlichen Fakultät der Universität Freiburg angenommenen Dissertation (1968).

[2]) Dr. Hans Hermann Bock, Institut für Mathematische Statistik der Universität Freiburg.

[3]) Große lateinische Buchstaben mögen – solange keine Verwechslung zu befürchten ist – sowohl für zufällige Vektoren als auch für deren Realisierungen stehen.

Figure 6: Leading page of the excerpt [2] of Bock's doctoral thesis in Metrika 18.

Tabelle der Bayesverfahren Φ^* bei einigen speziellen Annahmen bezüglich $q_{\mathfrak{A}}$, $\delta_i(\mathfrak{A})$, $\lambda_i(\mathfrak{A})$

	Parameter der a-priori-Verteilung			Zugehörige Bayesverfahren Φ^*
	$q_{\mathfrak{A}}$	$\delta_i(\mathfrak{A})$	$\lambda_i(\mathfrak{A})$	
I	$q_{\mathfrak{A}} \equiv q$	$\delta_i(\mathfrak{A}) \equiv \delta_i$ $\delta_i \in R^p$ fest, bekannt	$\lambda_i(\mathfrak{A}) = \dfrac{c}{n_i(\mathfrak{A})}$ $c > 0$ bekannt	$\sum_{i=1}^{m} n_i \lVert \bar{Y}_{A_i} - \bar{Y} \rVert^2 - \dfrac{1}{c+1} \sum_{i=1}^{m} n_i \lVert \bar{Y}_{A_i} - \delta_i \rVert^2 \to \max_{\mathfrak{A} \in \mathfrak{P}}$
II	$q_{\mathfrak{A}} \equiv q$	$\sum_{i=1}^{m} n_i \lVert \delta_i(\mathfrak{A}) \rVert^2 = \text{const.}$ $\delta_i(\mathfrak{A})$ bekannt		$\sum_{i=1}^{m} n_i \cdot \left\lVert \bar{Y}_{A_i} + \dfrac{1}{c} \delta_i(\mathfrak{A}) \right\rVert^2 \to \max_{\mathfrak{A} \in \mathfrak{P}}$
III	$q_{\mathfrak{A}} \equiv q$	$\delta_i(\mathfrak{A}) \equiv \delta_0$ $\delta_0 \in R^p$ fest, u. U. unbekannt	$\lambda_i(\mathfrak{A}) = \dfrac{c}{n_i(\mathfrak{A})}$ $c > 0$ u. U. unbekannt	$\sum_{i=1}^{m} n_i \cdot \lVert \bar{Y}_{A_i} - \bar{Y} \rVert^2 \to \max_{\mathfrak{A} \in \mathfrak{P}}$ äquivalent: $\sum_{i=1}^{m} \sum_{k \in A_i} \lVert Y_k - \bar{Y}_{A_i} \rVert^2 \to \min_{\mathfrak{A} \in \mathfrak{P}}$ „*ML*-Regel“
IV	$\delta_i(\mathfrak{A})$ beliebig, bekannt; i. a. mit $\mathfrak{A}$ variierend, speziell konstant bezüglich $\mathfrak{A}$	$\lambda_i(\mathfrak{A}) = c_i - \dfrac{1}{n_i(\mathfrak{A})}$ $c_i > 1$ bekannte Konstanten	$\sum_{i=1}^{m} n_i \lVert \bar{Y}_{A_i} - \bar{Y} \rVert^2 - \sum_{i=1}^{m} \dfrac{1}{c_i} \cdot \lVert \bar{Y}_{A_i} - \delta_i \rVert^2 \to \max_{\mathfrak{A} \in \mathfrak{P}}$	
V	$q_{\mathfrak{A}} = \beta \prod_{i=1}^{m} n_i(\mathfrak{A})^{p/2}$ β eine Normierungs-konstante	$\delta_i(\mathfrak{A}) \equiv \delta_0$ $\delta_0 \in R^p$ fest, u. U. unbekannt	$\lambda_i(\mathfrak{A}) = c - \dfrac{1}{n_i(\mathfrak{A})}$ $c > 1$ bekannte Konstante	$\sum_{i=1}^{m} \left(n_i - \dfrac{1}{c} \right) \cdot \lVert \bar{Y}_{A_i} \rVert^2 \to \max_{\mathfrak{A} \in \mathfrak{P}}$

124 H. H. Bock

Figure 7: Table of cluster criteria obtained for various prior assumptions (from Metrika 18).

Automatische Klassifikation

Theoretische und praktische Methoden
zur Gruppierung und Strukturierung von Daten
(Cluster-Analyse)

Von

Dr. rer. nat. Hans Hermann Bock
Technische Universität Hannover

Mit 54 Abbildungen

VANDENHOECK & RUPRECHT IN GÖTTINGEN

Studia Mathematica / Mathematische Lehrbücher

Herausgegeben von
Karl Peter Grotemeyer, Bielefeld
Dietrich Morgenstern, Hannover / Horst Tietz, Hannover

Band XXIV

ISBN 3-525-40130-2

Herstellung: Hubert & Co., Göttingen

DOI 10.20347/WIAS.REPORT.29

Figure 8: Title page of Bock's book with Vandehoeck and Ruprecht, 1974

Berlin 2017

Now, the cat was out of the bag. Bock had shown that statistical decision theory could be applied to compute cluster criteria. These didn't have to be guessed as done by earlier authors. By applying the theory to general normal models of clusters, the determinant criterion, too, could have been found. It was later derived from a slightly different statistical model by A.J. Scott and M.J. Symons [16], 1971, and by M.J. Symons [19], 1981, both in Biometrics. In both papers, Bock's work has unfortunately not been cited. The probabilistic theory of clustering has been shown to be extensible to elliptical and skewed distributions.

1.3 Cluster validation

Day [6] noted that local maxima of mixture likelihoods are not unique. The same is true for solutions of the classification model. In the k-means case, the solution with the least criterion is in most cases the desired one. An example with two "local" minima is shown in Fig. 10.

It is obvious that the solution with the smaller sum-of-squares criterion corresponds to the desired one. However, it even happens that the data set in hands is not clustered at all. Therefore, Bock [4] proposes four methods of cluster validation by significance tests. The first page of this publication is shown as Fig. 9. I report here on the last method in his paper. It uses the sum-of-squares criterion as a test statistic. As a main result, he determines its asymptotic distribution obtaining a maximum F-test. His theory uses the framework of homoscedastic mixtures with spherical components.

We are again given p-dimensional Euclidean data $x_1, \cdots, x_n$ and consider the test

H_0 : The data originates from a "unimodal" distribution
versus

H_1 : the data originates from $m > 1$ distinct (spherical) distributions.

Bock defines two statistics. The first one is Ward's criterion (1) or the trace of the "within" matrix,

$$g_n(\mathscr{C}) = \tfrac{1}{n}\mathrm{tr}W_n = \tfrac{1}{n}\sum_{j=1}^{m}\sum_{i\in C_j}\|x_i - \bar{x}_{C_j}\|^2.$$

The other one is the trace of the "between" matrix

$$b_n(\mathscr{C}) = \tfrac{1}{n}T_n - g_n(\mathscr{C}) = \tfrac{1}{n}\sum_{j=1}^{m}|C_j|\cdot\|\bar{x}_j - \bar{x}\|.$$

Here, T_n denotes the "total" matrix and the subscript n indicates the number of data points. Bock wishes a scale invariant test, since "often, in practice, only the type of the distribution of X_j can be specified (involving an unknown scale factor or standard deviation) ..." Therefore, he proposes as test statistic the quotient of between and within matrix,

$$k_n^* = \frac{b_n^*}{g_n^*} = \frac{T_n}{g_n^*} - 1, \tag{2}$$

Journal of Classification 2:77-108 (1985)

Journal of
Classification

On Some Significance Tests in Cluster Analysis

H. H. Bock

Technical University Aachen

Abstract: We investigate the properties of several significance tests for distinguishing between the hypothesis H of a "homogeneous" population and an alternative A involving "clustering" or "heterogeneity," with emphasis on the case of multidimensional observations $x_1, \ldots, x_n \in \mathbb{R}^p$. Four types of test statistics are considered: the (s-th) largest gap between observations, their mean distance (or similarity), the minimum within-cluster sum of squares resulting from a k-means algorithm, and the resulting maximum F statistic. The asymptotic distributions under H are given for $n \to \infty$ and the asymptotic power of the tests is derived for neighboring alternatives.

Keywords: Significance test; Homogeneity; Heterogeneity; Gap test; Minimum within-cluster sum of squares; Maximum F statistics; Asymptotic normal distribution.

1. Introduction

When a clustering algorithm is applied to a set of data, a classification of objects is obtained whether or not the data exhibit a true or "natural" grouping structure. This fact causes no problems if clustering is done for obtaining a practical (even if somewhat artificial) stratification of the given set of objects, e.g., for organizational purposes. However, if interest lies more in the recognition of an unknown clustering structure of the data (data analysis), an artificial clustering is not acceptable, and therefore the classes resulting from the algorithm must, in addition, be investigated for their relevance and their validity. Apart from descriptive, graphical, or exploratory methods, this task can be performed by using probabilistic models and suitable statistical significance tests.

Our approach is to consider a set of n p-dimensional observation points $x_1, \ldots, x_n$ in Euclidean space $\mathbb{R}^p$. We describe a series of statistical

Author's Address: Dr. H. H. Bock, Institut für Statistik und Wirtschaftsmathematik, Technical University Aachen, Wüllnerstr. 3, D-5100 Aachen, West Germany.

Figure 9: First page of Bock's 1985 publication in Journal of Classification.

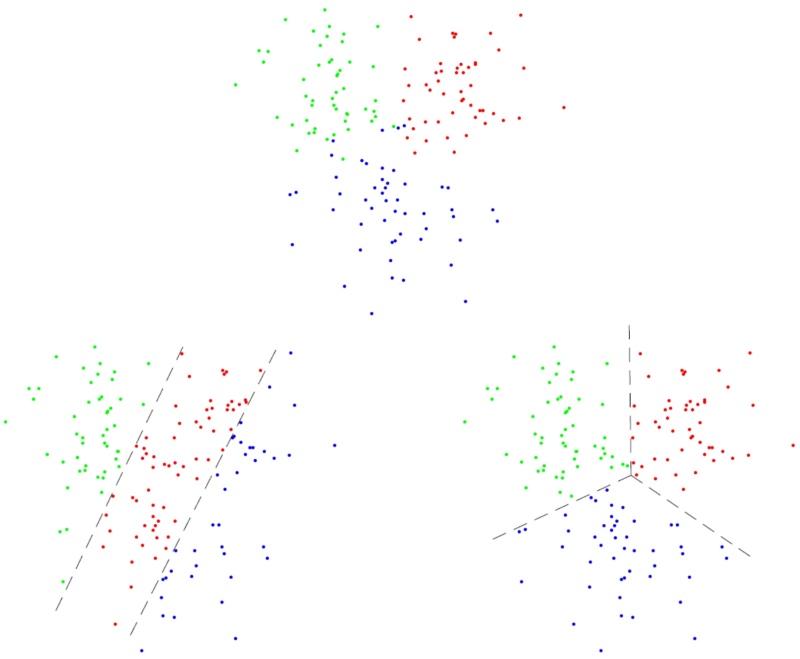


Figure 10: Top: clustered data set; bottom: two k-means solutions.

where g_n^* (b_n^*) denotes the minimum (maximum) value. Intuitively, k_n^* is large, if $\mathscr{C} = \{C_1, ..., C_m\}$ consists of m spherical clusters.

In order to use the test statistic (2), Bock needs the distribution of g_n^* under H_0, that is, the minimum of the sum-of-squares criterion over all partitions, and its asymptotic distribution. This is not an easy task. The minimum of Ward's criterion can also be represented in a different form, namely

$$\min_{\mathscr{C}} \sum_{j=1}^{m} \sum_{i \in C_j} \|x_i - \bar{x}_{C_j}\|^2 = \min_{\mathbf{z}, \mathscr{C}} \sum_{j=1}^{m} \sum_{i \in C_j} \|x_i - z_j\|^2 = \min_{\mathbf{z}} \sum_{i=1}^{n} \min_j \|x_i - z_j\|^2.$$

The right-hand side says that we must find m points $(z_1, \ldots, z_m) = \mathbf{z}$ such that $\sum_{i=1}^{n} \min_j \|x_i - z_j\|^2$ is minimum. This is called the best-location problem. The minimum partition $\mathscr{C}$ is the Voronoi decomposition to the minimum $\mathbf{z}$ and the minimum $\mathbf{z}$ consists of the mean vectors of the minimal clusters.

The best-location problem has also a continuous version. Instead of n points, he now considers their common distribution P_{X_1} and an arbitrary partition $\mathscr{B}$ of $\mathbb{R}^p$. For $\mathbf{z} = (z_1, \ldots, z_m)$, Bock [3], 1974, considers

$$g(\mathscr{B}, \mathbf{z}) = \sum_{1 \le j \le m} \int_{B_j} \|x - z_k\|^2 P_{X_1}(\mathrm{d}x).$$

and proves the following.

1.1 Theorem *(a) The following minima exist:*

$$g(\mathscr{B}) = \min_{\mathbf{z}} g(\mathscr{B}, \mathbf{z}), \quad g(\mathbf{z}) = \min_{\mathscr{B}} g(\mathscr{B}, \mathbf{z}), \quad g^* = \min_{\mathscr{B}, \mathbf{z}} g(\mathscr{B}, \mathbf{z}).$$

(b) We have

$$\min_{\mathbf{z}} g(\mathbf{z}) = \min_{\mathscr{B}} g(\mathscr{B}) = g^*.$$

Next, Bock takes recourse to the following theorem of consistency and asymptotic normality due to Pollard [14, 15] which had appeared just a few years earlier.

1.2 Theorem *Under regularity conditions on X_i, we have*

(a) Consistency: P-a.s.,

$$g_n^* = \tfrac{1}{n}\mathrm{tr}\, W_n^* = \min_{\mathbf{z}} \tfrac{1}{n} \sum_{i=1}^{n} \min_{1 \le j \le m} \|X_i - z_j\|^2$$
$$\underset{n\to\infty}{\longrightarrow} \min_{\mathbf{z}} \int \min_{1 \le j \le m} \|x - z_j\|^2 P_{X_1}(\mathrm{d}x) \ (= g^*).$$

(b) Consistency*: P-a.s., the minimal points $(z_1, ..., z_m)$ converge as $n \to \infty$ to*

$$\mathbf{z}^* = \operatorname*{argmin}_{\mathbf{z}} \int \min_{1 \le j \le m} \|x - z_j\|^2 P_{X_1}(\mathrm{d}x)$$

(c) Asymptotic normality of the m mean vectors:

$$\sqrt{n}(\mathbf{Z}_n - \mathbf{z}^*)) \underset{n\to\infty}{\longrightarrow} N_{0,\mathscr{I}} \quad \textit{(in distribution)}.$$

(d) Hence

$$\sqrt{n}\left(g_n^* - \min_{\mathbf{z}} \int \min_{1 \le j \le m} \|x - z_j\|^2 P_{X_1}(\mathrm{d}x)\right) \underset{n\to\infty}{\longrightarrow} N_{0,\tau^2} \quad \textit{(in distribution)},$$

where the variance τ^2 depends on a fourth moment of X_1.

As a consequence, Bock [4], 1985, obtains the following theorem on the distribution of the test statistic. It is the basis for his maximum F test. In passing, the theorem answers a conjecture of Hartigan's [10] about the asymptotic distribution of k_n^*. As above, g^* denotes the minimum best-location criterion and $k^* = \frac{\sigma^2}{g^*} - 1$.

1.3 Theorem *$\sqrt{n}(k_n^* - k^*)$ is asymptotically normally distributed with mean 0 and variance κ^2/g^*.*

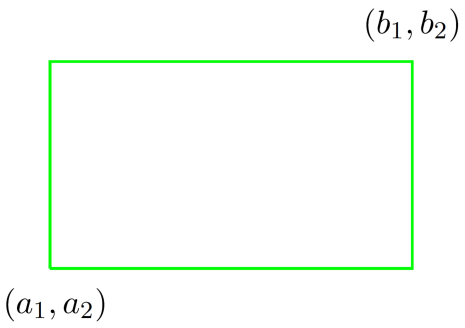


Figure 11: Hyperbox in R^2.

Therefore, the maximum F test for H_0 vs. H_1 reads:

(i) Specify a density that describes the hypothesis H_0 of homogeneity;

(ii) use it to determine the optimum partition $\mathscr{B}^* = (B_1^*, \ldots, B_m^*)$ of $\mathbb{R}^p$ in m clusters with its m centers and compute from here g^* und k^* and κ^2;

(iii) if $\sqrt{n}(k_n^* - k^*)$ lies in the rejection region of the asymptotic normal distribution $N_{0,\kappa^2/g^*}$, unimodality is rejected in favor of m clusters.

1.4 Symbolic data

Sometimes, one wishes to classify objects of higher complexity. Points in a symbolic data set are typically sets or more general objects. An example is

$$x_i = \left([20,25], \{\text{math}, \text{phys}, \text{chem}\}, \left(\begin{pmatrix}0.6\\ \text{DE}\end{pmatrix}, \begin{pmatrix}0.2\\ \text{NL}\end{pmatrix}, \begin{pmatrix}0.2\\ \text{JAP}\end{pmatrix}\right)\right)$$

This data item consists of the interval $[20,25]$, the three subjects “mathematics,” “physics,” and “chemistry,” and of three countries with a rating, each. In the multivariate context, a data item might consist of d intervals

$$x_i = \big([a_1, b_1], \cdots, [a_d, b_d]\big)$$

We associate with it the hyperbox (“box”, see Fig. 11)

$$R_i = [a,b] = \big([a_1, b_1] \times \cdots \times [a_p, b_p]\big)$$

We remain in the context of boxes. It is possible to define several dissimilarities of two such data items.

(a) The first one is the sum of the squared Euclidean distances of the "left, lower" and the "right, upper" corners:

$$\begin{aligned} d_v([a,b],[u,v]) &= \|(a_1,\cdots,a_p,b_1,\cdots,b_p)-(u_1,\cdots,u_p,v_1,\cdots,v_p)\|^2 \\ &= \|a-u\|^2+\|b-v\|^2 \end{aligned}$$

Note that the two objects $[a,b]$ and $[u,v]$ are equal, if and only if these corners coincide.

(b) Another dissimilarity is the Hausdorff distance between two boxes. Let K be such a box and let $d(a,K)=\min_{x\in K} d(a,x)$ be the distance of point a to K. The Hausdorff distance between the two boxes is

$$\begin{aligned} & d_H([a,b],[u,v]) \\ &= \text{Maximum of all distances between points from one set to the other set.} \\ &= \max_{c\in[a,b]} d\big(c,[u,v]\big) \vee \max_{w\in[u,v]} d\big(w,[a,b]\big) \end{aligned}$$

(c) Chavent and Lechevallier, 2002, propose a dissimilarity of Hausdorff's type:

$$d_1\big([a,b],[u,v]\big)=\sum_{k=1}^{p}|u_k-a_k|\vee|v_k-b_k|.$$

The Euclidean and the Hausdorff dissimilarities are illustrated in Fig. 12.

There is also a central box of a finite set of boxes. This is defined by

$$Z=\operatorname*{argmin}_{Q}\sum_{i=1}^{n} d(R_i,Q).$$

Here are two examples for $R_1=[a^{(1)},b^{(1)}],\ldots,R_n=[a^{(n)},b^{(n)}]$.

(a) In the Euclidean case, we have $Z=[\overline{a},\overline{b}]$.

(b) For Chavent and Lechevalier's dissimilarity, we find

$$Z=\prod_{k=1}^{p}[\mu_k-\lambda_k,\mu_k+\lambda_k],$$

where $\mu_k=\operatorname{med}\big(m_{1,k},\cdots,m_{n,k}\big)$, $\lambda_k=\operatorname{med}\big(\lambda_{1,k},\cdots,\lambda_{n,k}\big)$ and $m_{1,k}=(b_{1,k}+a_{1,k})/2$, $\lambda_{1,k}=(b_{1,k}-a_{1,k})/2$.

Figure 13 illustrates the centers for d_v and d_1.

Bock [5] is interested in establishing a probabilistic model for clustering symbolic data. To this end, he defines a probability density function on the set of all boxes. This, in turn, needs

Figure 12: Two dissimilarities between symbolic data items. Left: Euclidean, right: Hausdorff.

Figure 13: Central boxes: Left Euclidean, right Chavent and Lechevalier.

first a parametrization of a box. Let $M = (M_1, \ldots, M_p)$ be its center and let $L = (L_1, \ldots, L_p)$ be its midranges. Bock assumes that M is independent of L and that all midranges are independent. Moreover $M \sim N^{(p)}_{m,\sigma^2 I_p}$ spherically normal and $L_k \sim \Gamma(\alpha_k, \beta_k)$ has a Γ distribution.

Clustering of a (finite) subset of boxes proceeds iteratively and alternatingly by a k-means-type algorithm starting from an initial partition $\mathscr{C} = \{C_1, \ldots, C_m\}$.

(i) Compute the MLE's of the parameters of all clusters;

(ii) assign each box to a cluster according to the ML (or Bayesian) discriminant rule;

(iii) iterate (i) and (ii) until stationarity.

References

[1] Hans-Hermann Bock. *Statistische Modelle für die einfache und doppelte Klassifikation von normalverteilten Beobachtungen*. PhD thesis, University of Freiburg, Germany, 1968.

[2] Hans-Hermann Bock. Statistische Modelle und Bayessche Verfahren zur Bestimmung einer unbekannten Klassifikation normalverteilter zufälliger Vektoren. *Metrika*, 18:120–132, 1972.

[3] Hans-Hermann Bock. *Automatische Klassifikation*. Vandenhoeck & Ruprecht, Göttingen, 1974. In German.

[4] Hans-Hermann Bock. On some significance tests in cluster analysis. *J. Classification*, 2:77–108, 1985.

[5] Hans-Hermann Bock. Analyzing symbolic data. In Okada et al., editor, *Cooperation and Classification in Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 3–12, Heidelberg, 2009. Springer. Proceedings of the German-Japanese Workshops in Tokyo and Berlin.

[6] N.E. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56:463–474, 1969.

[7] A.W. Edwards and L.L. Cavalli-Sforza. A method of cluster analysis. *Biometrics*, pages 362–375, 1965.

[8] K. Florek, J. Lukaszewicz, J. Perkal, H. Steinhaus, and Zubrzycki S. Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum*, 2:282–285, 1951.

[9] John C. Gower. A comparison of some methods of cluster analysis. *Biometrics*, pages 623–628, 1967.

[10] John A. Hartigan. Distribution problems in clustering. In J. van Ryzin, editor, *Classification and Clustering*, pages 45–72. Academic Press, New York, 1977.

[11] Stuart P. Lloyd. Least squares quantization in PCM. Bell Labs memorandum, 1957.

[12] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L.M. LeCam and J. Neyman, editors, *Proc. 5th Berkeley Symp. Math. Statist. Probab. 1965/66*, volume I, pages 281–297, Berkeley, 1967. Univ. of California Press.

[13] Karl Pearson. Contributions to the theory of mathematical evolution. *Phil. Trans. Royal Soc. London, Series A*, 185:71–110, 1894.

[14] David Pollard. Strong consistency of k-means clustering. *Ann. Statist.*, 9:135–140, 1981.

[15] David Pollard. A central limit theorem for k-means clustering. *Ann. Statist.*, 10:919–926, 1982.

[16] A.J. Scott and M.J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27:387–397, 1971.

[17] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skrifter*, 5:1–34, 1948.

[18] H. Steinhaus. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci.*, 4:801–804, 1956.

[19] M.J. Symons. Clustering criteria and multivariate normal mixtures. *Biometrics*, 37:35–43, 1981.

[20] R. L. Thorndike. Who belongs in the family? *Psychometrika*, 18:267–276, 1953.

[21] Joe H. Ward, Jr. Hierarchical grouping to optimize an objective function. *J. Amer. Stat. Assoc.*, 58:236–244, 1963.

Part II

Papers of Talks

2 Probabilistic Variable Selection in Cluster Analysis

Gunter Ritter
Faculty of Informatics and Mathematics
University of Passau, Germany
ritter@fim.uni-passau.de

Introduction

Data sets in cluster analysis may cause problems for several reasons. There may be missing values, there may be outliers, or the data set may be big. Large size, in turn, can have two causes, many observations or many variables (or both). Whereas too many observations can be easily dealt with by random sampling, high dimension causes severer problems. This is the subject matter of this paper.

Sensitive clustering methods need in each cluster substantially more observations in each cluster than there are variables. Cluster characteristics cannot be well determined, otherwise. An extreme case violating this requirement is the so-called $p \gg n$ case, which means that the number of variables is much larger than data set size. Typical examples of high-dimensional data with a low number of observations are gene expression data of patients. The differentially expressed genes in a patient's tissue are the interesting ones; they make up only a small subset of all genes. The complementary probes on the microarray act like sensors; those related to the differentially expressed genes provide us with the variables relevant for clustering. In view of the present clustering task, the remaining probes yield just noise.

Fowlkes et al. [1] studied the effect of additional noise variables. Let us follow their ideas. Fig. 14 shows a data set clearly separated in five clusters. Any reasonable clustering algorithm should be able to detect the clusters given their number. If three noise variables are added, that is, the data points are shifted in three dimensions perpendicular to the drawing plane, then Scott and Symons' [8] determinant criterion

$$\tfrac{1}{2}\sum_{j=1}^{g} n_j \log\det S_j \tag{3}$$

yields a solution with 28 errors. (In Eq. (3), n_j is the size of cluster C_j, $S_j = \frac{1}{n_j}\sum_{i\in C_j}(x_i - \bar{x}_j)(x_i-\bar{x}_j)^\top$ is its scatter matrix, and $\bar{x}_j$ its mean vector.) If clusters are even better separated

DOI 10.20347/WIAS.REPORT.29 Berlin 2017

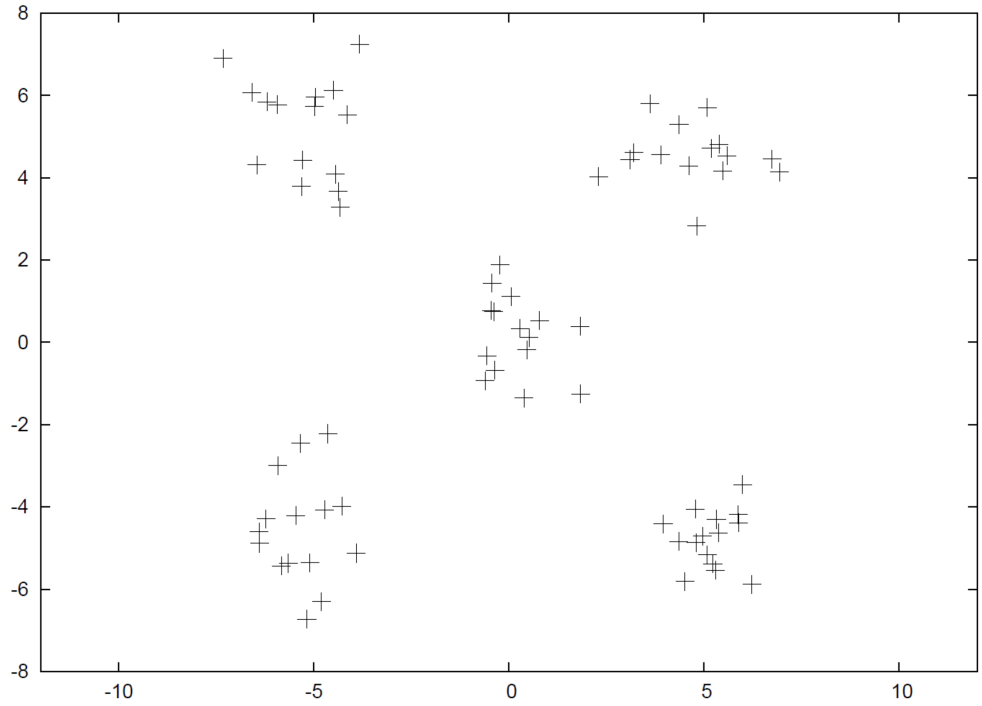


Figure 14: Well separated, two-dimensional data set.

as shown in Fig. 15 one still obtains fourteen errors, despite the fact that the original two-dimensional data set is unnaturally well separated.

If the original data set contains too many variables, as in the examples, variable selection methods are needed to reduce the dimension of the data set. Besides noise, the surplus variables may also be redundant. There are nowadays a number of variable selection methods. Raftery and Dean [5] propose a forward–backward method based on statistical testing by Bayes factors. They determine also the number of variables to be selected but apply their method only to low dimensional data sets. Tyler et al.'s [10] "invariant subspace selection" (ICS) compares different estimates of multivariate scatter matrices to reveal departure from elliptical symmetry. Hui and Lindsay [2] try to detect the largest white noise subspace returning its orthogonal complement. The last two methods select interesting, oblique subspaces. They have not explicitly been designed in view of clustering algorithms.

The following proposal has the flavor of Raftery and Dean's proposal but uses maximum likelihood estimation instead of testing. The basis is the notion of irrelevance as proposed by John et al. [3] and Koller and Sahami [4] which will next be described.

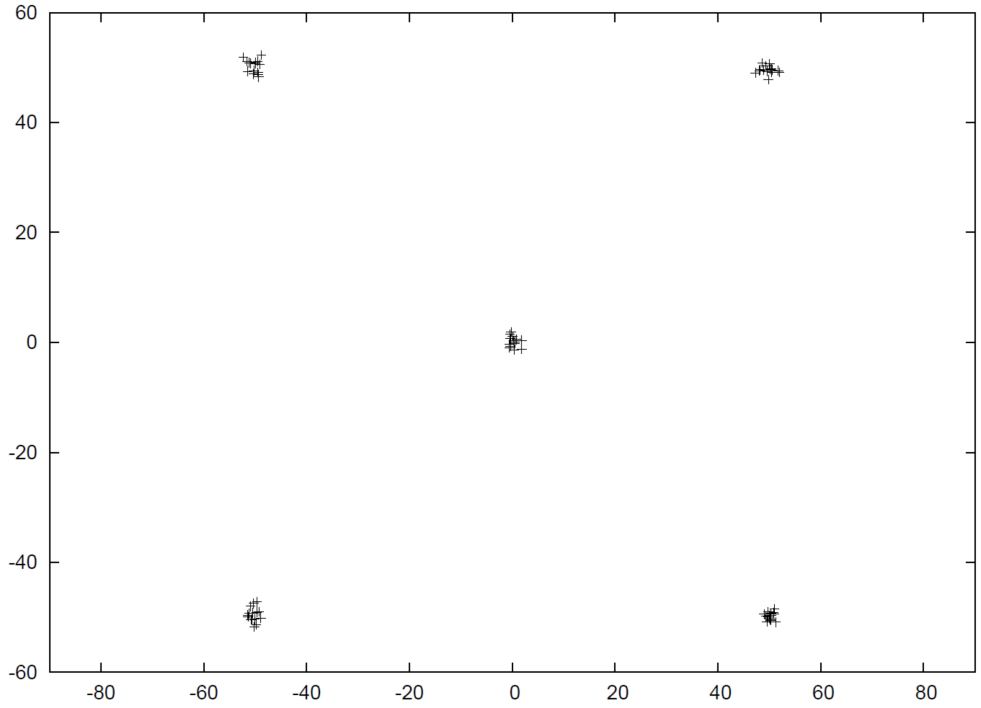


Figure 15: Extremely separated, two-dimensional data set.

2.1 Irrelevance in clustering

We first need some notation. We consider a random vector X in $\mathbb{R}^D$ distributed according to a g-component normal mixture,

$$\mu = \sum_{j=1}^{g} \pi_j N_{m_j,V_j}, \tag{4}$$

where $\pi_j > 0$, $\sum_{j=1}^{g} \pi_j$, are the mixing rates and the parameter pairs (m_j,V_j) of the normal components are pairwise distinct. The D coordinates in $\mathbb{R}^D$ will be numbered $1,2,\ldots,D$. The random vector X can be represented in the form $Y^{(L)}$, where $Y^{(1)},\ldots,Y^{(g)}$ are random vectors distributed as N_{m_j,V_j}, respectively, and where L is a random number in the interval $1\,..\,g$ with distribution $(\pi_1,\ldots,\pi_g)$. The number L has the meaning of a random assignment to a component (or class); it is assumed to be independent of all $Y^{(j)}$'s. We will also consider a (real-valued) data set $\mathbf{x} = (x_1,\ldots,x_n)$ with n observations of dimension D (that is, $x_i \in \mathbb{R}^D$, $1 \le i \le n$) drawn from n independent copies of $X_1 = Y_1^{(L_1)},\ldots,X_n = Y_n^{(L_n)}$ of $X = Y^{(L)}$. Realizations of the random assignments L_i will be denoted by ℓ_i, the class assignment of object i, and we write $\boldsymbol{\ell} = (\ell_1,\ldots,\ell_n)$.

Since μ is a mixture, the data set $\mathbf{x}$ is clustered according to the mixture components if they are sufficiently different. The aim is to detect these clusters. As an obstacle, we assume in

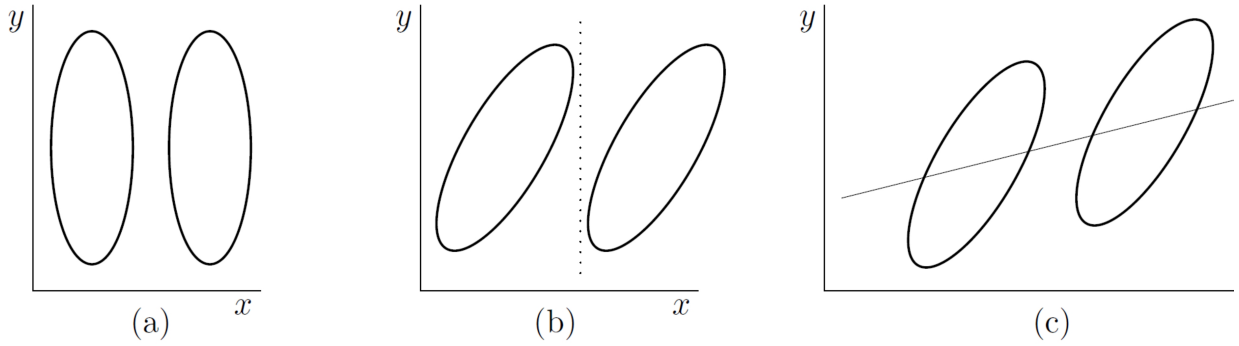


Figure 16: (a) y uninformative and irrelevant w.r.t. x; (b) y uninformative but relevant w.r.t. x; (c) y informative but irrelevant w.r.t. x.

addition that the dimension D is larger than actually needed to detect the clusters. This causes the problems already addressed in the Introduction. Another aim is therefore selection of a smaller subset of $d < D$ variables that contains the information on the cluster structure. The number d is assumed to be a priori given and must be chosen in relation to the number of observations and to the expected cluster sizes. In order to guarantee that scatter matrices of clusters w.r.t. d variables are almost surely nonsingular, we have to assume that each cluster has at least $d+1$ elements and that the restriction of the data set to any subset of d variables is in general position. This is almost surely satisfied if $P[X = x] = 0$ for all $x \in \mathbb{R}^D$. For $E \subseteq 1..D$, X_E (x_E) will be the restriction of X (x) to the entries in E.

Irrelevance of variables can be caused by redundancy and noise. John et al. [3] and Koller and Sahami [4] proposed a probabilistic model for irrelevance. Intuitively, a subset of variables, $E \subseteq 1..D$, is *irrelevant* w.r.t. to a disjoint subset F, $E \cap F = \emptyset$, if the union $E \cup F \subseteq 1..D$ does not contain more information w.r.t. the given clustering than F. More formally, we define the following.

2.1 Definition (a) The subset $E \subseteq 1..D$ is irrelevant w.r.t. to a disjoint subset $F \subseteq 1..D$, if L is conditionally independent of X_E given X_F, that is, for all j,

$$P[L = j \,|\, X_F, X_E] = P[L = j \,|\, X_F].$$

(b) The subset E is irrelevant if it is irrelevant w.r.t. its complement.

For illustration, some examples of irrelevance are shown in Fig. 16. All three mixtures are homoscedastic.

In Fig. 16(a), the ordinate y is clearly uninformative, that is, the ordinate contains by itself no cluster information, and it is also irrelevant w.r.t. the abscissa, x. In Fig. 16(b), the ordinate is uninformative but relevant w.r.t. the abscissa. That is, the ordinate improves the information

provided by the abscissa. It is sufficient to consider two observations on the dotted line. An observation at the bottom has a larger probability to belong to the right-hand population, one at the top is rather a member of the left population. Fig. 16(c) displays an example where the ordinate is informative but irrelevant w.r.t. the abscissa. The line connects the midpoints of the two normal populations with parameters $m^{(j)} = (m_x^{(j)}, m_y^{(j)})$, $j = 1,2$, and $V = \begin{pmatrix} v_x & v_{y,x} \\ v_{y,x} & v_y \end{pmatrix}$. Its slope is $\frac{m_y^{(2)} - m_y^{(1)}}{m_x^{(2)} - m_x^{(1)}} = \frac{v_{y,x}}{v_x}$, as can be easily derived from Definition 2.1.

Under mild assumptions, any data set has exactly one relevant subset of variables. The subset $F \subseteq 1..D$ of variables is called *structural* if no subset $\emptyset \neq C \subseteq F$ is irrelevant w.r.t. $F \setminus C$. The following theorem is due to Gallegos and Ritter and contained in Ritter [6].

2.2 Theorem *Let the real random variables X_i, $i \in 1..D$, have a strictly positive and continuous joint Lebesgue density $f_{(X_1,\ldots,X_D)}$. Then, there exists exactly one structural subset $F \subseteq 1..D$ with irrelevant complement.*

Normal mixtures allow us to deduce irrelevance from their parameters. To this end, it is favorable to represent a normal distribution as a regression. Any absolutely continuous distribution function f on $\mathbb{R}^D$ can be represented by its conditional distribution function $f(x_E \mid x_F)$ in the form

$$f(x) = f(x_E \mid x_F) \cdot f(x_F).$$

When $X = (X_F, X_E) \sim N_{m,V}$ is normal, the conditional distribution function has the representation

$$f(x_E \mid x_F) \sim X_E \mid [X_F = x_F] = m_{E|F} + G_{E|F} x_F + U_{E|F}.$$

Here $U_{E|F} \sim N_{0,V_{E|F}}$ is the residual random vector in $\mathbb{R}^E$. This is a reparametrization of $N_{m,V}$ by the conditional parameters $m_{E|F}$, $G_{E|F}$, $N_{0,V_{E|F}}$ and the parameters of X_F. Note that the number of real parameters of the regression model, that is, its dimension is $d_E + d_E \cdot d_F + \binom{d_E+1}{2} + \frac{(d_F+3)d_F}{2}$ (d_E and d_F are the sizes of E and $F = \complement E$, respectively, and the last term is the number of parameters of X_F). This is indeed the dimension of the normal model on $\mathbb{R}^D$, $\frac{(D+3)D}{2}$. The normal case allows the following characterization of irrelevance by the regression parameters of the components $j \in 1..g$, $G_{j,E|F}$, $m_{j,E|F}$, $V_{j,E|F}$. It will be the basis for the next section.

2.3 Theorem *(a) If X is a normal mixture (with nonsingular covariance matrix $\mathrm{V}X_F$), then the following statements are equivalent.*

(i) The subset E is irrelevant w.r.t. F;

(ii) the parameters $G_{j,E|F}$, $m_{j,E|F}$, and $V_{j,E|F}$ do not depend on j.

(b) In this case, these common parameters have the representations

(iii) $G_{E|F} = \mathrm{Cov}(X_E, X_F)(\mathrm{V}X_F)^{-1}$;

(iv) $m_{E|F} = m_E - G_{E|F} m_F$;

(v) $V_{E|F} = V_E - G_{E|F}\mathrm{Cov}(X_F, X_E)$.

2.2 Variable selection algorithm

We will use the notion of irrelevance introduced in Section 2.1 to extend Symons' [9] classical determinant criterion to variable selection. This well-known determinant criterion for general normal mixtures (4) is an extension of Scott and Symons' criterion (3) to clusters of arbitrary sizes. It reads

$$\frac{1}{2}\sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_j(\boldsymbol{\ell}) + n\mathrm{H}\big(\tfrac{n_1(\boldsymbol{\ell})}{n},\ldots,\tfrac{n_g(\boldsymbol{\ell})}{n}\big). \tag{5}$$

Here, $S_j(\boldsymbol{\ell})$ and $n_j(\boldsymbol{\ell})$ are the scatter matrix and size of the jth cluster of the assignment $\boldsymbol{\ell}$, respectively. The entropy $\mathrm{H}(p_1,\ldots,p_g) = -\sum_j p_j \log p_j$ accounts for unequal cluster sizes. The desired assignment $\boldsymbol{\ell}$ has a small value of criterion (5), but not necessarily the smallest one. A small value of the Scott-Symons criterion (3), that is, the left-hand side of Eq. (5), tends to equalize cluster sizes. Since the entropy is largest when cluster sizes are equal, it discourages small values of Eq. (5), favoring unequal cluster sizes unless the first term in Eq. (5) strongly insists on equally sized clusters. The entropy correction is, however, not just a heuristic idea but has been proved to be just right for equilibrating Scott and Symons' criterion (3).

Above, I wrote intentionally "... has a small value of the criterion " instead of "... has the least criterion." The solution with the smallest criterion is often not the desired one. First, it can be spurious having an unnaturally shaped, slim cluster, that is, a cluster with a very small eigenvalue of its shape matrix. Such a solution is unwanted in most cases. Second, the solution with the smallest criterion may even be wrong when all eigenvalues are decent. To obtain a reasonable solution, I rather prefer to use the so-called SBF plot; see Fig. 17. It is a plot of the negative log(HDBT ratio) vs. the log of criterion (5). The lower Pareto points of this plot belong to reasonable solutions. I usually select the one closest to the left lower corner; see Ritter [6], Chapter 4.

Schroeder [7] designed an iterative minimization algorithm for criterion (5). It can be interpreted as an iterative alternation of (normal) parameter estimation and Bayesian discriminant analysis until a stationary assignment has been found. The iteration becomes eventually stationary since it decreases the criterion and the criterion is bounded below since there are only finitely many partitions. I like to call this algorithm the k-parameters algorithm since it has as a special case the well-known k-means algorithm. Indeed, it is a simple exercise to show that Scott and Symons' criterion reduces to Ward's [11] sum-of-squares criterion if clusters are spherical of equal variance and if cluster sizes are equal. At the same time, Schroeder's algorithm collapses in this case to the k-means algorithm. The stationary solutions of the k-parameters algorithm are often called "local" minima although they are not local in any sense. A better word for local minimum would be *steady solution*. There is, however, an analogy to local maxima of mixture likelihoods.

It is well known that data sets possess in general many steady solutions. Therefore, the k-parameters algorithm has to be started from many initial points and the resulting steady

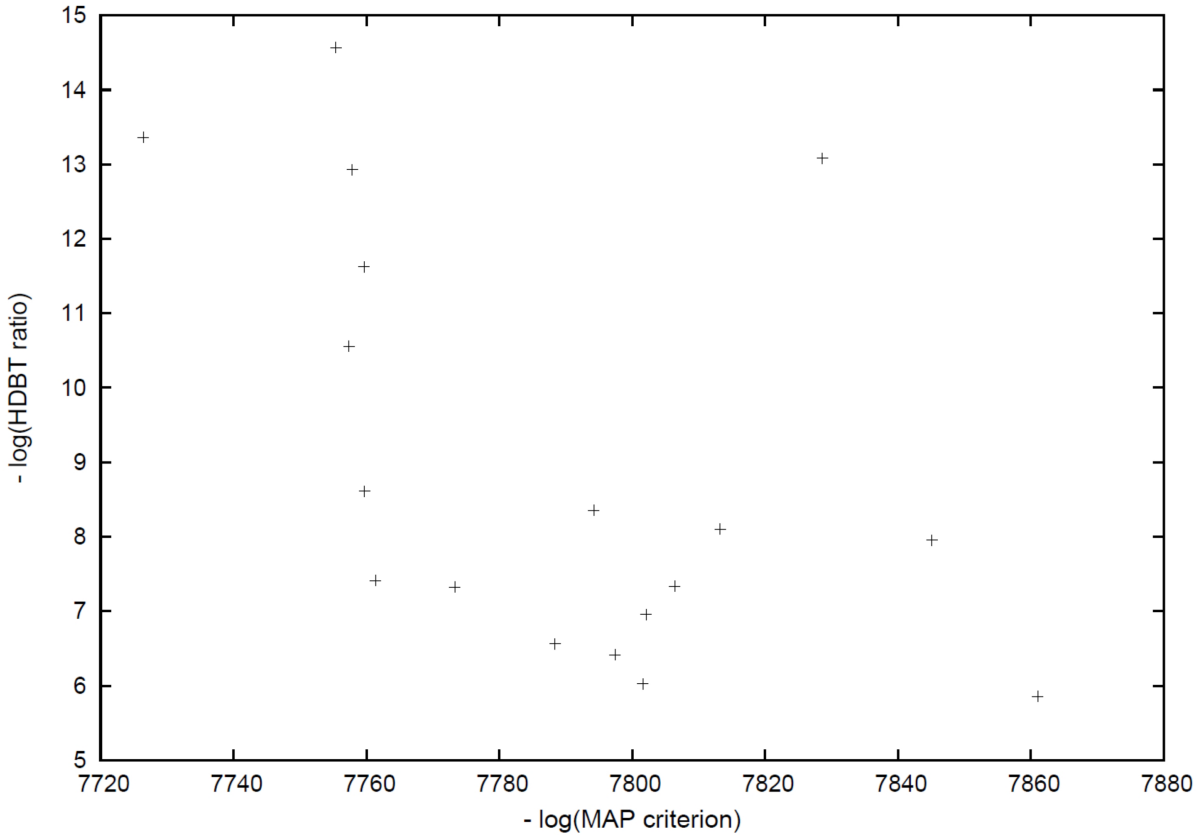


Figure 17: SBF plot.

solutions have to be validated in order to detect the desired one(s). This is the most difficult part of a cluster analysis. Another approach uses another method to generate a special initial solution for the k-parameters algorithm. This is, however, not the subject matter of this talk.

We will now extend Symons' criterion (5) to variable selection. The background of his method is the classification log-likelihood

$$\begin{aligned} f(\boldsymbol{\ell}, \mathbf{x}; \boldsymbol{\pi}, \mathbf{m}, \mathbf{V}) &= \sum_j \sum_{i:\, \ell_i = j} \log\left(\pi_j N_{m_j, V_j}(x_i)\right) \\ &= \sum_j \sum_{i:\, \ell_i = j} \log N_{m_j, V_j}(x_i) + \sum_j n_j(\boldsymbol{\ell}) \log \pi_j. \end{aligned} \tag{6}$$

Its maximization w.r.t. $\boldsymbol{\pi}$, $\mathbf{m}$, and $\mathbf{V}$ yields Symons' criterion (5). In view of variable selection, we now follow Ritter [6], Chapter 5, introducing the subset $F \subseteq 1..D$ of variables as an additional parameter. If $\complement F$ is irrelevant w.r.t. F, we obtain from Eq. (6) and from the results

of Sect. 2.1 the normal clustering-and-selection model

$$f(\boldsymbol{\ell},\mathbf{x};F,\boldsymbol{\pi},\mathbf{m},\mathbf{V}) = \sum_j \sum_{i:\,\ell_i=j} \log\left(\pi_j N_{m_j,V_j}(x_i)\right)$$

$$= \sum_{j=1}^{g} \sum_{i:\ell_i=j} \left(\pi_j \log N_{m_j,V_j}(x_{i,F})\right) + \sum_i \log N_{m_{E|F},V_{E|F}}(x_{i,E} - G_{E|F}x_{i,F}).$$

Its parameters are F, m_j, V_j, $m_{E|F}$, $V_{E|F}$ and $G_{E|F}$. Maximization w.r.t. $\boldsymbol{\pi}$, $\mathbf{m}$, and $\mathbf{V}$ yields a first form of the determinant criterion for clustering and selection

$$\tfrac{1}{2}\sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_{j,F}(\boldsymbol{\ell}) + n\mathrm{H}\left(\tfrac{n_1(\boldsymbol{\ell})}{n},\ldots,\tfrac{n_g(\boldsymbol{\ell})}{n}\right) + \tfrac{n}{2}\log\det S_{E|F}.$$

Here, $S_{j,F}(\boldsymbol{\ell})$ is the scatter matrix of the restrictions to F of the observations in cluster j and $S_{E|F}$ is the residual scatter matrix. By $\det S = \det S_F \cdot \det S_{E|F}$, $\log\det S_{E|F}$ differs from $-\log\det S_F$ by the constant $\log\det S$ and so, the final form of the determinant criterion for clustering and variable selection reads

$$\tfrac{1}{2}\sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_{j,F}(\boldsymbol{\ell}) + n\mathrm{H}\left(\tfrac{n_1(\boldsymbol{\ell})}{n},\ldots,\tfrac{n_g(\boldsymbol{\ell})}{n}\right) - \tfrac{n}{2}\log\det S_F. \tag{7}$$

Note that this criterion is completely affine equivariant. If an affine transformation A is applied to (7), then the first term is changed by $n\log\det A$ and the last term by $-n\log\det A$. In particular, units of measurement used for the observations are completely irrelevant. The idea is again finding small values of Eq. (7) w.r.t. $\boldsymbol{\ell}$ and F.

Besides model and criterion, the k-parameters algorithm, too, can be extended to an algorithm with variable selection. A proposal proceeds along the iteration

$$\left(\boldsymbol{\ell}^{(0)},F^{(0)}\right) \longrightarrow \left(\boldsymbol{\ell}^{(0)},F^{(1)}\right) \longrightarrow \left(\boldsymbol{\ell}^{(1)},F^{(1)}\right) \longrightarrow \ldots.$$

We obtain the following k-parameters algorithm with variable selection. It is a wrapper since variable selection uses the result of the clustering part of the algorithm. Generally, an assignment is admissible, if it allows estimation of the parameters of all clusters. In the present normal case this means regularity of scatter matrices.

2.4 Algorithm

// Input: Subset $F \subseteq 1\,..\,D$, $|F| = d$, admissible $\boldsymbol{\ell}$, and value of the criterion.
// Output: New quantities F_{new} and $\boldsymbol{\ell}_{\text{new}}$, with improved criterion *or* "stop."

1. *(Estimation)* Compute the sample mean vectors $\bar{x}_j(\boldsymbol{\ell})$ and scatter matrices $S_j(\boldsymbol{\ell})$, $1 \le j \le g$, and the total scatter matrix S.

2. *(Selection)* Minimize

$$h(F') = \sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_{j,F'}(\boldsymbol{\ell}) - n\log\det S_{F'} \quad (\le h(F))$$

w.r.t. F', $|F'| = d$. Denote the minimizer by F_{new}.

3. Use the quantities from step 1 to compute the MLE's of the regression parameters (G, m, V) w.r.t. $\boldsymbol{\ell}$ and the new subsets F_{new} and $E_{\text{new}} = \complement F_{\text{new}}$.

   Define the posterior probabilities

$$\begin{aligned} u_{i,j} = &\log n_j - \tfrac{1}{2}\log\det S_{j,F_{\text{new}}}(\boldsymbol{\ell}) \\ &- \tfrac{1}{2}(x_{i,F_{\text{new}}} - \bar{x}_{j,F_{\text{new}}}(\boldsymbol{\ell}))^\top S_{j,F_{\text{new}}}(\boldsymbol{\ell})^{-1}(x_{i,F_{\text{new}}} - \bar{x}_{j,F_{\text{new}}}(\boldsymbol{\ell})) \\ &- \tfrac{1}{2}(x_{i,E_{\text{new}}} - m - Gx_{i,F})^\top V^{-1}(x_{i,E_{\text{new}}} - m - Gx_{i,F}). \end{aligned}$$

4. *(Assignment)* Compute an admissible assignment $\boldsymbol{\ell}_{\text{new}}$ using a reduction step based on the statistics $u_{i,j}$.

5. *(Decision) If* F_{new} and $\boldsymbol{\ell}_{\text{new}}$ improve the criterion then return F_{new} and $\boldsymbol{\ell}_{\text{new}}$, *else* "stop".

If variables are independent, step 2 is easily obtained by sorting. Indeed, if scatter matrices are diagonal, the value $h(F')$ has the representation $\sum_{k\in F'}\left\{\sum_j n_j(\boldsymbol{\ell})\log S_j(\boldsymbol{\ell})(k,k) - n\log S(k,k)\right\}$, where $S_j(\boldsymbol{\ell})(k,k)$ ($S(k,k)$) are the diagonal entries in the kth row of $S_j(\boldsymbol{\ell})$ (S). In the independent case, it is therefore sufficient to compute the d smallest values k of $\sum_j n_j(\boldsymbol{\ell})\log S_j(\boldsymbol{\ell})(k,k) - n\log S(k,k)$. This is easily done by sorting.

It has been shown that the present algorithm is able to select a dozen meaningful variables from a thousand; see the gene expression example in Ritter [6], Chapter 6.

Discussion

Any variable selection method is fraught with some special risks. First, even noise variables often lead to clusterings, in particular, when there are not many observations. If a clustering of the irrelevant (noise) variables is stronger than the desired one of the relevant variables, then the former will probably be detected instead of the latter. Second, if the data contains clusterings in two respects, then it is not clear, which one has been detected and an unwanted clustering may result. For instance, if we have data of healthy and sick people of two ethnic populations and we decompose in two clusters, then it is not clear whether we have analyzed the ethnic properties or the health status. A cluster analysis can, in general, not be performed without knowledge of background and context.

References

[1] B.E. Fowlkes, R. Gnanadesikan, and J.R Kettenring. Variable selection in clustering. *J. Classif.*, 5:205–228, 1988.

[2] Guodong Hui and Bruce G. Lindsay. Projection pursuit via white noise matrices. *Sankhyā, Series B*, 72:123–153, 2010.

[3] George E. John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In Willian W. Cohen and Hyam Hirsh, editors, *Proceedings of the 11th International Conference on Machine Learning*, pages 121–129, San Mateo, CA, 1994. Morgan Kaufmann.

[4] Daphne Koller and Mehran Sahami. Toward optimal feature selection. In Lorenza Saitta, editor, *Machine Learning, Proceedings of the Thirteenth International Conference (ICML'96)*, pages 284–292, San Francisco, 1996. Morgan Kaufmann.

[5] Adrian E. Raftery and Nema Dean. Variable selection for model-based clustering. *J. Amer. Stat. Assoc.*, 101:168–178, 2006.

[6] Gunter Ritter. *Robust Cluster Analysis and Variable Selection*. Chapman & Hall/CRC, Boca Raton, London, New York, 2015. Monographs in Statistics and Applied Probability 137.

[7] Anne Schroeder. Analyse d'un mélange de distributions de probabilités de même type. *Revue de Statistique Appliquée*, 24:39–62, 1976.

[8] A.J. Scott and M.J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27:387–397, 1971.

[9] M.J. Symons. Clustering criteria and multivariate normal mixtures. *Biometrics*, 37:35–43, 1981.

[10] David E. Tyler, Frank Critchley, Lutz Dümbgen, and Hannu Oja. Invariant co-ordinate selection. *J. Royal Statist. Soc., Series B*, 71:549–592, 2009. With discussion and rejoinder.

[11] Joe H. Ward, Jr. Hierarchical grouping to optimize an objective function. *J. Amer. Stat. Assoc.*, 58:236–244, 1963.

3 The Multivariable Fractional Polynomial Approach, with Thoughts about Opportunities and Challenges in Big Data

Willi Sauerbrei
Institute for Medical Biometry and Statistics
University of Freiburg, Germany
wfs@imbi.uni-freiburg.de
and
Patrick Royston
2MRC Clinical Trials Unit at UCL
London, UK

Abstract

Data analysts are often faced with many covariates and a suitable model for explanation requires the selection of a subset of variables with a relevant influence on the outcome. For continuous variables it is important to determine a suitable function which fits the data well. We will introduce the basic concept and philosophy of the multivariable fractional polynomial (MFP) approach, which tackles both issues simultaneously. In the context of comparing two treatments we will introduce MFPI as an extension to investigate for potential interactions with continuous covariates. The approach avoids well known problems introduced by categorization. We will also introduce various opportunities and challenges of fractional polynomial modelling in Big Data. Furthermore, we will argue that treatment comparisons need to be based on well-designed randomized trials. In general, observational data do not allow to derive an unbiased estimate of the treatment effect, even if the sample size is very large.

Introduction

The number of covariates potentially included in a regression model is often too large and a more parsimonious model may have advantages. Several variable selection strategies (e.g. all-subset selection with various penalties for model complexity, or stepwise procedures) have been proposed for a long time (Sauerbrei, 1999). As there are few analytical studies about their properties, their usefulness is controversial. With continuous covariates the usual assumption of linearity may be violated. The multivariable fractional polynomial (MFP) approach simultaneously determines a functional form for continuous covariates and deletes uninfluential covariates (Royston and Altman, 1994; Sauerbrei and Royston, 1999; Sauerbrei et al., 2007a; Royston and Sauerbrei, 2008).

Continuous covariates are measured in most of the studies in the health sciences and MFP has become a popular approach for multivariable model building. For variable selection it uses backward elimination and for continuous covariates it checks whether a suitable (nonlinear) function from the class of fractional polynomials improves the fit significantly (Royston

DOI 10.20347/WIAS.REPORT.29 Berlin 2017

and Altman, 1994; Royston and Sauerbrei, 2008). The method also allows categorical and binary covariates. Extensions of MFP have been developed to look for interactions between continuous covariates and treatment (MFPI), between two continuous covariates (MFPIgen) and for interactions with time (non-proportional hazards, MFPT) in a Cox model (Royston and Sauerbrei, 2008; Sauerbrei et al., 2007b; Buchholz and Sauerbrei, 2011).

We have substantial experience in the analysis of real and simulated data with MFP, but restricted to 'larger' data sets. We will introduce key issues of MFP modelling and briefly discuss some opportunities and challenges when using fractional polynomial modelling in the context of 'big data', a highly relevant topic for the future. The phrase 'big data' is used for many different types of very large amounts of automatically collected data. Unfortunately, the concept of big data is not well-defined, since an essentially arbitrary dividing line seems to be imposed on the sample size, for no apparent reason. Nevertheless, the term seems to have stuck. In a recent Editorial, David Hand (2016) stresses the importance of distinguishing between two types of activity relating to big data. The first involves primarily data manipulation: sorting, searching, matching, and so on. Examples include online route finders and apps for updated status of bus and train traffic, with the associated issues addressed mostly by computer scientists and mathematicians. The second type of big data activity seeks to go beyond the data at hand, with the ultimate goals being either prediction of future data, or understanding of the mechanisms and processes that have generated the collected data. Achieving these goals will rely primarily on state-of-the-art statistical and machine learning methods. In addition, the method of data collection is relevant; briefly we may distinguish whether data come from a well-designed experiment (e.g. a randomized trial), a systematic collection (e.g. cancer registry) or whether they are 'found' data (e.g. internet poll). For a discussion see Keiding and Lewis (2016).

In this paper we have the second type of data in mind and as an application we will discuss key issues when comparing two treatments. Often, differences between effects of competing treatments are relatively small, but nevertheless relevant for patients. We will argue that data from a 'larger' randomized trial is required and that data from observational studies, even if the data set is 'Big' (very large), would not help to provide an unbiased estimate of treatment differences (Harford, 2014; Antes, 2015). We will also argue that the information from many RCTs is not fully exploited and discuss that MFPI should play a prominent role to investigate for potential interactions of a continuous covariate with treatment. Having 'Big Data' in mind and assuming that the selection of covariates and functional form for continuous covariates are an important part of the analysis, we will discuss opportunities and challenges of an analysis using MFP.

In this paper we have a 'larger' data set in mind, thoughts about big data are postponed to the specific subsection 6. In subsection 1 we discuss several key issues in variable selection. This is followed by subsection 2 on handling continuous covariates and the introduction of fractional polynomial modelling. The basic concept and philosophy of MFP modelling is introduced in subsection 3, followed by a short subsection on MFPI, the extension to investigate for interaction between a continuous and a binary covariate (subsection 4). MFPI can play

an important role when comparing two treatments (subsection 5). Before giving concluding remarks, in subsection 6 we discuss issues of MFP modelling in the context of big data. We have extensively published on the methodology and therefore details will not be given. We refer to the original papers, our book and the MFP website.

3.1 Model Building when Several Covariates are Available

In fitting regression models, data analysts are often faced with many covariates that may have an influence on an outcome variable. Consensus is that subject matter knowledge should generally guide model building, but it is often limited or at best fragile, making data-dependent model building necessary (Harrell, 2001). If the number of covariates is large, a parsimonious model involving a subset of the available covariates is often preferable (Sauerbrei, 1999). An aim of the analysis is the selection of covariates with more than a negligible influence on the outcome. In the health sciences the most popular methods for continuous, binary and censored survival data outcomes are normal-errors (linear) regression, logistic regression and Cox regression models. Issues and methods for variable selection are very similar among the three models mentioned. Usually, methods for variable selection and related issues have been developed and investigated for a normal-errors linear regression model and the methods, or at least their basic ideas, are commonly transferred to generalized linear models and to models for survival data. Sometimes additional problems, such as the definition of residuals or equivalents of R^2, exist. We refer to Andersen and Skovgaard (2010) for a text providing a useful unified treatment of regression models for different types of outcomes.

Relevant issues

In this part of subsection we assume that 'linearity' is a suitable assumption for the effect of a continuous covariate and our main emphasis is on models for explanation (interpretation). We have more traditional methods for variable selection (e.g. backward elimination) in mind. There have been several recent developments in the literature on variable selection but we know of no strong argument favoring replacement of backward elimination with another procedure in the MFP algorithm (see subsection 3). Some of our arguments are hardly defensible for 'small' sample sizes and high-dimensional data, such as -omics data. Such situations are implicitly excluded. Under our assumptions we consider the following issues as the most relevant to model selection: Aim (model for prediction or for explanation), model complexity, model stability, incorporating the model uncertainty concept, selection bias and shrinkage of regression coefficients as a potential way to correct for it. For more details see our website `http://mfp.imbi.uni-freiburg.de/`.

Aim of the model and model complexity

Many different aims are possible when developing a multivariable model and the specific aim has an influence on the suitability of a chosen approach. For a detailed discussion see section 2.4 in Royston and Sauerbrei (2008). In many analyses the most important distinction is between models aiming to derive a suitable covariate and models aiming to identify factors which seem to help explaining the value of an outcome. For a discussion see the paper entitled 'To explain or to predict' by Shmueli (2010). He illustrates that these phrases mean different things in different disciplines and mentions relevant distinctions and practical implications for explanatory and predictive modeling. For example, in the social sciences the term explanatory model is used nearly exclusively for testing causal theory. Although we agree with Shmueli that our approach to derive a model would be better called descriptive modeling, we will proceed with the better known (in the health sciences) term 'explanatory model'.

For stepwise variable selection procedures, the significance level (to be chosen by the analyst) is the key user-adjustable setting that influences model complexity. For details on stepwise procedures and a discussion of the close relationship between the significance level and the information criteria AIC and BIC see section 2.6 in Royston and Sauerbrei (2008). Deriving explanatory models is the main aim in this paper. There are several arguments that simpler models are preferable for such situations (Sauerbrei, 1999; Royston and Sauerbrei, 2008 section 2.9.4).

Model complexity, model stability and model uncertainty

Model complexity, model stability and model uncertainty are three different issues of data-dependent model building. However, they are closely related. A more complex model (in this context, a model including more covariates) is usually less stable as it almost invariably includes several covariates which have only a 'weak' effect on the outcome (Sauerbrei and Schumacher, 1992; Sauerbrei et al., 2015). When selecting a specific model, the uncertainty of the selection process is (usually) ignored. To improve models for prediction, the model uncertainty concept was introduced some 20 years ago (Chatfield, 1995, Draper, 1995). A predictor and its variance are estimated by averaging predictors from many (unstable) models. Usually the Bayesian framework is used for model selection and assessment of model uncertainty (Bayesian model averaging; Hoeting et al., 1999). Extending an approach by Buckland et al. (1997), Augustin et al. (2005) suggested using the bootstrap to handle model uncertainty. In contrast to the Bayesian approach which uses Occam's razor to reduce the number of models, Augustin et al. (2005) proposed using a screening step to eliminate covariates with at most a weak effect. Obviously, the number of models included in the second part for model averaging is severely reduced. For a detailed illustration see the example in Sauerbrei et al. (2015). In subsection 3 we will describe the MFP approach to select covariates and functional relationships for continuous covariates. We have also conducted some investigations in the context of function stability (Royston and Sauerbrei, 2003).

Variable selection and shrinkage

Concerning approaches for variable selection, the situation is very confusing. Triggered by the problem of identifying a small number of relevant covariates in a high-dimensional data, many procedures have been proposed recently. However, the number of helpful comparisons between strategies is limited. There is agreement that variable selection will cause biases in estimates of regression parameters and many of the more recent strategies combine variable selection with shrinkage in a regularized approach. For an overview of techniques see Hastie et al. (2009) and several issues are also discussed in Schumacher et al. (2012). In the context of low-dimensional data van Houwelingen and Sauerbrei (2013) assessed whether post-selection two-step approaches using global shrinkage proposed by van Houwelingen and Le Cessie (1990) or parameterwise shrinkage (PWSF, (Sauerbrei, 1999)) can improve selected models. They also compared results to models derived with the LASSO procedure (Tibshirani, 1996), probably the most popular approach to combine variable selection and shrinkage in a one-step approach. Concerning prediction ability the performance of backward elimination (BE) with a suitably chosen significance level was not worse compared to the LASSO and BE models selected were much sparser, an important advantage for interpretation and transportability. It could be shown that the PWSF approach compares favourably to global shrinkage. It was summarized that BE followed by PWSF is a suitable approach when variable selection is a key part of data analysis, provided that the amount of information in the data is not 'too small'.

In the context of using the MFP procedure to derive a multivariable model data-dependently, it was noted that regression parameter estimates of FP functions are biased and may need to be shrunken (Sauerbrei and Royston, 1999). The PWSF approach was considered as one potential way to handle this issue. However, for covariates which are either highly correlated or associated with regard to contents, such as several parameters describing a nonlinear FP2 function, the approach has weaknesses. For such cases the methodology was extended by so-called 'joint shrinkage factors', a compromise between global and parameterwise shrinkage (Dunkler et al., 2016).

3.2 Continuous Covariates

Continuous covariates are often encountered in life. We measure age, weight, blood pressure and many other things. In medicine, such measurements are often used to assess risk or prognosis or to select a therapy. However, the question of how best to extract useful information from continuous covariates is an important challenge (Rosenberg et al., 2003), in the multivariable context interrelated with the selection of covariates for inclusion in a model. In a short summary, topic group 2 of the STRATOS (STRengthening Analytical Thinking for Observational Studies) initiative states (Sauerbrei et al., 2014):

"In practice, multivariable models are usually built through a combination of (i) a priori inclusion of well-established 'predictors' of the outcome of interest and (ii) a posteriori selection of additional variables, based often on arbitrary, data-dependent procedures and criteria such as statistical significance or goodness-of-fit measures. There is a consensus that all of the

many suggested model building strategies have weaknesses (Miller, 2002) but opinions on the relative advantages and disadvantages of particular strategies differ considerably. The effects of continuous predictors are typically modeled by either categorizing them (which raises such issues as the number of categories, cutpoint values, implausibility of the resulting step-function relationships, local biases, power loss, or invalidity of inference in case of data-dependent cutpoints) (Greenland, 1995) or assuming linear relationships with the outcome, possibly after a simple transformation (e.g. logarithmic). Often, however, the reasons for choosing such conventional representation of continuous variables are not discussed and the validity of the underlying assumptions is not assessed.

To address these limitations, statisticians have developed flexible modeling techniques based on various types of smoothers, including fractional polynomials (Royston and Altman, 1994; Royston and Sauerbrei, 2008) and several 'flavors' of splines. The latter include restricted regression splines (Harrell, 2001; Boer, 2001), penalized regression splines (Wood, 2006) and smoothing splines (Hastie and Tibshirani, 1990). For multivariable analysis, these smoothers have been incorporated in generalized additive models."

To categorize or to model?

For continuous covariates, a simple and popular approach is to assume a linear effect, but the linearity assumption may be questionable. To avoid this strong assumption, researchers often apply cutpoints to categorize the covariate, implying regression models with step functions. This simplifies the analysis and may or may not simplify interpretation of results. It seems that the usual approach in clinical and psychological research is to dichotomize continuous covariates, whereas in epidemiological studies it is customary to create several categories, often four or five, allowing investigation of a crude dose-response relationship. However, categorization discards information and raises several critical issues such as how many cutpoints to use and where to place them (Altman et al., 1994; Royston et al., 2006). Sauerbrei and Royston (2010) illustrate several critical issues by investigating prognostic factors in patients with breast cancer. As a more suitable approach to analysis, they propose to model continuous covariates with fractional polynomials (FP). See Royston and Sauerbrei (2008) for a monograph on this topic and the related website `http://mfp.imbi.uni-freiburg.de/`.

Fractional polynomials

Class of FP functions

The class of fractional polynomial (FP) functions is an extension of power transformations of a covariate. For most applications FP1 and FP2 functions are sufficient.

FP1: $\beta_1 x^{p1}$

FP2: $\beta_1 x^{p1} + \beta_2 x^{p2}$

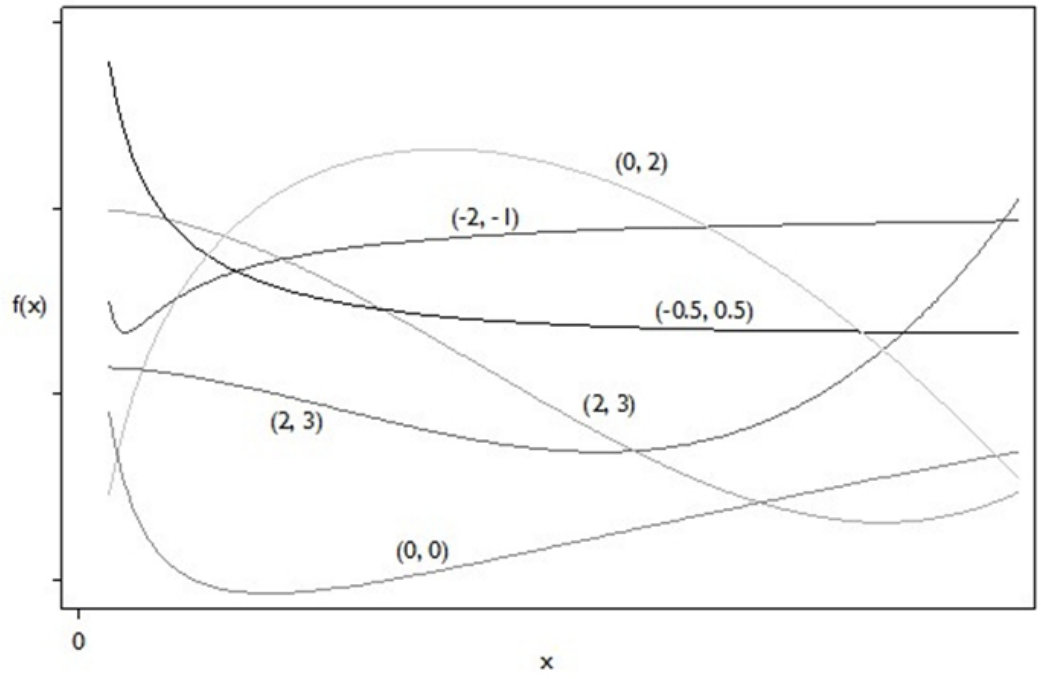


Figure 18: Various shapes of FP2 functions with different power terms $p1$ and $p2$.

For the exponents $p1$ and $p2$ a set **S={-2, -1, -0.5, 0, 0.5, 1, 2, 3}, with $\mathbf{0} = \log \mathbf{x}$** was proposed. For $p1 = p2 = p$ ('repeated powers') an FP2 function is defined as $\beta_1 x^p + \beta_2 x^p \log x$. This defines 8 FP1 and 36 FP2 models. The values $p1 = 1,\ p2 = 2$ define the quadratic function. The class of FP functions seems to be small, but it includes very different types of shapes (Fig. 18). General FPm functions are well-defined and straightforward, but will not be discussed here as they are rarely used. FP3 or more complex FP functions may improve the fit in some cases (particularly in a univariate analysis), but in the multivariable context, which is the main issue here, we are not aware of any relevant example. Occasionally they also find a use as effective approximations to intractable mathematical functions (Royston and Altman, 1997).

Selecting an FP function

A suitable function should fit the data well, and also be simple, interpretable and generally usable. To assess whether a covariate has a significant effect, the FP function selection procedure (FSP) starts by comparing the best fitting allowed FP (often FP2) function of a continuous covariate x with the null model (Royston and Sauerbrei, 2008 section 4.10). If significant, the procedure proceeds by comparing FP functions with a 'simple' (usually linear) default function. Using FSP the default function is often selected. More complex FP

functions are chosen only if they fit the data much better (based on a significance criterion), which implies that sample size (effective sample size in survival data) plays an important role. Modifications are required in 'big data', see subsection 6.

Before starting to select a suitable function, the analyst must decide on a nominal p-value (α) and on the degree (m) of the most complex FP model allowed. Typical choices in medicine are $\alpha = 0.05$ and FP2 ($m = 2$). In the following we describe FSP when FP2 is chosen. It is straightforward to adapt the procedure for use with other FP degrees. Based on minimizing the deviance (minus twice the maximized log likelihood), the best FP1 and best FP2 function are determined. The following test procedure assumes that the null distribution of the difference in deviances between an FPm and an FP ($m-1$) model is approximately central χ^2 on two degrees of freedom. For details see section 4.9.1 of Royston and Sauerbrei (2008). The FP function is determined for the variable x using the following closed test procedure:

1 Test the best FP2 model for x at the α significance level against the null model using four d.f. If the test is not significant, stop, concluding that the effect of x is 'not significant' at the α level. Otherwise continue.

2 Test the best FP2 for x against the default (usually a linear function) at the α level using three d.f. If the test is not significant, stop, the final model being the default. Otherwise continue.

3 Test the best FP2 for x against the best FP1 at the α level using two d.f. FP2 selects two power terms and estimates two corresponding parameters, therefore 4 d.f.; correspondingly FP1 has 2 d.f., giving a difference of two d.f. If the test is not significant, the final model is the best FP2, otherwise the final model is the best FP1. End of procedure.

Note that the α level for the selection of the FP function can be different to the significance level of backward elimination. If α = 1 in the latter then x is always selected and step 1 is redundant. Using the flavor of a closed test procedure ensures that the overall type 1 error is close to the nominal significance level. For some results concerning type 1 error and power we refer to simulation studies described in section 4.10.5 of our book.

3.3 MFP: an Approach to Multivariable Model-building with Several Continuous Covariates

MFP is an approach to multivariable model-building which retains continuous covariates as continuous, finds non-linear functions if sufficiently supported by the data, and removes weakly influential covariates by backward elimination (BE). The main issues of the approach arise from the two key components: variable selection with backward elimination and selection of an FP function to model non-linearity.

The MFP algorithm - basic concept

Like backward elimination, the MFP algorithm starts with all candidate covariates entered as linear terms (the 'full model') and investigates whether any covariates can be eliminated. However, for each of the continuous covariates the FSP is used to check whether a non-linear function fits the data significantly better than a linear function. After a first cycle some covariates will often be eliminated and for some continuous covariates a better fitting non-linear function may have been determined. The algorithm starts a second cycle, but the new starting model now has fewer covariates (as some were eliminated) and perhaps non-linear functions for some of the continuous covariates. In the second cycle all covariates are reconsidered (even if they were not significant at the end of the first cycle) and the FSP is used again to determine the 'best' fitting FP function (it may be different because other 'adjustment' covariates are in the model). This yields the result of cycle 2 which is the starting point for cycle 3. In most cases the model does not change anymore in cycle 3 or 4 and the algorithm stops with the final MFP model.

Important is the order of 'searching' for model improvement by better fitting non-linear functions. Obviously, mismodelling the functional form of a covariate with a strong effect is more critical than mismodelling the functional form of a covariate with a weak effect. The order is determined by ascending p-values from likelihood ratio tests for elimination from the full model. Covariates with a small p-value are considered first. Boxes 6.1 and 6.2 in Royston and Sauerbrei (2008) illustrate the algorithm in an example. Most often 0.05 is used as the significance criteria for both variable elimination and function selection, however, these two important parameters for variable and function selection can (and should) be flexibly chosen by the analyst. Depending on the aim of an analysis more or less stringent significance criteria may be preferable.

MFP modelling - philosophy and related matters

For a detailed description of the algorithm and some relevant issues see Chapter 6 in Royston and Sauerbrei (2008). In the discussion of it, we consider in detail four relevant issues (1 - Philosophy of MFP; 2- Function Complexity, Sample Size and Subject-Matter Knowledge; 3- Improving Robustness by Preliminary Covariate Transformation; 4- Conclusion and Future). Our thoughts about these issues are summarized in a table entitled 'Towards recommendations for model building by selection of variables and functional forms for continuous predictors in observational studies, under the assumption of Tab 1.3.' This table is adapted from Sauerbrei et al. (2007a), where we expressed thoughts about our philosophy of MFP modelling:

"Issues such as model stability, transportability and practical usefulness need more attention in model development. The latter are all connected with the often neglected criterion of external validation. Increasing their importance will result in models that are built with the aim to get the big picture right instead of optimizing specific aspects and ignoring others. With

a good model building procedure, the analyst should be able to detect strong factors, strong non-linearity for continuous variables, strong interactions between variables and strong non-proportionality in survival models. With such a model one is less concerned about failing to include variables with a weak effect, failing to detect weak interactions or failing to find some minor curvature in a functional form of a continuous covariate. Such a model should be interpretable, generalizable and transportable to other settings. In contrast to results from spline techniques, which are often presented as a function plot, an FP function is a simple formula allowing general usage. Our aims agree closely with the philosophy of MFP and its extensions for interactions (Royston and Sauerbrei, 2003) and time-varying effects (Sauerbrei et al., 2007b). Modifications that may improve the usefulness of MFP are combination with shrinkage and a more systematic check for overlooked local curvature."

In a large simulation study comparing MFP with various spline approaches, we provided some evidence for the conclusions given in the table of recommendations', but further simulation studies are needed (Binder et al, 2013). It is planned to conduct them in topic group 2 of the STRATOS initiative 'Selection of variables and functional forms in multivariable analysis' (Sauerbrei et al, 2014).

3.4 Extension of MFP to Investigate for Interactions

Given the enormous amount of resources spent on conducting a large clinical trial, it is surprising that greater efforts are not made to try to extract more information from clinical trials data. In the context of potential interactions between continuous covariates and treatment, we have argued for the use of MFPI (multivariable fractional polynomials - interaction) for such investigations (Royston and Sauerbrei, 2004; Sauerbrei and Royston, 2007). Unfortunately, dichotomization is still the 'standard', even though most of the well-known problems of categorization mentioned above transfer to analyses for interactions. The key ideas of MFPI are: first, MFPI estimates for each treatment group a fractional polynomial function representing the prognostic effect of the continuous covariate of interest, optionally adjusting for other covariates. Second, the difference between the functions for the treatment groups is calculated and tested for significance. The testing is done through an analysis of interaction between treatment and the FP function. A plot of the difference (e.g., log hazard ratio) against the covariate, together with a 95% CI, is termed a 'treatment-effect plot'. A treatment-effect plot for a continuous covariate not interacting with treatment would be a straight line parallel to the x-axis, whereas a treatment-covariate interaction would be indicated by a non-constant line, often increasing or decreasing monotonically. For more details see subsection 6 in our book (Royston and Sauerbrei, 2008). In a recent simulation study we were able to illustrate striking advantages of MFPI over methods based on dichotomization or categorization (Royston and Sauerbrei, 2013; Royston and Sauerbrei, 2014). Based on these results, we slightly changed our recommendation for the most suitable approach (our new default). For details see the website or the latter paper.

3.5 Opportunities of MFPI when Comparing Treatments

By re-analyzing data from an MRC randomized trial in patients with renal cancer, we illustrated additional opportunities to investigate for interactions of a continuous covariate with treatment (Royston et al., 2004). In Fig. 19 we show Kaplan-Meier estimates in all patients and in patients defined by 4 subgroups based on white cell count (WCC) values. These subgroups are motivated by the treatment effect function for WCC (top right). Using MFPI we investigate ten continuous covariates as potential modifiers of the treatment effect. Nine covariates did not exhibit any important interaction, but for WCC the test for interaction was significant at the 1% level. We use Kaplan-Meier plots in subpopulations as check of the derived treatment effect function.

The five plots of Kaplan-Meier estimates show that the proportional hazard assumption of the Cox model is acceptable in all populations and we estimated treatment effects in each of the groups. The estimated hazard ratio (HR: Interferon to MPA; 95% confidence interval) in all patients is 0.75 (0.60 - 0.93), which clearly shows the benefit of interferon. However, in subgroups defined by increasing values of WCC we observe increasing estimates agreeing with the treatment effect function and the impression from the plots for subgroups (I: 0.53 (0.34 - 0.83), II: 0.69 (0.44 - 1.07), III: 0.89 (0.57 - 1.37), IV: 1.32 (0.85 - 2.05)).

There is a large effect favoring interferon in group I (very low WCC values). The advantage disappears for patients with higher WCC values. Analyses in subgroups support the estimated treatment effect function.

Concerning the interpretation of results from MFPI analyses, we need to distinguish between prospectively planned analyses and a retrospectively conducted search for markers which may have an influence on the effect of treatment. Results from a retrospective search need to be seen as hypothesis generation, requiring validation in new data. For hypothesis generation we recommend using small p-values (e.g. 0.01), otherwise larger p-values may be acceptable. In any case, we strongly recommend checking estimated treatment effect functions by conducting analyses in subgroups.

3.6 Analyzing Big Data with MFP - on Opportunities and Challenges

So far we have no experience analyzing 'Big Data' with MFP or more generally with FP methodology. In the following we will consider two very different 'big data' situations and point to potential opportunities and challenges when using FPs for the analysis.

Large(r) sample size

Having a large sample size offers many opportunities for MFP analyses but also raises several issues of our test-based FP function selection procedure. Obviously FSP needs to be adapted because a very large sample size would (nearly) always result in selecting the most

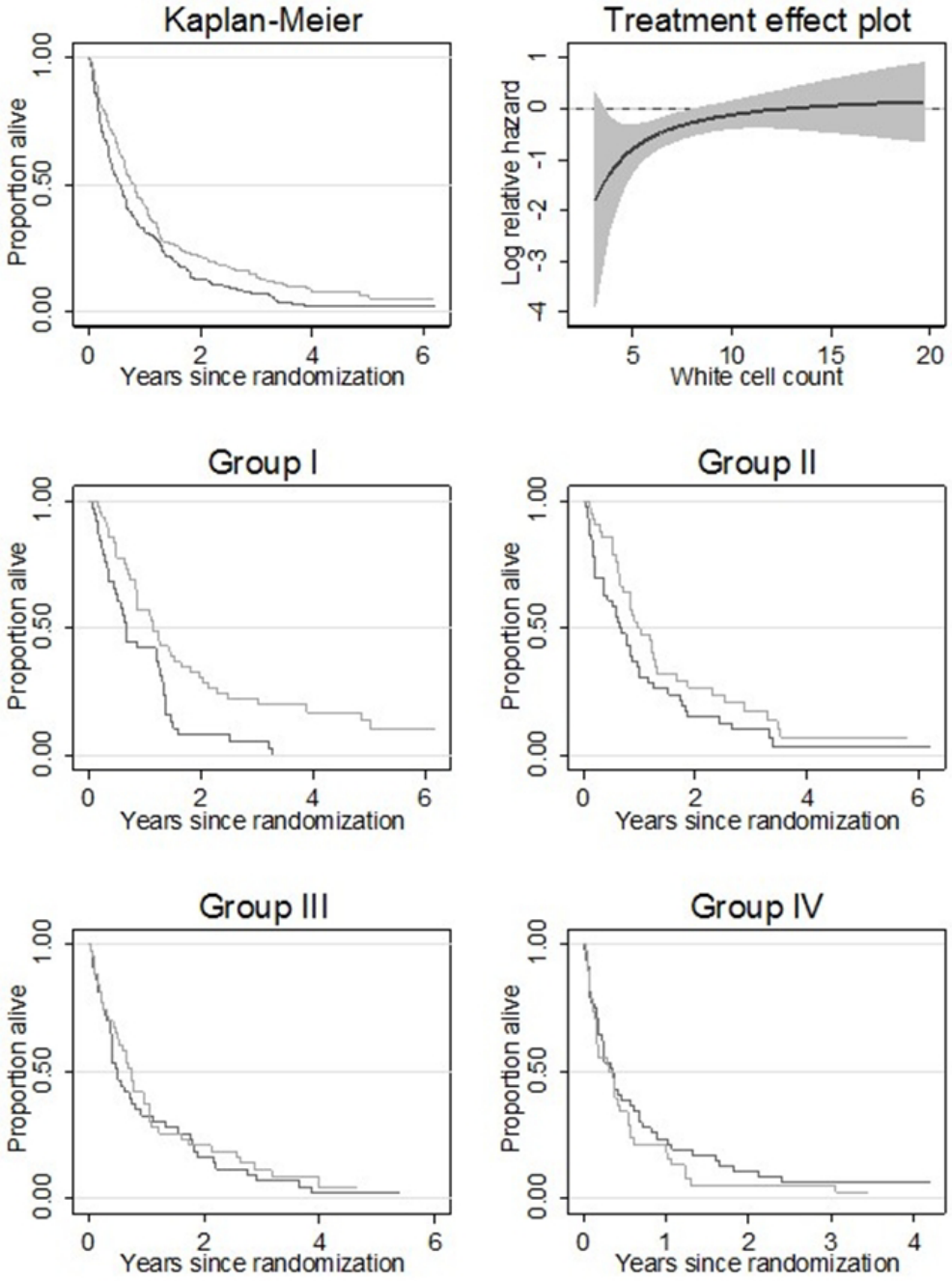


Figure 19: Plots of Kaplan-Meier estimates of overall survival probability for patients treated with interferon (pale) or MPA (dark) and estimated treatment effect function (with 95% CI) comparing the treatment effect dependent on white cell count (WCC), the only significant covariate interacting with treatment (top right). The four Kaplan-Meier plots (middle and bottom) show survival estimates in subgroups determined by WCC values.

DOI 10.20347/WIAS.REPORT.29 Berlin 2017

complex allowed FPm (typically FP2) function.

Careful consideration of whether FP2 functions should be allowed is one simple way to handle this issue. For example, restriction to the FP1 class could be suitable if subject matter knowledge provides a strong argument that a function should be monotonic. In that case the best fitting transformation of the eight functions would be selected. In a similar way FSP could be modified to select the best of the thirty-six FP2 functions if a non-monotonic function would be suitable.

Another approach would be to use the area between curves (ABC) criteria to replace significance tests in the FSP. ABC was proposed by Govindarajulu et al. (2007) to quantify the distance between smoothed curves and later adapted to quantify the distance between two curves estimating time-varying effects in the Cox model for survival data (ABCtime, Buchholz et al., 2014). In a procedure similar to FSP, distances between best FP2, best FP1 and the linear function could be considered. However, further work on a suitable metric to compare two curves is needed. What is a relevant ABC value to conclude that the best FP2 fits 'substantially better' than the linear function or the FP1 function? This issue needs experience in real studies and in simulations.

In the context of MFP modelling it is also important to adapt the variable selection part of MFP. One simple possibility is to choose the BIC (Bayes Information Criterion; Schwarz, 1978) as the criterion for model selection. The penalty constant of BIC is log(n), which may help to restrain the selection of many significant covariates with a very small effect. BIC or extremely small p-values such as 0.000001 may also be used in FSP for the selection of an FP function. A different line would to try adapting the variable selection (backward elimination) part by using ideas from the change-in-estimates approach (Greenland et al., 1989). That may also help for categorical covariates, also needing adaption for the case of very large sample sizes.

However, practical experience is needed to see whether these ideas are sufficient to adapt the current MFP procedure to handle the problem of variable and function selection in very large data sets. Very large sample sizes offer also many new possibilities for MFP methodology. For potential interactions between two continuous covariates we have proposed MFPIgen as an extension of MFPI (see section 7.11 in Royston and Sauerbrei, 2008). However, to conduct such an analysis a 'large' sample size is needed. For 'very' large sample sizes an adaption as discussed above may be required.

Having very large datasets allows the analyst to partition the data and give the often neglected model validation aspect much more weight. To get some ideas about external validation of a 'derivation' model, data partitioning is often done, even with 'medium sized' data sets. In very large data sets natural partitions may be available (e.g. three hospitals with large datasets each) and a partition could be possible without the severe disadvantage of losing power, which is often low anyway. See van Houwelingen (2000) for related discussions. Related is the possibility of dividing the data into several (well-defined) subpopulations, conduct an (MFP) analysis in each of them and summarize results in a meta-analysis. Using 'big data' from nine SEER registries, we proposed a new approach for the meta-analysis of

functions (Sauerbrei and Royston, 2011).

Very large number of covariates and small sample size

This situation is becoming more and more relevant in the health sciences, often called 'omics' research. This term encompasses multiple molecular disciplines that involve characterization of global sets of biological molecules such as DNAs, RNAs, proteins and metabolites (IoM (Institute of Medicine), 2012). Typical sample sizes are between 100 and 500 (for survival data the effective sample size, the number of events, is often much smaller) and the number of covariates range from several hundreds to several hundred thousand. Obviously, deriving a 'suitable' model is a challenge. 'Traditional' statistical modelling approaches cannot be used and many strategies have been adapted and developed during the last years. Considering a preliminary covariate screening step, various methods of regularization and the combination of variable selection and shrinkage play a key role. However, usually it is assumed that the effect of a continuous covariate is linear. We could imagine that consideration of the eight functions from the FP1 class could improve some of the models. After a pre-selection of covariates (say selection of the top 500) it would be easy to consider (in univariate analyses) whether any of the seven non-linear FP1 functions provides a much better fit compared to a linear function. The p-value or the ABC criterion may be used for the comparison.

To identify extreme values in omics data, Boulesteix et al. (2011) used a simple pre-transformation, originally proposed to improve robustness of MFP models (Royston and Sauerbrei, 2007), and compared gene rankings derived from the original and the transformed values. For some datasets they could identify striking differences in the gene rankings, caused by altering single observations. The approach could be extended to consider the best FP1 function as the pre-transformation and an extension to multivariable models should be possible.

Concluding Remarks

We have provided a brief overview of multivariable model building based on fractional polynomials for modelling continuous covariates. We have concentrated on the MFP approach which combines backward elimination as a strategy for variable selection with the selection of a suitable function from a well-defined class of fractional polynomials. The aim of a multivariable model has a substantial influence on the suitability of a model building procedure (Shmueli, 2010). Different strategies can produce very different models, but predictors from different models are often (very) similar (Sauerbrei et al., 2015). In the health sciences models for explanation play a more important role and we have such models in mind in our discussion. For the variable selection part we have discussed model complexity as the key issue and shrinkage as a potential way to correct for bias introduced by data dependent modelling. The complexity of a BE model can be easily controlled by the significance level and

we use it as the key parameter for both parts of MFP.

Comparing two (or more) treatment strategies is one of the most important investigations in the health sciences. From a statistical point of view the popular phrase 'individualized treatment' implies interactions of treatment with 'several' patient characteristics. So far interactions with a continuous covariate are usually investigated by categorizing (dichotomizing) the continuous covariate and investigate for treatment differences in subgroups. In the context of prognostic factors, risk factors and many others, the severe disadvantages introduced by categorization have been well known for many years (Altman et al., 1994, Royston et al., 2006). Obviously, most of the problems transfer to models investigating for interactions. For a more detailed discussion see (Hingorani et al., 2013). One of their recommendations reads *"Standards in statistical analysis of prognosis research should be developed which address the multiple current limitations. In particular, continuous variables should be analysed on their continuous scale and non-linear relationships evaluated as appropriate."*

The MFP procedure was extended to MFPI as a more suitable way to investigate for interactions between a binary treatment (extension for categorical covariates are straightforward) and a continuous patients characteristic. In an example we have demonstrated that MFPI analyses can help to identify prognostic factors which interact with treatment, in some areas in medicine they are called predictive factors. To report relevant details of MFP and MFPI analyses is straightforward, another important factor of our approaches. The importance of transparent reporting and reproducible research has become a key issue in medical research. As software for MFP and MFPI is generally available (originally all routines have been programmed by Patrick Royston in Stata, for details see the website) it should be possible to reproduce an analysis, provided the data are publicly available.

So far we have no experiences using MFP and MFPI in the context of 'Big Data'. We have outlined some potential chances and problems in using our approaches. However, the key issues depend on the specific problem and the way data were collected. In the health sciences there are many promises related to 'Big Data' but it is obvious that more data will not solve every problem (Antes, 2015). Very large sample sizes can be helpful to reduce the random error and increase power, but potential biases are the more relevant in many analyses. Consequently, for investigations to compare treatments and to search for treatment modifying factors, we have used the data from a randomized trial. Concerning data quality, Hand (2016) points out that 'large does not necessarily mean good, useful, valuable or interesting. Big does not necessarily mean accurate or comprehensive'. With this short overview we aim to illustrate that fractional polynomial methodology can be used sensibly for many analyses requiring modelling of continuous covariates.

References

Altman, D.G., Lausen, B., Sauerbrei, W. and Schumacher, M (1994): Dangers of using 'Optimal' cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer*

Institute, 86: 829–835

Andersen, P.K. and Skovgaard, L.T. (2010): *Regression with Linear Predictors*. Springer, New York

Antes, G. (2015): *A new Science(ability)?* Lab Times Online

Augustin, N., Sauerbrei, W. and Schumacher, M. (2005): The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modelling 5: 95–118*

Binder, H., Sauerbrei, W. and Royston, P. (2013): Comparison between splines and fractional polynominals for multivariable model-building with continous covariates: a simulation study with continous response. *Statistics in Medicine 32: 2262–2277*

Boer, C. de (2001): *A Practical Guide to Splines*. revised edn. Springer, New York

Boulesteix A.-L., Guillemot V. and Sauerbrei W. (2011): Use of pretransformation to cope with extreme values in important candidate features. *Biometrical Journal 53(4): 673–688*

Buchholz, A. and Sauerbrei, W. (2011): Comparison of procedures to assess non-linear and time- varying effects in multivariable models for survival data. *Biometrical Journal 53(2): 308–331*

Buchholz, A., Sauerbrei, W. and Royston, P. (2014): A measure for assessing functions of time-varying effects in survival analysis. *Open Journal of Statistics 4: 977–998*

Buckland, S.T., Burnham, K.P. and Augustin, N.H. (1997): Model selection: an integral part of inference. *Biometrics 53: 603–618*

Chatfield, C. (1995): Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society, Series B 158: 419–466*

Draper, D. (1995): Assessment and propagation of model selection uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B 57: 45–97*

Dunkler, D., Sauerbrei, W. and Heinze, G. (2016): Global, Parameterwise and Joint Post-Estimation Shrinkage. *Journal of Statistical Software 69: 8*

Govindarajulu, U.S., Spiegelman, D., Thurston, S.W., Ganguli, B. and Eisen, E.A. (2007): Comparing Smoothing Techniques in Cox Models for Exposure-Response Relationships. *Statistics in Medicine 26: 3735–3752*

Greenland, S. (1989): Modeling and variable selection in epidemiologic analysis. *American Journal for Public Health 79(3): 340–349*

Greenland, S. (1995): Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology (Cambridge, Mass.) 6 4: 450–454*

Hand, D.J. (2016): Editorial: 'Big data' and data sharing. *Journal of the Royal Statistical Society, Series A 179, 3: 629–631*

Harford, T. (2014): Big data: are we making a big mistake? *Financial Times*

Harrell, F.E. (2001): *Regression modeling strategies, with applications to linear models, logistic regression, and survival analysis*. Springer, NewYork

Hastie, T.J. and Tibshirani, R. (1990): *Generalized Additive Models*. Chapman & Hall, London

Hastie, T.J., Tibshirani, R. and Friedman, J. (2009): *The Elements of Statistical Learning*. 2nd edn. Springer, New York

Hingorani, A.D., van der Windt, D., Riley, R.D., Abrams, K., Moons, K.G.M., Steyerberg, E.W., Schroter, S., Sauerbrei, W., Altman, D.G., Hemingway, H. for the PROGRESS Group (2013): Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *British Medical Journal 346*

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999): Bayesian model averaging: A tutorial. *Statistical Science 14: 382–417*

IOM (Institute of Medicine) (2012): *Evolution of Translational Omics: Lessons Learned and the Path Forward*. The National Academies Press, Washington, DC

Keiding, N. and Louis, T.A. (2016): Perils and potentials of self-selected entry to epidemiological studies and surveys. *J.R.Statistical Society, Series A 2: 319–376*

Miller, A. (2002): *Subset Selection in Regression*. Taylor & Francis: Boca Raton, Florida

Rosenberg, P.S., Katki, H., Swanson, C.A., Brown, L.M., Wacholder, S. and Hoover, R.N. (2003): Quantifying epidemiologic risk factors using nonparametric regression: model selection remains the greatest challenge. *Statistics in Medicine 22: 3369–3381*

Royston, P. and Altman, D.G. (1994): Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with disc.). *Applied Statistics 43: 429–467*

Royston, P. and Altman, D.G. (1997): Approximating statistical functions by using fractional polynomial regression. *The Statistician 46: 411–422*

Royston, P., Altman, D.G. and Sauerbrei, W. (2006): Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine 25: 127–141*

Royston, P. and Sauerbrei, W. (2003): Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Statistics in Medicine 22: 639–659*

Royston, P. and Sauerbrei, W. (2004): A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine 23: 2509–2525*

Royston, P. and Sauerbrei, W. (2007): Improving the robustness of fractional polynomial models by preliminary covariate transformation: a pragmatic approach. *Computational Statistics and Data Analysis 51: 4240–4253*

Royston, P. and Sauerbrei, W. (2008): *Multivariable Model-Building - A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley.

Royston, P. and Sauerbrei, W. (2013): Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis. *Statistics in Medicine 32(22): 3788–3803*

Royston, P. and Sauerbrei, W. (2014): Interaction of treatment with a continuous variable: simulation study of power for several methods of analysis. *Statistics in Medicine 33: 4695–4708*

Royston, P., Sauerbrei, W. and Ritchie, A. (2004): Is treatment with interferon-alpha effective in all patients with metastatic renal carcinoma? A new approach to the investigations of interactions. *British Journal of Cancer 90: 794–799*

Sauerbrei, W. (1999): The use of resampling methods to simplify regression models in medical statistics. *Applied Statistics 48: 313–329*

Sauerbrei, W., Abrahamowicz, M., Altman, D.G., le Cessie, S. and Carpenter, J. on behalf of the STRATOS initiative (2014): STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. *Statistics in Medicine 33: 5413–5432*

Sauerbrei, W., Buchholz, A., Boulesteix, A.-L. and Binder, H. (2015): On stability issues in deriving multivariable regression models. *Biometrical Journal 57: 531–555*

Sauerbrei, W. and Royston, P. (1999): Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *Journal of the Royal Statistical Society, Series A 162: 71–94*

Sauerbrei, W. and Royston, P. (2007): Modelling to extract more information from clinical trials data: on some roles for the bootstrap. *Statistics in Medicine 26: 4989–5001*

Sauerbrei, W. and Royston, P. (2010): Continuous Variables: To Categorize or to Model? In: Reading, C. (Ed.): *The 8th International Conference on Teching Statistics- Data and Context in statistics education: Towards an evidence based society*. International statistical Institute, Voorburg

Sauerbrei, W. and Royston, P. (2011): A new strategy for meta-analysis of continuous covariates in observational studies. *Statistics in Medicine 30(28): 3341–3360*

Sauerbrei, W., Royston, P. and Binder H. (2007a): Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine 26: 5512–5528*

Sauerbrei, W., Royston, P. and Look, M. (2007b): A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biometrical Journal, 49: 453–473*

Sauerbrei, W. and Schumacher, M. (1992): A Bootstrap Resampling Procedure for Model Building: Application to the Cox Regression Model. *Statistics in Medicine 11: 2093–2109*

Schumacher, M., Holländer, N., Schwarzer, G., Binder, H. and Sauerbrei, W. (2012): Prognostic Factor Studies. In: Crowley, J., Hoering, A. (Eds): *Handbook of Statistics in Clinical Oncology*. Third Edition, Chapman and Hall/CRC, 415–470

Shmueli, G. (2010): To explain or to predict? *Statistical Science 3: 289–310*

Tibshirani, R. (1996): Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B 58(1): 267–288*

Van Houwelingen, H.C. (2000): Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine 19: 3401–3415*

Van Houwelingen, H.C. and le Cessie, S. (1990): Predictive value of statistical models. *Statistics in Medicine 9: 1303–1325*

Van Houwelingen, H.C. and Sauerbrei, W. (2013): Cross-validation, shrinkage and variable selection in linear regression revisited. *Open Journal of Statistics 3: 79–102*

Wood, S. (2006): *Generalized Additive Models*. Chapman & Hall/CRC, New York

4 On Estimating Pricing Models from End-Consumer Internet Car-Configuration Data

Tino Fuhrmann, Marvin Schweizer, Andreas Geyer-Schulz
Information Services and Electronic Markets
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
Tino.Fuhrmann@student.kit.edu,
Marvin.Schweizer@student.kit.edu,
Andreas.Geyer-Schulz@kit.edu, and
Peter Kurz
TNS Deutschland GmbH
München, Germany
Peter.Kurz@tns-infratest.com

Abstract

In this contribution we report on our first attempts of extracting a pricing-model from an anonymous end-consumer Internet car configurator data set made available from TNS Infratest for a data mining competition of the special interest group for data analysis of the German Classification Society (GfKl e.V.) in Karlsruhe on 20.-21. November 2015. In this report, we concentrate on the simplest possible rational pricing model – a linear part-worth utility function. We introduce a new data-transformation for product configuration data in general: the elimination of "irrational" product configuration types. We combine this transformation with an elimination of configuration types which are price outliers. Our second contribution is the analysis of the null space of the pricing model in a post-processing phase to improve the interpretation of the pricing model.

Introduction and Motivation

"A product configurator is a software-based expert system that supports the user in the creation of product specifications by restricting how predefined entities (physical or non-physical) and their properties (fixed or variable) may be combined." (A. Haug [3, p. 19])

Modern product configurators are the car industry's response to increased global competition, because they enable mass customization at an industrial scale [8]: *"The customer should get what he wants, when he wants it at an attractive price."* Product configurators enable the customer to build his own product autonomously – even if the product is complex. Figure 20 shows that product configurators play a key role across several functional areas of a company: Empirical configuration data improves e.g. strategic product portfolio planning, offer generation in the operative sales process, production planning, and, last but not least, the pricing of product lines. Researchers at Sawtooth Software Inc. investigated product

DOI 10.20347/WIAS.REPORT.29 Berlin 2017

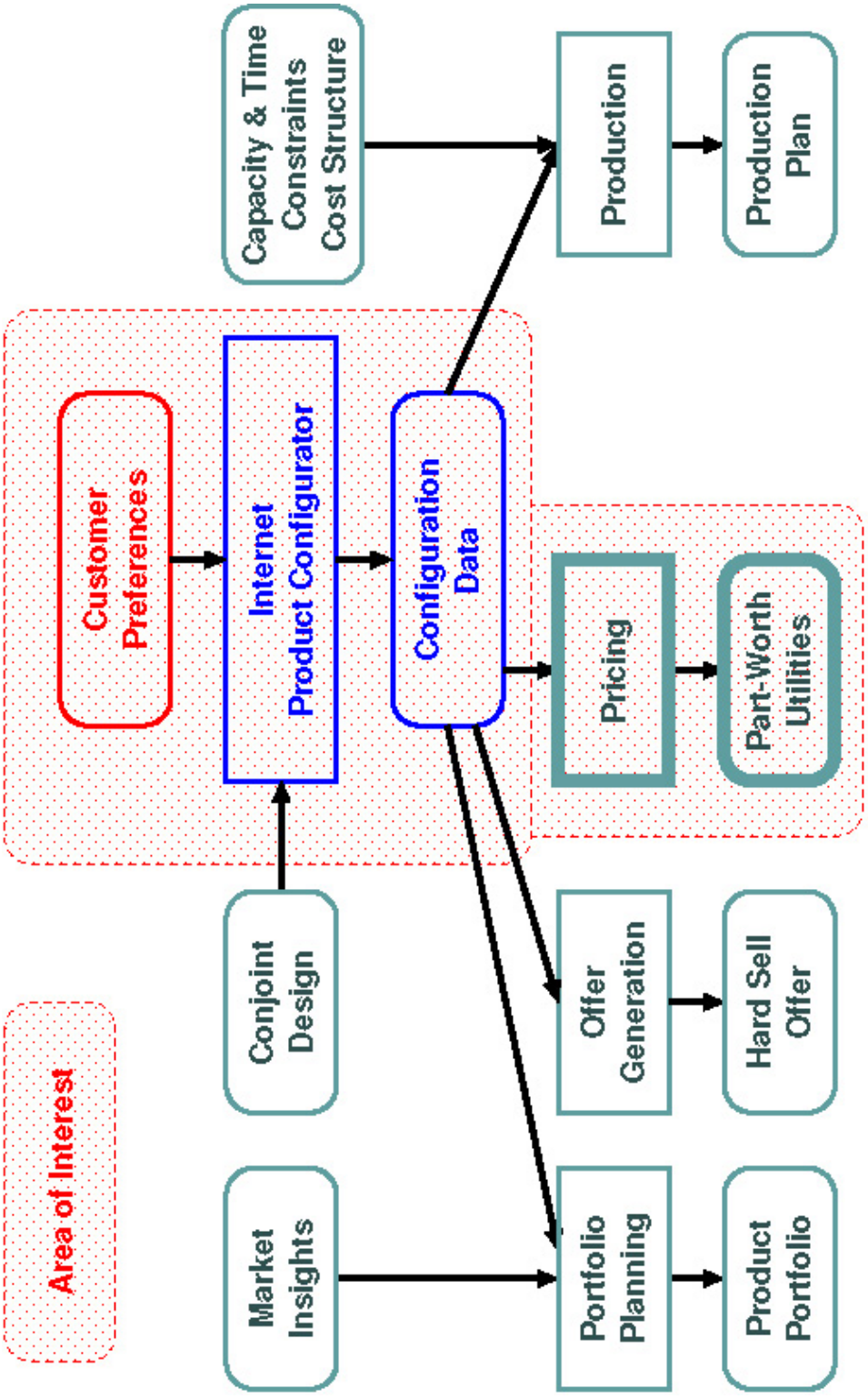


Figure 20: Pricing and Product Configurators

configurators as part of adaptive choice-based conjoint analysis as early as 2006 (see [4], [9], and [7]).

The car industry has reacted to this strategic challenge only recently. In 2013 the international benchmark study on mass customization companies of Walcher and Piller [10] did not yet contain a single car configurator. However, the Configurator Database Project (as of December 29th, 2016) listed 87 end-consumer Internet car configurators with all global players (and their major brands) present. Despite the intensive use of car configurators by the car industry, academic research on datasets of end-consumer car configurators is practically non-existent, because of the lack of publicly available datasets of this type. In a recent survey on consumer decision-making and configuration systems (see [5]), the main emphasis is on consumers' behavioral deviations from rationality and their causes.

While we may safely assume that each global player knows his own pricing models (and considers them a strategic secret), it is nevertheless interesting to investigate methods of extracting pricing models from large end-user Internet car configuration data sets and to know the limits of these methods. In addition, the assessment of the quality and information content of such Internet data sets remains an open problem.

Our contribution is structured as follows: In Subsect. 4.1 we describe the end-consumer car configuration data set used in this investigation. Next, we introduce the basics of linear part-worth utility functions and their estimation by weighted least-squares (WLS) in Subsect. 4.2. In the next two Subsects. (4.3 and 4.4) we introduce the data transformations used in preprocessing and the analysis of null space of the models used for computing a canonical model representation. We discuss first results in Subsects. 4.2, 4.3, and 4.4, respectively. In Subsect. 4.5 we discuss the results and limitations of the pre- and postprocessing transformations introduced.

4.1 The Car Configurator Data Set

The preprocessing of the original data set of TNS Infratest (collected from 473 819 respondents, 3 days from the first half of 2012 with 962 799 configurations) is described in [2] and reduces the data set by a lossless transformation to a data set of 943 (weighted) configuration types with 112 binary variables and, in addition, frequency (weight), price, line, and engine type. In the following, we use the preprocessed data set with the 112 binary attributes grouped for easier reference. Since we will concentrate on the Sports Line, we indicate all attributes which are observed in the configuration types of the Sports Line as bold:

1 6 attribute groups with mutually exclusive attributes (only one attribute in a group can be set to 1):

   1.1 4 model lines (**Sports Line**, Modern Line, Luxury Line, No Line).

   1.2 9 engine types, (**1, 2, 3, 4, 5, 6, 7, 8, 9**). We assume that engine types 1 to 4 are petrol engines, and engine types 5 to 9 are diesel engines.

1.3 12 color variants: **Hematite grey metallic, sparkling bronce metallic, alpine white, black saphire metallic, deep sea blue metallic, blue water metallic, peacock blue metallic, glacier silver metallic, orion silver metallic, mineral white metallic, black**, and **crimson red metallic**.

1.4 11 trim variants (3 observed): **Aluminum with fine longitudinal grain with accent strip in milky glass look, fine wood burr walnut with accent strip in chrome**, **aluminum with fine longitudinal grain with red accent strip**, fine wood burr walnut with black accent strip, high polish cashmere silver with accent strip in milky glass look, aluminum with fine longitudinal grain with black accent strip, fine wood fine line anthracite with intarsia and accent strip in chrome, aluminum with fine longitudinal grain and black accent strip high polish black with red accent strip, matt satin silver, and fine wood fine line porous structured with accent strip in milky glass look.

1.5 16 cushion (interior upholstery) variants (5 observed): **Fabric leather combination oyster, leather Dakota black with red contrasting seam, leather Dakota coral red with black contrasting seam, fabric Imola anthracite with red contrasting seam, leather Dakota black II**, leather Dakota Everest grey with black contrasting seam, leather Dakota Veneto beige I, leather Dakota Veneto beige II, fabric leather combination anthracite, fabric Imola anthracite with grey contrasting seam, leather Dakota oyster with contrasting seam in dark oyster, leather Dakota black I, leather Dakota saddle brown, leather Dakota black with contrasting seam in dark oyster, fabric Salome saddle brown anthracite, fabric anthracite.

1.6 24 rim variants (5 observed): **17 inch alu basis II, 17 inch alu sport II, 17 inch alu luxury II, 18 inch alu sport III, 18 inch alu luxury III**, 18 inch alu basis II, 18 inch alu luxury I, 17 inch alu sportI, 18 inch alu modern III, 17 inch alu modern II, 18 inch alu basis I, 17 inch alu luxury I, 17 inch alu basis III, 16 inch alu basis II, 18 inch alu modern I, 18 inch alu sport I, 18 inch alu luxury II, 16 inch alu basis I, 17 inch alu modern I, 18 inch alu sportII, 18 inch alu basis III, 16 inch steel basis, 17 inch alu basis I, and 18 inch alu modern II.

The attributes **model line** and **engine type** are used as a priori segmentation attributes for identifying iso-price segments of configuration types.

41 attributes of of the 76 binary attributes in this group are not observed for configuration types of the Sports Line.

2 36 attributes which can be combined (any subset of attributes can be set to 1) structured as follows:

2.1 4 packages: sport, **comfort, storage,** and **light interieur**.

2.2 2 types of transmission: **four wheel drive** and **automatic transmission**.

2.3 8 driving assistants: **cruise control with braking function, cruise control with stop go function, parking assistant, rear view camera, lane change warning, lane departure warning, road sign recognition,** and **head up display**.

2.4 8 attributes for steering, light, and chassis: **adaptive chassis with lowering**, sport leather steering wheel, **variable sports steering, performance leather steering wheel, xenon light, adaptive cornering light, glass sunroof,** and **sun protection blind**.

2.5 9 attributes for convenience, security, etc.: **seat heating for front seats**, sports seats for front seats, **electric seat adjustment, lumbar support for front seats, climate control, alarm system, arm rest for front seats, comfort access,** and **hitch**.

2.6 5 attributes for navigation, media, and communication: **navigation system business, hifi system, dvd changer, mobile phone prep with bluetooth usb,** and **digital radio**.

The 3 attributes sports package, sport leather steering wheel, and sports seats for front seats are not configured in the configuration types of the Sports Line.

Configuration types of the Sports Line have 68 binary attributes, 35 belong to the 6 groups of mutually exclusive attributes. For four groups of these attributes (Color, Rims, Cushions, Trims) we know the part-worths from the setup of a conjoint experiment partially contained in the data, but we do not use them. The second group of attributes contains 33 attributes which can be combined. For the second we do not know the part worths. The technical constraints of the car configurator are unknown.

4.2 Estimating a Linear Part-Worth Utility Function

The theory of choice in micro-economics and statistical utility theory formalize a general, axiomatic and normative model how rational decision-makers should act. Rational behavior is captured by the axioms of expected utility theory (EUT) introduced by John von Neumann and Oskar Morgenstern in 1944 [6, Chapter 3, pp. 15-31] and compatible with linear utility functions.

The simplest rational pricing model is a linear (part-worth) utility function $U(C)$:

$$U(C) = pw_0 + \sum_{c_j \in C} pw_j \cdot c_j$$

where the constant pw_0 is the part-worth (base price) of the configuration, C denotes the set of attributes describing the configuration and $c_j \in \{0,1\}$ the j-th attribute in C and pw_j the part-worth of the j-th attribute. Under the assumptions that the base price pw_0 is for a

car configuration without configured attributes and that the presence of the j-th attribute in a configuration ($c_j = 1$) is more valuable than its absence ($c_j = 0$), all part-worths should be positive: $pw_j \geq 0, \quad \forall j$. We assume that $U(C)$ at least equals the *price* a consumer is willing to pay for a car with configuration C: $U(C) = price$.

For the estimation of the part-worth utilities and the base price(s) of car configurations from the data set we use the following linear regression model:

$$\mathbf{price} = \mathbf{C} \cdot \mathbf{pw} + \mathbf{u}$$

where the dependent variable **price** is an $N \times 1$ vector, $\mathbf{C}$ is an $N \times J$ regression matrix (each line represents a car configuration), **pw** is the $J \times 1$ parameter vector (of part-worths), and **u** is an $N \times 1$ vector, N is the number of car configurations, J the number of boolean attributes of a car configuration. $\mathbf{C}_{i.}$ denotes the i-th line of $\mathbf{C}$ and is a $1 \times J$ vector. Since we concentrate only on car configurations of the Sports Line, there are 5 attribute groups with mutually exclusive attributes. We suppress the constant and this implies that we have one default configuration for each engine (the most important attribute). In each of the 4 attribute groups color, interior upholstery, trims, and rims one variable must be configured. This implies that we can only estimate the part-worths of $n-1$ attributes in an attribute group of n attributes. The last attributes are part of the default configuration. We use a completely specified model, because we want to extract as much information from the dataset as possible. However, this approach implies that $C^T C$ is not of full rank, because some attributes are linear dependent and others are not observed. We deal with this complication in Subsect. 4.4.

However, by moving from car configurations to car configuration types whose number we represent as T, we can reduce the computational effort considerably (by three orders of magnitude) because $T << N$ for our dataset. This implies that we move from minimizing the residual sum of squares

$$RSS(\mathbf{pw}) = \sum_{i=1}^{N} (\mathbf{price}_i - \mathbf{C}_{i.} \cdot \mathbf{pw})^2$$

of car configurations to minimizing the weighted sum of squares of car configuration types

$$WSS(\mathbf{pw}, \mathbf{W}) = \sum_{i=1}^{T} \mathbf{W}_{ii} (\mathbf{price}_i - \mathbf{C}_{i.} \cdot \mathbf{pw})^2$$

with $\mathbf{C}$ now representing the car configuration types and $\mathbf{w}_{ii}$ (a diagonal element of the diagonal weight matrix $\mathbf{W}$) the number of times the i-th car configuration type has been observed in the data set. This simply means, we solve the weighted least squares problem $(\mathbf{price} - \mathbf{C} \cdot \mathbf{pw})^T \mathbf{W} (\mathbf{price} - \mathbf{C} \cdot \mathbf{pw})$ by

$$\hat{\mathbf{pw}} = (\mathbf{C}^T \mathbf{W} \mathbf{C})^{-1} \mathbf{C}^T \mathbf{W} \cdot \mathbf{price}$$

Note, in this contribution, we use weighted least squares for parameter estimation in order to replace the computation of the $\mathbf{C}^T\mathbf{C}$ matrix for car configurations by the computation of the $\mathbf{C}^T\mathbf{W}\mathbf{C}$ matrix of configuration types to reduce the computation effort. We do not try to deal with heteroscedasticity by reweighting as suggested e.g. in [1, Chap. 4.5] and [11].

4.3 Preprocessing: The Elimination of Irrational and of Price Outlier Configuration Types

4.3.1 The Elimination of Irrational Configuration Types

But are end-consumers designing their own car in a rational manner? Obviously not, as the comparison of the attributes of two configuration types of the iso-price segment in Table 2 shows.

Table 2: The Configuration Types for Sports Line, Engine 2 of the Iso-Price Segment at 35 300 Euro: B a subset of A

Configuration Type	A	B
Color:	Orion Silver Metallic	Orion Silver Metallic
Rims:	17 Inch Alu Sport II	17 Inch Alu Sport II
Cushions:	Fabric Imola Anthracite with Red Contrasting Seam	Fabric Imola Anthracite with Red Contrasting Seam
Trims:	High Polish Black with Red Accent Strip	High Polish Black with Red Accent Strip
	parking assistant	parking assistant
	lane change warning	lane change warning
	dvd changer	dvd changer
	xenon light	

Iso-price segments are defined by choices between car configurations of the same model line and engine type with the same price under the assumption that an attribute configured adds value to a car configuration. The comparison of the attribute sets of the configurations in an iso-price segment allows us to analyze deviations from rationality, because of the axiom that a consumer always prefers more (the value provided by an additional attribute) to less. In the whole data set, 17% of the consumers have configured car configurations which are proper subsets in an iso-price segment. We call these configurations *irrational*.

When estimating rational pricing models from product configuration data, the elimination of irrational configurations is – as far as we know – a new data transformation which takes care of irrational behavior. Figure 21 shows a first, naive filter algorithm for implementing this data transformation.

1 For each iso-price segment in data set do

1.1 Perform a subset comparison operation between all pairs of configuration types in an iso-price segment and build a list of all subset configuration types found.

1.2 Flag all configuration types which are proper subsets as irrational.

2 Delete all irrational configuration types from data set.

Figure 21: A Naive Filter Algorithm for the Elimination of Irrational Configurations

This algorithm identifies 91 configuration types of the 416 configuration types (with 220 514 configurations) of the Sports Line and leaves a total of 325 rational configuration types (with 179 545 configurations (81%)). The effects of this transformation on the weighted residuals of a linear path worth utility model can be seen in line 3 of Table 3 and in the 3rd boxplot of Fig. 22 on the right hand side, both labelled *Rational*.

4.3.2 The Elimination of Price Outlier Configuration Types

It is well known that linear regression results are sensitive to outliers. The boxplot of configuration prices of Fig. 22 shows that all configuration types with a configuration price higher than 55000 Euro should be considered as outliers. By checking the residual errors of the configuration types we have verified that the price outliers are also the outliers in the boxplot of the weighted residuals.

Elimination of all configuration types with a price above 55000 Euro should improve the estimates of the linear part-worth utility function. The effects of this transformation on the weighted residuals of a linear path worth utility model can be seen in line 2 of Table 3 and in the 2nd boxplot of Fig. 22 on the right hand side, both labelled *No Outliers*.

4.3.3 The Effects of the Transformations on Weighted Residuals

Figure 22 on the right hand side and Table 3 allow us to compare the effects of the two data transformations and their joint effect on the residuals and the weighted residuals of the linear part-worth utility functions of Subsect. 4.2. We see that the joint effect of both data transformations eliminates most of the outliers of the residuals and leads to a more symmetric distribution of the residuals.

4.4 Postprocessing: Analyzing the Null Space of the Model

Unfortunately, not all parameters of a linear part-worth utility function can be estimated. In R, these parameters are flagged with NA (Not Available). We distinguish the following cases:

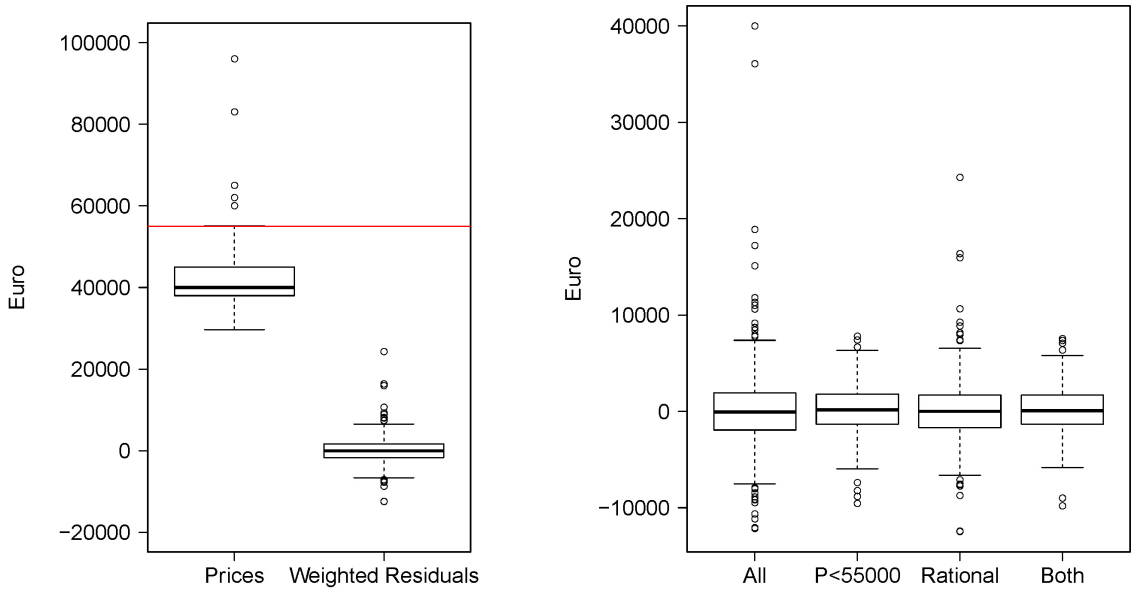


Figure 22: Boxplot of Configuration Prices and Residuals (left). Outliers have prices above $55\,000$ Euros. Boxplot of Residuals after Transformations (right).

Table 3: Effects of Transformations on Weighted Residuals of Sports Line Configuration Types

Configuration Types	Min	1Q	Median	3Q	Max
All	-385 345	-36 153	-1 145	36 614	768 514
No Outliers	-282 421	-26 019	2 973	34 423	205 543
Only Rational	-396 824	-34 695	0	37 919	517 244
Both: No Outl. & Only Rat.	-287 586	-26 993	1 523	30 758	181 377

1 Some attributes of a car configuration type of a line have not been selected by consumers. These attributes remain unobserved. The $\mathbf{C}^T\mathbf{W}\mathbf{C}$ matrix does not have full rank and these attributes form one part of the null space of the model. In addition, in our data set the unobserved attribute j has the property that $\sum_{i=1}^{T} \mathbf{C}_{i,j} = 0$. For configuration types of the Sports Line we have identified 44 attributes of this type which have been reported in Subsect. 4.1.

2 The rest of the null space are attributes which are linear dependent on other attributes. The structure of this linear dependency must be analyzed completely. We treat this case in the following.

Mathematically, the existence of linear dependent attributes implies that the weighted least squares problem does not have a unique solution, but a set of equivalent solutions exist. The complete set of solutions can be represented completely as a canonical basis together with a set of linear change of basis operators. Equivalent means equivalent with regard to the optimization criterium of the regression problem.

For product configuration data not all attributes in a group of mutually exclusive attributes can be identified: At least one attribute of such a group must be configured in each configuration and, therefore, a linear dependency with the constant of the regression model exists. The pricing model's constant is interpreted as the price of a default configuration for which the default attributes (one for each group of mutually exclusive attributes) which we can not estimate are set. The prices of the other attributes in such a group indicate the cost of replacement of the default attribute by the other attributes of the group. The signs of these relative prices depend on the choice of the default configuration. In the car configuration data set, 6 groups of such mutually exclusive attributes exist: model lines, engine types, colors, interior upholstery, trims, and rims.

In order to make part-worth utilities easily interpretable and comparable, we define a **canonical product configuration** as the configuration type with a set of mutually exclusive (must-be-configured) default attributes of lowest price. From a mathematical point of view, the canonical product configuration is the canonical basis. Relative to the default attribute, all other part-worths of attributes of such a group of mutually exclusive attributes are always positive.

Weighted linear regression in R as implemented by `lm` uses a deterministic algorithm which assigns variables to the basis in the sequence in which they are listed in the model specification of `lm`. We start with a regression model specification with the independent variables in arbitrary order. For each group of mutually exclusive variables, we check the signs of the parameters. If negative signs exist, we choose the variable with the most negative parameter and we exchange this variable with the last variable of the group.

For example, for the color attributes, we get the parameters shown in Table 4: **Crimson Red Metallic** is the color of the default car configuration. Only one color attribute (**Mineral White Metallic**) is slightly significant. And **Deep Sea Blue Metallic** is the color attribute with the most negative value.

Table 4: Estimation of Parameters for Color Attributes. Significance Code: . = 0.1.

Attribute	β	Std. Error	t-value	$P(>\mid t \mid)$	Sign.
HematiteGreyMetallic	324.85	1 041.72	0.312	0.755420	
SparklingBronceMetallic	-100.24	1 032.38	-0.097	0.922725	
AlpineWhite	635.43	1 003.40	0.633	0.527129	
BlackSaphireMetallic	44.73	1 089.34	0.041	0.967280	
DeepSeaBlueMetallic	-1 043.27	1 042.96	-1.000	0.318120	
BluewaterMetallic	673.50	1 114.82	0.604	0.546298	
PeacockBlueMetallic	743.47	1 392.23	0.534	0.593801	
GlacierSilverMetallic	1 776.48	1 164.84	1.525	0.128485	
OrionSilverMetallic	-844.81	1 094.14	-0.772	0.440765	
MineralWhiteMetallic	2 525.75	1 467.92	1.721	0.086541	.
Black	944.33	1 406.65	0.671	0.502622	
CrimsonRedMetallic	NA	NA	NA	NA	

To obtain the canonical parameters of the color attributes we moved the attribute **Deep Sea Blue Metallic** to the last position of the color attributes. Compare the parameter estimates of the color attributes shown in Table 4 with the canonical solution shown in Table 5 and observe how signs and significance of the part-worths change.

However, linear dependencies can be more complicated: For the group of rims, we have discovered three groups of linear dependencies by permutation of the model specifications: For all configuration types of engines 3, 4, 7, 8, and 9 only the rim X18InchAluLuxury has been selected and is linear dependent on the engine attribute. For engines 2 and 6, only the rims X17InchAluLuxuryII and X17AluBasisII have been selected and they are linear dependent. The same dependency exists between the rims X18InchAluSport III and X17InchAluSport II for engines 1 and 5.

At the moment, we have only analyzed the linear dependencies of the mutually exclusive attributes.

4.5 The Canonical Model After Both Transformations

The canonical model (and all equivalent models) are highly significant and explain more than 99 percent of the variance: The residual standard error is 59940 on 253 degrees of freedom (DF), R^2 is 0.997 and the adjusted R^2 is 0.996. The F-statistic is 1361 on 61 and 253 DF with a p-value less than $2.2e-16$.

The 9 canonical default configurations (one for each engine type) of the Sports Line have the color **Deep Sea Blue Metallic**. Their interior upholstery is **Fabric Leather Combination Oyster** with trims configured as **Aluminium with Fine Longitudinal Grain with Accent**

Strip in Milky Glass Look. Rims differ between engines: For engines 1 and 5, we have **X18InchAluSport III**, for engines 2 and 6, **X17InchAluLuxuryII** and for engines 3, 4, 7, 8, 9: **X18InchAluLuxury**. These attributes are the non-identified attributes of the canonical car configuration. The prices of the canonical default configurations are typeset in bold in Table 5. They range from 30367 Euro for the default configuration of engine 1 to 47218 Euro for the default configuration of engine 9.

The parameter estimates of the part-worth utilities of the canonical model for the attributes with mutually exclusive attributes are shown in column Both of Table 5.

The estimates for all other attributes are shown in column Both of Table 6. In the attribute groups of *Driving Assistants* and *Convenience, Security, ...* we find 10 attributes of the 12 attributes with negative signs. This indicates that the model of a simple linear part-worth utility function does not completely explain the unknown pricing strategy embedded in the product configurator and that further analysis is required.

Conclusion

In this contribution we have presented the preprocessing method of the elimination of irrational configuration types (without reweighting) for product configuration data sets. In addition, we have shown that a partial recovery of a pricing model from product configuration data is possible with the restriction that one attribute of each group of mutually exclusive attributes can not be estimated for regression models whose constants capture the price of the default configuration. In addition, we have made progress in the analysis of the null space of regression models for complex product configuration data: We have introduced the concept of a canonical configuration as the least price configuration (in the sense that its default attributes have the lowest price in their group of mutually exclusive attributes) and we have shown how this configuration can be found with the help of permutations of the model specification. A potential improvement for the elimination of irrational configuration types is finding a proper reweighting scheme of rational configuration types.

References

[1] Cameron, A. C. and Trivedi, P. K. (2005): Weighted Least Squares. In: Microeconometrics. Methods and Applications. Cambridge University Press, 81–85.

[2] Fuhrmann, T., Schweizer, M., Geyer-Schulz, A. and Kurz, P. (2016): Mining Consumer-Generated Product-Configuration Data. *Archives of Data Science, Series A*, forthcoming.

Table 5: Canonical Parameter Estimation (CPE) of Part-Worth Attribute Utilities for Sports Line's Configuration Types. The 6 attribute groups with exclusive attributes. Prices of default configurations in bold. Significance Codes (only model Both): $***=0.001$, $**=0.01$, $*=0.05$, $.=0.1$.

Topic	Attributes	All	Rational	$P<55000$	Both	Sign.
Engines	Engine 1	30 444	30 333	30 072	**30 367**	***
	Engine 2	33 620	33 273	33 022	**33 458**	***
	Engine 3	40 187	39 575	39 223	**39 377**	***
	Engine 4	43 514	43 398	43 483	**43 352**	***
	Engine 5	33 680	32 890	33 360	**33 535**	***
	Engine 6	34 160	34 017	34 252	**34 602**	***
	Engine 7	39 578	39 121	38 123	**38 841**	***
	Engine 8	46 456	45 500	43 413	**44 091**	***
	Engine 9	54 116	52 006	46 426	**47 218**	***
Color	DeepSeaBlueMetallic	NA	NA	NA	NA	
	OrionSilverMetallic	-754	-620	240	198	
	SparklingBronceMetallic	865	542	1 288	943	
	CrimsonRedMetallic	-1 414	-1 105	976	1 043	
	BlackSaphireMetallic	1 102	865	1 360	1 088	
	HematiteGreyMetallic	943	1 096	1 398	1 368	*
	AlpineWhite	2 421	2 086	2 132	1 679	**
	BluewaterMetallic	4 703	2 535	2 137	1 717	*
	PeacockBlueMetallic	2 231	1 627	2 365	1 787	
	Black	2 798	2 132	2 197	1 988	
	GlacierSilverMetallic	3 131	2 799	3 421	2 820	**
	MineralWhiteMetallic	1 886	1 379	4 411	3 569	**
Interior	Fabric Leather Combination Oyster	NA	NA	NA	NA	
	Leather Dakota (LD) Black II	982	1 386	424	460	
	LDB with Red Contrasting Seam	-186	216	371	480	
	LD Coral Red with Black Contrasting Seam	959	1 493	1 347	1 498	***
	Fabric Imola Anthracite with Red Contrasting Seam	1 811	1 963	2 510	2 555	**
Trims	Aluminum with Fine Longitudinal Grain (AFLG) with Accent Strip in Milky Glass Look	NA	NA	NA	NA	
	ALFG with Red AccentStrip	-1 228	-479	-31	146	
	Fine Wood Burr Walnut with Accent Strip in Chrome	-13	-33	670	568	
Rims	X17InchAluLuxuryII	NA	NA	NA	NA	
	X17InchAluBasisII	2 638	2 121	1 885	1 719	***
	X18InchAluSportIII	NA	NA	NA	NA	
	X17InchAluSportII	2 293	2 025	1 244	1 414	.
	X18InchAluLuxuryIII	NA	NA	NA	NA	

Table 6: CPE of Part-Worth Attribute Utilities for Sports Line's Configuration Type. Attribute Combinations. Significance Codes (only model Both): $***=0.001$, $**=0.01$, $*=0.05$, $.=0.1$.

Attributes	All	Rational	$P<55000$	Both	Sign.
Packages					
Storage package	679	1 492	42	348	
Comfort package	782	1 545	759	1 156	**
Light package interior	3 394	3 297	835	1 452	**
Transmission					
Automatic transmission	1 346	1 402	825	1 060	.
Four wheel drive	3 316	1 878	2 049	1 450	*
Driving Assistants					
Head up display	-6 456	-4 151	-3 319	-2 432	*
Rear view camera	-525	-1 145	-1 898	-1 858	**
Lane change warning	-2 356	-3 119	-1 720	-1 553	.
Cruise control with stop go function	-184	839	-801	-605	
Cruise control with braking function	1 442	139	5	-578	
Parking assistant	398	99	342	-31	
Road sign recognition	-329	857	943	674	
Lane departure warning	2 422	1 326	3 391	2 895	**
Steering, Light, Chassis, ...					
Variable sports steering	-4 742	-4 525	-1 696	-2 255	**
Sun protection blind	8 343	7 206	-469	35	
Xenon light	901	748	617	365	
Performance leather steering wheel	1 409	1 495	945	1 109	**
Glass sunroof	1 105	1 832	1 064	1 471	**
Adaptive cornering light	-672	213	1 621	1 588	**
Adaptive chassis with lowering	1 554	91	2 707	1 669	**
Convenience, Security, ...					
Lumbar support for front seats	-622	-1 302	-1 545	-1 158	*
Electric seat adjustment	186	-466	-755	-777	
Alarm system	283	394	-531	-271	
Seat heating for front seats	-520	97	-616	-247	
Comfort access	-466	-608	177	57	
Arm rest for front seats	-109	-394	430	222	
Climate control	1 348	827	800	589	
Hitch	713	1 817	1 541	2 412	***
Navigation, Media, and Communication					
Hifi system	-415	46	-72	-122	
Digital radio	2 084	2 272	58	14	
Mobile phone prep with bluetooth usb	-1 071	-1 312	502	84	
Navigation system business	770	775	1 362	1 058	**
DVD changer	2 653	3 346	2 061	2 224	***

[3] Haug, A. (2007): Representation of Industrial Knowledge – as a Basis for Developing and Maintaining Product Configurators. PHD Thesis, Department of Manufacturing Engineering & Management, Technical University of Denmark. Lyngby.

[4] Johnson, R., Orme, B. and Pinnell, J. (2006): Simulating Market Preference with Build Your Own Data. In: Sawtooth Software, Inc. (Ed.): *Proceedings of the Sawtooth Software Conference 2006*, vol. 12, Sequim, Washington, 239–253.

[5] Mandl, M., Felfernig, A. and Teppan, E. (2014): Consumer Decision-Making and Configuration Systems. In: Felfernig, A., Hotz, L., Bagley, C. and Tiihonen, J. (Eds.): *Knowledge-Based Configuration: From Research to Business Cases*, Morgan Kaufman, Waltham, 181–190.

[6] Morgenstern, O. and Neumann, J. von (1990): *Theory of Games and Economic Behavior*. Princeton Univ. Press, Princeton.

[7] Orme, B. K. and Johnson, R. M. (2008): Testing Adaptive CBC: Shorter Questionnaires and BYO vs. Most Likelies. Tech. rep., Sawtooth Software, Inc., 530 W. Fir St. Sequim, WA 98382.

[8] Pine, B. J. (1999): *Mass Customization: The New Frontier in Business Competition*. Harvard Business School Press, Harvard.

[9] Rice, J. and Bakken, D. G. (2006): Estimating Attribute Level Utilities from Design Your Own Product Data. In: Sawtooth Software, Inc. (Ed.): *Proceedings of the Sawtooth Software Conference 2006*, vol. 12, Sequim, Washington, 229–238.

[10] Walcher, D. and Piller, F. (2013): *The Customization 500 – An International Benchmark Study on Mass Customization*. Lulu Inc., Raleigh.

[11] White, H. (1980): A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, **48**(4), 817–838.

5 GenPCCA: Markov State Models for Non-Equilibrium Steady States

Konstantin Fackeldey
Technical University Berlin
Strasse des 17 Juni 135, 10623 Berlin, Germany
fackeldey@zib.de
and
Marcus Weber
Zuse Institute Berlin (ZIB)
Takustr. 7, 14195 Berlin, Germany
weber@zib.de

Abstract

For equilibrium systems Markov State Models (MSM) are a powerful tool for grouping states according to a metastability criterion. Given a reversible Markov chain, in MSM the eigenvalue structure of the underlying Markov chain is exploited for detecting metastable sets, such that the dynamics of a system in a high dimensional space can be described by the entries of a small transition probability matrix. Considering Non-Equilibrium Steady States the underlying Markov chain is no longer reversible and thus the eigenvalue structure, being the backbone for MSM can no longer be employed. To overcome this, we present a novel MSM method (GenPCCA) being capable to find a low dimensional description of even non reversible Markov processes by using a Schur decomposition instead of using eigen vectors. We show the performance of GenPCCA on networks for gene expression.

Introduction

In a Markov State Model (MSM) an underlying stochastic process is described by a transition matrix between clustered states. Under the assumption, that the stochastic process has a metastable behavior, i.e. the system stays mostly in a metastable set of states and switches only rarely between these sets, the stochastic process can be described by a low dimensional basis. In literature several concepts for the description of the clusters can be found (e.g. Sarich and Schütte (2014), Bowman et al (2014)). The coarse graining procedure of the stochastic process onto clustered states can also be interpreted as a Galerkin projection. One fundamental assumption in MSM is that the stochastic process is reversible allowing for spanning the low dimensional space of the clusters by eigen vectors. It has been understood, that this assumption is quite realistic for equilibrium states, i.e. states in a thermal equilibrium. However, for Non-equilibrium steady states (NESS) is non-reversible.

In Fill (1991) the eigenvalue bounds of the mixing rates for reversible Markov chains have been extended to non- reversible chains by reversiblizing the non-reversible matrix. Based on this clustering methods for non-reversible processes Runolfsson and Ma (2007), Huisinga et al (2004) but also other approaches Jacobi (2010), Sarich and Schütte (2014) have been developed.

DOI 10.20347/WIAS.REPORT.29 Berlin 2017

In order to tackle this problem, the authors of Fritzsche et al. (2007) proposed - and further developed by Tifenbach (2011) - to replace the eigenvalue problem by a singular value decomposition and using the singular values and singular vectors. However, in Jacobi (2010) it is claimed that the singular vectors do not have the relevant sign structure to identify the metastable states, thus it is not preserving the dynamical structure of the Markov chain. Nevertheless, this method has been applied in Tjakra (2013) in the context of identifying the collective variables.

In this article we propose a novel clustering method (GenPCCA) aiming at grouping states of a Markov chain by their transition behavior on the basis of a Schur decomposition.

It turns out that this novel method offers a powerful analysis of the Markov chain which also includes the identification of coherent subsets and the freedom of regarding an arbitrary initial distribution of states. Thus this novel method covers a broader class of applications by including non reversible Markov chains. Since this method is a generalization of PCCA+ towards non-reversible processes it is named GenPCCA (Generalized Perron Cluster Analysis).

5.1 Non-Equilibrium Steady States

Let a homogeneous Markov chain in a finite state space $\Gamma = \{1, \ldots, N\}$ be given by $\{X_i, i \in \mathbb{N}\}$ with the transition matrix

$$P = (p_{ij})_{i,j=1,\ldots,N} \qquad p_{ij} = \mathbb{P}(x_{t+1} = i | x_t = j).$$

In MSM we seek for a projection $G : \mathbb{R}^N \to \mathbb{R}^n$ such that the states $(x_i)_{i=1,\ldots,N}$ are clustered into collection of metastable states $(C_\alpha)_{\alpha=1,\ldots,n}$ where

$$\mathbb{P}(C_\alpha | C_\alpha) \approx 1,$$

meaning that the system process stays long in metastable subsets C_α and rarely switches between the sets.

The dynamics of the system can then be described by a low-dimensional projection of P, i.e. a matrix

$$G(P) = P_c = (p^c_{\alpha\beta})_{\alpha(i),\beta(j)=1,\ldots,n}$$

where $n \ll N$.

In the case of a metastable Markov chain, the state space can be decomposed into metastable subsets building the low dimensional space.

In order to guarantee that P_c inherits the correct dynamical behavior of the underlying Markov chain, it has to meet the Chapman-Kolmogorov equation, i.e.

$$(G(P))^k = G(P^k). \tag{8}$$

In general, the projection $G(P)$ of a Markov chain is not Markovian does not meet the semi-group property given by (8), and thus the stochastic process induced by the $n \times n$ transition matrix $P_c = G(P)$ between the clusters is in general not a Markov process.

Markovianity can be guaranteed by claiming on the projection G:

- *invariant subspace condition:* there exists a matrix $X \in \mathbb{R}^{N\times n}$ (for a suitable choice of n) which meets
$$PX = X\Lambda \tag{9}$$
for $\Lambda \in \mathbb{R}^{n\times n}$

- *orthogonality relation*
$$X^T D_\eta X = I_{n\times n}, \tag{10}$$
where $D_\eta = \mathrm{diag}(\eta_1, ..., \eta_N)$ and $\Lambda \in \mathbb{R}^{n\times n}$, i.e. the X are spanning an n dimensional invariant subspace of P.

With the *invariant subspace condition* (9) and the *orthogonality condition* (10) the projection $G(P)$ is given by
$$G(P) = (C^T D_\eta C)^{-1}(C^T D_\eta P C) \tag{11}$$
where $C = X\mathscr{A}$ for a suitable transformation matrix $\mathscr{A}$ which we specify Section 5.2 .

In other words Conditions (9) and (10) of a projection G are sufficient for (8). We remark, that a singular valued decomposition of P does not meet (9) and consequently a Galerkin projection leads to a projection error Sarich and Schütte (2014).

The eigen vectors of the dominant eigenvalues of P, i.e. the eigenvalues close to one, which are typically used in MSM meet the invariant subspace condition and the orthogonality relation. Thus in a reversible MSM construction, the metastable subsets can be described by the span of the eigen vectors corresponding to eigenvalues close to 1.

For equilibrium statistical mechanics, Markov State Models (MSM) have celebrated quite great success. Möller-Levet (2003), Shumway (2003), Vlachos et al. (2003), Li (2001), Deuflhard (2000), Deuflhard and Weber (2005) by employing the eigenvalue structure.

However many biological phenomena can be found to relax towards a steady flux, these methods are referred to as Non-equilibrium Steady States (NESS). Typical examples are systems driven by time dependent or non-conservative external forces.

5.2 Markov State Models for Non Equilibrium Steady States

In Non Equilibrium Steady States, the system under consideration does not reach thermal equilibrium state but a steady state. This fact precludes in general MSM since then then it can no longer be guaranteed that the eigenvalues are real valued. Consequently the metastable sets can not be described by the eigen vectors of the eigenvalues close to one, since the eigenvalues can not be arranged in an order and a projection based on complex eigen vectors would lead to a complex matrix $G(P) = P_c$.

In the foregoing section the orthogonality relation in the context of eigen vectors was realized by assuming that the underlying process is reversible. By resigning the reversibility of the underlying Markov chain, an interpretation of a transition matrix in terms of unconditional transition probabilities is not possible since then the eigen vectors do not meet the invariance condition (9) and the subspace condition (10) in general. Moreover the spectrum of its corresponding transition matrix is in general not real but complex.

We thus take advantage of a Schur decomposition. Instead of using eigenvalues we employ Schur vectors. Let therefore $\widetilde{X}$ be n Schur vectors of $\widetilde{P} = D_\eta^{0.5} P D_\eta^{-0.5}$, then we have

$$\begin{aligned} & \widetilde{P}\widetilde{X} = \widetilde{X}\Lambda \\ \iff \quad & D_\eta^{0.5} P D_\eta^{-0.5}\widetilde{X} = \widetilde{X}\Lambda \\ \iff \quad & P D_\eta^{-0.5}\widetilde{X} = D^{-0.5}\widetilde{X}\Lambda \\ \iff \quad & PX = X\Lambda, \quad X = D_\eta^{-0.5}\widetilde{X}. \end{aligned} \tag{12}$$

We have thus shown, that a Schur decomposition meet the invariant subspace condition (9) and the orthogonality condition (10). Consequently $G(P)$ is given by (11) with Schur vectors X. This allows us to state the following

5.1 Theorem *Let $G(P)$ be given by* (11)*, i.e.*

$$G(P) = (C^T D_\eta C)^{-1}(C^T D_\eta P C),$$

where X are the Schur vectors according to (12) *and $C = X\mathscr{A}$ and D_η be some initial distribution of the Markov chain, then*

$$(G(P))^k = G(P^k).$$

Proof:

$$\begin{aligned} G(P) &= (C^T D_\eta C)^{-1}(C^T D_\eta P C) \\ &= (\mathscr{A}^T X^T D_\eta X \mathscr{A})^{-1}(\mathscr{A}^T X^T D_\eta P X \mathscr{A}) \\ &= (\mathscr{A}^T X^T D_\eta X \mathscr{A})^{-1}(\mathscr{A}^T X^T D_\eta X \Lambda \mathscr{A}) \\ &= (\mathscr{A}^T \mathscr{A})^{-1}(\mathscr{A}^T \Lambda \mathscr{A}) \\ &= \mathscr{A}^{-1}\Lambda\mathscr{A}, \end{aligned}$$

such that $G(P)$ meets the desired criterion:

$$(G(P))^k = (\mathscr{A}^{-1}\Lambda\mathscr{A})^k = \mathscr{A}^{-1}\Lambda^k\mathscr{A} = G(P^k).$$

□

5.2 Remark Note that the Markov chain in Theorem 5.1 does neither need to be aperiodic nor irreducible. Moreover - in contrast to the reversible case - the initial distribution η has not to be the stationary distribution. Theorem 5.1 may also be interpreted as commutativity between propagation in time (k steps) and discretization G, which is a desired property for long term predictions.

In the real Schur decomposition the matrix Λ is an upper triangle matrix with possibly 2×2-blocks on

its diagonal

$$\begin{pmatrix} a_{11} & z_1 & * & \dots & \dots & \dots & \dots & * \\ \bar{z}_1 & b_{11} & * & \dots & \dots & \dots & \dots & \vdots \\ 0 & 0 & \ddots & \dots & \dots & \dots & \dots & \vdots \\ \vdots & \dots & \dots & a_{ii} & z_i & \dots & \dots & \vdots \\ \vdots & \dots & \dots & \bar{z}_i & b_{ii} & \dots & \dots & \vdots \\ \vdots & \dots & \dots & \dots & \dots & \ddots & * & * \\ \vdots & \dots & \dots & \dots & \dots & 0 & a_{nn} & z_n \\ 0 & \dots & \dots & \dots & \dots & 0 & \bar{z}_n & b_{nn} \end{pmatrix}.$$

The remaining problem is, that an arrangement of the Schur decomposition in descending order (of eigenvalues) is no longer possible. In Brandts (2002) it has been proposed to arrange the Schur-values according to a absolute distance to a given target value z. For the reversible case $z = 1$ should be chosen, to guarantee that P_C is close to unit matrix allowing for a clustering into metastable states (the eigenvalues of P_C correspond to these selected values).

For the non reversible case, however, we can apply another method by arranging the Schur-values according to the distance from the unit circle. In this case P_C will have eigenvalues close to the unit circle and, thus, will be similar to a permutation matrix, which can be seen as a clustering of states in the sense of coherent sets Froyland and Padberg-Gehle (2014). This feature of GenPCCA will be shown in the section of illustrative examples below.

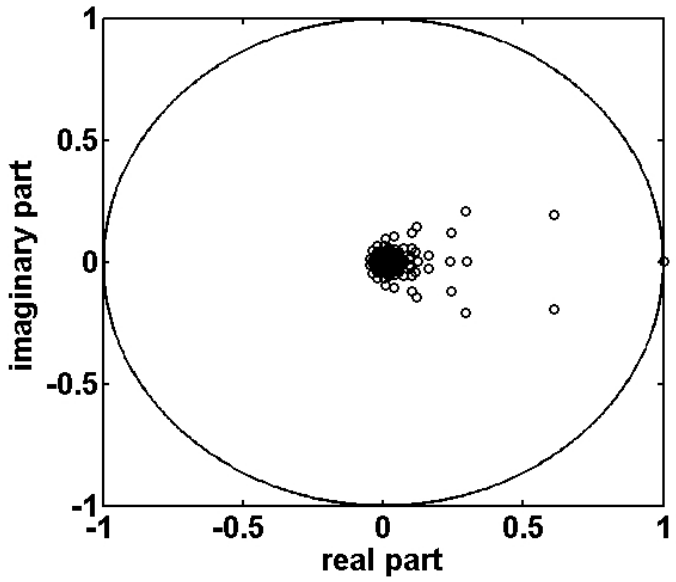


Figure 23: Spectrum of the transition matrix P of the ODE system.

So far we have not yet explained how to obtain the matrix $\mathscr{A}$ from Theorem 5.1. In the framework of GenPCCA, this step is identical to PCCA+ Weber (2005) and Deuflhard and Weber (2005). The

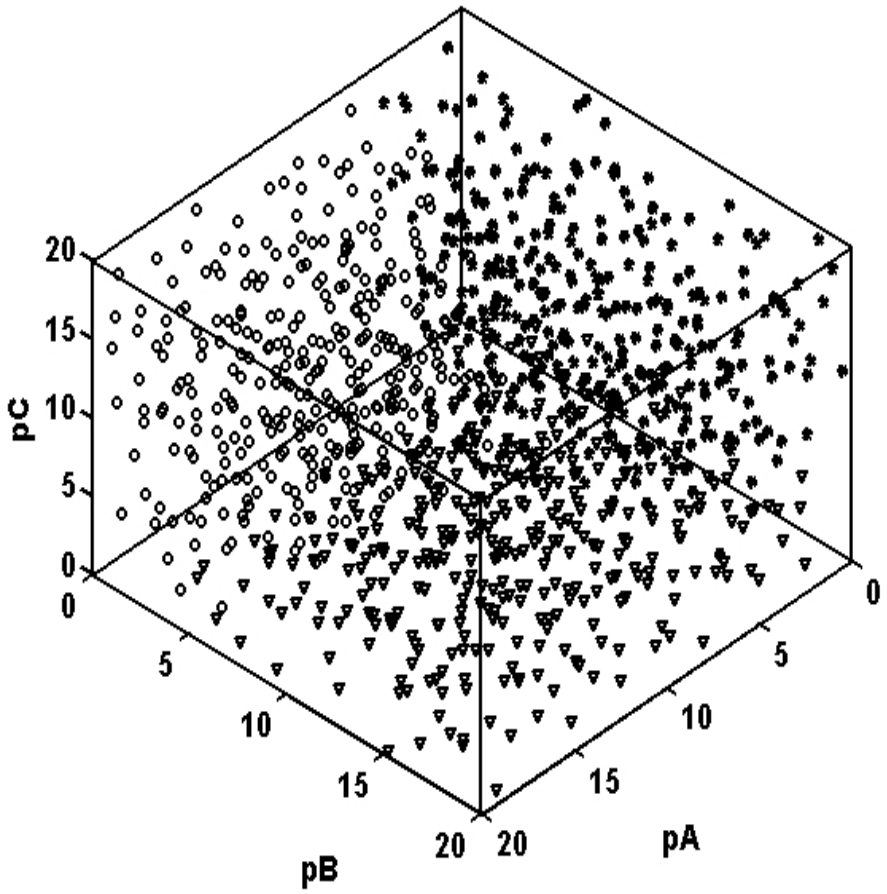


Figure 24: Coordinate system showing the concentrations of the three protein species. These are the 1000 starting points of the ODE system. Each of the thousand points is assigned to one of the three colored regions (metastable regions) by using GenPCCA. It can be clearly seen, that the points with two high (=20) and one low protein concentration (=0) are the centers of the metastable regions.

problem of finding the matrix $\mathscr{A}$ can be converted to an optimization problem. More precisely, GenPCCA finds a transformation matrix $\mathscr{A}$ mapping the column vectors of Schur vectors X, spanning the invariant subspace, to the basis $C = X\mathscr{A}$ used for the projection $G(P)$. Finding an optimal $n \times n$-basis transformation matrix $\mathscr{A}$ is the aim of this algorithm. As input the matrix X of the invariant subspace is needed. The output of GenPCCA is the above mentioned matrix of membership vectors C. The column vectors of both matrices, X and C, span the same subspace. Thus, GenPCCA provides an invariant subspace projection of P, such that the subspace spanning vectors C have an interpretation in terms of membership vectors. The selection of $\mathscr{A}$ is realized by a convex maximization problem Weber (2005), Deuflhard and Weber (2005).

5.3 Example Gene Expression

In Elowitz and Leibler (2000) a regulatory network for gene expression in Escherichia coli is proposed explaining the interplay between the genes TetR, λcI and LacI (see Elowitz and Leibler (2000) for more details). The main part of this network is the repressilator consisting of three genes, where each one of the genes produces a protein which represses the transcription (production of mRNA) of one of the other two genes. This is a typical example of a NESS due to the cyclic cascades of protein expression, where standard MSM method would fail since in the example the detailed balance condition (reversibility) is violated. However, this system meets the balanced condition, given by

$$\sum_i \pi_i P_{ij} = \pi_j.$$

such that we can use GenPCCA. The kinetics of this model can be explained by a system of six differential equations: Let us denote p_A/m_A as the concentrations of LacI, p_B/m_B as the concentrations of Tet1R and p_C/m_C as the concentrations of cI We then have three differential equations for the concentrations m_A, m_B, m_C of the mRNA and three for the concentrations p_A, p_B, p_C of the proteins,

$$\begin{aligned}
\frac{dm_A}{dt} &= -m_1 + \frac{\alpha}{1+p_C^n} + \alpha_0 & \frac{dp_A}{dt} &= -\beta(p_A - m_A) \\
\frac{dm_B}{dt} &= -m_B + \frac{\alpha}{1+p_A^n} + \alpha_0 & \frac{dp_B}{dt} &= -\beta(p_B - m_B) \\
\frac{dm_C}{dt} &= -m_C + \frac{\alpha}{1+p_B^n} + \alpha_0 & \frac{dp_C}{dt} &= -\beta(p_C - m_C),
\end{aligned} \tag{13}$$

where $\alpha = 298.2$ transcriptions per second, $\beta = 1/5$ the ratio of protein decay rate to the mRNA decay rate, $\alpha_0 = 0.03$ is the growth constant and $n = 2$ is a Hill coefficient.

For computing the system of ODEs we took a Niederreiter sequence Lidl and Niederreiter (1984), Niederreiter (1988) of 1000 starting points in the six dimensional space of the ODE system. We multiplied each component of the Niederreiter sequence with 20, to have starting values in the interval $[0, 20]$. In each of these points we simulated the ODE for 1.5 seconds (`ode45` Mathworks (2012)). For computing the entries of a stochastic transition matrix P the endpoints of each simulations have been assigned to the starting points by using an exponential assignment function such that

$$P_{ij} = \frac{\exp(-0.2 \cdot \|x_{\text{start},i} - x_{\text{end},j}\|)}{\sum_{k=1}^{1000} \exp(-0.2 \cdot \|x_{\text{start},i} - x_{\text{end},k}\|)}, \qquad i, j = 1, ..., 1000.$$

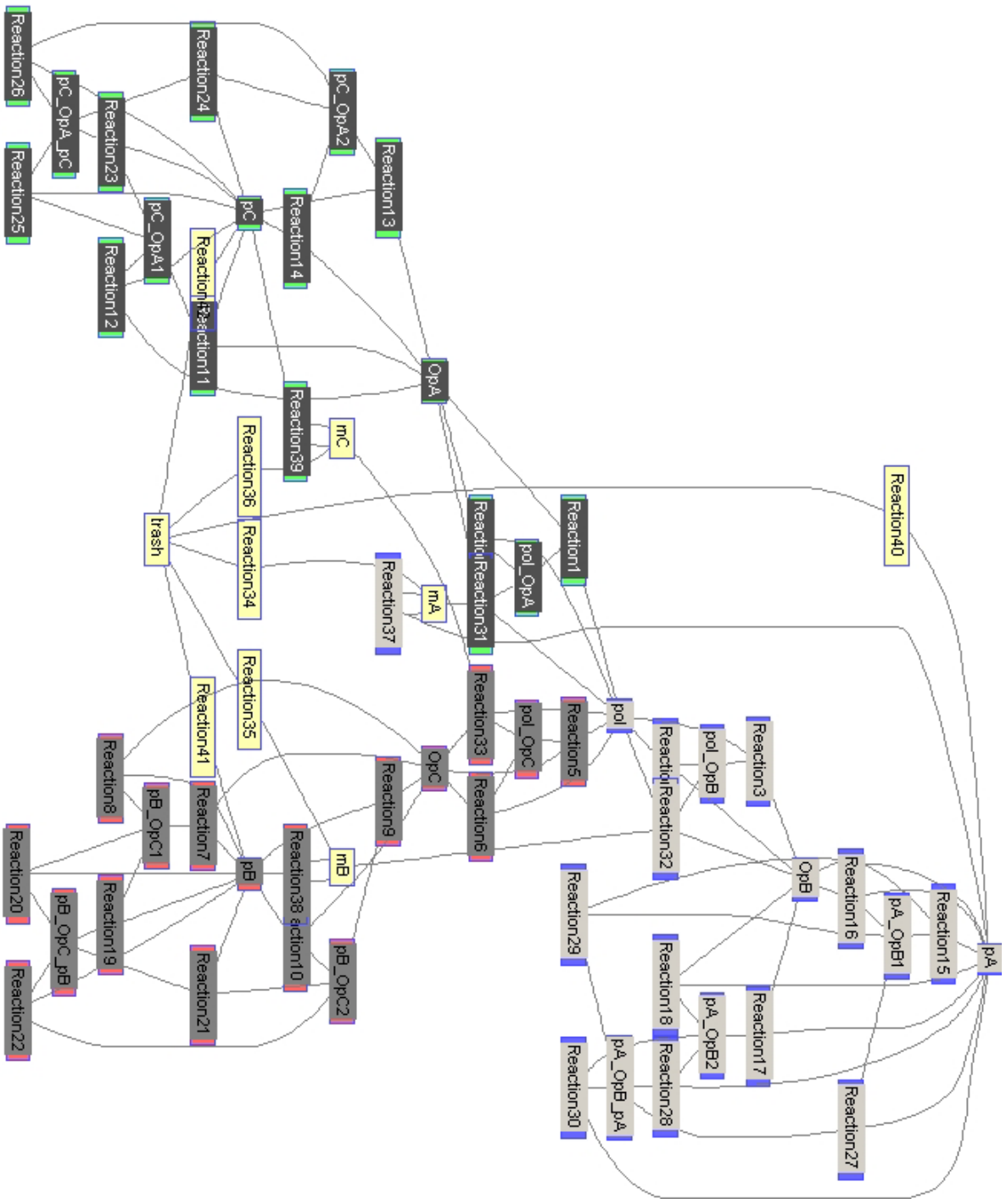


Figure 25: Clusters colored in the network.

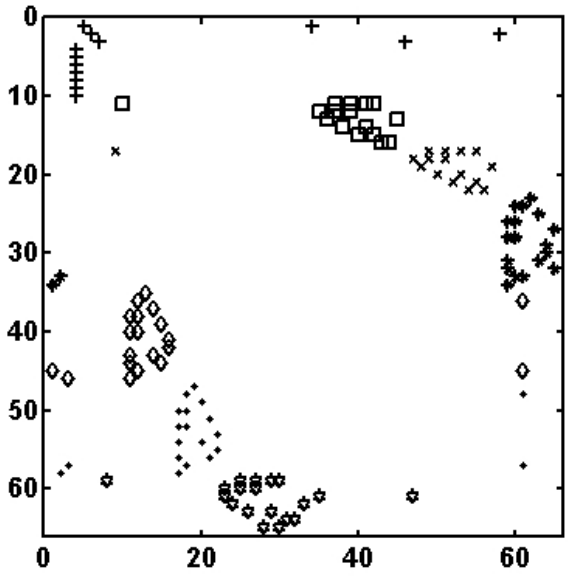


Figure 26: Permutating and coloring the non-zero entries of the adjacency matrix g according to the assignment to the seven clusters detected by GenPCCA.

Then in a next step we computed the spectrum of P (Fig.23 shows the spectrum of P of the ODE system). Since the underlying process is a NESS, the spectrum has also complex eigenvalues. The leading eigenvalues (absolute values) are $\lambda_1 = 1.0000, \lambda_2 = 0.6083 \pm 0.1959i, \lambda_3 = 0.2921 \pm 0.2081i, \lambda_4 = 0.2994$. We selected 3 eigenvalues to be the dominating ones, i.e., closest to the unit circle. Note, that despite the NESS, the Perron Frobenius Theorem for stochastic matrices still holds, such that $\lambda_1 = 1$ is clearly algebraically and geometrically simple.

Computing the corresponding Schur decomposition according to (12) where η is the stationary distribution of P, leads to a 1000×3-matrix X which has been used for the GenPCCA algorithm. The corresponding projection of P to a 3×3-transition matrix $G(P)$ is given as:

$$G(P) = \begin{pmatrix} 0.7404 & 0.2521 & 0.0075 \\ 0.0190 & 0.7330 & 0.2480 \\ 0.2318 & 0.0249 & 0.7433 \end{pmatrix}.$$

One can clearly see that the process has a cyclic structure, i.e., it is a NESS. Three different states can be identified: Approximately 75% of the transitions show a metastable behavior whereas $\approx 25\%$ enter in a cyclic manner into one of the other metastable regions. Thus, the matrix $G(P)$, can be used to detect the cyclic flow of the system. The Schur vectors can be used to identify the corresponding regions in the state space of the ODE system. The clustering of the starting points is shown in Fig. 24.

The above ODE system introduced by Elowitz and Leibler (2000) can also be considered as a reaction network. This kind of modeling is adopted in the SimBio Toolbox of MATLAB Mathworks (2012). The package `oscillograph` includes a 65×65-adjacency matrix g which corresponds to the edges of the reaction network (cp. Fig. 25).

Besides the above gene expression reactions, a node 'trash' has been introduced, representing the dissipated entities. The matrix P is generated by normalizing the matrix g via rescaling its rows. Using GenPCCA with the matrix of the leading 7 Schur values, we built the 7×7-matrix $G(P)$. The corresponding seven clusters are mapped to the non-zero entries of g (cp Fig 26). GenPCCA obvioulsy detected the four parts of the reaction network which belong to TetR, λcI, LacI and the trash (cp Fig 25). Additionally, GenPCCA can separate the products from the reactions in the protein production parts (cp. colored adjacency matrix in Fig. 26). Note that in this case P has a ßinkrepresenting a non-equilibrium process disallowing to use conventional MSM.

Conclusions

In contrast to existing MSM our novel method does neither need a reversible stochastic process nor the stationary density of the process in order to meet the invariance (9) and orthogonality condition (10). GenPCCA thus broadens the area of applications of MSM.

Acknowledgment. The work of Marcus Weber has been done for the DFG Collaborative Research Center 765.

References

Bowman, G., Pande, V. and Noé, F. (2014): *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, Springer Berlin Heidelberg

Brandts, H.J. (2002): Matlab code for sorted real schur forms. *Num Lin Alg App, 9(3):249–261*

Deuflhard, P.,Huisinga, W., Fischer, A. and Schütte, C. (2000): Identification of almost invariant aggregates in reversible nearly uncoupled markov chains. *Linear Algebra and its Applications, 315(1-3):39 – 59*

Deuflhard, P. and Weber, M. (2005): Robust perron cluster analysis in conformation dynamics. *Linear Algebra and its Applications, 398(0):161 – 184* Special Issue on Matrices and Mathematical Biology

Elowitz, M. and Leibler, S..(2000): A synthetic oscillatory network of transcriptional regulators. *Nature, (403):335–338*

Fill, J. (1991): Eigenvalue bounds on convergence to stationarity for nonreversible markov chains, with an application to the exclusion process. *Ann. Appl. Probab., 1(1):62–87*

Fritzsche, D., Mehrmann,V. Szyld,D.B. and Virnik, E. (2007): An SVD approach to identifying metastable states of Markov chains *ETNA, 29:46–69*

Froyland, G. and Padberg-Gehle, K. (2014): Almost-invariant and finite-time coherent sets: directionality, duration, and diffusion. *Ergodic Theory, Open Dynamics, and Coherent Structures, volume 70 of Proceedings in Mathematics and Statistics, pages 171–216*

Huisinga, W., Meyn,S. and Schütte, C. (2004): Phase transitions and metastability in markovian and molecular systems. *Ann. Appl. Probab., 14(1):419–458*

Jacobi,M.N. (2010): A robust spectral method for finding lumpings and meta stable states of non-reversible Markov chains *ETNA, 37:296–306*

Li,C.,Biswas,G.,Dale,M. and Dale,P. (2001): Building models of ecological dynamics using hmm based temporal data clustering - a preliminary study *Advances in Intelligent Data Analysis volume 2189 of Lecture Notes in Computer Science, pages 53–62. Springer Berlin Heidelberg*

Lidl R. and Niederreiter, H. (1984): Finite Fields *Cambridge University Press, Cambridge*

Mathworks (2012): Matlab simbiology version 4.2

Möller-Levet, S., Klawonn,F., Cho,K.H. and Wolkenhauer,O. (2003): Fuzzy clustering of short time-series and unevenly distributed sampling points. *LNCS, Proceedings of the IDA2003, pages 28–30. Springer Verlag*

Niederreiter, H. (1988): Low-discrepancy and low-dispersion sequences. *Journal of Number Theory, (30):51–78*

Runolfsson, T. and Ma, Y (2007): Model reduction of nonreversible markov chains. *Decision and Control, 2007 46th IEEE Conference on, pages 3739–3744*

Shumway, R.H. (2003):Time-frequency clustering and discriminant analysis. *Stat Prob Let, 63(3):307–314*

Sarich, M. and Schütte, C. (2014): Utilizing hitting times for finding metastable sets in non-reversible markov chains. *Technical Report 14-32, ZIB, Takustr.7, 14195 Berlin*

Tjakra, J.D., Bao,J., Hudon, N. and Yang, R. (2013): Analysis of collective dynamics of particulate systems modeled by markov chains. *Powder Technology, 235(0):228 – 237*

Tifenbach, R. (2011): On an SVD-based Algorithm for Identifying Meta-stable States of Markov Chains. *ETNA, 38:17–33*

Vlachos,M. and Lin, J., Keogh,E. and Gunopulos, D. (2003): A wavelet-based anytime algorithm for k-means clustering of time series. *In Proc. Workshop on Clustering High Dimensionality Data and Its Applications, pages 23–30*

Weber, M.(2006): Meshless Methods in Conformation Dynamics. *PhD thesis, Freie Universität Berlin*

Liao, T.W. (2007): A clustering procedure for exploratory mining of vector time series. *Pattern Recogn., 40(9):2550–2562*

6 Football and the Dark Side of Cluster Analysis

Christian Hennig
Department of Statistical Science
University Collage London (UCL)
Great Britain
c.hennig@ucl.ac.uk, and
Serhat Akhanli
Department of Statistical Science
University Collage London (UCL)
Great Britain
serhat.akhanli.14@ucl.ac.uk

Abstract

Using a dataset of football player performance data, we discuss exemplarily different decisions by the user that are required for dissimilarity definition and clustering, namely representation, transformation, standardisation and variable weighting.

6.1 A principle for data preprocessing

"The dark side of cluster analysis" refers to the fact that clustering and mapping multivariate data are strongly affected by preprocessing decisions such as variable transformations ("data cleaning" belongs to preprocessing but is not treated here). The variety of options is huge and guidance is scant. This paper treats, in condensed form, some of the issues, using a dataset of football players performance data.

The framework here is the design of a dissimilarity measure, used for multidimensional scaling and dissimilarity-based clustering.

Clustering and mapping are unsupervised; decisions cannot be made by optimising cross-validated prediction quality. Neither is it a convincing rationale to transform data to standard distributional shapes such as the Gaussian. A more general discussion is given in [4].

General principle: Data should be preprocessed in such a way that the resulting effective distance between observations matches how distance is interpreted in the application of interest.

Corollary: Different ways of data preprocessing are not objectively "right" or "wrong"; they implicitly construct different interpretations of the data.

Data driven principles such as optimising stability or "clusterability" are suspicious: can the data decide on their own how they should be interpreted?

DOI 10.20347/WIAS.REPORT.29 Berlin 2017

6.2 Overview of decisions

Here is an overview of decisions that need to be made.

Representation: decisions about how to represent the relevant information in the variables properly; this may involve excluding variables, defining new variables summarising or framing information in better ways, and certain kinds of "interpretation-based" (as opposed to data-based) standardisation.

Transformation: variables should be transformed in such a way that the resulting differences match appropriate "interpretative distances" adapted to the meaning of the variables and the specific application.

Standardisation: variables should be standardised in such a way that a difference in one variable can be traded off against the same difference in another variable when aggregating variables for computing dissimilarities.

Weighting: some variables may be more important/relevant than others - weighting is about appropriately matching the importance of variables.

Mathematically, both standardisation and weighting are multiplications by a constant, but the rationales are quite different.

Variable selection and **dimension reduction** are special cases of representation and weighting.

Defining indexes summarising information guided by interpretation is an alternative to data-based dimension reduction.

6.3 Basic ingredients

The framework here is the **construction of a dissimilarity measure** by aggregating variable-wise distances, e.g.,

Gower aggregation ([3])
$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{p} w_i d_i(x_{1i}, x_{2i})$

Euclidean aggregation
$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{i=1}^{p} w_i d_i(x_{1i}, x_{2i})^2}$.

Alternatives exist but are not treated here.

Mapping is then done by Kruskal's nonparametric multidimensional scaling ([2]) and clustering by "partitioning around medoids" ([1]).

The general principle above also applies to the choice of mapping and clustering method, but not treated here.

6.4 Football players dataset

Football players characterised by 125 variables taken from whoscored.com (we have data on > 2000 players but use only 75 prominent ones from the 2014/15 season for illustration).

Variables:

12 position variables (binary) - indicating where a player can play.

Age, height, weight (ratio scale numbers)

Subjective data: Man of the match, media ratings

Appearance data of player and team, number of appearances, minutes played

Count variables (top level): goals, tackles, shots, passes, fouls, clearances etc.

Count variables (lower level): subdivisions such as shots by body parts, type of pass (long, corner, freekick etc.), successful/failed etc.

Aim: to provide a mapping and classification of players that can be used by managers and clubs looking for players.

6.5 Representation

- Standardise count variables by number of minutes played.
- Top level/lower level count variables:
  use total count (per 90 minutes),
  use percentages on lower level, e.g.,
  shots 5.5, shots right foot 3.8, left foot 0.8, header 0.9,
  use shots 5.5,
  percentages right foot 69, left foot 14, header 16
  percentage profiles are complementary information to total.
- For binary position data use “geco coefficient” ([5]) based on aggregating “geographical distance” for every position to closest position of other player.
- We decided to not use subjective variables.
  It’d be legitimate to use them - this is a decision about what meaning the results should have.

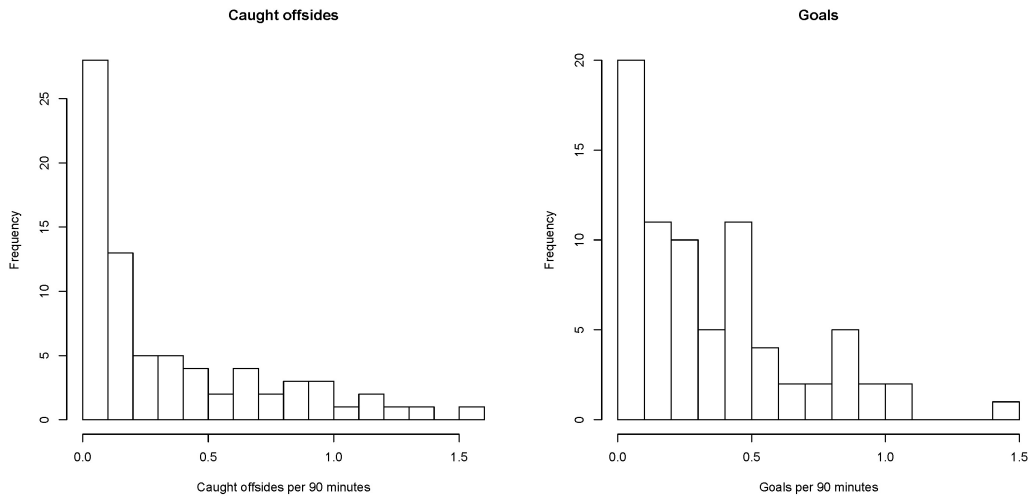


Figure 27: Histogram of offsides (left hand side) and goals (right).

6.6 Transformation

Some variables are very skewly distributed.

There is more variation at the upper end, which suggests that “interpretative distance” between large values should be transformed down. But goals count (approximately) linearly in football.

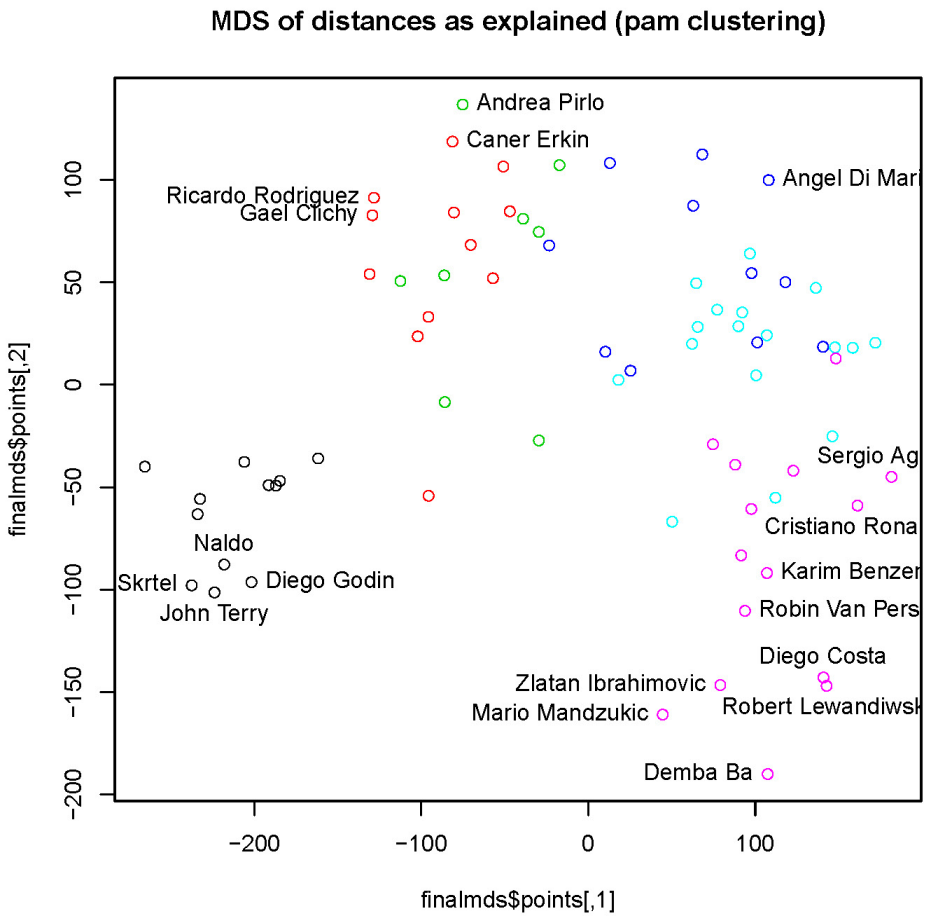
So we use a concave transformation for “Caught offsides”, but none for “Goals”.

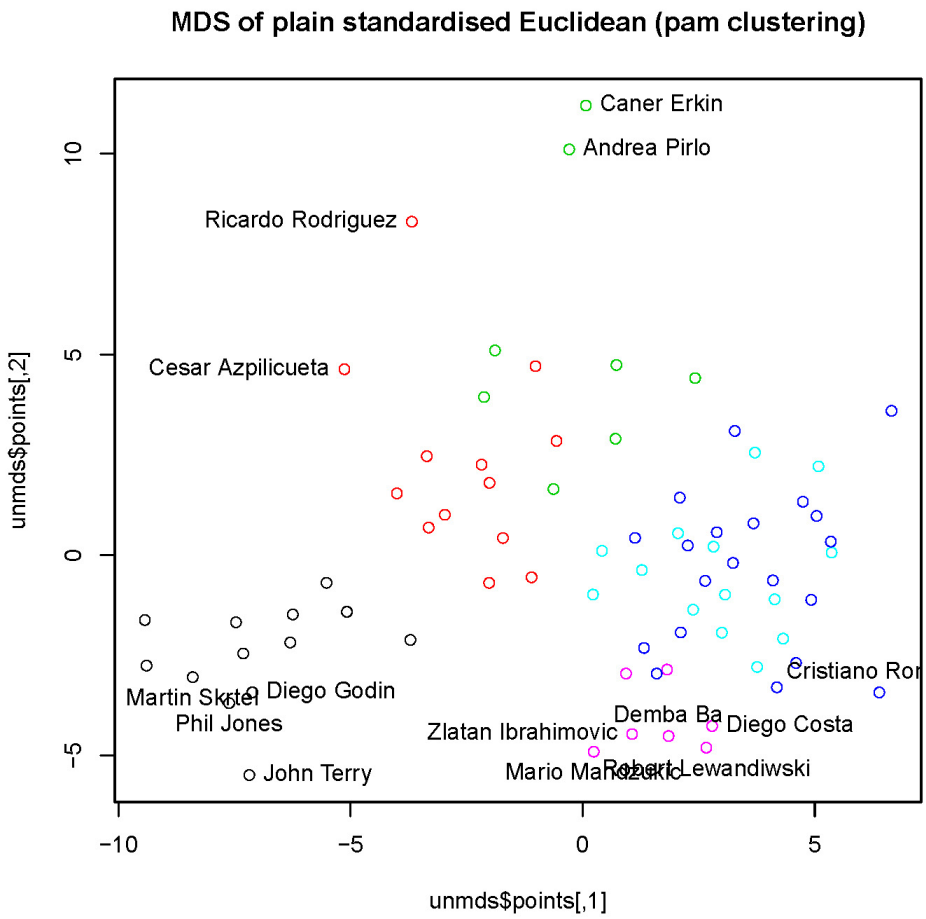
We decide between $\log(x+c)$ or $\sqrt{x+c}$ and about the value of c by looking at what it does to the values and what seems appropriate (subjective football expertise). We explore by *sensitivity analysis* what difference it makes. (We decided to use the plain square root here; details are omitted in this paper.)

Transformations: data dependent?
Unfortunately, researchers have no clear formal idea about “interpretative distance”. The rationale of transformation is matching “interpretative distance” independently of the data. But researchers may need to look at data for having a clearer idea about “interpretative distance”.

6.7 Standardisation

Percentage variables, player age, goals, passes per 90 minutes don’t have compatible variation. Standardisation is needed.


MDS of distances as explained (pam clustering)
Andrea Pirlo
Caner Erkin
Angel Di Mari
Ricardo Rodriguez
Gael Clichy
Sergio Ag
Cristiano Rona
Karim Benzer
Robin Van Pers
Diego Costa
Zlatan Ibrahimovic
Robert Lewandiwsk
Mario Mandzukic
Demba Ba
Naldo
Skrtel
Diego Godin
John Terry
finalmds$points[,2]
finalmds$points[,1]
100
50
0
−50
−100
−150
−200
−200
−100
0
100



But different percentages at same level (shots left, right, header) should be standardised by pooled variance, because variations are compatible and relative sizes should be preserved. Bigger variation should have bigger implicit weight.

Standardisation should not destroy implicit weighting by variance, where appropriate.

6.8 Weighting

Variable weights are useful if some variables seem more important than others. This influences the meaning of the results.

Weighted percentage distributions as “one variable”, e.g., left foot, right foot, header shots are each weighted 1/3; we think about this as giving groups of variables that together formalise a certain aspect of the information unit weight.

Correlation, shared information: if variables are correlated because of *redundant* information (e.g., percentages adding to 100), weight them down.
If variables with *complementary meanings* are correlated, there is no reason not to give them full weight.

Results

"External validation" by football knowledge: Erkin and Pirlo are quite different, but in the same cluster in plain Euclidean solution. Rodriguez and Clichy are expected to be similar, which they are with distances constructed here.

References

[1] Kaufman, L. and P. J. Rousseeuw (1990): *Finding Groups in Data*. Wiley, New York.

[2] Cox, T. F. and M. A. A. Cox (1994): *Multidimensional Scaling*. Chapman and Hall, Boca Raton.

[3] Gower, J. C. (1971): A general coefficient of similarity and some of its properties. *Biometrics 27*, 857–874.

[4] Hennig, C. (2015): Clustering Strategy and Method Selection. In C. Hennig, M. Meila, F. Murtagh, and R. Rocci (Eds.), *Handbook of Cluster Analysis*, Chapman & Hall/CRC, Boca Raton FL, Chapter 31, pp. 703–730.

[5] Hennig, C. and B. Hausdorf (2006): Design of dissimilarity measures: a new dissimilarity measure between species distribution ranges. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Ziberna (Eds.), *Data Science and Classification*, Springer, Berlin, pp. 29–38.

[6] Hennig, C. and T. F. Liao (2013): Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification (with discussion). *Journal of the Royal Statistical Society, Series C 62*, 309–369.

7 On the Practical Relevance of Modern Machine Learning Algorithms for Credit Scoring Applications

Gero Szepannek
Stralsund University of Applied Sciences, Germany
gero.szepannek@fh-stralsund.de

Abstract

Although many new algorithms like e.g. support vector machines, boosting, random forests or neural networks have been proposed in the recent past logistic regression does still represent the gold standard in industrial praxis.

Benchmarking studies show the general superiority of flexible learning techniques that are able to detect complex structures. These studies typically restrict to the evaluation of one or several performance measures (like misclassification rate) and ignore further aspects of practical feasibility.

In this paper a critical investigation of pros and cons of modern machine learning techniques with respect to business requirements and their practical relevance is worked out. An exemplary case study based on credit scoring using random forests is executed.

Introduction

Although many new algorithms like e.g. support vector machines, boosting, random forests or neural networks have been proposed in the recent past there are several reasons why logistic regression does still represent the gold standard in industrial praxis:

1 Logistic regression is widely taught.

2 There are many software implementations of logistic regression available.

3 The resulting models are stable and no further parameter tuning is necessary/possible.

4 The results are easy to interpret.

While the first two reasons are historical and currently under change the third one refers to the necessity an additional parameter specification that is appropriate to the specific data situation. A wrong parameter specification will lead to suboptimal predictive power of the resulting model and can thus be considered as an operative risk from an economic point of view.

On the other hand, many benchmarking studies (cf. e.g. Szepannek et al., 2008, Szepannek et al., 2010, Lessmann et al., 2013) have shown some general superiority of modern flexible learning techniques in terms of quantitative performance measures, especially for complex data structures.

Figure 28 (left) shows a typical example in the credit scoring context: the relationship between age and default rate is nonlinear. A linear model like logistic regression will not be appropriate here. For

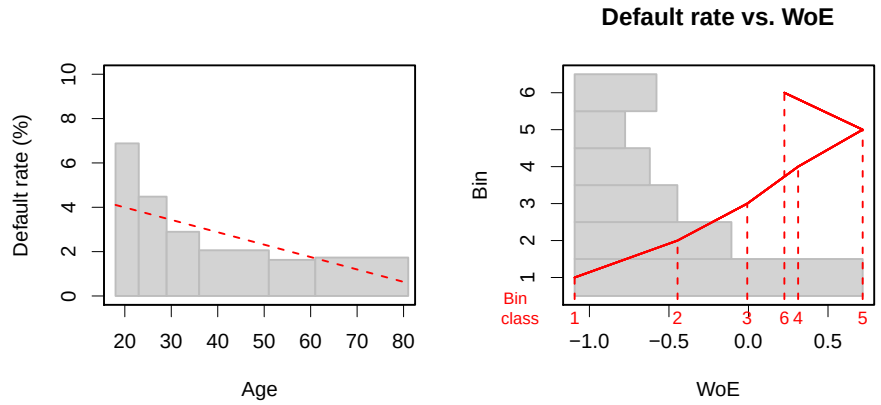


Figure 28: Example of a nonlinear relationship (left), binning and WoE transform (right).

this reason industrial praxis in credit scoring is a pre-binning of the data by the analyst. The binned variables are further used for modelling as dummy variables or transformed to **weights of evidence** (WoEs) where a new numeric variable

$$X_{WoE} = \log\left(\frac{f(x|Y=1)}{f(x|Y=0)}\right)$$

is created from the original variable X (cf. Fig. 28, right). A property of this transform is its (univariate) linearity in the logit of Y (cf. e.g. Szepannek, 2011). Such a pre-binning of the data not only allows to model nonlinear relationships using logistic regression and but also guarantees an implicit plausibility check of the data by the analyst which reduces operational risk.

But pre-binning still does not take into account nonlinear relationships between explanatory variables and moreover the manual nature of the process does still require some kind of additional automatized pre-processing if the number of variables is very large. In order to compare different methods academic benchmark studies typically set up automatized benchmark experiments and the comparison with an industrial use case is not the scope. As a consequence most benchmark studies will be biased towards algorithms that allow for a higher degree of automatization.

For this reason in this document logistic regression models both with and without pre-binning are constructed as a the baseline for comparisons. It has to be stated that manual pre-binning denotes some loss in scientific rigor which is accepted in order to increase practical relevance of the results.

An experiment is set up based on real world data from the credit scoring business context. Random forests (Breiman, 2001) are selected as they represent one of the most popular modern machine learning techniques. In contrast to typical benchmark studies not only the performance of an optimally tuned parameter set is analyzed but also the performance over the whole range of parameter combinations. As a consequence the study concentrates on random forests and no further classifiers are investigated.

In Subsec. 7.1 the parameters of random forests are discussed and in Subsec. 7.2 the experiment is presented. Several questions are analyzed to test the practical benefit of modern machine learning techniques:

- Potential benefit vs. the risk of a performance decrease
- Investigation of random forest parameters
- Cost benefit considerations: time to invest
- Evaluation of the improvements from a business perspective.
- The effect of the analyst
- Interpretation of the model

The results are given in Subsec. 7.3 and finally a summary is given in Subsec. 7.4.

7.1 Random Forests and Parameter Tuning

Random Forests (Breiman, 2001) are designed to avoid shortcomings w.r.t. bias and variance of either small or large single decision trees: a set of large trees (with small bias) is built on bootstrap samples. The variance of the final classifier is reduced by averaging the predictions of all trees. In addition, the single trees are further uncorrelated by allowing only for random subset of variables at each single split of each tree (cf. e.g. Segal, 2006). The **number of trees** as well as the **number of variables** that are considered for each split are two main parameters for random forest construction and optimized in most benchmark studies. A larger number of trees will generally increase computation time and improve performance but saturate. The number of randomly chosen variables should depend on the dimension of the data set: if it is too large the existence of dominant variables may lead to very similar trees and reduce the bootstrapping effect. On the other hand, selecting too few variables may be dangerous if the data consists of a large percentage of purely noisy variables w/o any information on the class.

As an extension to many studies the current experiment does not restrict on these two parameters but includes several additional parameters (for an overview see Boulesteix, 2012), namely the **minimum node size** and the **maximum number of terminal nodes**. Both parameters control the depth of single trees within the forest and should be chosen according to the philosophy of large unbiased trees.

In addition the bootstrap samples can be chosen either with or without **replacement**. A corresponding sample size w/o replacement is of $0.632\times$ the sample size. Strobl et al. (2007) describe the bias of variable importance for sampling with replacement. Finally, samples can be **balanced** w.r.t. the classes as tree splitting criteria typically are affected by the class proportions (cf. e.g. Bischl et al., 2016, Brown and Mues, 2012, Crone and Finlay, 2012). For this reason both balanced and unbalanced samples are investigated for random forest construction.

7.2 Case Study

An issue of credit scoring research is the general lack of real world data as they are confidential in general. For this study the freely available German Credit Data (Hoffmann, 1994) from the UCI

Additional debtors	Number of credits
Age	Other instalment plans
Amount	**People liable**
Credit history	**President residance**
Duration	Property
Duration employment	Purpose, product
Foreign worker	Savings
Housing, type of residence	Gender, family status
Instalment rate	Status checking account
Job	Telephone available

Table 7: Variables of the German credit data set.

Machine learning repository (Newman et al., 1998) is used. Based on one single data set the results should rather be considered as a case study without claim for generality. Bischl et al. (2016) try to overcome this issue by collecting a large number of data sets from other domains but transferability to the credit scoring context remains questionable.

The data consists of 1000 observations in two classes (default vs. non-default) which is quite few compared to real world applications. There are 20 explanatory variables. The prior probability of default is 0.3 which does not reflect typical unbalance of real world scoring data. In contrast to decision trees logistic regression is quite insensitive to class unbalance (Bischl et al., 2016, Brown and Mues, 2012, Crone and Finlay, 2012).

Table 7 summarizes the variables, **numeric** variables are in bold. Typical for business applications the variables are not all metric but most data comes along in categories further emphasizing the analyst's role in model building as the categories are w/o any order but have to be interpreted w.r.t. their meaning from a business point of view.

As a baseline logistic regression models are built with and without pre-binning of the variables. Binning is done based on manual plausibility checks of the splits generated by univariate decicsion trees with varying complexity parameters (Therneau and Atkinson, 1997). For the binned variables separate regression models are built either using dummies or WoEs.

The **Gini coefficient** $G = 2(AUC - 0.5)$ is used as a performance measure as it represents the most popular statistic to evaluate the performance of credit scoring systems in practise. Given the comparatively small sample size 10 fold cross validation is used for performance evaluation. Good practice would consist in an additional inner CV loop for parameter optimization of random forests (cf. e.g. Szepannek et al., 2010, Bischl et al., 2016). The focus of this study slightly differs from typical benchmark experiments as the entity of all models with different parameters is of interest rather than only the optimally parameterized one. For this reason no parameter optimization on an additional inner loop has been done here.

A total of 2304 random forests has been investigated based on an exhaustive parameter grid of the parameters given in table 8 (the default parameters are in bold, for the number of terminal nodes there is no restriction in the default parameterization of random forests, corresponding to a value of 500 in the setting).

In order to investigate its relevance for business a test on the achieved improvement is implemented according to Henking et al. (2006) using approximate normal distribution of the AUC with standard

Tuning Parameter	Values
Number of trees	$\{20, 50, 200, \mathbf{500}, 1000, 2000\}$
Number of variables | Split	$\{2, \mathbf{4}, 8, 16\}$
Min. node size	$\{\mathbf{1}, 5, 20, 50\}$
Max. number terminal nodes	$\{5, 10, 20, 50, 100, \mathbf{500}\}$
Sampling with replacement	$\{\mathbf{yes}, no\}$
Balanced class sizes	$\{yes, \mathbf{no}\}$

Table 8: Parameters for random forest optimization.

Pre-inning	Dummies	WoEs
Yes	57.13	57.80
No	59.06	**59.75**

Table 9: Results of logistic regression models.

error:

$$\hat{\sigma}_{AUC} = \sqrt{\frac{AUC(1-AUC) + (N_D - 1)(q_1 - AUC^2) + (N_{ND} - 1)(q_2 - AUC^2)}{N_D N_{ND}}}$$

and $q_1 = \frac{AUC}{2-AUC}$, $q_2 = \frac{2AUC^2}{1+AUC}$ as well as $N_{(N)D}$ the number of (non-)defaults in the sample.

7.3 Results

Logistic Regression: Table 9 shows the results of the logistic regression models. Both preliminary binning and WoE pre-transform of the binned data improve performance. Whereas improvement by binning might result from allowing to model nonlinearities the additional gain of using WoEs may be a consequence of the small data set as the use of dummy characteristics increases the degrees of freedom of the subsequent regression model.
Random forests and their parameterization: The best random forest model achieves a Gini coefficient of 61.53 which is no significant improvement (p value: 0.294). The optimal parameters are identical to the default ones except from the number of variables offered at each split being $2 < 4$ ($=$ floor of the $\sqrt{\cdot}$ of the number of variables in the data set). Figure 29 shows the frequency of model performance over all parameter combinations.

It can be seen that the relative improvement by using random forest models is quite small whereas the potential loss can be dramatic for a bad parameter set. Only 7.6% of the models improve logistic regression performance but 19.5% even led to significant performance decrease. This underlines the risk of blindly applying modern machine learning algorithms, instead a thorough understanding of the methodology is required. Interestingly, just using default parameterization already results in a (slight) performance increase (60.04, p value 0.465). In order to prevent from misinterpreting the results it has to be remarked that the frequencies in Figure 29 result from an experimental design and do not reflect a distribution as the figure might suggest. The results are further biased by the attempt to investigate a quite exhaustive parameter grid whereof some combinations turn out to be avoidable, leading to an analysis of the impact of the single parameters. Figure 30 shows the performance as a function of the parameters and levels in an OFAT sense: Mainly, the considerations of Subsec. 7.1 are confirmed as both small (flat) base single trees as well as a number of trees that is too small

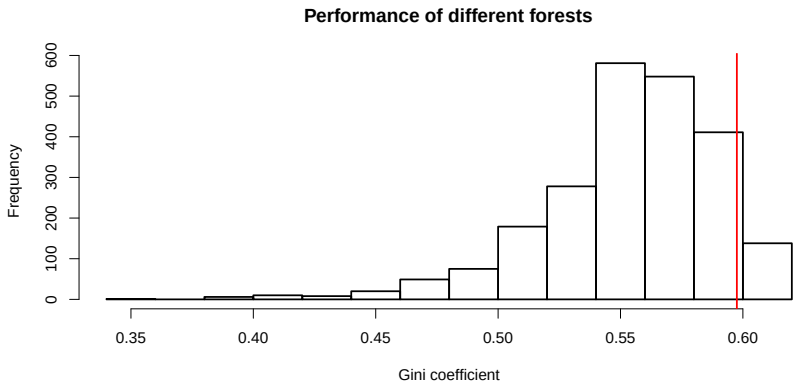


Figure 29: Results over all random forests (red line: logistic regression baseline).

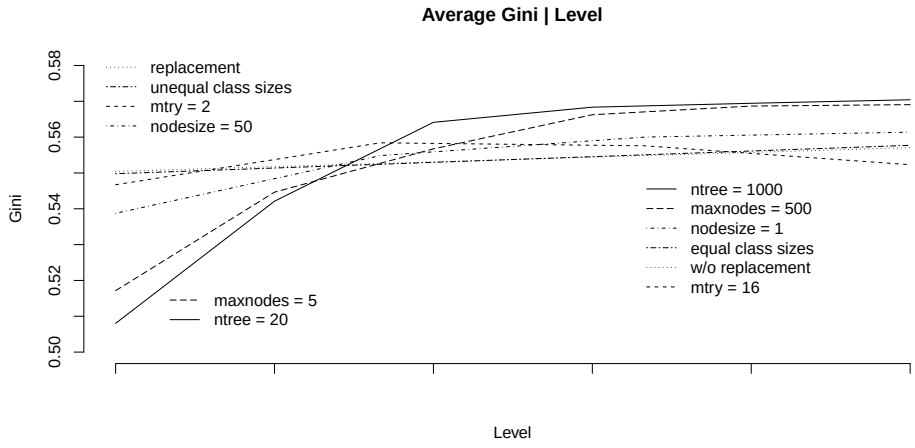


Figure 30: OFAT analysis of the performance by parameter levels.

result in strong performance decrease. Remarkably a number of four variables that is offered at each split on average gives the best results but the optimal model is obtained for two variables only. As a summary, some parameter values can be discarded for both theoretical considerations and empirical evidence.

In conclusion, the observed benefit of using random forests is relatively small and no significant improvement is obtained. But one may consider a business case of a population with 5% defaulters in total and a sum of 2bn. EUR of annual funding. In this case an improvement of only $+1\%$ rejected defaulters in will lead to a profit increase of 1mn EUR per year. The ROC curves for visual analytics as provided by many statistical software packages will not even allow to recognize any difference in the corresponding graphs.

Cost vs. benefit analysis: Different to academic research, time-to-market is typically an important business consideration which also holds for the credit scorecard modelling process. For this reason

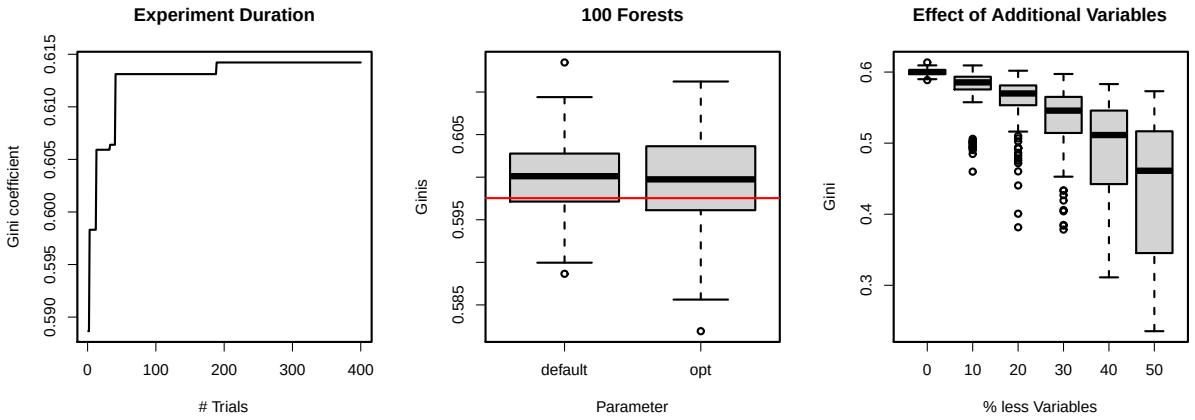


Figure 31: Performance vs. optimization time (left) and variance of forests (center) and performance decrease for lack of variables (right).

and based on the observed results from the last paragraph a parameter space (according to the previous grid but reduced by the identified implausible values for the tree and forest size) has been set up and a random search of additional 10000 forests in this parameter space has been set up in order to check the increase in performance with duration of the optimization process. (Please note that a naive random search can be speed up by intelligent search algorithms like genetic algorithms or iterated F-racing, cf. e.g. Bischl et al., 2016). Figure 31 (left) shows that the optimum is reached after only 400 iterations and not improved anymore which clearly indicates that the costs related to the additional parameter optimization process are comparatively low.

Astonishingly, it can be seen from the graph that the optimal performance from the first experiment (61.53) is not reached again within 10000 additional forests. This phenomenon is due to the ”randomness of random forests”: as the forests do depend on bootstrap samples two random forest models based on the same parameters will not be identical (see also Schäfer, 2006). Figure 31 (center) shows the performance variability of 100 random forests models for both the default parameters as well as the ”optimal” parameter set: the performance of the first experiment is not reached again, which again underlines the importance of the second test loop for parameter optimization in order to avoid overfitting (cf. Subsec. 7.2).

The higher variance of the best parameterization from experiment 1 can be explained by the additional randomness that results from selecting only two instead of four variables at each split which can increase or decrease performance depending on the selected variables. **A consequence, model selection should consist in taking into account both expectation and variance of the model performance estimation.** In order to win a competition like kaggle one may accept a higher variance to obtain increased upper performance quantiles whereas in a risk management context lower **quantiles of model performance estimation** will more relevant.

Model tuning vs. integration business knowledge: Improving models may not only be achieved by improving statistical modelling but also by identification of additional important explanatory variables. As the used data represents a real world example we can analyze what happened if some variables

were not available (to simulate the improvement that can obtained by new variables). Figure 31 (right) shows the loss in performance if a randomly selected percentage of variables were not available for model building. Please note the difference: as opposed to the random forest parameter the variables are not just removed for single splits within a tree but for the whole forest, here. The graph outlines the importance of identifying predictive characteristics: the benefit of an additional variable is much higher than the additional benefit obtained by model optimization. This step has to be considered as an important factor and emphasises the importance of a proper integration of business knowledge into the model building process.

Understanding the model: Finally, risk models not only have to be communicated to several directions (management, employees as well as its results towards the customers), there are also regulatory constraints often based on a natural mistrust towards black box algorithms. It should be remarked at this point that a **properly validated** and demonstrated superiority in model performance will lead to better estimates of risk which denotes one of the central targets of regulation.

Often the interpretability of regression coefficients (score points) are considered as advantageous to understand the key drivers of a model and to allow for its validation. The concept of variable importance (cf. e.g. Strobl et al., 2007) can be used to quantify the relevance of a variable within any classification model. Moreover the decrease in variance importance on out-of-time samples can be used for validation purposes and thus a decomposition of the black box.

7.4 Summary

The paper aims to bridge a gap between academic research concerning modern machine learning techniques and their business relevance for credit scoring applications. Random forests are investigated as one of today's most popular machine learning algorithms. The results are compared to logistic regression in a realistic setting. Several practical aspects are discussed like parameter tuning, business relevance of the improvements as well as cost vs. benefit aspects. In summary, the obtained benefits were comparatively small, but still of potentially large monetary impact. In addition the identification of new predictive characteristics has been demonstrated to be of great importance underlining the relevance of business knowledge integration in the modeling process.

As an important result, the simultaneous investigation of expectation and variance of classifier performance estimation has been worked out to be appropriate for model selection in a risk context potentially leading to a focus on quantiles.

Finally, variable importance has been presented as a tool to remove doubts concerning black box like algorithms w.r.t. regulatory constraints or for model validation purposes and improve the estimation of risks.

References

Bischl, B., Kuehn, T. and Szepannek, G.: On Class Imbalance Correction for Classification Algorithms in Credit Scoring. In Luebbecke, L., Koster, A., Letmathe, P., Madlenerm, R., Peis, B. and Walther, G. (eds.): *OR 2014, 37–43*, Springer, Heidelberg.

Boulesteix, A., Janitza, S., Kruppa, J. and König, I. (2012): Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics. *Technical Report 129/2012*, Dept. Statistics, LMU Munich.

Breiman, L. (2001) Random Forests. *Machine Learning 45 (1)* 5–32.

Brown I., Mues C. (2012): An Experimental Comparison of Classification Algorithms for Imbalanced Credit Ccoring Data Sets, *Expert Systems with Applications 39 (3)*, 3446–3453.

Crone S., Finlay S. (2012): Instance Sampling in Credit Scoring: an empirical study of sample size and balancing, *International Journal of Forecasting 28 (1)*, 224–238.

Henking, A., Blum, C. and Fahrmeir, L. (2006): *Kreditrisikomessung. Statistische Grundlagen, Methoden und Modellierung*, Springer, Berlin.

Hoffmann, H. (1994): German Credit Data Set (Statlog) `http://archive.ics.uci.edu/ml/datasets/Statlog+\%28German+Credit+Data\%29`.

Lessmann S., Seow H., Baesens, B. and Thomas, L. (2013): Benchmarking State-of-the-art Classification Algorithms for Credit Scoring: A Ten-year Update. `http://www.business-school.ed.ac.uk/waf/crc\_archive/2013/42.pdf`.

Newman, D., Hettich, S., Blake, C. and Merz, C. (1998): UCI Repository of Machine Learning Database. `http://www.ics.uci.edu/~mlearn/MLRepository.html` University of California, Irvine, Dept. of Information and Computer Sciences.

Schäfer, M. (2006): Random Forests: A Case Study, Talk at 28. AG DANK, 27.10.2006, Dortmund.

Segal, M. (2004): Machine Learning Benchmarks and Random Forest Regression. *escholariship Univerity of California, Center for Bioinformatics and Molecular Biostatistics*. `http://escholarship.org/uc/item/35x3v9t4`.

Strobl, C., Boulesteix, A., Zeileis, A. and Hothorn, T. (2007): Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics* 8–25.

Szepannek, G. (2011): Vortransformation in der Kreditantrags-Scoremodellierung. Talk at Data Mining Anwendertag, Heidelberg, `http://www.sas.com/reg/offer/de/datamining\_2011?page=download`.

Szepannek, G., Gruhne, M., Bischl, B., Krey, S., Harczos, T., Klefenz, F., Dittmar, C. and Weihs, C. (2010): Perceptually based Phoneme Recognition in Popular Music. In Locarek-Junge, H., Weihs,C. (eds.) *Classification as a Tool for Research*, Springer, Heidelberg, 751–758.

Szepannek, G., Schiffner, J., Wilson, J., Weihs, C. (2008): Local Modelling in Classification. In: Perner, P. (ed.): *Advances in Data Mining: Medical Applications, E-Commerce, Marketing, and Theoretical Aspects* Springer LNAI 5077, Berlin, 153–164.

Therneau, T., Atkinson, E. (1997): An Introduction to Recursive Partitioning using RPART Routines. *TR 61, Mayo Foundation*, `http://www.mayo.edu/hsr/techrpt/61.pdf`.

8 Finding Groups in Compositional Data - Some Experiments

Hans-Joachim Mucha
Weierstrass Institute for Applied Analysis and Stochastics (WIAS)
Mohrenstr. 39
10117 Berlin
E-Mail: mucha@wias-berlin.de and
Tatjana Mirjam Gluhak
Institut für Geowissenschaften
Johannes Gutenberg-Universität
Geomaterial- & Edelsteinforschung
Johann-Joachim-Becher-Weg 21, 55128 Mainz
E-Mail: gluhak@uni-mainz.de

Abstract

In archaeometry, the chemical composition of oxides of objects is measured, and often the results are presented in percentages. Then, the so-called compositional data analysis should be applied as "the only one valid statistical analysis". This paper is concerned with finding groups (clusters) in (strict) compositional data, that is, nonnegative data with row sums equal to a constant, usually 1 in case of proportions or 100 in case of percentages. It reports about some experiments. Without loss of generality, the cluster analysis of observations of compositional data is considered, where the row profiles contain parts of some whole. Distance functions between profiles and appropriate clustering methods are recommended. Nowadays, besides oxides, a much greater number of trace elements can be measured by new innovative technical equipments. Usually, these measurements are in ppm (parts per million) or ppb (parts per billion) which causes additional problems for compositional data analysis. Applications to archaeometry are presented.

8.1 Introduction and Motivation

In archaeology, the aim of clustering is to find groups in data such as proveniences of glass objects or pottery. The motivating example is taken from Baxter & Freestone (2006) where the complete original data matrix $\mathbf{Z} = (z_{ij})$ is published as it is analyzed hereafter. It is compositional data with nonnegative elements $z_{ij} \geq 0$, $(i = 1,2,\ldots,I, j = 1,2,\ldots,J)$. Each archaeological object (observation) is characterized by $J = 11$ variables, the contents of oxides in %. The sum in each row is exactly 100%. This dataset of colorless Romano-British vessel glass contains two groups. Group 1 consists of 40 cast bowls with high amounts of Fe_2O_3. Group 2 also consists of 40 objects: this is a collection of facet-cut beakers with low

DOI 10.20347/WIAS.REPORT.29 Berlin 2017

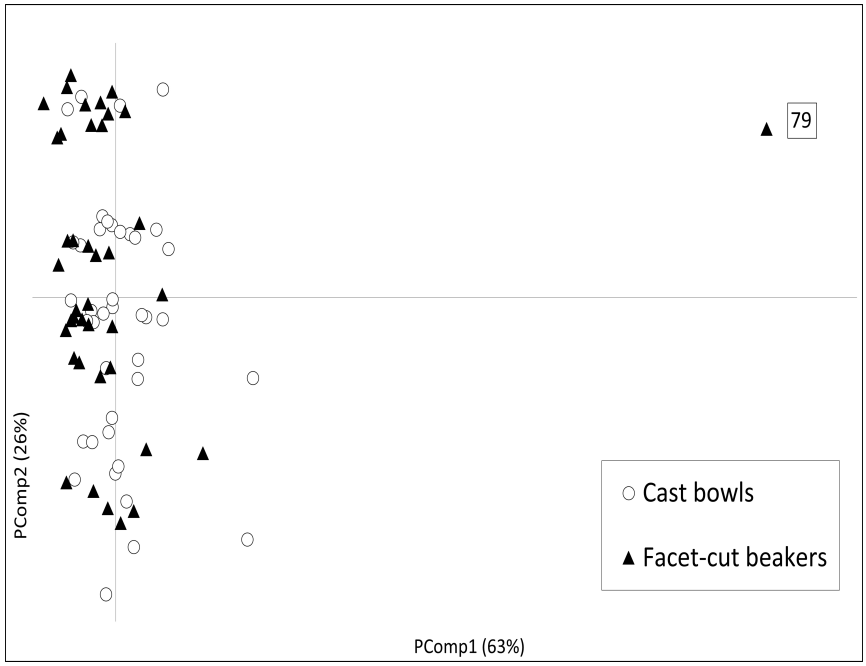


Figure 32: PCA plot of log-ratio transformed data (Romano-British vessel glass: here one Sb_2O_3 value of the dataset was changed slightly from 0.08 to 0.0001 (observation "79")).

Al_2O_3. Cluster analysis is unable to find the true classes for log-ratio transformed data: 35 objects were wrongly classified, see Fig. 2 and Table 2 in Mucha et al. (2008). This is an error rate of about 44 %. Generally, log-ratio compositional data analysis (Aitchison 1986) means: instead of the original data matrix $\mathbf{Z}$ use $\mathbf{X}$ with elements

$$x_{ij} = \log(z_{ij}/g(\mathbf{z}_i)) \ , \tag{14}$$

where $g(\mathbf{z}_i) = (z_{i1}z_{i2}\ldots z_{iJ})^{1/J}$ is the geometric mean of the ith observation. This transformation is restricted to strictly positive values $z_{ij} > 0$ which is an essential drawback in applications to archaeological and geological sciences. Fig. 32 shows another drawback of log- ratio data analysis: it is an outlier "producing" technique. The PCA plot with the class membership of the objects is based on log-ratio transformed data (14). In addition, to become apparent, the following experiment was done: We change only one value of Sb_2O_3 very slightly from 0.08 to 0.0001 (observation "79"). (Then the corresponding values of this row are trimmed to sum up strictly to 100%). As a result, the observation "79" becomes a very strong outlier. However it is already a possible outlier when doing log ratio analysis of the original data, see Fig. 2 in Mucha et al. (2008). Obviously, besides the zero value problem, compositional data analysis seems to be inappropriate for positive values near 0. Already Baxter and Freestone (2006) criticized that Aitchison argued that all others transformations are "meaningless" and "inappropriate" for compositional data. The authors presented the failure of PCA for different data sets based on the log-ratio transformation.

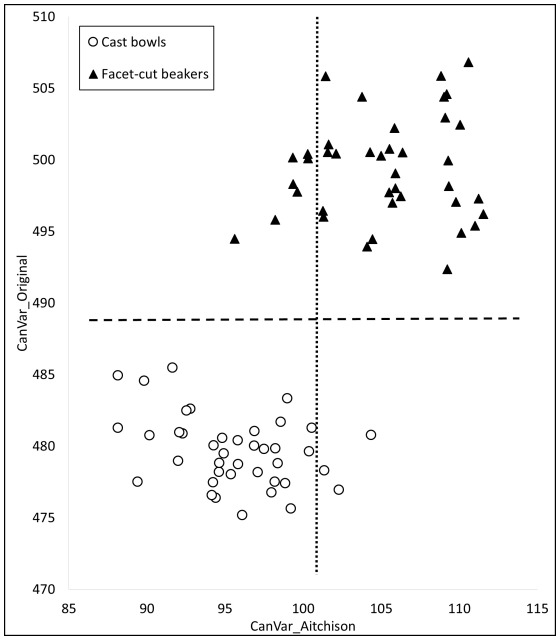


Figure 33: Scatterplot of the canonical discriminant variable of log-ratio transformed data vs. the canonical discriminant variable of the original Romano-British vessel glass data.

Another experiment: What about the performance of the well-known linear discriminant analysis based on log-ratio transformed data referred to (14)? It is poor: Four “cast bowls” and seven “facet-cut beakers” are wrongly classified by the estimated linear classifier. The two classes cannot satisfactory be separated by the corresponding canonical discriminant variable “CanVar_Aitchison”. Fig. 33 shows a visual comparison with the canonical discriminant variable of the original data. The latter separates the true classes without error, see the horizontal broken line in between the two class-wise distributions. This is different to the vertical dotted line which visualizes the classification result of “CanVar_Aitchison” mentioned already above. In the linear discriminant analysis of the original data, the F-value of about 610 (degrees of freedom are 11 and 68) is quite high. It is highly significant.

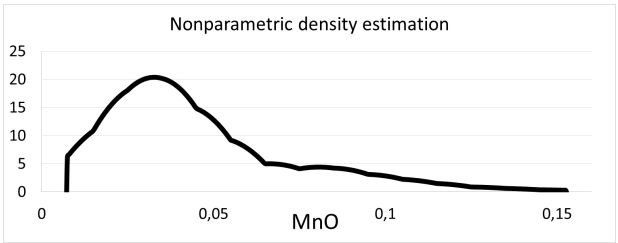


Figure 34: The variable MnO is very skew-symmetric (data: Romano-British vessel glass).

8.2 Finding Clusters in Compositional Data

We consider here the simplest model-based Gaussian clustering method. It seems to be an appropriate (practical) model for clustering highdimensional datasets (i.e., $I \approx J$). Let $\mathscr{C} = \{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_I\}$ denote the finite set of observations. Further, let $\mathscr{P} = \{\mathscr{C}_1, \ldots, \mathscr{C}_K\}$ of $\mathscr{C}$ be the partition we are looking for. The minimization of the sum of squares (SS) criterion (i.e., derived from the simplest model-based Gaussian clustering model)

$$V_K(\mathscr{P}) = \sum_{k=1}^{K} \mathbf{W}_k, \tag{15}$$

is equivalent to the minimization of

$$V_K(\mathscr{P}) = \sum_{k=1}^{K} \frac{1}{|\mathscr{C}_k|} \sum_{i \in \mathscr{C}_k} \sum_{h \in \mathscr{C}_k, h>i} d_{ih}, \tag{16}$$

where

$$d_{ih} = d(\mathbf{x}_i, \mathbf{x}_h) = \|\mathbf{x}_i - \mathbf{x}_h\|^2 \tag{17}$$

is the squared euclidean distance between two observations $\mathbf{x}_i$ and $\mathbf{x}_h$, and $\mathbf{W}_k$ is the usual estimate of the within-cluster covariance matrix Σ_k of cluster $\mathscr{C}_k$. The corresponding distance matrix is $\mathbf{D} = (d_{ih})$, $(i = 1, 2, \ldots, I, h = 1, 2, \ldots, I)$. It is symmetric, and it is additive, i.e., it can be expressed by the sum of multiple (variable-wise) distance matrices $\mathbf{D}^{(j)}$:

$$\mathbf{D} = \sum_{j=1}^{J} q_j \mathbf{D}^{(j)}, \tag{18}$$

where the element $d_{ih}^{(j)} = (x_{ij} - x_{hj})^2$ is the squared euclidean distance with respect to variable j, and $q_j = 1$ is the standard "weight" of variable j. Eq. (18) becomes more general by taking into account nonnegative weights of variables $q_j \geq 0$. For example, $q_j = 1/s_j^2$ means standardisation of variable j to standard deviation equal to 1, where $s_j > 0$ is the standard deviation of the original variable j. Greenacre (1988) uses the weights $q_j = x_{++}/x_{+j}$ for clustering the row profiles (i.e., the vectors of relative frequencies) of a contingency table. Here x_{+j}/x_{++} is the jth element of the average row profile. Then, the elements of $\mathbf{D}$ in (18) are the Chi-square distances between profiles (i.e., between vectors of proportions that sum up strictly to 1). So far, clustering of vectors of proportions in compositional data analysis can also be considered within the framework of correspondence analysis.

Concerning standardisation and weighting of variables see Sect. 6 in this volume. Another common approach in (robust) statistics is to transfer the values for every variable to their rank values. The advantage is that important practical problems are managed in the twinkling of an eye, namely the scaling problem and outlier problem. However, one loses a lot of information. The logarithmic transformation is very common in geochemistry because the

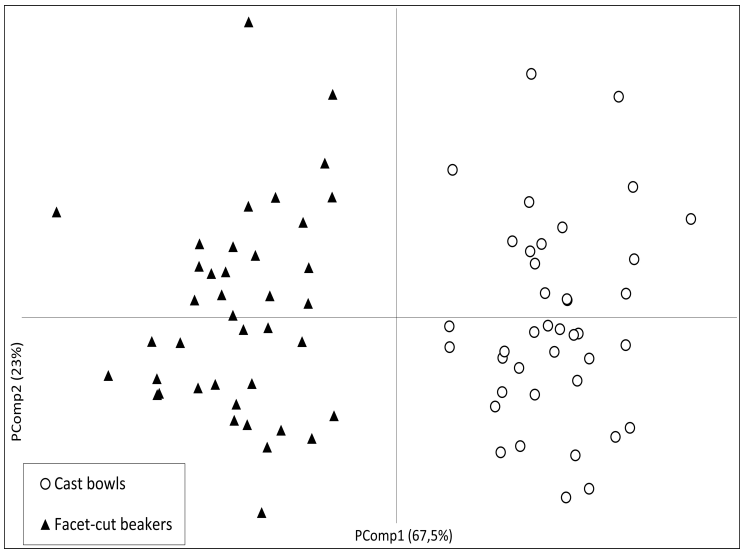


Figure 35: PCA plot of log-transformed data (Romano-British vessel glass).

distributions of the chemical elements are often skew-symmetric with a long tail on the right side (see, for example, Fig. 34). The general logarithmic transformation is

$$x_{ij} = \log(z_{ij} + c_j), \tag{19}$$

where $c_j > 0$ is an appropriate constant with regard to variable j. Hereafter, $c_j = 1$, $j = 1, 2, \ldots, J$, is used. As a result, Fig. 35 shows the PCA plot of log-transformed data. The hierarchical Ward's clustering finds the true classes without error. It minimizes the criterion (16), that is, the SS-criterion (15). Fig. 36 shows univariate and bivariate densities for the first two principal components. Another visualization of the bivariate density is shown in Fig. 36. The two classes, "cast bowls" and "facet-cut beakers", look well separated.

The additive (variable-wise) techniques such as standardisation, weighting, logarithmic transformation, and transformation to rank order data, are different to the multivariate log-ratio transformation (14). The latter takes all the data into account when transforming a single variable.

8.3 Application: Cluster Analysis of Basalt Ground Stone Tools from El-Wad

There is an ongoing study to establish a geochemical-mineralogical characterization of different basaltic rocks used as processing tools during the Natufian culture of the later parts of the Epipalaeolithic period (c. 15,000-11,500 Cal BP) in the southern Levant (Gluhak and Rosenberg 2013). It focuses on tools found in various sites attempting to define raw material

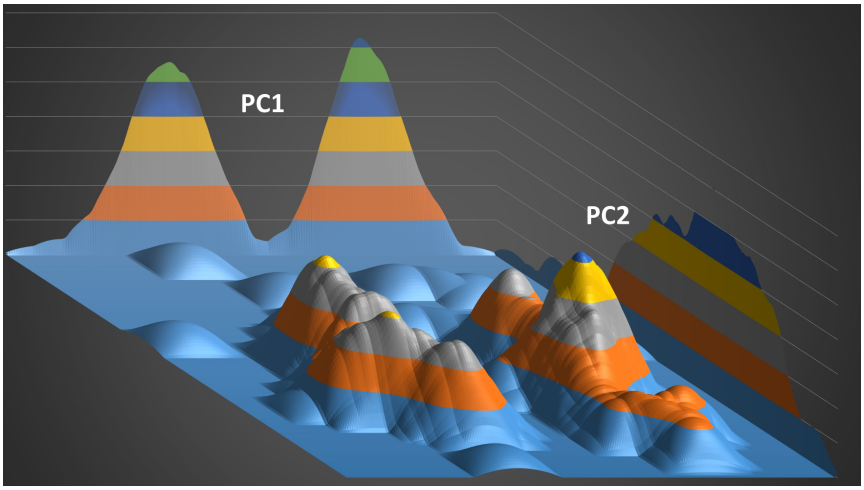


Figure 36: Nonparametric univariate and bivariate density estimation of the first two principal components of log-transformed data (Romano-British vessel glass).

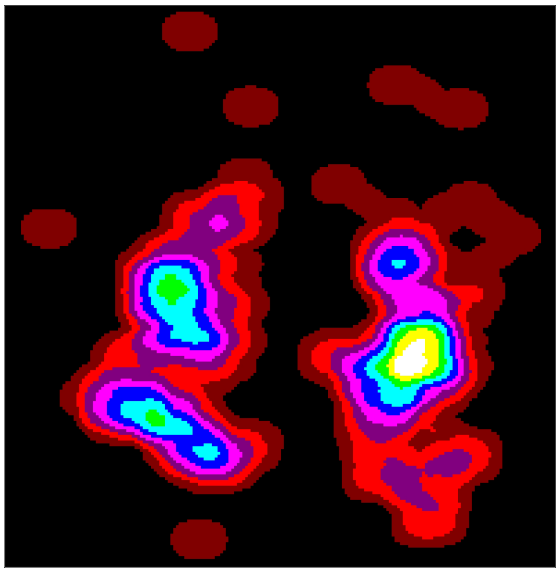


Figure 37: Several cuts of the nonparametric bivariate density estimation of the first two principal components (see Fig. 36).

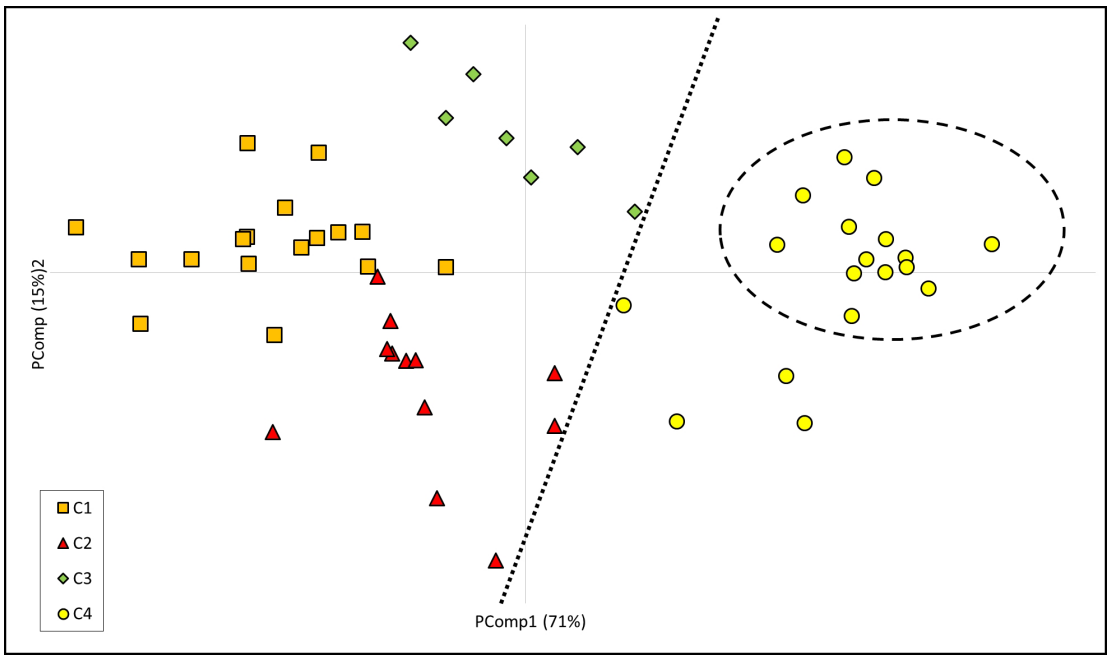


Figure 38: PCA plot of log-transformed basalt data showing the result of clustering (see also Fig. 39).

variability between tool types as well as to pursue their provenance. Here we analyze only a small subset of $I = 55$ tools found in el-Wad (Carmel Mountains). Altogether 39 oxides and chemical elements were measured. In the following, only the 10 oxides are under investigation, i.e., $J = 10$. The logarithmic transformation (19) with $c_j = 1,\ j = 1,2,\ldots,j$, is used.

Fig. 38 shows the corresponding PCA plot. Here the result of Ward's cluster analysis into 4 clusters is presented. In addition, the circle surrounds a very stable sub-cluster (see also Fig. 39: The corresponding Jaccard values of stability τ are underlined and in bold). The dotted line separates the data into two clusters which constitute the stablest partition obtained. Fig. 39 shows the result of the investigation of stability of Ward's clustering via the bootstrap resampling technique. Here the degree of stability of an individual cluster is measured by the Jaccard similarity measure τ which can achieve a maximum value equals 1. The latter allows the investigation of stability of both the individual clusters (Hennig 2007) and the partitions. The latter is based on an averaged Jaccard value such as

$$\tau^* = 1/K \sum_{k=1}^{K} \tau_k,$$

where K is the number of clusters in the partition. The partition into two clusters has the highest averaged Jaccard value $\tau^* = 0.93$, see at the bottom line in Fig. 39.

Jaccard measure and averaged Jaccard (bottom)														
Cluster	#	τ	#	τ	#	τ	#	τ	#	τ	#	τ	#	τ
1	11	0,878	12	0,859	12	0,84	12	0,761	17	0,711	29	0,726	36	**0,94**
2	1	0,37	5	0,81	5	0,777	5	0,677	12	0,67	7	0,475	18	**0,9**
3	5	0,831	10	0,691	12	0,747	12	0,737	7	0,752	18	**0,91**		
4	10	0,685	2	0,379	7	0,873	7	0,866	18	**0,9**				
5	2	0,449	7	0,851	14	**0,95**	18	0,864						
6	7	0,782	14	**0,96**	4	0,623								
7	14	**0,95**	4	0,746										
8	4	0,777												
τ*		0,811		0,823		0,831		0,796		0,771		0,756		**0,93**

Figure 39: Stability of individual clusters of Ward's clustering (see also Fig. 38). High values of stability are in bold.

Outlook

The stability of the clusters of basaltic rocks looks quite different. Cluster analyses of all tools from all sites based on all variables (10 oxides and 29 trace elements) are postponed to the future. Then the pivotal question has to be answered: Can we find stable groups in such (mixed) data by applying "usual" statistical clustering? To answer this, further investigations are necessary where variable selection in cluster analysis (Mucha and Bartel 2016) seems the favorite way to do this. And, besides the simplest model-based Gaussian clustering criterion (15), the logarithmic sum-of-squares criterion (Mucha 2009) seems to be also an appropriate criterion for this data, namely minimizing

$$V_K^*(\mathscr{P}) = \sum_{k=1}^{K} |\mathscr{C}_k| \log \mathrm{tr} \frac{\mathbf{W}_k}{|\mathscr{C}_k|} \ , \tag{20}$$

or, equivalently

$$V_K^*(\mathscr{P}) = \sum_{k=1}^{K} |\mathscr{C}_k| \log(\sum_{i \in \mathscr{C}_k} \sum_{h \in \mathscr{C}_k, h>i} \frac{1}{|\mathscr{C}_k|^2} d_{ih}), \tag{21}$$

where d is the squared euclidean distance (17). Both criteria (15) and (20) look for clusters of spherical shape. However, the more general criterion (20) can also detect clusters of different volumes. As the SS-criterion (15) is, it is also additive because of its analogon (21) which is based on pairwise distances (17). Again, the latter can be generalized by variable weighting (see eq. (18)).

References

AITCHISON, J. (1986): *The Statistical Analysis of Compositional Data*. Chapman and Hall, London.

BAXTER, M. J. and FREESTONE, I. C. (2006): Log-ratio Compositional Data Analysis in Archaeometry. *Archaeometry*, 48, 511–531.

GLUHAK, T. M., ROSENBERG, D. (2013): Raw material variability and provenance before the neolithic revolution: studies of natufian ground stone tools from Eynan and El-wad (Israel). *Metalla*, Sonderheft 6, 196–199.

GREENACRE, M. J. (1988): Clustering the Rows and Columns of a Contingency Table, *Journal of Classification*, 5, 39–52.

HENNIG, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis*, 52, 258–271.

MUCHA, H.-J. (2009): ClusCorr98 for Excel 2007: Clustering, Multivariate Visualization, and Validation. In: Mucha, H.-J. and Ritter, G. (Eds.) *Classification and Clustering: Models, Software and Applications*, Report 26, WIAS, Berlin, 14–40.

MUCHA, H.-J., and BARTEL, H.-G. (2016): Bottom-up variable selection in cluster analysis using bootstrapping: A proposal. In: A.F.X. Wilhelm, H.A. Kestler (Eds.): *Analysis of Large and Complex Data*, Springer, Cham, 2016, pp. 125–135.

MUCHA, H.-J., DOLATA, J., and BARTEL, H.-G. (2008): Effects of data transformation on cluster analysis of archaeometric data. In: Ch. Preisnach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (Eds.): *Data Analysis, Machine Learning and Applications*, Springer, Berlin, 681–688.

Part III
Data Analyses

9 Comparison of the Results of the Competition

Hans-Joachim Mucha

Introduction

Every now and then in the past more than 20 years, data analysis experiments were performed at the meeting. Beginning with the meeting 2005 at Infratest in Munich, Germany, the data analyses were transformed to a competition with more contributions. As introduced at the autumn meeting 2013 at UCL in London, it runs as a competition with real-world data sets and book prizes for the winners. This time, the following book prizes were donated by WIAS:

(a) Gunter Ritter (2015): Robust Cluster Analysis and Variable Selection. Chapman & Hall/CRC Monographs on Statistics & Applied Probability;

(b) Vladimir Spokoiny and Thorsten Dickhaus (2015): Basics of Modern Mathematical Statistics, 18 of Springer Texts in Statistics, Springer;

(c) Christian Hennig, Marina Meila, Fionn Murtagh, and Roberto Rocci (2015): Handbook of Cluster Analysis. Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

9.1 The “Chemsensor” dataset for competition

There is a plethora of existing data analysis methods and the idea is to apply them to various data sets with different characteristics in order to learn about their strengths and weaknesses. For this purpose, a real life data set was issued about two months before the event: The bioaerosol data for unsupervised classification (clustering) competition were provided by Dr. Chistian Hennig.

Table 10: Confusion matrix of the true class vs. result of linear discriminant analysis, see Sarantaridis et al. (2012). The columns show which class the events were assigned to by cross validation.

Class	BGP	BSS	JSS	BWP	Total
BGP	**29**	0	1	0	30
BSS	1	**22**	7	0	30
JSS	1	10	**19**	0	30
BWP	5	3	3	**19**	30
Total	36	35	30	19	120

DOI 10.20347/WIAS.REPORT.29 Berlin 2017

Table 11: True class membership and cluster analysis results. In addition, the original numbering is shown as it is in Sarantaridis et al. (2012)

No.	Class	P5	P6	P7	P3	No. Orig.	No.	Class	P5	P6	P7	P3	No. Orig.
1	4	1	5	1	3	107	61	3	1	6	3	3	68
2	4	1	5	1	3	108	62	1	3	6	3	2	30
3	2	2	3	2	1	60	63	3	2	3	2	3	75
4	3	2	3	2	1	66	64	4	1	1	5	3	96
5	2	1	6	3	2	53	65	3	4	3	2	2	64
6	3	1	6	3	2	67	66	3	4	3	2	3	73
7	4	3	5	1	2	100	67	4	3	1	5	3	91
8	3	2	3	2	3	62	68	1	3	6	6	2	11
9	2	4	4	2	2	50	69	1	3	6	3	1	29
10	4	1	6	4	3	105	70	3	4	3	2	3	88
11	4	1	6	4	3	118	71	2	4	4	4	1	42
12	4	1	2	1	2	109	72	4	2	4	4	3	115
13	4	2	1	5	3	99	73	2	1	6	3	2	55
14	3	4	4	2	2	85	74	1	3	5	7	2	16
15	2	2	4	2	3	43	75	3	1	5	5	3	72
16	3	2	3	2	3	71	76	4	1	2	1	3	119
17	1	3	6	6	2	4	77	4	5	2	3	2	94
18	3	2	3	2	1	81	78	4	1	2	1	1	98
19	1	1	5	7	2	6	79	1	3	6	3	2	3
20	4	2	4	2	3	102	80	3	4	4	2	2	70
21	1	4	4	4	1	18	81	4	1	6	3	2	106
22	4	1	1	5	3	104	82	1	3	5	7	3	19
23	2	2	4	2	2	45	83	4	3	1	5	3	93
24	2	2	3	2	1	35	84	1	3	6	6	2	5
25	3	2	3	2	3	82	85	2	1	5	7	2	33
26	4	1	6	1	3	120	86	2	2	4	2	3	46
27	2	2	3	2	1	36	87	3	2	3	2	3	86
28	3	2	3	2	2	90	88	2	3	5	7	3	51
29	3	3	6	6	2	61	89	3	2	3	2	3	80
30	3	1	6	6	2	63	90	1	3	6	3	3	14
31	1	3	6	3	2	10	91	3	3	2	6	2	65
32	4	1	2	3	1	95	92	2	1	5	7	3	31
33	4	1	2	1	3	116	93	4	2	4	2	3	97
34	3	2	3	2	2	87	94	1	1	5	3	3	15
35	4	1	6	4	3	110	95	3	2	3	2	2	69
36	2	2	3	2	2	57	96	1	3	2	3	1	26
37	2	1	6	3	1	54	97	1	3	6	6	2	2
38	4	1	1	5	3	101	98	1	3	6	6	2	20
39	4	1	1	5	2	112	99	1	3	6	6	2	22
40	1	3	6	6	2	27	100	4	5	2	1	2	117
41	1	3	6	3	1	9	101	1	3	6	6	3	21
42	3	2	3	2	1	78	102	3	2	3	2	1	84
43	2	2	4	2	1	44	103	1	3	2	3	2	8
44	1	3	6	6	2	25	104	2	2	3	2	3	40
45	4	1	6	4	2	113	105	2	3	5	7	2	47
46	2	2	3	2	2	37	106	1	3	6	6	3	13
47	1	3	6	3	2	12	107	2	1	6	3	1	56
48	3	2	3	2	3	89	108	1	3	6	6	2	7
49	3	1	5	3	3	79	109	2	2	4	2	3	41
50	2	2	3	2	2	49	110	2	2	3	2	1	52
51	2	2	3	2	2	48	111	1	3	5	7	2	17
52	4	1	4	4	3	92	112	1	1	6	3	2	1
53	3	1	5	7	3	74	113	2	2	3	2	2	58
54	3	2	3	2	2	83	114	1	1	6	1	2	24
55	2	1	5	7	2	39	115	4	1	2	3	1	103
56	2	4	3	2	2	32	116	3	1	3	2	1	77
57	4	5	1	5	3	111	117	1	1	2	3	1	28
58	1	3	6	6	3	23	118	3	2	3	2	1	76
59	2	4	4	2	3	38	119	2	2	3	2	3	34
60	2	2	3	2	2	59	120	4	2	1	5	3	114

Table 12: True class (rows) vs. result "P5" of G. Szepannek, see Table 16 in Sect. 10.

Class	1	2	3	4	5	Total
BGP	5	0	**24**	1	0	30
BSS	7	**17**	2	4	0	30
JSS	7	16	2	**5**	0	30
BWP	**19**	5	3	0	3	30
Total	38	38	31	10	3	120

Table 13: True class (rows) vs. result "P6" of G. Ritter, see Sect. 11.

Class	1	2	3	4	5	6	Total
BGP		3		1	5	**21**	30
BSS			13	**8**	5	4	30
JSS		1	**20**	2	3	4	30
BWP	**9**	8		4	3	6	30
Total	9	12	33	15	16	35	120

The "Chemsensor" dataset was collected for testing a new method to detect bioaerosol particles based on gaseous plasma electrochemistry. The presence of such particles in air has a big impact on health, but monitoring bioaerosols poses great technical challenges. Sarantaridis et al. (2012) attempted to tell several different bioaerosols apart based on voltage changes over time on eight different electrodes when particles passed a premixed laminar hydrogen/oxygen/nitrogen flame. Four biological particles were tested: Bermuda grass pollen (BGP), Bermuda smut spores (BSS), Johnson smut spores (JSS) and Black walnut pollen (BWP). These define the "true classes" in the dataset. Each class comprised of 30 single events, giving a total of 120 events (particulates, observations) for analysis. An event was defined as the response from a single biological particle interacting with the detector, composed of 8 time series (from 8 electrodes) of potential difference values lasting 4 ms. In turn, every time series consisted of 301 voltage points/values, so that for each event a 2408-dimensional vector was constructed. These were provided to the participants of the competition.

Sarantaridis et al. (2012) carried out supervised classification after reducing the dimensionality by summarizing the relevant information in the time series in seven characteristic features, namely maximum voltage in series, minimum voltage in series, maximum voltage change caused by electrode, difference between final and initial voltage, length of positive change caused by the electrode, length of negative change caused by the electrode and time point at which the more extreme maximum was achieved. This therefore yielded 7*8=56 variables. These were not provided to the participants of the competition in order to give participants the option to use their own methods of dimension reduction. In fact, this is data for supervised classification. Even with sophisticated methods of dimension reduction and classification it is a difficult task to derive a good classifier. Sarantaridis et al. (2012) wrongly classified 31 observations using linear discriminant analysis on the subset of 56 variables. This corresponds to an error rate of about 26%, see Table 10.

Table 14: True class (rows) vs. result “P7” of R. Schachtner, see Sect. 12.

Class	1	2	3	4	5	6	7	Total
BGP	1		12	1		**12**	4	30
BSS		20	4	1			**5**	30
JSS		**22**	3		1	3	1	30
BWP	**9**	2	4	6	9			30
Total	10	44	23	8	10	15	10	120

Table 15: Confusion matrix of the true class (rows) vs. result “P3” of M. Weber.

Class	1	2	3	Total
BGP	5	**19**	6	30
BSS	**8**	14	8	30
JSS	6	11	13	30
BWP	3	7	**20**	30
Total	22	51	47	120

9.2 Comparison of the results of the competitors

The dataset accompanied by a shortened description without the reference was sent to the participants. However, it was risky to deliver such a description. In addition, the original observations (Sarantaridis et al. (2012)) comes already grouped (class 1: 1–30, 2: 31–60, 3: 61–90, 4: 91–120). I was afraid that competitors searched the web for additional information such as number of clusters and their cardinality. So, I rearranged the observations randomly before sending them to the participants.

Four data analysis experts took part in the competition: Gero Szepannek, Gunter Ritter, Reinhard Schachtner, and Marcus Weber. Table 11 shows the cluster analysis results, namely the partitions “P5”, “P6”, “P7”, and “P3”, respectively. The corresponding confusion matrices are shown in Tables 12, 13, 14, and 15. None of the competitors found out the true number of clusters. I was a little bit happy about this because it removed my worries that the competition was unfair. The “Chemsensor” dataset can be downloaded from the website
`http://www.wias-berlin.de/workshops/dank2016/committee.jsp`.

References

SARANTARIDIS, D., HENNIG, C. and CARUANA D. J. (2012): Potentiometric Tomography in Flames for Bioaerosol detection. Chemical Science 3, 2210–2216.

10 Clustering of Time Series Data

Gero Szepannek
Stralsund University of Applied Sciences, Germany
gero.szepannek@fh-stralsund.de

Abstract

A brief summary of the AG DANK 2016 data analysis task of clustering time series data is given. The important steps include the choice and discussion of an appropriate preprocessing, dimensionality reduction as well as clustering algorithm and model selection.

Introduction

In tradition of earlier fall meetings of the AG DANK the meeting was accompanied by a data set provided to the participants in order to compare and discuss results and methodology for its analysis: For 120 particulates (observations), there are eight time series of voltage values, corresponding to eight electrodes. Every time series consists of 301 voltage measurements. Values had been transformed by subtracting the first time point so that this is zero for each of the eight electrodes, because according to the chemists this is not informative (Sarantaridis et al., 2012).

The task consisted in the unsupervised identification of groups (i.e. clusters) of particulates where the true groups were not known in advance. Comparison with the true groups (classes) is used as a criterion to evaluate the results of the participants.

10.1 Methodology

Preprocessing Technique: An important subtask of cluster analysis consists in an appropriate preprocessing of the data (cf. e.g. Roever and Szepannek, 2005, Weihs and Szepannek, 2008). The emphasis of preprocessing is evident as no supervisor is available to evaluate results but the evaluation of models has to be done based on the input variables itself. A description of several preprocessing strategies is given in Hennig and Liao (2013).

For time series typically it is not meaningful to consider each time point as a separate variable which would fail for time shifts. Often time series are mapped to the frequency domain. In the research field of automatic speech recognition an additional *cepstral analysis* (subsequent inverse Fourier transform) is commonly used to extract relevant spectral information: harmonics appear as periodic signals in the spectrum and thus via high coefficients in the *cepstrum*. In consequence only the amplitudes of lower cepstral coefficients are used for further modeling (Davis and Mermelstein, 1980).

Figure 40 shows the original data for one electrode. There is no periodic structure in the data visible. Cepstral preprocessing is done using `R` package `tuneR` (Ligges, 2016) for each of the eight electrodes separately although its appropriateness questionable for the time series under consideration as well as a subsequent unit variance scaling of the coefficients.

10.2 Cluster Algorithms

Cluster Algorithms: As there are eight vectors of cepstral coefficients an intuitive idea might consist in a simultaneous dimensionality reduction and clustering. The ORCLUS algorithm (Aggarwal and Yu, 2000) as implemented in the `R` package `orclus` (Szepannek, 2013a) performs alternating PCA and observation clustering. In order to determine both an appropriate subspace dimension and number of clusters the sparsity coefficient can be used (Szepannek, 2011).

Figure 41 (left) shows an example of cluster specific subspaces as identified by the ORCLUS algorithm. In a simulation study orclus results turned out to be sensitive to the parameterization (Szepannek, 2011). Clusters as in Fig. 41 (left) can be interpreted as normal mixtures with a cluster-specific correlation structure. For this reason an alternative approach consists in the popular model based MCLUST (Fraley and Raftery, 2012). Here, BIC allows for a model selection between different candidate results. Figure 41 (right) shows BIC of different models depending on the covariance structure and cluster number.

10.3 Results and Discussion

The result of MCLUST using cepstral preprocessing showed the highest accuracy among several competing solutions presented on the workshop, although the graphical analysis of the original time series does not support its appropriateness. Table 16 shows the results of the clustering compared with the true classes. The BIC criterion did not identify the correct number of (four) classes here but cluster 5 of the solution does contain only 3 observations. The accuracy of 54% can be considered as an improvement of the Bayes prior (i.e. 25%) but in terms of prediction the result is not satisfactory. Especially class 2 and 3 are not separated by the clustering.

Class	1	2	3	4	5	sum
BGP	5	0	**24**	1	0	30
BSS	7	**17**	2	4	0	30
JSS	7	16	2	**5**	0	30
BWP	**19**	5	3	0	3	30
sum	38	38	31	10	3	120

Table 16: Confusion matrix of true classes (rows) vs. clusters (columns) of the optimal solution using MCLUST.

Generally, unsupervised learning is able to identify a structure in the input variables which is typically done by maximization of a predefined criterion but there is no guarantee that this structure is related to a supervisor (Szepannek, 2013b). An uprising question for this data set will be the analysis of signal and noise with respect to the true classes for the variables under investigation. A supervised learning approach might serve to complete the picture.

I would like to acknowledge Prof. A. Geyer-Schulz from IISM at Karlsruhe Institute for Technology for interesting discussions on John Tukey and Cepstral analysis during the meeting.

References

Aggarwal, C. and Yu, P. (2000): Finding Generalized Projected Clusters in High Dimensional Spaces. *ACM SIGMOD* 29(2), 70–81.

Davis, K. and Mermelstein, P. (1980): Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. In: *IEEE Transactions on Acoustics Speech and Signal Processing* 28 (4), 357–366.

Fraley, C. and Raftery, A. (2012): MCLUST Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. *Technical Report no. 597*, Department of Statistics, University of Washington.

Hennig, C. and Liao, T. (2013): How to find an appropriate clustering of mixed type variables with applications to socio-economic stratification. *JRSS C*, 62, 309–369.

Ligges, U. (2016): `tuneR`: Analysis of Music and Speech. `R` package version 1.3-1. `http://CRAN.R-project.org/package=tuneR`.

Roever, C. and Szepannek, G. (2005): Application of a genetic algorithm to variable selection in fuzzy clustering. In C. Weihs and W. Gaul (eds): *Classification - The Ubiquitous Challenge*, 674–681, Springer.

Sarantaridis, D., Hennig, C. and Caruana, D. (2012): Bioaerosol detection using potentiometric tomography in flames, *Chemical Science* 3, 2210–2216.

Szepannek, G. (2011): ORCLUS Subspace Clustering using R, Talk at 33. AG DANK, 11.11.2011, Düsseldorf.

Szepannek, G. (2013a): `orclus`: Subspace Clustering. `R` package version 0.2-5. `http://CRAN.R-project.org/package=orclus`.

Szepannek, G. (2013b): Contribution to the discussion of the paper of Hennig, C. and Liao, T.: How to find an appropriate clustering of mixed type variables with applications to socio-economic stratification. *JRSS* C, 62, 309–369.

Weihs, C. and Szepannek, G. (2008): Distances in Classification. *Transactions on Case-Based Reasoning* 2, 3–14.

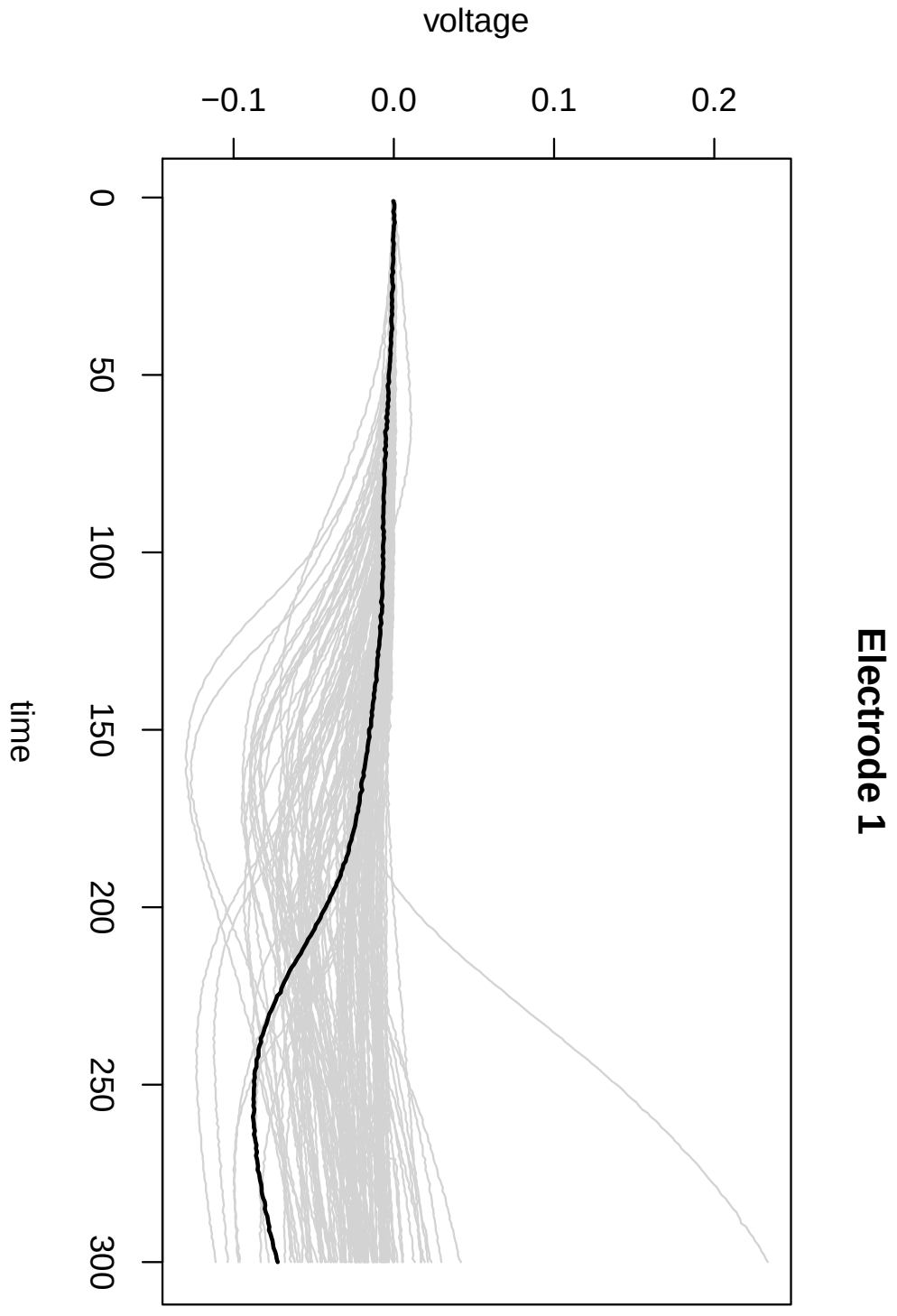


Figure 40: Time series of the first electrode for all 120 particulates.

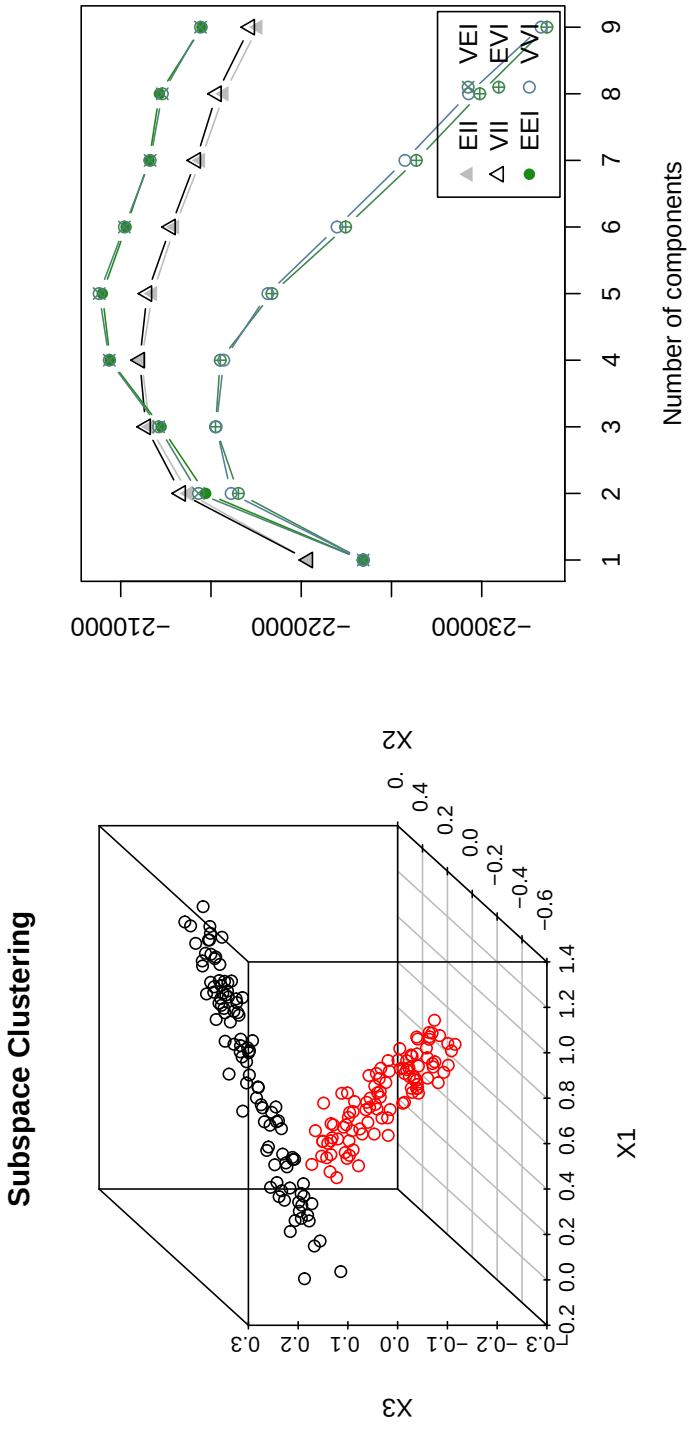


Figure 41: Example of clusters within an associated subspace (left) and MCLUST model selection (right).

11 Probabilistic Analysis of "Chemsensor"

Gunter Ritter
Faculty of Informatics and Mathematics
University of Passau, Germany
ritter@fim.uni-passau.de

"Chemsensor" is a classical numerical data set with 120 lines of length 2408, each. Each line consists of eight concatenated time series produced by eight sensors; see Figure 42. This suggests to treat the observations as objects in the sense of pattern recognition.

Obvious landmarks (or characteristic points) are the spikes and the endpoints visible in the plots of the eight time series. I used as features the locations and values of the spikes of the curves and their final values. This resulted in a new data set of 28 variables. Since there are only 120 observations, this number is still too large. Therefore, the probabilistic clustering and selection algorithm described in Ritter [2], Chapter 5, was applied to detect eight relevant variables. See also my contribution "Probabilistic variable selection in cluster analysis" in the present volume. The number eight is just a guess (and maybe still too large for just 120 observations). The algorithm is based on the classification and selection likelihood

$$\tfrac{1}{2}\sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_{j,F}(\boldsymbol{\ell}) + n\mathrm{H}\left(\tfrac{n_1(\boldsymbol{\ell})}{n},\ldots,\tfrac{n_g(\boldsymbol{\ell})}{n}\right) - \tfrac{n}{2}\log\det S_F.$$

Here, $n\ (=120)$ is the size of the data set, $\boldsymbol{\ell}\colon 1\,..\,n \to 1\,..\,g$ is a cluster assignment, g is the number of clusters, $n_j(\boldsymbol{\ell})$ is the size of cluster j w.r.t. assignment $\boldsymbol{\ell}$, $F \subseteq 1\,..\,120$ is a subset of variables, $S_{j,F}$ is the scatter matrix of cluster j w.r.t. the variables in F, and S_F is the total scatter matrix w.r.t. F.

I finally fed the eight selected variables to a program that implements the determinant criterion and a Gaussian hierarchical method for initial solutions; see Fraley [1]. The algorithm needs a cluster number as input parameter. I chose six, since this number yielded a partition with reasonably separated clusters. Figure 43 displays the related SBF plot. It shows the criteria and HDBT ratios of all "local" minima found for six clusters. I selected the solution closest to the left lower corner as the favorite one. Pairwise plots of its clusters are presented in Figure 44.

References

[1] Chris Fraley. Algorithms for model-based Gaussian hierarchical clustering. *SIAM J. Scient. Comp.*, 20:270–281, 1998.

[2] Gunter Ritter. *Robust Cluster Analysis and Variable Selection*. Chapman & Hall/CRC, Boca Raton, London, New York, 2015. Monographs in Statistics and Applied Probability 137.

DOI 10.20347/WIAS.REPORT.29 Berlin 2017

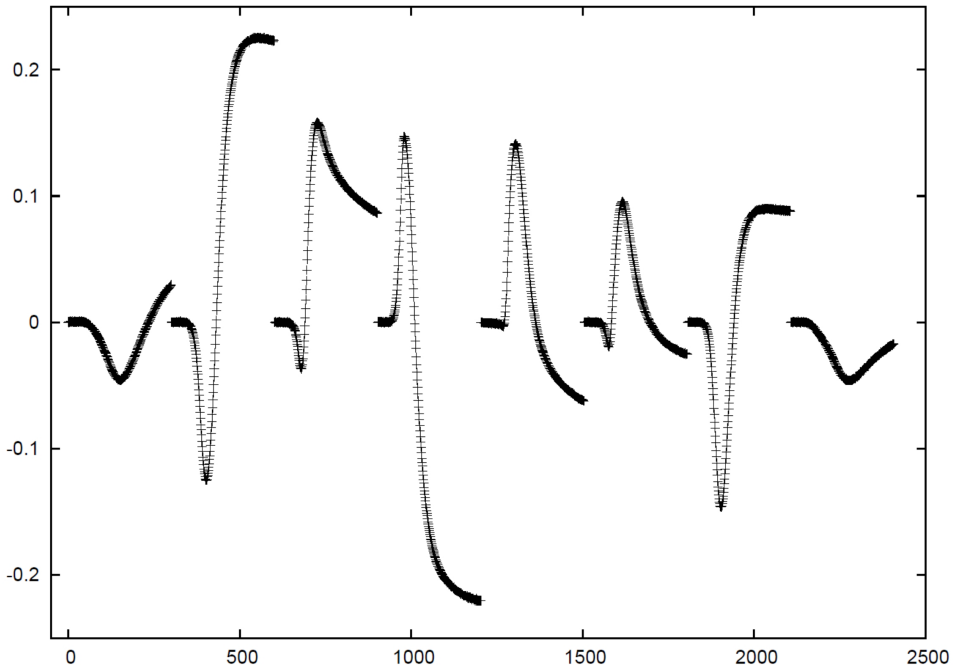


Figure 42: One line in the data set.

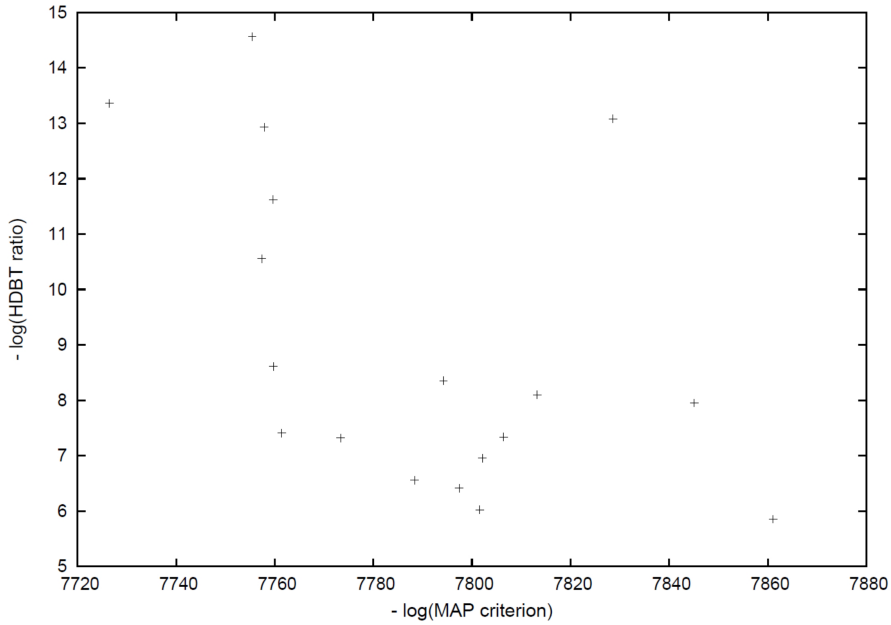


Figure 43: SBF plot

Figure 44: Pairwise plots of the six clusters.

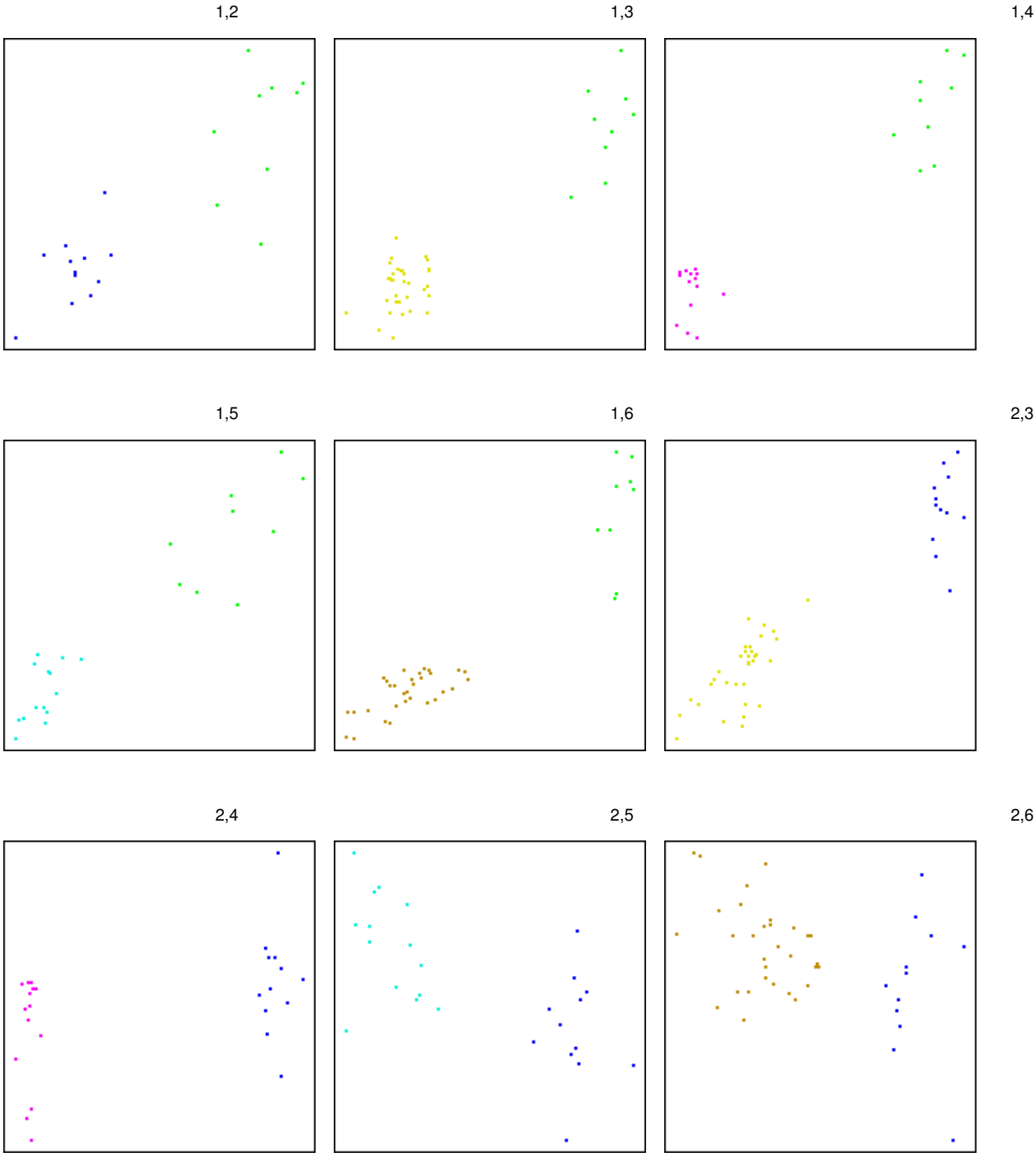
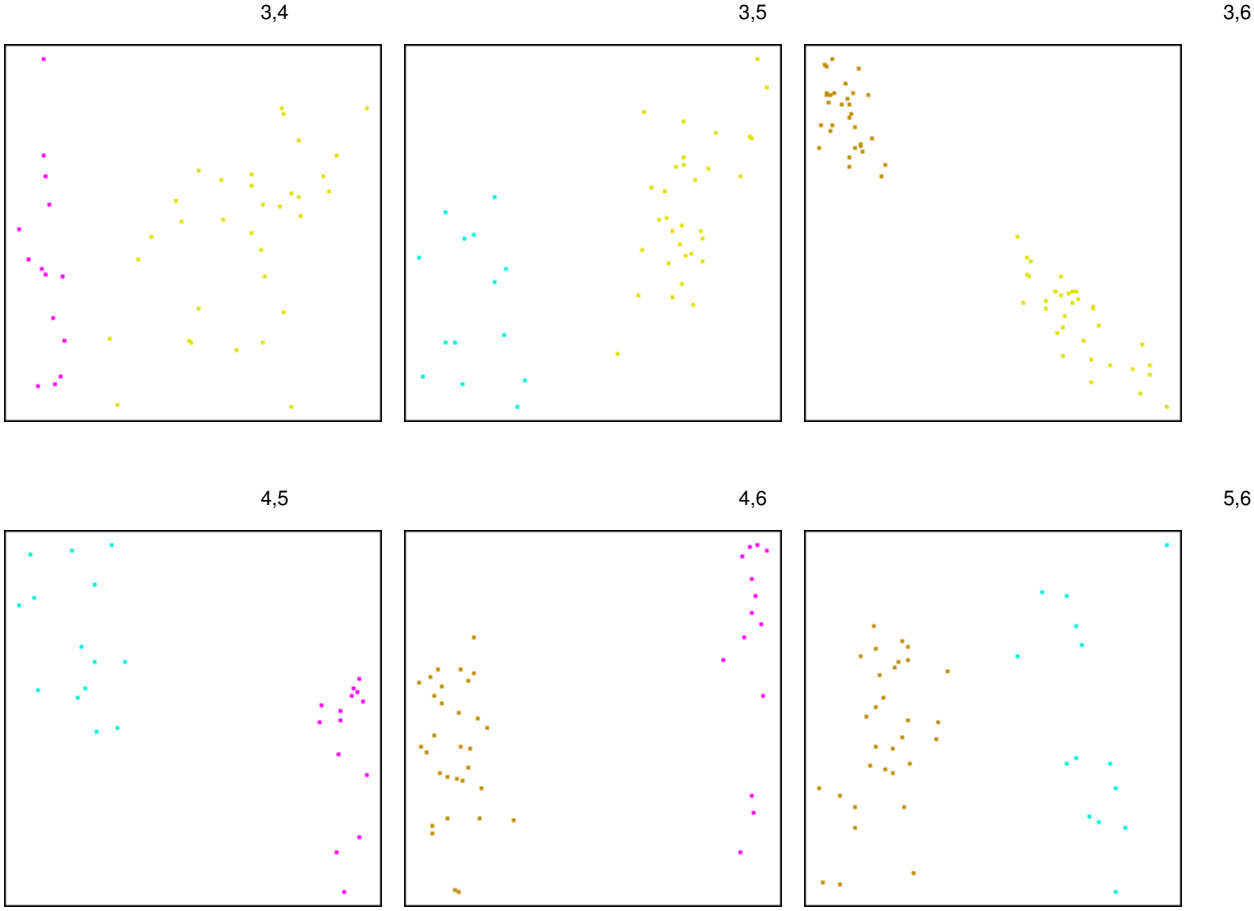


Figure 44: Pairwise plots of the six clusters, continued.



12 Analysis of the Flame Plasma Electrochemical Sensor Dataset

Reinhard Schachtner, Gerhard Pöppel and Thomas Siegert
Infineon Technologies AG Regensburg, Germany
reinhard.schachtner@infineon.com

Abstract

We give a short summary of our approach to the exploratory analysis of the flame plasma electrochemical sensor data set which constitutes this year's competition data of the AG DANK autumn meeting in Berlin 2016. After a short description of the techniques applied we motivate our decision in favour of the finally chosen cluster solution.

12.1 Inspection of the raw data

The investigated data consists of $120 \times 8 \times 301$ measurement values comprising 120 particulates, measured at 8 electrodes at 301 consecutive time points. According to the dataset description, values have been transformed subtracting the first time point so that this is zero for each of the eight electrodes.

We started our analyses by visual inspection of the data. For each of the 8 electrodes each particulate shows a typically smooth curve starting at zero (see figure 45 for data examples). We noticed a few measurements to be accompanied by small noise (presumably stemming from the measurement procedure) and one particulate (case 46) to contain a discontinuity. All together, the mentioned irregularities which can be typical for real world measurements were found to be at an ignorable low level and we did not consider any kind of outlier removal procedure necessary here.

12.2 Clustering Methods

Throughout this contribution, we utilize the clustering algorithms available in the free statistics software R [4] under the hierarchical clustering function `hclust`. This choice seems reasonable for two reasons:

1 The number of underlying clusters was not given in advance, hence a dendrogram representing the overall clustering structure is expected to reveal this information to the data analyzer.

2 The function `heatmap` offers the possibility to display the re-ordered raw data together with a cluster dendrogram. Hence, a given data partition can be easily checked for plausibility by human inspection.

In the present case no such thing like a ground truth, i.e. the true cluster memberships, was provided in advance. Hence, the respective data partitions generated by various hierarchical clustering

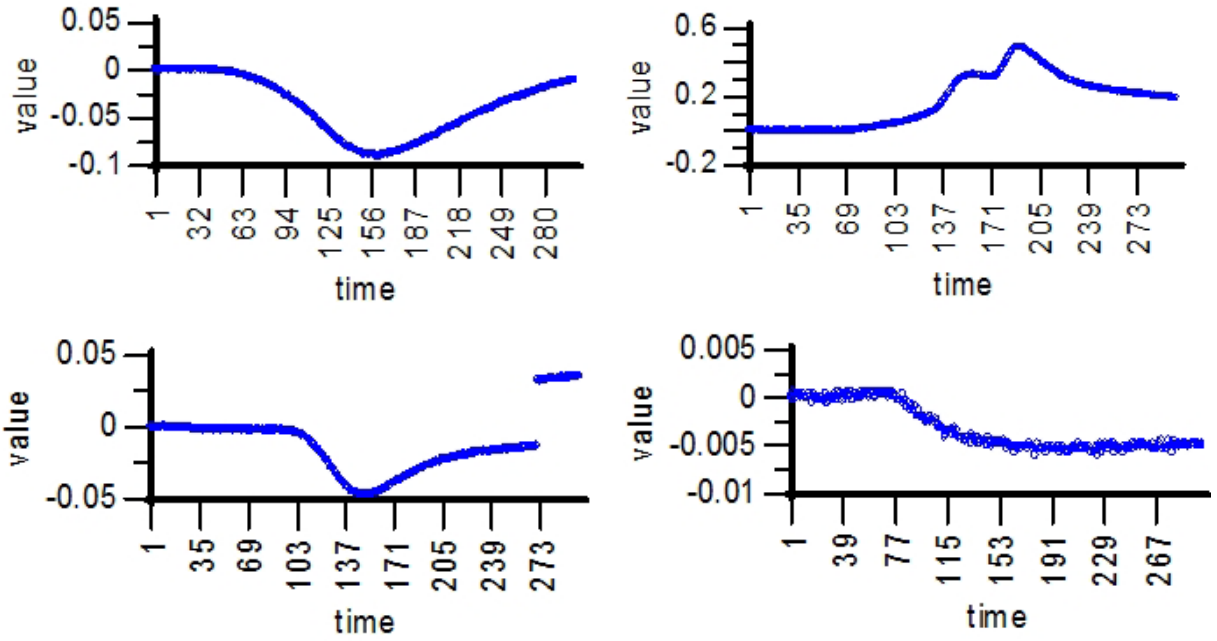


Figure 45: top: Two typical instances of the dataset: 1 particulate in 1 electrode, measured at 301 time points bottom left: case 46 has a discontinuity, bottom right: measurement accompanied by small noise

algorithms were evaluated with respect to their plausibility by human perception. We also applied various data preprocessing variants like scaling and centering, as well as approaches to reduce the dimensionality as described in section 12.3.

Summarizing, Ward's agglomerative hierarchical clustering method turned out to produce the most comprehensible data partitions. It is interesting to notice that there exist several algorithms claiming to implement Ward's method as described in [7], but can lead to different clustering results (see [3] for a discussion).

12.3 Data Preprocessing

Data Transformations

Depending on the physical quantities to be measured, measurement equipment sometimes needs calibration which can cause offsets and scaling effects in the data. We therefore investigated the impact of scaling and centering operations on the dataset by subtracting the mean value for every particulate and electrode and applying a z-standardization respectively. The observed dendrogram structures did not show remarkable simplifications in these cases.

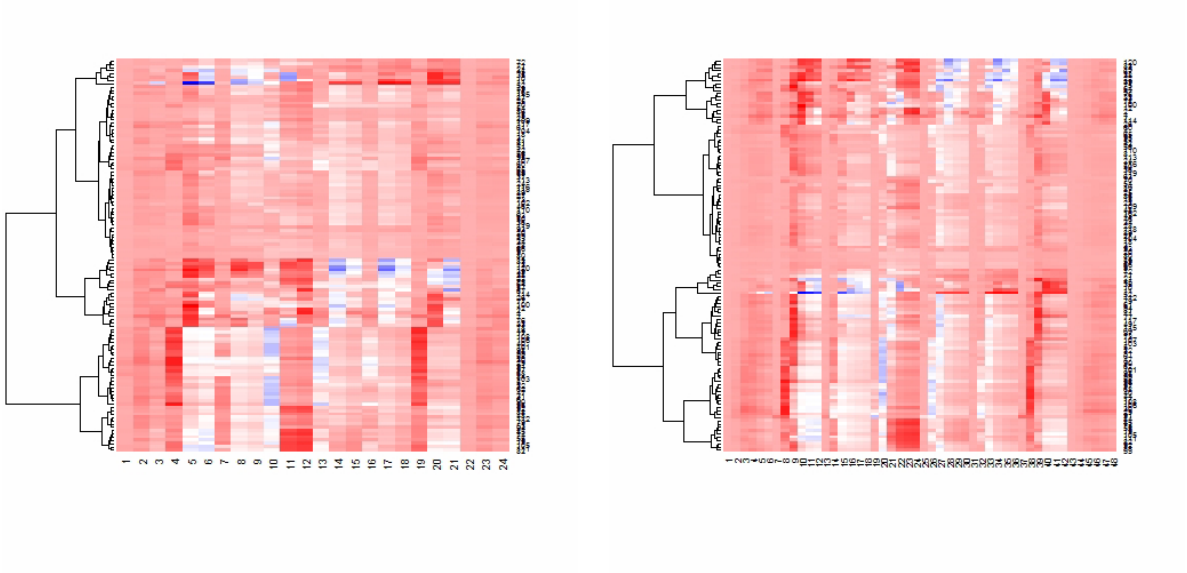


Figure 46: Hierarchical clustering of two compressed versions of the dataset
left: 301 time points averaged into 3 groups, right: 301 time points averaged into 6 groups per electrode and particulate. Small values are displayed in red, large values in blue

Dimensionality Reduction

Being confronted with the setting of $n = 120$ observations in $p = 2403$ dimensions the data analyst might feel unhopefully surrendered to the curse of dimensionality. In order to improve the relation between variables and observations, dimensionality reduction approaches were applied to the data. A simple way of compression of the investigated data set can be done by taking averages over time where the 301 time points per particulate and electrode are grouped into $2, 3, \ldots$ time ranges and the group averages are taken as input to the cluster algorithm (see figure 46 for examples).

Again, it turned out that the overall clustering structure did not improve remarkably compared to the original, uncompressed data.

Matrix and Tensor Factorization

In the analysis of multidimensional biomedical data sets it is common practice to apply so called exploratory matrix factorization (EMF) techniques like Principal Component Analysis (PCA), Independent Component Analysis (ICA) or Nonnegative Matrix Factorization (NMF) seeking uncorrelated, statistically independent or strictly non-negative features which characterize the data sets under study and serve as discriminative features for classification purposes (see [2] and references therein). Under certain assumptions which need to be verified in the respective applications, the EMF approaches typically yield a low-rank representation of the original datasets which can be used as input for a clustering algorithm, thus avoiding the problem of clustering in high dimensions. If the data contains non-negative values only, non-negative tensor factorization approaches can be applied (see e.g. [1] for a survey on non-negative matrix and tensor factorization). We discuss one such approach in more

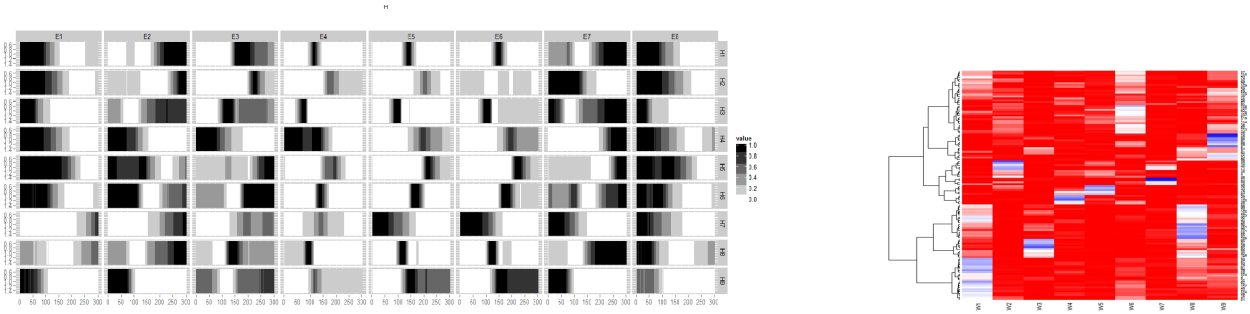


Figure 47: Entries of the basis tensor $h_{ljk} \geq 0$ (left) and clustered weight matrix $w_{il} \geq 0$ (right) resulting from a non-negative tensor factorization of the transformed electrochemical sensor data set into $L = 9$ components using the method from [6]

detail.

Assume that the entry (i,j,k) of a non-negative data array $x_{ijk} \geq 0$ can be represented as a function of L underlying components

$$x_{ijk} = f\left(\sum_{l=1}^{L} w_{il} h_{ljk}\right) \tag{22}$$

where an element of the weight matrix $w_{il} \geq 0$ is related to the contribution of component l on constituent i, while an element of the basis tensor $h_{ljk} \geq 0$ describes component l on electrode j at time k.

In order to apply the non-negative tensor factorization algorithm described in [6] to the electrochemical sensor data set, the $120 \times 8 \times 301$ entries need to be transformed into the range between $[0, 1]$ so that it can be interpreted as something like a normalized physical intensity. In figure 47 an example decomposition of the transformed data into $L = 9$ components is given. The 120×9 weight matrix (right) represents the data in a space spanned by the $9 \times 8 \times 301$ basis tensor (left), and was used as input for the clustering procedure described in section 12.2.

Note that the data representation in the form of eq. (22) implies the assumption that the data is generated as a certain superposition of underlying sources. Moreover, the choice of an optimal number of underlying basic components L is not straight forward, and different representations lead to different cluster partitions of the data. The clustering structure shown in figure 47 is not notably simpler than in original space and no additional information on the data generating process was available to justify the superposition assumption.

12.4 Proposed Solution

Since the participants of the competition were requested to submit one single cluster partition before the workshop started, we had to restrict ourselves in favour of one solution among the many diverse variants obtained during our analyses.

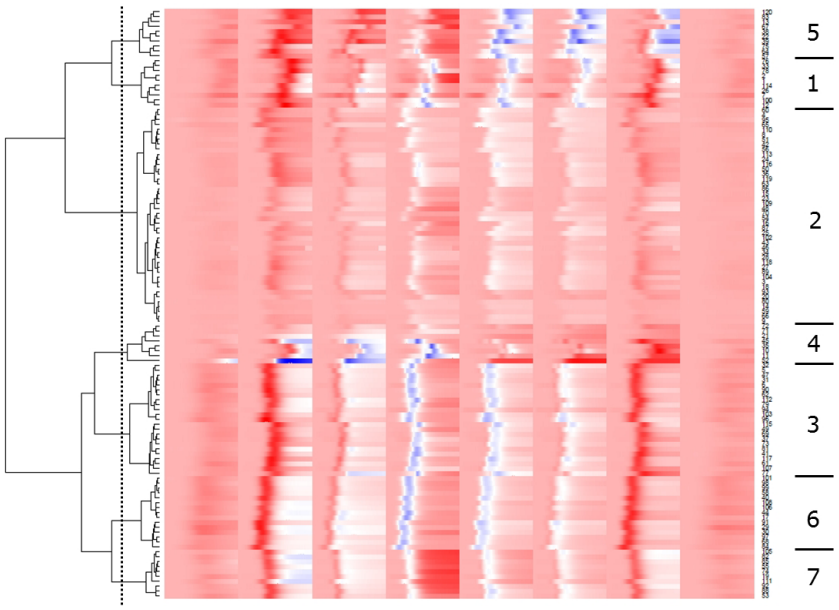


Figure 48: Proposed clustering solution. The dashed vertical line on the left hand side represents the chosen clustering partition, the numbers on the right hand side denote the seven cluster labels. Small values are red, large values are blue.

As explained in Subsection 12.2 the visual appearance of a clustering structure served as a criterion for its quality. Neither of the evaluated data manipulations discussed in section 12.3 showed an obvious advantage over the clustering structure obtained via untransformed data, and no additional knowledge on the data generating process was given which can give rise for the necessity of a certain data preprocessing. Hence, we decided to chose the solution depicted in figure 48 which can be obtained via the hierarchical Ward's clustering on untransformed data to be our contribution to the data analysis competition.

We decided that an intersection of the dendrogram at 7 clusters leads to a data partition which is well comprehensible by human perception (see Table 11, column "P7", for the cluster memberships). Note for example that case 52 in cluster 4 which constitutes one extra cluster in other cluster variants follows a similar trace as its cluster mates from cluster 4 but with higher intensity in both up and down directions.

In an attempt to cluster validation, we give a visualization of the determined clustering solution in figure 49. It turns out that it is possible to find a low-dimensional data representation in a supervised fashion which exhibits the seven clusters as distinct data clouds well-separated from each other.

Therefore, the original data was first transformed via PCA. In this ill-posed small n large p setting, at most 120 linearly independent principal components can be extracted from the data. The first 113 extracted principal components corresponding to the numerically nonzero eigenvalues of the correlation matrix together with the cluster labels were fed into a Manova module, essentially performing a linear discriminant analysis (LDA). The computations for principal component analysis and linear discriminant analysis described in figure 49 were done using Cornerstone software (camLine GmbH).

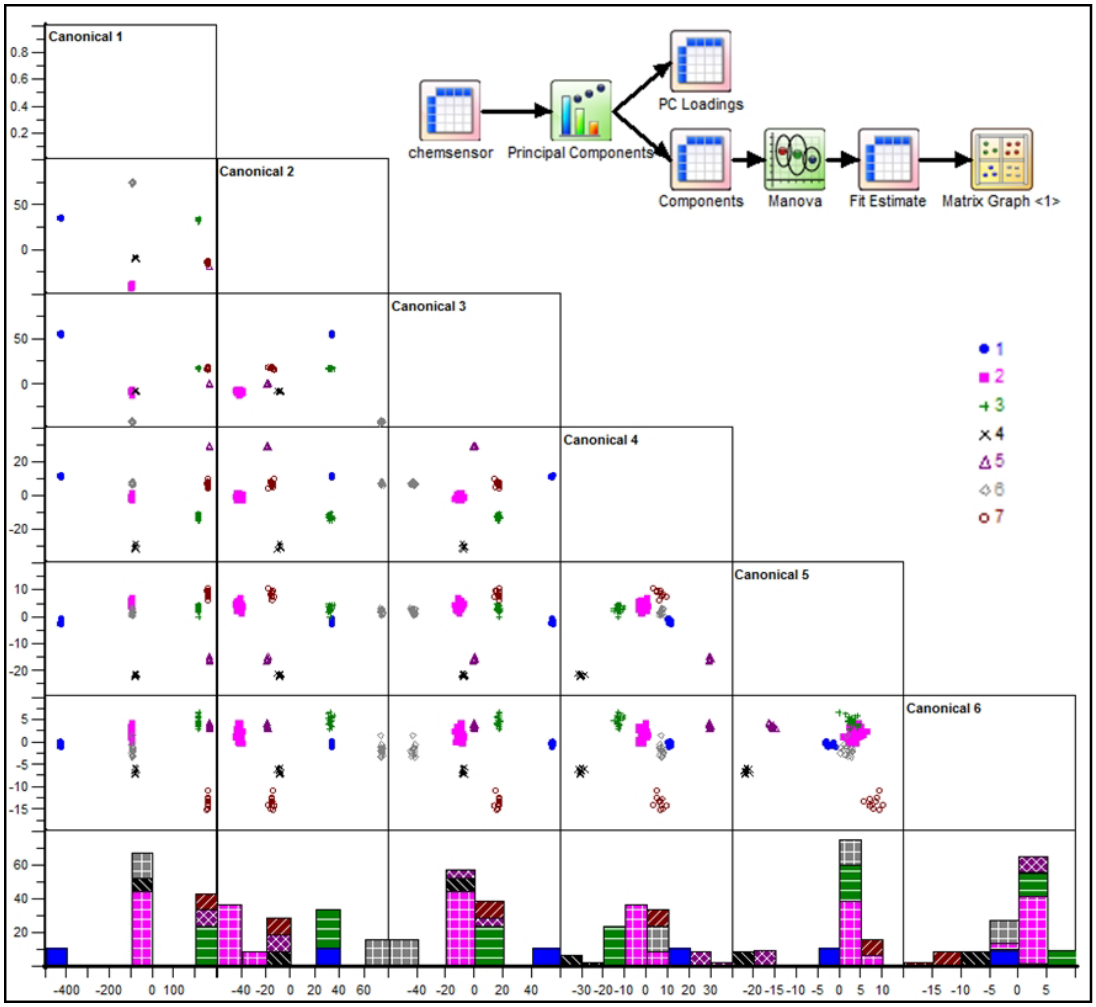


Figure 49: Visualization of the proposed clustering solution Table 11, column “P7”, in terms of a linear discriminant analysis (LDA) on the first 113 principal components of the data.

Discussion

After the discussion of all proposed solutions during the workshop, the unravelled ground truth turned out to be 4 different classes of particles [5].

It is interesting to notice that a clustering structure which seems reasonable (see the representations in figures 48 and 49) has comparatively little in common with the truly underlying partition.

A possible explanation of this behavior for real world data sets is that the physical measurement process or measurement equipment can induce additional structure to the data (beneath other influence factors possibly not considered to be relevant during the data recording).

As an example, consider data analysis in a semiconductor fabrication environment, where one is often concerned with yield detractors originating in the manufacturing process. In this case the data stems from measurement equipment which evaluates the performance of the fabricated integrated circuits. Typically, multiple devices are tested concurrently on the same test equipment on different test sites to improve testing throughput. This procedure is called multisite testing. In consequence, the test equipment can generate additional site-specific characteristics in the measurement data, (i.e. offsets between the respective test sites) overlaying the data properties originating from the manufacturing processes.

References

[1] Cichocki, A., Zdunek, R., Phan, A.-H., and Amari, S.: Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation, *John Wiley & Sons*, 2009

[2] Lang, E.W., Schachtner, R., Lutter, D., Herold, D., Kodewitz, A., Blöchl, F., Theis, F.J., Keck, I., Gorriz, J. M., Gomez-Vilda, P. and Tome, A. M. : Exploratory Matrix Factorization Techniques for Large Scale Biomedical Data Sets, *Bentham Science Publishers, Hilversum*, 22, pp. 26-47, 2011

[3] Murtagh, F. and Legendre, P.: Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 31, 2014

[4] R Core Team (2014). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria
http://www.R-project.org/

[5] Sarantaridis, D., Hennig, C. and Caruana, D. J.: Bioaerosol detection using potentiometric tomography in flames, *Chemical Science* ,3 , pp. 2210-2216, 2012.

[6] Siegert, T., Schachtner, R., Pöppel, G. and Lang, E.W.: A Nonnegative Tensor Factorization Approach for Three-Dimensional Binary Wafer-Test Data, *Proceedings of the 15th IEEE International Conference on Machine Learning and Applications*, 2016

[7] Ward, J.H.: Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association*, 58, 301, 1963

13 List of participants

List of participants of the 38th AG DANK autumn meeting, Nov. 18th to 19th 2016, WIAS, Berlin

- ◯ **Baier**, Daniel, Prof. Dr., Universität Bayreuth
- ◯ **Dickhaus**, Thorsten, Prof. Dr., Universität Bremen
- ◯ **Erlekam**, Franziska, Dr., Zuse Institute Berlin (ZIB)
- ◯ **Fischer**, Bernd, Dr., German Cancer Research Center (DKFZ), Heidelberg
- ◯ **Fuhrmann**, Tino, Universität Karlsruhe
- ◯ **Geyer-Schulz**, Andreas, Prof. Dr., Universität Karlsruhe
- ◯ **Hennig**, Christian, Dr., University College London (UCL), Department of Statistical Science, Great Britain
- ◯ **Kurz**, Peter, Dipl.-Math., TNS Infratest, München
- ◯ **Lausen**, Berthold, Prof. Dr., University of Essex, Mathematical Sciences, Colchester, Great Britain
- ◯ **Mucha**, Hans-Joachim, Dipl.-Math., WIAS Berlin
- ◯ **Müller-Funk**, Ulrich, Prof. Dr., Universität Münster
- ◯ **Niknejad**, Amir, Prof. Dr., Zuse Institute Berlin (ZIB)
- ◯ **Osterloh**, Kurt, BAM Berlin
- ◯ **Rese**, Alexandra, Dr., Universität Bayreuth
- ◯ **Ritter**, Gunter, Prof. Dr., Universität Passau
- ◯ **Röhrl**, Norbert, Dr., Universität Stuttgart
- ◯ **Sauerbrei**, Willi, Prof. Dr., Universität Freiburg
- ◯ **Schachtner**, Reinhard, Dr., Infineon Technologies AG, Regensburg
- ◯ **Schweizer** , Marvin Luca Raphael Connor, Universität Karlsruhe
- ◯ **Sperber**, Wolfram, Dr., FIZ Karlsruhe
- ◯ **Szepannek**, Gero, Prof. Dr., Fachhochschule Stralsund
- ◯ **Tabelow**, Karsten, Dr., WIAS Berlin
- ◯ **Weber**, Marcus, Dr., Zuse Institute Berlin (ZIB)
- ◯ **Wilhelm**, Adalbert, Prof. Dr., Jacobs University Bremen

DOI 10.20347/WIAS.REPORT.29 Berlin 2017