Maastricht University | universiteit hasselt

2014 | Faculty of Sciences

DOCTORAL DISSERTATION

Challenges in Cluster Analyses for Longitudinal Data

Doctoral dissertation submitted to obtain the degree of
Doctor of Science: Statistics, to be defended by

Liesbeth Bruckers

Promoter: Prof. Dr Geert Molenberghs | UHasselt / tUL

D/2014/2451/53

universiteit hasselt | I-BioStat | KU LEUVEN

Interuniversity Institute for Biostatistics
and statistical Bioinformatics

Preface

This dissertation puts together a number of research topics in the area of model-based cluster analysis. The topics are all inspired by consultancy collaborations.

In 1996, I joined the Center for Statistics as a statistical consultant. I graduated that year from the Master of Statistics program and was confident I had sufficient statistical training to tackle a wide range of consultancy projects, in terms of study designs, types of data, and research questions to be investigated.

The first project I got involved in was a collaboration between SPIL and two research groups of Hasselt University, i.e., Health Psychology represented by Prof. dr. Jan Vinck and Biostatistics. SPIL is the platform of all psychiatric initiatives in Limburg. The platform aims at underpinning policy guidelines and reorganisation of psychiatric care, through data exchange and transparent communication among the initiatives. For each patient, seeking help, a wealth of information is collected. Of course, the main objective is to determine the most appropriate and effective therapy for the patient. But at an aggregated level, this information unfolds the landscape of mental health in Limburg. To draw the map of psychiatric care, a large set of items, registered for all patients, was used. Items relating to the geographical setting, patient characteristics, his/her social situation, diagnostic information, type of care needed, and items relating to the available expertise. To increase the efficiency of the mental health care, it is desirable to create wards where patients with similar behavior and in need of similar treatment are grouped. A group of patients that could benefit from such a re-location, were patients whose behaviour was disturbing in such a way that it disrupts the working of the whole ward. These are the so-called PDB (persistent disturbing behavior) patients. Despite the numerous items registered, the label 'PDB' was not available in the data file as such. According to medical caretakers, the PDB patients were quite diverse in terms of their diagnostic classification, the way they (mis)behaved, the measures needed to adjust this behavior, the stage setting of the wards, etc. In short, the PDB group was not a homogenous group of patients. Rather, it was believed that subtypes of PDB existed.

From a statistical point of view, this problem seemed not that hard to be investigated. Cluster analysis was the way to go. But immediately it became clear that the methods, with which I had familiarized myself during my training, were not fully addressing all aspects of the data. PDB is not a static condition, nor is the care that patients need or receive fixed. To capture this, repeated measurements data was used and a technique to group similar profiles was applied. To reduce the amount of information, the set of items characterizing PDB was reduced into one PDB-score, by means of logistic regression. Simplifying the multivariate repeated measurements setting to a univariate one. It is probably this research project, rooted in the womb of SPIL, that triggered my curiosity in cluster analysis, and hence was the start of the research presented in this work.

The following years, I was confronted with a variety of projects where interest was in revealing natural groupings in the data, groupings for which no labels (yet) were present in the data. Each of these studies had its own statistical challenges to be tackled. Data resulting from animal experiments can be high dimensional, in the sense that the evolution of a large number of parameters is monitored. We used multivariate repeated EEG observations to cluster rats with similar evolutions. Studies in rats are well controllable. As long as the animals do not break out or die, a balanced and complete data structure can be obtained. Studies in humans, on the other hand, generally result in unbalanced data with complex patterns of missing data. Standard cluster analysis assumes that the fact that an observation is missing, does not hold information about the measurement itself, given all data. Presumably this is not true for the 'Aortic Abdominal Aneurysms' (AAA) study. The increase in the artery's diameter of patients with an AAA is monitored carefully, since rupture of the artery is likely to be a death warrant. In case of a fast growing artery, the patient is removed from the study, resulting in missing observations for the diameter. The informative missingness in the data should be considered when performing a cluster analysis. The 'Heart Failure' (HF) study combines a number of challenges. Heart failure patients are being monitored by means of electric devices. Daily measurements for blood pressure (systolic and diastolic), heart rate, and weight are collected. This results in a huge set of observations — daily measurements for 6 markers during a period of 6 months — of which part is missing.

It is clear that the topics addressed in the thesis are very applied; the goal is to better support applied research in human and medical sciences.

Dankwoord

Bij deze wil ik iedereen bedanken die de laatste jaren, rechtstreeks of onrechtstreeks, bijgedragen heeft tot het tot standkomen van deze doctoraatsthesis.

Mijn promotor Prof. dr. Geert Molenberghs ben ik zeer dankbaar voor alle kansen die ik gekregen heb binnen het Centrum voor Statistiek. Mijn eerste jaren als consultant waren zeer leerrijk. De ervaringen opgedaan in het buitenland tijdens het bijwonen van cursussen, workshops en congressen zijn onbetaalbaar. De eerste inhoudstabel van dit doctoraat is trouwens opgesteld in februari 2004, tijdens ons gezamelijk bezoek aan Havanna, Cuba. Zonder je uitstekende en enthousiaste begeleiding, je waardevolle input en goede raad, al die jaren lang, was dit nooit gelukt. Dank je wel, Geert.

SPIL, Maastricht Universitair Medisch Centrum, Janssen Pharmaceutica en het hartcentrum van het Jessa Ziekenhuis ben ik dankbaar voor het aanleveren van interessante gegevens en onderzoeksvragen. Dank jullie wel voor de waardevolle discussies en fijne samenwerking.

In de loop der jaren heb ik via mijn werk als consultant contacten gelegd met een groot aantal onderzoekers uit verschillende disciplines. Elk van deze samenwerkingen heb ik als zeer positief ervaren, maar twee samenwerkingen wil ik in het bijzonder vernoemen. Er wordt nergens zo hartelijk gelachen als op SPIL-overlegmomenten, dank je wel Miet Smeets en alle SPIL-partners voor de reeds 18-jarige samenwerking. De deadlines voor het steunpunt Milieu en Gezondheid zijn altijd krap, maar toch kijk ik steeds uit naar de maandelijkse overlegmomenten. Dank aan alle steun-punters voor de fijne samenwerking, maar in het bijzonder aan Elly Den Hond. Dank je wel, voor je voortdurende stimulans en peptalks bij de moeilijke momenten zowel voor het steunpunt-gerelateerde werk als voor mijn doctoraatswerk.

Verder wil ik alle collega's van Censtat, huidige en ex-collega's, bedanken voor de aangename samenwerking. Dank jullie wel voor de leuke koffiepauzes. Er zijn heel wat collega's en bureaugenoten gepasseerd de afgelopen jaren. Het is onmogelijk iedereen apart te bedanken, maar velen van jullie zijn ondertussen zoveel meer dan collega's. Helena, Kris, Ivy, Hilde, Tina en Francesca voor jullie toch een speciale dank je wel voor de hilarische momenten, de gezellige etentjes en het vele gebak.

Net zo belangrijk was de stimulans van vrienden. Dank jullie wel voor de bemoedigende woorden, de interesse en de leuke uitstapjes. Dank aan de 'verjaardag-club' voor de fijne avonden. Bart, dank je wel voor de vele filmpjes en buikpijn-momenten van het lachen. Dank ook aan mijn salsa-kameraden. De kilometers die wij samen afgelegd hebben op de dansvloer zijn ontelbaar, de uren die we nadien zijn blijven plakken net zo. Dank je wel Kris, Veronique, Kathleen, Linda en Eddy voor de broodnodige ontspanning.

Mijn laatste woord van dank is voor mijn familie. Mama, papa zonder jullie onvoorwaardelijk steun en belangstelling was dit werk niet mogelijk geweest. Bedankt voor de kansen die jullie mij hebben gegeven. Mama, je kan je vraag 'moet je nog lang leren?' opbergen. Het zit er eindelijk op :-) Aan mijn broers, zus en schoonzussen, dank jullie wel voor de leuke activiteiten elke zondagnamiddag. De filmpjes, fietstochtjes en terrasjes waren perfect om mijn hoofd leeg te maken.

Iedereen van harte bedankt!

Liesbeth Bruckers
Diepenbeek, 12 december 2014

List of Publications

Material covered in this dissertation:

L. Bruckers, L., Molenberghs,G., Poncelet, J., Brouns, K., Cuypers, W. Slaets, H., and Vanheyst, I. (2000). Identificatie en inschatting van de omvang van de groep patiënten met persisterend storend gedrag. *Acta Hospitalia*, **41**, 21–30.

Deboel, T., Bruckers, L., Rogiers, G., Dox, E., Slaets, H., Molenberghs, G., and Goeyvaerts w. (2009). Een kwantitatieve kijk op persisterend storend gedrag bij psychiatrische patiënten. *Acta Hospitalia*, **49**.

Bruckers, L., Serroyen, S., Molenberghs, G., Slaets, H., Goeyvaerts, W. (2010). Latent class analysis of persistent disturbing behaviour patients by using longitudinal profiles. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59**, 495–512.

Serroyen, J., Bruckers, L., Rogiers, G., and Molenberghs, G. (2010). Characterizing persistent disturbing behavior using longitudinal and multivariate techniques *Journal of applied statistics*, **37**, 341–355.

Hellenthal, F.A., Pulinx, B., Bruckers, L., Molenberghs, G., Kleinveld, H.A., Welten, R., van Dieijen-Visser, M.P., Wodzig, W.K., and Schurink, G.W.H. (2014). Prediction of Abdominal Aortic Aneurysm. *European Journal of Vascular and Endovascular Surgery. Submitted.*

Bruckers, L., Molenberghs, G., Pulinx, B., Hellenthal, and Schurink G. (2014) Cluster Analysis for Repeated Data with Dropout: Sensitivity Analysis Using a Distal Event. *Journal of Biopharmaceutical Statistics.Accepted, under review*

Bruckers, L., Molenberghs, G., Drinkenburg, P., and Geys H. (2014). A Cluster Algorithm for Multivariate Longitudinal Data. *Journal of Biopharmaceutical Statistics.Accepted, under review*

Bruckers, L., Molenberghs, G., Verbeke, G., and Geys, h. (2014). Detecting Infuential Observations in a Model-Based Cluster Analysis. *Statistical Methods in Medical Research.Accepted, under review*

Bruckers, L., Molenberghs, G., and Dendale, P. (2014) Clustering Multiply Imputed Multivariate High-Dimensional Longitudinal Profiles. *To be submitted.*

Additional publications:

Croes, K., De Coster, S., De Galan, S., Morrens, B., Loots, I., Van de Mieroop, E., Nelen, V., Sioen, I., Bruckers, L., Nawrot, T., Colles, A., Den Hond, E., Schoeters, G., van Larebeke, N., Baeyens, W., and Gao, Y. (2014). Health effects in the Flemish population in relation to low levels of mercury exposure: from organ to transcriptome level. *International Journal of Hygiene and Environmental Health*, **217**, 239–247.

Vrijens, J., Leermakers, M., Stalpaert, M., Schoeters, G., Den Hond, E., Bruckers, L., Colles, A., Nelen, V., Van den Mierop, E., Van Larebeke, N., Loots, I., and Baeyens, W. (2014). Trace metal concentrations measured in blood and urine of adolescents in Flanders, Belgium: Reference population and case studies Genk-Zuid and Menen. *International Journal of Hygiene and Environmental Health*, **217**, 515–527.

Remy, S., Govarts, E., Bruckers, L., Paulussen, M., Wens, B., Den Hond, E., Nelen, V., Baeyens, W., van Larebeke, N., Loots, I., Sioen, I., and Schoeters, G. (2014). Expression of the sFLT1 gene in cord blood cells is associated to maternal arsenic exposure and decreased birth weight. *PLOS ONE*, **9**, 1–11.

Croes, K., Colles, A., Koppen, G., De Galan, S., Vandermarken, T., Govarts, E., Bruckers, L., Nelen, V., Schoeters, G., Van Larebeke, N., Denison, M.S., Mampaey, M., and Baeyens, W. (2013). Determination of PCDD/Fs, PBDD/Fs and dioxin-like PCBs in human milk from mothers residing in the rural areas in Flanders, using the CALUX bioassay and GC-HRMS. *Talanta*, **113**, 99–105.

Den Hond, E., Paulussen, M., Geens, T., Bruckers, L., Baeyens, W., David, F., Dumont, E., Loots, I., Morrens, B., de Bellevaux, B.N., Nemery, B., Nelen, V., Schoeters, G., Van Larebeke, N., and Covaci, A. (2013). Biomarkers of human exposure to personal care products: Results from the Flemish Environment and Health Study (FLEHS 2007-2011). *Science of The Total Environment*, **463**, 102–110.

Vundelinckx, B., Dierickx, C., Bruckers, L., and Dierickx, C.H. (2012). Functional and radiographic medium-term outcome evaluation of the Humerus Block, a minimally invasive operative technique for proximal humeral fractures. *Journal of Shoulder and Elbow Surgery*, **21**, 1197–1206.

Kicinski, M., Viaene, M., Den Hond, E., Schoeters, G., Covaci, A., Dirtu, A., Nelen, V., Bruckers, L., Croes, K., Sioen, I., Baeyens, W., Van Larebeke, N., and Nawrot, T.S. (2012). Neurobehavioral function and low-level exposure to brominated flame retardants in adolescents: a cross-sectional study. *Environmental Health*, **11**:86.

Morrens, B., Bruckers, L., Den Hond, E., Nelen, V., Schoeters, G., Baeyens, W., Van Larebeke, N., Keune, H., Bilau, M., and Loots, I. (2012). Social distribution of internal exposure to environmental pollution in Flemish adolescents. *International Journal of Hygiene and Environmental Health*, **215**, 474–481.

Schoeters, G., Den Hond, E., Colles, A., Loots, I., Morrens, B., Keune, H., Bruckers, L., Nawrot, T., Sioen, I., De Coster, S., Van Larebeke, N., Nelen, V., Van de Mieroop, E., Vrijens, J., Croes, K., Goeyens, K., and Baeyens, W. (2012). Concept of the Flemish human biomonitoring programme. *International Journal of Hygiene and Environmental Health*, **215**, 102–108.

Croes, K., Colles, A., Koppen, G., Govarts, E., Bruckers, L., Van de Mieroop, E., Nelen, V., Covaci, A., Dirtu, A.C., Thomsen, C., Haug, L.S., Becher, G., Mampaey, M., Schoeters, G., Van Larebeke, N., and Baeyens, W. (2012). Persistent organic pollutants (POPs) in human milk: A biomonitoring study in rural areas of Flanders (Belgium). *Chemosphere*, **89**, 988–994.

Bruckers, L and Smeets, M. (2012). Wie doet beroep op onze Limburgse GGZ-voorzieningen ? *Overspil: informatieblad van het overlegplatform 'Samenwerking Psychiatrische Initiatieven Limburg'*, **19**, 4–7.

Schoeters, G., Den Hond, E., Colles, A., Loots, I., Morrens, B., Bruckers, L., Sioen, I., Van Larebeke, N., Nelen, V., Van De Mieroop, E., Vrijens, J., Croes, K., Baeyens, W., and Covaci, A. (2012). The Flemish Environment and Health Study (FLEHS) Second Survey (2007-2011): Establishing Reference Values for Biomarkers of Exposure in The Flemish Population. *Biomarkers and Human Biomonitoring Volume 1 : Ongoing Programs and Exposures, Knudsen, L., Merlo, D.F. (Ed.)*, 135–165.

Keunen, E., Truyens, S., Bruckers, L., Remans, T., Vangronsveld, J., and Cuypers, A. (2011). Survival of Cd-exposed Arabidopsis thaliana: Are these plants reproductively challenged? *Plant Physiology and Biochemistry*, **49**, 1084–1091.

Croes, K., Van Langenhove, K., Den Hond, E., Bruckers, L., Colles, A., Koppen, G., Loots, I., Nelen, V., Schoeters, G., Nawrot, T., Van Larebeke, N., Denison, M.S., Vandermarken, T., Elskens, M., and Baeyens, W. (2011). Quantification of PCDD/Fs and dioxin-like PCBs in small amounts of human serum using the sensitive H1L7.5c1 mouse hepatoma cell line: Optimization and analysis of human serum samples from adolescents of the Flemish human biomonitoring program FLEHS II. *Talanta*, **85**, 2484–2491.

Den Hond, E., Dhooge, W., Bruckers, L., Schoeters, G., Nelen, V., van de Mieroop, E., Koppen, G., Bilau, M., Schroijen, C., Keune, H., Baeyens, W., and van Larebeke, N. (2011). Internal exposure to pollutants and sexual maturation in Flemish adolescents. *Journal of Exposure Science and Environmental Epidemiology*, **21,** 224–233.

Dhooge, W., Den Hond, E., Koppen, G., Bruckers, L., Nelen, V., van de Mieroop, E., Bilau, M., Croes, K., Baeyens, W., Schoeters, G., and van Larebeke, N. (2011). Internal exposure to pollutants and sex hormone levels in Flemish male adolescents in a cross-sectional study: associations and dose-response relationships. *Journal of Exposure Science and Environmental Epidemiology*, **21**, 106–113.

Morrens, B., Loots, I., and Bruckers, L. (2011). Programme flamand de biomonitoring sur les adolescents : les gradients sociaux observés diffèrent d'un polluant á l'autre. *Education Santé, hors série*, 10–13.

Dhooge, W., Den Hond, E., Koppen, G., Bruckers, L., Nelen, V., Van De Mieroop, E., Bilau, M., Croes, K., Baeyens, W., Schoeters, G., and Van Larebeke, N. (2010). Internal exposure to pollutants and body size in Flemish adolescents and adults: Associations and dose-response relationships. *Environment International*, **36**, 330–337.

Harries, R., Lawson, S., and Bruckers, L. (2010). Assessment of microcalcifications with limited number of high-precision macrobiopsies. *European Journal of Cancer Prevention*, **19**, 374–378.

Govarts, E., Den Hond, E., Schoeters, G., and Bruckers, L. (2010). Determinants of Serum PCBs in Adolescents and Adults: Regression Tree Analysis and Linear Regression Analysis. *Human and Ecological Risk Assessment*, **16**, 1115–1132.

Vankrunkelsven, P., Kellen, E., Lousbergh, D., Cloes, E., Op de Beeck, L., Faes, C., Bruckers, L., Mertens, R., and Buntinx, F. (2010). De verkoop van producten voor hormonale subsitutietherapie: evolutie van borstkankerincidentie tussen 1992-2008. *Tijdschrift voor geneeskunde*, **66**, 25–30.

Lousbergh, D., Buntinx, F., Rummens J., Op de Beeck, L., Vanden Brande, J., Dhollander, D., Kellen, E., Faes, C., Bruckers, L., Cloes, E., Lathouwers, D., Meekers, E., and Hensen, K. (2010). Tien jaar kanker in de provincie Limburg (1996-2005): incidenties, trends en voorspellingen. *Tijdschrift voor geneeskunde*, **66**, 3–10.

Lousbergh, D., Buntinx, F., Rummens, J., Op de Beeck, I., Vanden Brande, J., Dhollander, D., Kellen, E., Hensen, K., Faes, C., Bruckers, L., Lathouwers, D., Meekers, E., and Cloes, E. (2010). Tien jaar kanker in de provincie Limburg (1996-2005): de belangrijkste individuele lokalisaties. *Tijdschrift voor geneeskunde*, **66**, 11–18.

De Bock, K., Bruckers, L., and Coune, Y. (2010). Onderzoek naar no-show : Wie stuurt zijn kat? *Zorgwijzer*, **12**, 26–27.

Vankrunkelsven, P., Kellen, E., Lousbergh, D., Cloes, E., Op de Beeck, L., Faes, C., Bruckers, L., Mertens, R., Coebergh, J.W., Van Leeuwen, F.E., and Buntinx, F. (2009). Reduction in hormone replacement therapy use and declining breast cancer incidence in the Belgian province of Limburg. *Breast Cancer Research and Treatment*, **118**, 425–432.

Croes, K., Baeyens, W., Bruckers, L., Den Hond, E., Koppen, G., Nelen, V., Van de Mieroop, E., Keune, H., Dhooge, W., Schoeters, G., and Van Larebeke, N. (2009). Hormone levels and sexual development in Flemish adolescents residing in areas differing in pollution pressure. *International Journal of Hygiene and Environmental Health*, **212**, 612–625.

Den Hond, E., Govarts, E., Bruckers, L., and Schoeters, G. (2009). Determinants of polychlorinated aromatic hdrocarbons in serum in three age classes-Methodological implications for human biomonitoring. *Environmental Research*, **109**, 495–502.

Bilau, M., De Henauw, S., Schroijen, C., Bruckers, L., Den Hond, E., Koppen, G., Matthys, C., Van De Mieroop, E., Keune, H., Baeyens, W., Nelen, Vera., Van Larebeke, N., Willems, J.L., and Schoeters, G. (2009). The relation between the estimated dietary intake of PCDD/Fs and levels in blood in a Flemish population (50-65 years). *Environment International*, **35**, 9–13.

Koppen, G., Den Hond, E., Nelen, V., Van De Mieroop, E., Bruckers, L., Bilau, M., Keune, H., Van Larebeke, N., Covaci, A., Van De Weghe, H., Schroijen, C., Desager, K., Stalpaert, M., Baeyens, W., and Schoeters, G. (2009). Organochlorine and heavy metals in newborns: Results from the Flemish Environment and Health Survey (FLEHS 2002-2006). *Environment International*, **35**, 1015–1022.

Robaeys, G., Nevens, F., Starkel, P., Colle, I., Van Eyken, P., Bruckers, L., Van Ranst, M., and Buntinx, F. (2009). Previous Intravenous Substance Use and Outcome of Liver Transplantation in Patients With Chronic Hepatitis C Infection. *Transplantation Proceedings*, **41**, 589–594.

De Coster, S., Koppen, G., Bracke, M., Schroijen, C., Den Hond, E., Nelen, V., de Mieroop, E., Bruckers, L., Bilau, M., Baeyens, W., Schoeters, G., and van Larebeke, N. (2008). Pollutant effects on genotoxic parameters and tumor-associated protein levels in adults: a cross sectional study. *Environmental Health*, **7**:26.

Kellen, E., Zeegers, M.P., Bruckers, L., and Buntinx, F. (2008). The Investigation of A Geographical Cluster of Bladder Cancer. *Acta Clinica Belgica*, **63**, 313–320.

Abatih, E., Van Oyen, H., Bossuyt, N., and Bruckers, L. (2008). Variance estimation methods for health expectancy by relative socio-economic status. *European Journal of Epidemiology*, **23**, 243–249.

Bilau, M., Matthys, C., Baeyens, W., Bruckers, L., De Backer, G., Den Hond, E., Keune, H., Koppen, G., Nelen, V., Schoeters, G., Van Larebeke, N., Willems, J.L., and De Henauw, S. (2008) Dietary exposure to dioxin-like compounds in three age groups: Results from the Flemish environment and health study. *Chemosphere*, **70**, 584–592.

Keune, H., Loots, I., Bruckers, L., Bilau, M., Koppen, G., van Larebeke, N., Schoeters, G., Nelen, V., and Baeyens, W. (2008) Monitoring environment, health and perception: an experimental survey on health and environment in Flanders, Belgium. *International Journal of Global Environmental Issues*, **8**, 90–111.

Schroijen, C., Baeyens, W., Schoeters, G., Den Hond, E., Koppen, G., Bruckers, L., Nelen, V., Van De Mieroop, E., Bilau, M., Covaci, A., Keune, H., Loots, I., Kleinjans, J., Dhooge, W., and Van Larebeke, N. (2008). Internal exposure to pollutants measured in blood and urine of Flemish adolescents in function of area of residence. *Chemosphere*, **71**, 1317–1325.

Keune, H., Loots, I., Bruckers, L., Bilau, M., Koppen, G., van Larebeke, N., Schoeters, G., and Nelen, V. (2008). Monitoring environment, health and perception: an experimental survey on health and environment in Flanders, Belgium. *International Journal of Global Environmental Issues*, **8**, 90–111.

Kellen, E., Zeegers, M.P., Dirx, M., Houterman, S., Droste, J., Lawrence, G., Bruckers, L., Molenberghs, G., Joniau, S., Buntinx, F., and Truyers, C. (2007). Occurrence of both bladder and prostate cancer in five cancer registries in Belgium, The Netherlands and the United Kingdom. *European Journal of Cancer*, **43**, 1694–1700.

Robaeys, G., De Bie, L., Wichers, M.C., Bruckers, L., Nevens, F., Michielsen, P., Van Ranst, M., and Buntinx, F. (2007). Early prediction of major depression in chronic hepatitis C patients during peg-interferon alpha-2b treatment by assessment of vegetative-depressive symptoms after four weeks. *World Journal of Gastroenterology*, **13**, 5736–5740.

Vandeloo, M., Bruckers, L., and Janssens, J. (2007). Effects of lifestyle on the onset of puberty as determinant for breast cancer. *European Journal of Cancer Prevention*, **16**, 17–25.

Lousbergh, D., Cloes, E., Op de Beeck, L., Rummens, J.L., Vanden Brande, J., Faes, C., Bruckers, L., Molenberghs, G., Dhollander, E., Kellen, E., Hensen, K., Lathouwers, D., Meekers, E., and Buntinx, F. (2007). Ten years of cancer in the Belgian Province of Limburg.

Robaeys, G., Van Vlierberghe, H., Mathei, C., Van Ranst, M., Bruckers, L., and Buntinx, F. (2006). Similar compliance and effect of treatment in chronic hep-

atitis C resulting from intravenous drug use in comparison with other infection causes. *European Journal of Gastroenterology & Hepatology*, **18**, 159–166.

Van Vlierberghe, H., Leroux-Roels, G., Adler, M., Bourgeois, N., Horsmans, Y., Brouwer, J., Colle, I., Delwaide, J., Brenard, R., Bastens, B., Henrion, J., de Vries, R.A., de Galocsy, C., Michielsen, P., Nevens, Y., Robaeys, G., and Bruckers, L. (2003). Daily induction combination treatment with alpha 2b interferon and ribavirin or standard combination treatment in naive chronic hepatitis C patients. A multicentre randomized controlled trial. *Journal of Viral Hepatitis*, **10**, 460–466.

Tibaldi, F., Bruckers, L., Van Oyen, H., Van der Heyden, J., and Molenberghs, G. (2003). Statistical software for calculating properly weighted estimates from Health Interview Survey data. *Social and Preventive Medicine*, **48**, 269–271.

Robaeys, G., Van Vlierberghe, H., Mathei, S., Van Ranst, M., Bruckers, L., and Buntinx, F. (2003). Compliance and effect of treatment for chronic hepatitis C (CHC) in intravenous drug users (IVDUs). *Journal of Hepatology*, **38**, 165–165.

Hellings, N., Gellin, G., Medaer, R., Bruckers, L., Palmers, Y., Raus, J., and Stinissen, P. (2002). Longitudinal study of antimyelin T-cell reactivity in relapsing-remitting multiple sclerosis association with clinical and MRI activity. *Journal of Neuroimmunology*, **126**, 143–160.

Janssens, J., Peeters, L., Joossens, J.V., Hongenaert, A., Van Elsen, A., Bonte, J., Gourdin, P., Nagels, I., De Thibault De Bousinghe, L., Molenberghs, G., Vinck, J., Bruckers, L., Renard, D., Servaty, J., and Tafforeau, J. (2002). Obesity and alcohol consumption: alcohol drinking habits in Belgium and body mass index. *Cerevisia: Belgian Journal of brewing and biotechnology*, **27**, 99–106.

Tibaldi, F., Demarest, S., Van Oyen, H., Tafforeau, J., Bruckers, L., Molenberghs, G., and Van Steen, K. (2002). Changing strategies in the organization of the Belgian Health Interview Survey 2001. *Archives of Public Health*, **60**, 275-294.

Hens, N., Bruckers, L., Arbyn, M., Aerts, M., and Molenberghs, G. (2002). Classification Tree Analysis of Cervix Cancer Screening in the Belgian Health Interview Survey 1997. *Archives of Public Health*, **60**, 275–294.

Janssens, J., Van Hecke, E., Geys, H., Bruckers, L., Renard, D., and Molenberghs, G. (2001). Pesticides and mortality from hormone-dependent cancers. *European Journal of Cancer Prevention*, **10**, 459–467.

Renard, D., Bruckers, L., Molenberghs, G., Vellinga, A., and Van Damme, P. (2001). Repeated-measures models to evaluate a hepatitis B vaccination program. *Statistics in Medicine*, **20**, 951–963.

Janssens, J., Bruckers, L., Joossens, J.V., Molenberghs, G., Vinck, J., Renard, D., and Tafforeau, J. (2001). Overweight, obesity and beer consumption. Alcohol drinking habits in Belgium and body mass index. *Archives of Public Health*, **59**, 223–238.

Steen, R., Vuylsteke, B., DeCoito, T., Ralepeli, S., Fehler, G., Conley, J., Bruckers, L., Dallabetta, G., and Ballard, R. (2000). Evidence of declining STD prevalence in a South African mining community following a Core-Group intervention. *Sexually Transmitted Diseases*, **27**, 9–11.

Vellinga, A., Van Damme, P., Bruckers, L., Weyler, J., Molenberghs, G., and Meheus, A. (1999). Modelling long-term persistence of hepatitis B antibodies after vaccination. *Journal of Medical Virology*, **57**, 100–103.

Janssens, J., Bruckers, L., Molenberghs, G., Michiels, L., Staelens, Y., Deleu, M., and Raus, J.(1999). Prognosis of very young breast cancer patients. *Women and Cancer*, **1**, 34–41.

Janssens, J., Shapira, N., Debeuf, P., Michiels, L., Putman, R., Bruckers, L., Renard, D., and Molenberghs, G. (1999). Effects of soft drink and table beer consumption on insulin response in normal teenagers and carbohydrate drink in youngsters. *European Journal of Cancer Prevention*, **8**, 289–295.

Wuyts, F., De Bodt, M., Molenberghs, G., Bruckers, L., andBelgian Study Group on Voice Disorders (1996). Research Work of the Belgian Study Group on Voice Disorders 1996: Results. *Acta Oto-Rhino-Laryngologica Belgica*, **50**, 331–341.

Contents

List of Figures

List of Tables

Chapter 1

Overview of the Dissertation

Data clustering is a useful technique to explore multivariate data for a structure of *natural* groupings, for which there is no prior information available in the data. The groups exhibit different relationships among the multivariate outcomes. Clustering has also been used to identify outliers, and to suggest hypotheses concerning relationships. The term *data clustering* first appeared in an article published in 1954 with anthropological data and one of the most popular and simple clustering algorithms, K-means, was first published in 1955. The problem of organizing observations into sensible groupings is prevalent in many disciplines, as is clear from the number of applied and methodological articles published. Clustering has successfully been used.

- In transcriptomics clustering is used to build groups of genes with related expression patterns, known as coexpressed genes. Often such groups contain functionally related proteins or genes that are co-regulated.

- Based on multivariate data from surveys and test panels, market researchers use cluster analysis to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers.

- In mathematical chemistry, 3000 chemical compounds were clustered in the space of 90 topological indices based on structural similarity.

- In the fields of plant and animal ecology, clustering is used to describe as well as to make spatial and temporal comparisons of communities of organisms in heterogeneous environments.

- In astronomy, stars are classified in categories based on their light intensity and surface temperature.

- One of the principal research questions in community care is whether use of resources by different patient groups is aligned with a patient's clinical and psychosocial needs. However a patient's diagnosis alone, e.g., of schizophrenia, cannot explain differences in severity of illness among patients or their variable needs for resources. To better map a patient's use of resources and satisfaction with the care and a patient's needs, a cluster analysis of patients with schizophrenia was performed. The analyses revealed four clusters that differ on the basis of the severity of psychopathology, disability, and family burden (Lora et al., 2001).

Depending on the field, cluster analysis is also known as Q-analysis, typology, clumping, numerical taxonomy, unsupervised learning, learning by observation, etc. Cluster analysis is not one specific algorithm, but refers to the general task to be solved, i.e., grouping observations in such a manner that observations similar to each other, according to some measure of distance between data, are grouped together, while dissimilar observations belong to different groups. A multiplicity of clustering algorithms has been proposed in the literature; they differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Fraley and Raftery (1998) suggest dividing the clustering methods into two main groups: hierarchical and partitioning methods. Han and Kamber (2001) suggest categorizing the methods into three additional main categories: density-based methods, model-based clustering, and grid-based methods. Chapter 2 contains a compendious introduction to clustering and some clustering algorithms.

Although standard hierarchical and partitioning cluster algorithms are widely used and have shown to be effective, they are less adequate when clusters need to be discovered in data exhibiting complex structures. This is for example the case for repeated measurements data, for spatially obtained observations, and studies using clustered sampling. The standard techniques assume observations to be independent random realizations of some statistical model and similarity metrics are used to deal with sets of observations. Generally these techniques are not applicable when the data consists of profiles, as is the case for all case studies considered in this dissertation (see Chapter 3). The directionality information, contained in the profiles, is discarded in the cluster procedure. No restriction is placed on the mean structure via covariates or otherwise. Furthermore, traditional clustering algorithms require a fixed-dimensional size which is usually not the case for longitudinal studies. Measurement times may be unequally spaced within an individual and may differ across individuals resulting in highly unbalanced structures.

In these settings, model-based clustering methods, such as finite-mixture models, have advantages compared to non-probabilistic techniques. Finite-mixture models have a long history in statistics. They have been used to model population heterogeneity, generalize distributional assumptions, and lately, for providing a convenient yet formal framework for clustering and classification. In a model-based approach each cluster is mathematically represented by a parametric distribution, like a Gaussian or a Poisson. The data is described by the mixture of the distributions that optimizes the fit between the data and the specified model. Clusters are then defined as observations coming most likely from the same distribution. These models can be applied to directional data and allow for a parsimonious representation of the mean by putting restrictions on the model. The methodology conveniently handles missing and irregularly spaced measurements. Finally, the uncertainty for cluster-membership assignment of each observation is naturally quantified via the posterior probabilities. Finite-mixture models and model-based clustering are briefly introduced in Chapter 4. For illustrative purposes the Persistent Disturbing Behavior Data, previously analyzed by Serroyen et al. (2010), will be used. Serroyen et al. (2010) executed a conventional cluster analysis on cross-sectional data for this study, we will use the repeated measurements to reveal latent subgroups in the data.

Although cluster analysis has been used extensively in applied sciences and was a topic of many methodological papers, there are still a number of open and controversial topics. Some of these issues are specific to model-based clustering, e.g., via finite-mixture models, and some are common to a variety of clustering algorithms.

- Evaluating if a certain clustering algorithm is appropriate or not is a problematic and controversial issue. In fact Bonner (1964) was the first to argue that there is no universal definition for what is a good clustering. The evaluation remains mostly in the eye of the beholder. Nevertheless, several evaluation criteria have been developed in the literature. These criteria are usually divided into two categories: internal and external.

- Cluster performance can be optimized by excluding variables that are uninformative and irrelevant. The identification of variables with more discriminating power than others has been discussed by Raftery and Dean (2006).

- The decision about the number of clusters/components K is often equivocal. The optimal choice of K seeks a balance between maximum conglomeration of the data using a single cluster, and maximum accuracy by assigning each data point to its own cluster. The percentage of variance explained as a function of the number of clusters, the average silhouette of the data, cross-validation, and information criterion approaches when a likelihood function is involved, have been used to determine K.

- Melnykov (2013) discusses some problems typically encountered with model-based clustering. Estimation in finite-mixture models is conveniently done by maximum-likelihood estimation, implemented via the expectation-maximization algorithm. But the likelihood function may be unbounded, e.g., because of singular covariance matrices in the case of Gaussian mixtures with heterogeneous dispersions. The likelihood function can have spurious solutions, with the parameter vector solution lying close to the boundary of the parameters space. The EM algorithm is iterative and its performance can depend severely on particular starting points because the likelihood function often has numerous local maxima. Several approaches have been considered in the literature to find reasonable starting points (e.g., obtain an initial partition by K-means), but so far no recommendations for global initialization are available.

In this dissertation, we address a number of aspects and limitations of cluster analysis that received less attention in the statistical literature, so far. These topics are most naturally addressed in a model-based cluster approach. Cluster analysis becomes challenging when the dimensionality of the data increases. In this work, two chapters are devoted to this topic. Chapter 5 proposes an algorithm to detect similar patterns in a multivariate longitudinal data setting. Clustering of repeated measurements data, by means of finite-mixture models, has successfully been demonstrated. However, the methods described in the literature, are applicable in situations where observations need to be grouped based on the evolution over time of a limited set of responses, in general one and at maximum three. In a multivariate repeated measurements setting, where various outcomes are measured simultaneously over time, these cluster methods break down and easily run into computational problems due to an increase in the number of parameters to be estimated. To find clusters that are unique in the evolutions over time for the set of outcomes as well as in the correlation structure, an algorithm based on pseudo-likelihood estimation is presented in Chapter 5. The algorithm is inspired by work of Fieuws and Verbeke (2008), they used a bivariate joint-modelling approach to perform a discriminant analysis.

Modern data collection techniques permit observations to be densely sampled over a continuum, usually time, again enhancing the dimensionality of the data, even if only one response is monitored. In such situations, the observed data is seen as a realisation of a smooth underlying process. This type of data is referred to as functional data and methodology is available to describe and model dependencies between these so-called functional data curves. In general, a smoother is applied and the curse of dimensionality is circumvented by reducing the dimension of the data, e.g., via a principal component analysis, prior to statistical analysis. Population heterogeneity in the functional curves is then reflected as heterogeneity in the reduced data. Jacques and Preda (2013) demonstrated a cluster analysis for functional data when the data consists of bivariate functions. We apply this method to a setting where each observation is characterized by six curves and demonstrate that the reduced information about the dependencies in the data, allows to reveal homogenous groups (Chapter 6). The data reduction technique, which is an essential building block in the procedure of Jacques and Preda (2013), restricts its applicability to rectangular complete data structures. Records with missing values are discarded in the analyses. However, missing data are almost unavoidable in longitudinal studies. In Chapter 6, we present a solution to circumnavigate this problem. We propose to use the technique of multiple imputation to create a set of m complete data sets and subsequently execute a ensemble clustering to summarize the resulting m partitions into a final cluster result. Ensemble clustering refers to the situation in which a number of different partitions have been obtained for a particular dataset and it is desired to find a single (consensus) clustering which is a better fit in some sense than the existing clusterings. An ensemble can be obtained, for example, by varying the parameters of the clustering algorithm, by resampling or reweighting the set of objects, or by employing several different clustering algorithms. To our knowledge, ensemble clustering has not yet been used to summarize partitions obtained after multiple imputation. Chapter 6 introduces the procedure of Jacques and Preda (2013) and basic concepts of ensemble clustering. The usefulness of an assemblage of multiple imputation and ensemble clustering is illustrated for the functional heart failure data and by means of simulations.

In fact, multiple imputation assumes that the missing-data mechanism is missing at random (MAR). MAR allows the mechanism describing missingness to depend on covariates and observed outcomes but, given these, not further on unobserved outcomes. When the parameter space describing the measurement and missingness process equals the product of the individual parameter spaces inference can be based on the marginal observed data density only (ignorability). But MAR is a rather restrictive assumption and it can never definitively be excluded that missingness further depends on unobserved outcomes. This more general missing mechanism is referred to as missing not at random (MNAR). Under MAR and ignorability, a maximum-likelihood analysis will produce unbiased estimates. But biased estimates are obtained when the missingness mechanism is MNAR. The information about the response contained in it being observed or missing, can be incorporated in the statistical analysis by jointly modeling the measurement and missingness process. Depending on the chosen factorization of the joint likelihood, the model is a so-called selection model or pattern-mixture model. These models do not necessarily induce the same conclusions. The inherent difficulty is that they rely on unverifiable assumptions and that the data alone is inconclusive in voting for the best model. Chapter 7 applies a finite-mixture model to the measurement process and a series of missing-data models to the missingness process. These type of models have been used by Muthén et al. (2011), to model non-ignorable dropout in the STAR*D antidepressant trial. The authors focussed on the comparison of the results of the different models in terms of the number of clusters and the average cluster-specific profiles. We will execute a similar exercise for the abdominal aorta aneurysm (AAA) dataset, but we complement the model comparison with an evaluation of the sensitiveness of the posterior probabilities and the final classification of the subjects. Notable differences are seen in the results, but choosing between the models is very hard. Muthén et al. suggest that external information, associated with the latent grouping in the data, can assist in a sensible ranking of the different models. Such an exercise was also implemented for the AAA study and described in Chapter 7.

Noise and outliers affect the estimates of the parameters describing the components of a finite-mixture model, and as such the posterior probabilities and partition of the observations. Different methods have been described in the literature to investigate this. Cheng and Milligan (1996), for example, define an influential observation as an observation that, when removed from the data, leads to a different partitioning. However, interest could be in quantifying the influence of a single observation, not only on the final partitioning, but also on the parameter estimates describing the components. Chapter 8 deals with this topic. Influence in the mixed model has been investigated by Lesaffre and Verbeke (1998). They use a local influence analysis to assess the effect of perturbations from the usual assumptions in the mixed model. We obtained local influence diagnostics for finite-mixture models using case-weight perturbations. For each observation, this results is a set of influence diagnostics, of which one measures the influence on the mixture probabilities. However, even if no change in cluster-membership is observed the influence on the posterior probabilities can still be substantial. In Chapter 8, we show how local influence diagnostics obtained for finite-mixture models allow to quantify the influence of an observation on the posterior probabilities of all other observations, without refitting the finite-mixture model.

Finally, Chapter 9 summarizes the most important findings and sketches some topics that are still open for future research.

Chapter 2

Brief Resume of Cluster Analysis

Cluster analysis (Johnson & Wichern, 2007) refers to a collection of procedures that attempt to determine natural groupings (or clusters/classes/components) of objects (observations, events) in a population. The unsupervised classification is based on information found in the data describing the objects and their relationships. The term 'cluster' does not have a precise definition, but often a similarity-based definition is employed: *a cluster groups objects such that objects in the same cluster are more similar to each other (in some sense or another) than to objects in another cluster.*

Cluster analysis is often mixed-up with supervised classification, discriminant analysis or decision analysis. Although the techniques are related they serve different purposes. In a cluster analysis, a set of unlabeled objects is organized into similar groups. Whereas the aim of discriminant analysis is to label a new object (or to assign it to a group), based on a set of labeled objects.

While clustering is used to reveal natural groupings in the data, it also has been demonstrated to be valuable in detecting incorrect class-labels, outliers, errors, and bias.

In this chapter, we briefly introduce the building blocks of a cluster analysis. Jain and Dubes (1988) describe the following steps in a typical cluster analysis: object representation, choice of an object proximity measure, choice of a clustering algorithm, and evaluation of the obtained grouping. Each of these steps will be elaborated upon in the following paragraphs, without the intention of being complete. For details of specific clustering methods we refer to, for example, Johnson & Wichern (2007).

Object representation refers to the vector of measurements (also called features, attributes, variables) for each object used as input for the clustering algorithm. Since a cluster analysis is not able to differentiate between relevant and irrelevant features, data is sometimes preprocessed. Identifying a (effective) subset of the collected features to be used in the clustering algorithm is feature selection. Transforming the available features into a lower-dimensional space, e.g., via a principal component analysis, is feature extraction. The purpose of feature selection and extraction is to remove noise in the data and to obtain an interpretable final classification. Standardizing the features, such that clustering is not dominated by the feature with the widest range, is also part of the choice of the object representation.

Indispensable to most clustering methods is a quantification of the **similarity** between two objects. The similarity measure can be a metric or based on a probability distribution. Often a distance measure on the feature space is used. Among the most popular metrics are the Euclidean distance, Manhattan distance, Pearsons' correlation for continuous features, Spearman's rank correlation, Kendall's Tau for ordinal features, simple matching coefficient and Jaccard coefficient for binary features. Kullback-Leibler divergence and mutual information are (dis)similarity measures based on a probability distribution. The choice of the (dis)similarity measure should consider the features type and scale, the desired interpretation of similarity (e.g., proximity or association), sensitivity to outliers, and underling distributional assumptions for the features.

Many **clustering algorithms** have been described to discover groups of 'similar' objects. A distinction between hierarchical and partitional clustering methods can be made. The methods can differ in terms of the number of clusters an observation can be assigned to, i.e., hard versus fuzzy clustering. Furthermore, the methods can differ in how the features are used.

Hierarchical clustering techniques result in a nested series of partitions, called a tree or dendogram. The hierarchical clustering can proceed in a agglomerative or divisive way. The agglomerative approach starts with each object as a separate cluster, and successively merges the two nearest clusters together. This is done repeatedly until all objects are in one cluster. Divisive clustering considers all objects as a cluster and iteratively splits this cluster. The distance between two clusters can be constructed in multiple ways. Single linkage, complete linkage and average linkage define the distance between clusters as the minimum, maximum and respectively average distance between any two objects of the clusters. Ward's method, on the other hand, uses the increase in sum of squares when merging two clusters as the distance measure.

Partitional techniques produce a single partition of the objects into disjoint clusters, by optimizing a criterion function. For these methods, the number of clusters, k, needs to be specified. The solution to the optimization problem can be solved by enumerating all possible ways of dividing the objects into k clusters. The best split is chosen by evaluating its ‘goodness’ by an objective function. Except for small data sets, this exhaustive search is computationally not feasible. Many algorithms have been described in the literature to find solutions of the objective function. The most popular one is probably the k-means algorithm. This algorithm attempts to minimize the sum of the squared distances between the objects and their cluster centers, by iteratively reallocating objects to the clusters until convergence. This algorithm is a simple local search technique. More advanced stochastic methods like evolutionary algorithms, genetic algorithms, and simulated annealing explore the solution space more efficiently.

Model-based clustering is also based on optimizing an objective function. Model-based clustering fits the data to a mathematical model. Often it is assumed that the data are generated by a mixture of underlying distributions, each described by a set of parameters. The clustering algorithm then attempts to find the best estimates of the parameters by maximizing the log-likelihood function.

Density-based clustering identifies regions in which the density of objects/points is exceeding a threshold value (e.g., the minimum number of objects in the neighbourhood). Regions with a low density of objects indicate clusters of noise or clusters of outliers.

Hard or crisp clustering algorithms allocate each observation to exactly one cluster. *Fuzzy or soft* methods assign an observation to multiple clusters, with a degree of membership. Fuzzy clustering thus allows uncertainty in the clustering task and is useful when clusters are not well separated.

Most clustering methods are *polythetic* in nature, meaning that the features are used simultaneously in the process. *Monothetic* methods sequentially introduce features in the cluster process.

Clustering algorithms define clusters that are unknown a priori. Moreover, if no natural grouping is present an artificial structure is imposed. **Cluster validity** methods inspect aspects such as the optimal number of clusters, the fit of the final partition to the data, consistency and robustness of the partition when re-sampling the data, etc. The quality of a cluster result is measured by external (with respect to an a-priori structure), internal validity indices, and by the homogeneity of clusters and the separation between them.

Clustering is a descriptive technique. The solution is not unique and it strongly depends on the choices made. When deciding upon the clustering algorithm and similarity measure, the dimensionality of the data in terms of number of objects and features, and the type and scale of the features are determining factors. Hierarchical methods make no assumptions about the data distribution, are applicable to any feature type, and do not require specification of the number of clusters. However, in the presence of a lot of noise, overlapping clusters or clusters of different shape and size, these methods perform poorly. Partitional methods are attractive for large data sets, but need the specification of the number of clusters. These methods work well for isolated, compact, spherical clusters. Model-based clustering is an appealing alternative to these heuristic methods. Since, the underlying framework is probabilistic the choice of the optimal number of clusters comes down to model selection. Model-based clustering also allows to take into account complex design features, and to model outliers and missing values explicitly. The work presented in this dissertation is situated in the domain of model-based clustering, therefore an introduction to and application of model-based clustering is given in Chapter 4.

Chapter 3

Motivating Case Studies

In this chapter, we introduce the longitudinal studies that gave rise to the research covered in this dissertation. The main data sets have a common application domain, i.e., human health studies. But each study exhibits specific challenges.

The Persistent Disturbing Behavior (PDB) data set, presented in Section 3.1, will be used in Chapter 4, to introduce and demonstrate model-based clustering applied to univariate longitudinal data.

The Electro-Encephalogram (EEG) data in rats, described in Section 3.2, requires the development of a clustering algorithm for multivariate longitudinal data. This is the topic of Chapter 5. This data is also used in Chapter 8 to investigate the sensitivity of a finite-mixture model for longitudinal data to aberrant data points.

The Abdominal Aortic Aneurysm (AAA) data (Section 3.3) and Heart Failure (HF) data (Section 3.4) are both characterized by the presence of missing data. For the AAA data, a sensitivity analysis will be performed by fitting MNAR-models in the context of a cluster analysis (Chapter 7). The technique of multiple imputation will be used for the HF data. Chapter 6 delineates how a final partition of the data can be obtained after multiple imputation. Furthermore, the HF data will also be used to illustrate a clustering method for functional data.

3.1 Persistent Disturbing Behaviour Data (PDB)

3.1.1 Background

Mental health care institutions in Belgium are confronted with a group of chronic therapy resistant patients, which is problematic in that neither scientific definitions, theory, nor a legal framework is in place. These patients cannot be treated satisfactorily with current therapies and medication. Their behaviour is disturbing in the sense that living together in their natural environment, or even in a hospital ward, is extremely difficult. Given that their disease systems are unstable, and that their behavior is persistent over time, intensive 24-hour supervision is required. This condition is referred to as *persistent disturbing behavior* (PDB).

The Belgian mental health care system is clearly not accommodating to this group. The patients are predominantly found in psychiatric hospitals and psychiatric nursing homes. Psychiatric hospitals are defined as non-residential institutions for intensive specialist care. As the PDB group needs a prolonged stay in such a setting, a psychiatric hospital is not the optimal environment. In addition, a 1996 law states that a psychiatric nursing home is intended for patients with stabilized chronic psychiatric conditions. While the law does not specify the meaning of stabilized condition, it is generally understood that PDB patients are not stable. We therefore have to conclude that mental health care does not explicitly accommodate the PDB group.

With respect to the PDB group four important questions can be raised. First, how can the PDB group be distinguished from related but different groups, such as patients with acute or short-term disturbing behaviour. Second, because a clear definition is not available, the size of the PDB group is unclear. Third, it is conceivable that the PDB group consists of a number of subgroups that can be usefully distinguished. Finally, it is not clear in which residential setting such patients should be accommodated.

To rectify this situation, legislative work is needed. Before this can be done, one first needs to properly define the PDB group and perform a quantitative analysis, formulating answers to the aforementioned questions.

In the following sections, we will in turn introduce (1) the Minimal Psychiatric Data (MPD) registry system used by the Belgian mental health care institutions, (2) a cross-sectional pilot study set up in 1998 to estimate the size of the PDB group and to identify MPD items discriminating between PDB and non-PDB patients (Bruckers et al. (2000), and (3) research by Serroyen et al. (2010) using the data of the pilot study and registry data to develop a longitudinal PDB marker.

3.1.2 The Psychiatric Registry Data

For every patient admitted to a residential psychiatric care setting in Belgium specific data are registered. This registration system was made mandatory in 1996 for psychiatric hospitals and in 1998 for the psychiatric nursing homes by the federal Ministry of Public Health and is called Minimal Psychiatric Data (MPD). The entire set of data is extensive, organized in a number of modules (medical admission, medical treatment, intermittent discharge, medical discharge).

A description of the goals of the registration system and a detailed overview of all items included in the registry can be found at 'http://www.mpg.uhasselt.be/nl/default.htm'. The MPD registration data has to be send to the Ministry of Public Health twice a year. Data for the period January-June has to be uploaded in September, data for the period July-December in February of the next year.

3.1.3 Pilot Study, 1998

In 1998, a cross-sectional pilot study was set up in the psychiatric hospitals and the psychiatric nursing homes in the Belgian province of Limburg to (1) estimate the size of the PDB group and (2) explore factors that discriminate between PDB and non-PDB patients. More information about the study and the results can be found in Bruckers et al. (2000).

Patients were screened by an interdisciplinary team and classified by expert opinion as PDB when the team judged that living together with the patient is hard and that s/he needed continuous supervision. The persistence dimension was approached by restricting attention to patients residing in chronic-patient wards within psychiatric hospitals or in psychiatric nursing homes. Patients residing in one of these wards had in general already had intensive therapy in an acute ward and, in case of a psychiatric nursing home, also a long stay in a chronic ward.

To keep the burden on the field workers as low as possible, it was decided to include a sample of wards and to rely, as far as possible, on existing information, rather than initiating further data collection. In November 1998, a number of wards were screened for PDB behaviour. Information on 611 patients was obtained. For 189 patients the interdisciplinary team judged that their behavior was persistent disturbing. This information was supplemented with relevant MPD items (registered in the second semester of 1998). Based on these MPD items, a function discriminating between

PDB and non-PDB patients was developed. For each patient, this analysis results in a score, quantifying the probability to be PDB. When this probability exceeds a threshold value the patient was classified as PDB.

The functional form of the discriminant function, as obtained from logistic regression, for the patients admitted in a psychiatric hospital is:

$$\begin{aligned} \text{logit}(\text{PDB}_{ij}) &= -4.81 + 1.73 \cdot \text{Aggr.A}_{ij} + 0.62 \cdot \text{Aggr.P}_{ij} + 0.33 \cdot \text{Suicid}_{ij} \\ &\quad +0.47 \cdot \text{Appear}_{ij} + 0.40 \cdot \text{Respect}_{ij} - 0.03 \cdot \text{Age}_i + 1.81 \cdot \text{Sex}_i \\ &\quad -1.50 \cdot \text{DDAC}_i + 0.56 \cdot \text{Schizo}_i - 2.32 \cdot \text{Resid}_i. \end{aligned} \tag{3.1}$$

The index i refers to the patient, whereas j is the measurement occasion (second semester 1998) within a patient. Strictly speaking the index j is not needed in this expression. The construction of the discriminant function is based solely on MPD data collected for the second semester of 1998. The purpose of the index j will become clear in Section 3.1.4 where a longitudinal version of the PDB score is introduced. The predictive covariates in (3.1) have the following meaning: 'Aggr.A' stands for aggression towards oneself (auto-aggression), 'Aggr.P' for aggression against other people, 'Suicid' for suicide danger, 'Appear' for appearance, 'Respect' for respect for others, 'Age' stands for age (in years) in 1998, 'Sex' is a binary indicator for a patient's sex with the reference category being the female patients, 'DDAC' stands for the diagnostic class Delirium, Dementia, Amnestic and Cognitive disorders, 'Schizo' for the diagnostic class schizophrenia, and 'Resid' for the residual diagnostic class (so-called *V-codes*, a technical term for conditions that are the focus of clinical attention without being considered disorders). The scoring of these items is explained further on in this section. That sex and age, for example, are included in the score might be met with surprise. Such concern would be warranted when a purely behavioural approach is envisaged. However, the goal here is to put forward rules that classify a patient as belonging to the PDB group as accurately as possible. To complicate matters, in some of the analyses, such as the one reported here, the aim is to do this at a single point in time. Of course, then only essentially 'disturbance'-oriented information is available, rather than reliable longitudinal 'persistence' information. We believe that this offers an extra motivation to employ background covariates.

For the psychiatric nursing home patients, the functional form of the discriminant function is:

$$\begin{aligned} \text{logit}(\text{PDB}_{ij}) &= -6.39 + 1.26 \cdot \text{Aggr.A}_{ij} + 1.15 \cdot \text{Aggr.O}_{ij} + 0.65 \cdot \text{Asoc}_{ij} \\ &\quad +1.21 \cdot \text{Separ}_{ij} + 0.70 \cdot \text{Social}_{ij} + 0.59 \cdot \text{Respect}_{ij} \\ &\quad -0.85 \cdot \text{Retar}_i, \end{aligned} \tag{3.2}$$

with the same abbreviations as in (3.1) and in addition ‘Aggr.O’ standing for aggression against objects, ‘Asoc’ for anti-social attitude, ‘Separ’ for need for separation or isolation, ‘Social’ for socially unacceptable behaviour and ‘Retar’ stands for the diagnostic class mental retardation.

The *mental signs and symptoms*, i.e., ‘Aggr.A,’ ‘Aggr.P,’ ‘Aggr.O,’ and ‘Asoc,’ are direct indications for whether or not a patient’s disturbing behaviour contributed to his/her admission or extension of the medical stay. These problems are rated on a three-point scale, ranging from 0 to 2, indicating to which extent the problem is contributing to the admission of the patient or to the extension of the medical stay. A score of 0 indicates that the problem was not contribution at all, 1 indicates that the problem was present but not the reason for the admission or extension, a score of 2 indicates that the problem is the direct cause for the admission or extension of the stay.

Indirect indications for the degree of disturbance are captured by the items referring to preventive suicidal supervision, ‘Suicid,’ and the need for separation, ‘Separ.’ For each item, a binary variable was created, indicating whether or not the activity was performed at least once during the treatment period (semester).

‘Respect’, ‘Appear’, and ‘Social’ belong to a set of *patient functioning* items that describe the interaction between the patient and his/her immediate surroundings, as well as the relationship between the patients and their co-residents. Each item is scored on four-point scale, ranging from 1—4. When no limitations are observed in a patient’s functioning the item will be equal to 0, whereas a score equal to 4 implies serious limitations.

The diagnostic classes ‘DDAC’, ‘Schizo’, ‘Retar’, and ‘Resid’ are yes/no indicators relating to specific diagnostic classes. These diagnostic classes are constructed according to a consensus document, designed by the Limburg collaborative network in psychiatry, SPIL, summarizing the diagnostic DSM-IV codes in 11 classes (SPIL-RPL 1997, Munson 2001).

The logistic-regression-based classification, presented in (3.1) and (3.2), turned out to have good discriminative power. The screening status and the classification status agree for about 80% of the screened patients. The ROC c statistic (Agresti 2002), which quantifies the discriminative ability, equals 0.85 for the psychiatric hospitals and 0.88 for the psychiatric nursing homes. Bruckers et al. (2000) observed that the sensitivity and specificity for the psychiatric hospitals (psychiatric nursing homes) were 77.2% (71.9%) and 78.7% (85.5%), respectively. These values were attained for a cutoff value of 0.40 (0.28) for the hospitals (nursing homes) on the logit-score.

An important conclusion from the pilot study was that, following such a discriminant rule, 35.5% of the patient population in a psychiatric hospital might belong to the PDB group, with a similar figure (32.1%) for the psychiatric nursing homes. The corresponding 95% confidence intervals for the size of the PDB group are [198; 242] and [100; 135].

Of course, these findings have to be taken with some caution. First, as stated before, the data used for the analysis constituted a pilot sample of PDB patients and controls, 611 in total, intended to build the classification rule from. Hence, its use lies in the ability to compare both groups, rather than in being representative for a larger population. Second, and more important, the discriminant function focuses on the disturbance aspect, neglecting the persistence. Patients residing in one of these wards in general already had intensive therapy in an acute ward and, in case of a psychiatric nursing home, also a long stay in a chronic ward. But of course, whether or not the group considered to be PDB in 1998 indeed was chronic in their disturbing behaviour is questionable. The fact that these patients are staying in long-stay wards only indicates that we are dealing with chronic disease statuses, not necessarily that the disturbing behaviour is persistent. It is possible that the patient was going through an acute phase of disturbing behaviour, something hard to disentangle based on information localized in time. Serroyen et al. (2010) addressed this point (Section 3.1.4).

It is clear that the study should be seen as a pilot. Nevertheless, it is important to know whether the group is sufficiently large to warrant specific components of care. Even allowing for some overestimation, it is clear that the group is large enough to render its consideration as a single, monolithic group impractical for organization. This is particularly challenging, given the disturbing character of the disorder, necessitating special small-scale care units.

3.1.4 Longitudinal PDB Score

By making use of the repeated measurements collected within the psychiatric registry data, Serroyen et al. (2010) investigated the persistence aspect of the PDB group. The authors also performed a cluster analysis based on the 1998 data, to initiate identification of subgroups within the PDB group.

The discriminant function described in Section 3.1.3 was developed based on data registered in the second part of 1998. But in fact, the items which make up the discriminant score have been recorded twice annually since 1996 for the psychiatric hospitals and since 1998 for the psychiatric nursing homes. The score can thus be calculated at the other registration occasions as well, thus producing a longitudinal profile per patient. The minimal psychiatric registry was put in place only at the second semester of 1996 in psychiatric hospitals and at the first semester of 1998 in psychiatric nursing homes. Given that 1996 was the year that the registration system started, it is prudent not to put too much trust in the data for this semester. This semester will be ignored in subsequent analysis. Furthermore, for the first semester of 1997, no data are available, owing to the start of the registration system. End 2000, the legal registration framework changed. Therefore attention was restricted to the period 1997.2–2000.1 for the psychiatric hospitals and 1998.1–2000.1 for the psychiatric nursing homes. Figure 3.1 shows individual profiles of the PDB scores for 20 randomly selected PDB patients, 10 residing in a psychiatric hospital and 10 in a psychiatric nursing home.

Serroyen et al. (2010) employed linear mixed models to study the evolution of the mean discriminant function, for the PDB and non-PDB groups. They conclude that the evolutions are different for PDB and non-PDB patients and that a non-linear structure emerges for the non-PDB group. Differences in the variance and correlation structure of the two groups give more insight in the persistent dimension of the PDB patients. Relatively more heterogeneity is seen among PDB patients, opening perspectives for further subdivision. The correlation structure, for the patients residing in a psychiatric hospital, is subtly different between both groups. The PDB group is roughly of a first-order autoregressive type, showing relatively large correlations between adjacent measurements (around 0.75), which decreases with increasing time lag, dropping to about 0.35. Thus, the PDB group exhibits a chronic behavior from the beginning, with fluctuations happening in the long run rather than immediately. The non-PDB group correlation structure is closer to compound symmetry, amended by the fact that the correlations increase towards later times. This may suggest there is an unstable, acute phase at the beginning of the study. For the psychiatric nursing homes, the picture emerging from the estimated correlation structures is different. Both are relatively close to compound-symmetry, with a common correlation around 0.65. This is plausible from a field work point of view, because these patients are almost by definition of a chronic type.

A K-means cluster analysis, using Gower's distance measure (Gower 1971), suggests the presence of two clusters in the PDB group (data of 1998). Cluster #1 (n= 91) appears to consist of PDB patients with higher scores on the ordinal variables mobility, recognition of persons, notion of time, initiative, socially unacceptable behavior, respect for others, and conflicts, as compared to cluster #2 (n= 98). This indicates that the patients in cluster #1 show more pathological behavior.

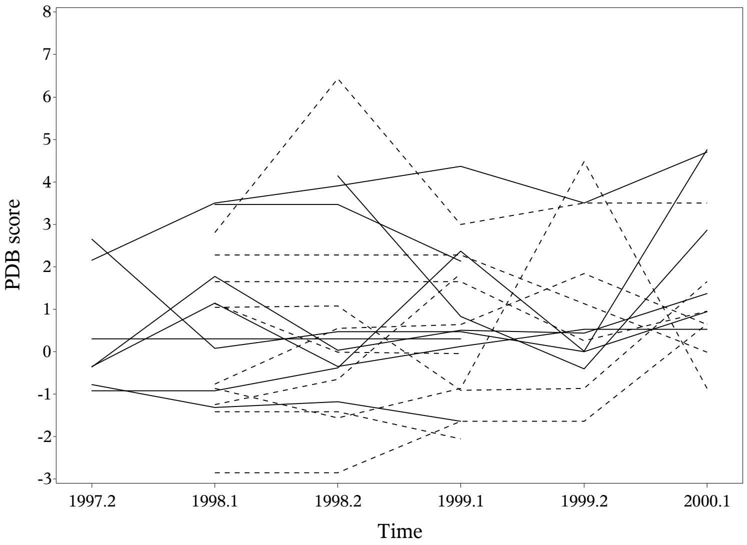


Figure 3.1: *Random sample of individual PDB-score profiles for 20 PDB patients, 10 residing in a psychiatric hospital (full line) and 10 residing in a psychiatric nursing home (dashed line).*

A major issue with conventional clustering is that it starts from cross-sectional data, thus focusing on similarity at one point in time. However, patients exhibiting the same characteristics, the same behaviour at one point in time can still evolve, and diverge, in a multitude of ways. In Chapter 4, the cluster analysis presented by Serroyen et al. (2010) is refined by making use of the longitudinal nature of the data. This is done using conventional linear mixed models and so-called growth-mixture and latent-class growth models.

3.2 Electro-Encephalogram Data in Rats (EEG)

The aim of EEG studies is to characterize the effects of psychotropic drugs on cortical brain activity, on the basis of spectral electro-encephalograms. An EEG study in rats, conducted at Janssen Pharmaceutica (Belgium), is used. Although the brain waves of rats and humans are observed in comparable frequency bands, not all functionalities are the same. There are, however, more similarities than differences, making experiments measuring the electrical brain activity in rats very interesting to study the effect of psychoactive agents on the activity of human brains.

Depending on the frequency measurements range, the brain activity is referred to as delta activity (below 4 Hz per second), theta activity (4–7.5 Hz per second), alpha activity (8-12.5 Hz per second), beta activity (13–30 Hz per second), and gamma activity (above 30 Hz per second). With the exception of the delta activity, each frequency range is further refined in low and high activity (e.g., $\alpha_1, \alpha_2, \ldots$). Delta activity is normally seen in babies or in adults in slow-wave sleep. Theta activity is seen in children or during drowsiness or arousal in adults. Alpha waves are seen when a person is alert in a relaxed way, closing the eyes. Beta activity (low amplitude) is often associated with active, busy, or anxious thinking and active concentration. Rhythmic beta waves are linked with pathological or drug-related causes. Gamma waves are related with strong mental activity like solving problems, fear, and awareness.

The EEG study includes 10 psychoactive agents at 4 different doses, including a placebo dose. To each compound, 32 rats were randomly assigned, 8 per dose group. The compounds included in the study are: (1) Psychostimulants: Amphetamine, Nicotine; (2) Antidepressant: Buprorion; (3) Cholinesterase inhibitors: Donepezil, Galantamine, Tacrine; (4) Anti-epileptics: Lamotrigine, Valproate; and (5) NMDA receptor antagonists: Memantine, PCP. Cholinesterase inhibitors are used to treat moderate to severe dementia of the Alzheimer's type. The anti-epileptics listed are used in the treatment of mania. The NMDA receptor antagonists are used for different purposes. Memantine is used to treat moderate and severe dementia of the Alzheimer's type and in that view could be listed with the cholinesterase inhibitors. PCP in low to moderate doses acts as a stimulant, whilst at higher doses it has a sedative effect.

Forty-five minutes after administration of the psychoactive agent, the brain signals of the rats in active wake state are monitored every 15 minutes during 1.5 hours, at six different positions in the brain (left and right frontal, left and right parietal, left and right occipital). For each rat, 9 activity profiles are obtained, at the 6 different positions in the brains.

Chapter 5 introduces a clustering algorithm for multivariate longitudinal data, as generated in the EEG study. To illustrate the clustering algorithm, we focus on the frequencies obtained at the left prefrontal cortex. So we are facing 9-variate longitudinal profiles. To be able to compare the results with analyses done in the past, we only include the placebo and the highest dose level. This reduces the data set to 160 rats of which 139 have follow-up data (73 in the placebo and 66 rats in the highest dose level).

To visualize the data, the individual longitudinal profiles for the 9 frequency mea-

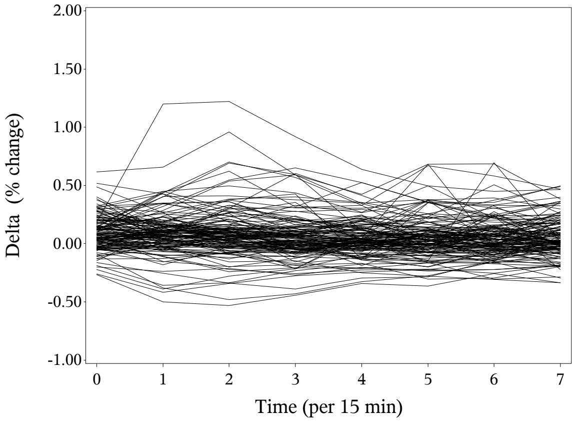


Figure 3.2: *Individual profiles for the delta frequencies (time = 0 is 45 min after baseline) – EEG study.*

surements are given in Figures 3.2 — 3.4. The response of interest is the percentage change with respect to the measurement at baseline Y_{ib} (administration of the drug): $Y_{ij}^{'} = 100(Y_{ij} - Y_{ib})/Y_{ib}$. At baseline all percentage changes are by definition equal to zero. The graphical display therefore excludes the baseline data. In graphical displays and in the statistical models, time zero refers to the first measurement obtained after administering the drug (i.e., after 45 minutes). Heterogeneity is seen in all waves, some rats have a decrease in the frequency while for others an increase is obtained as an effect of the drug. For some waves extreme profiles are seen, such as for the α_1 wave. This heterogeneity is of course induced by administrating 10 different drugs at different dose levels. When applying the proposed clustering algorithm, this information will not be taken into account. The goal of the analyses is to see if, within the set of 139 rats, it is possible to identify subpopulations that are homogeneous in the growth parameters for the 9 waves. The information about the compounds and doses will later be used to assess whether the identified groups are meaningful.

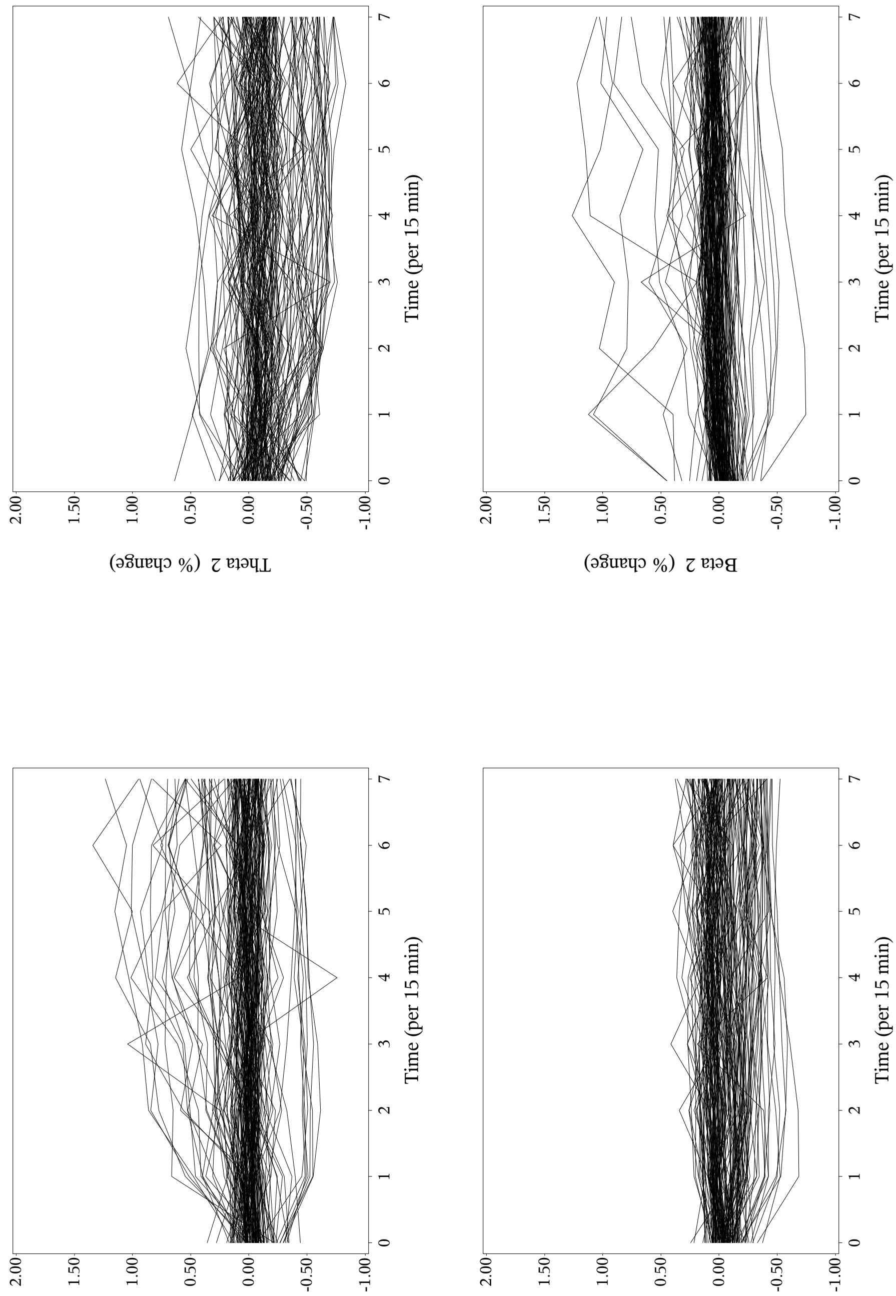


Figure 3.3: *Individual profiles for the theta and beta frequencies (time =0 is 45 min after baseline) – EEG study.*

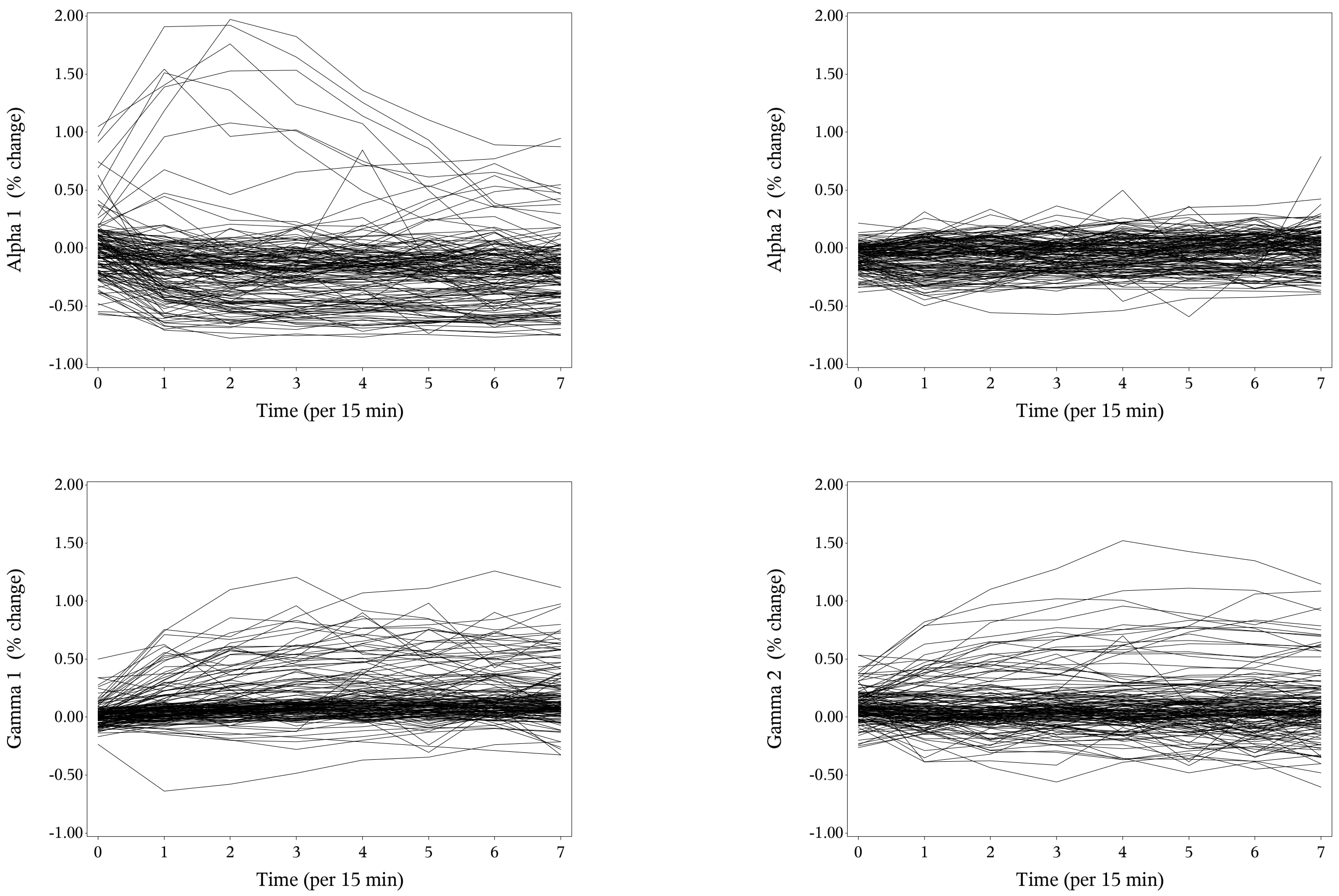


Figure 3.4: *Individual profiles for the alpha and gamma frequencies (time =0 is 45 min after baseline) – EEG study.*

3.3 Abdominal Aortic Aneurysm Data (AAA)

Abdominal aortic aneurysm (also known as AAA) is a localized dilatation of the abdominal aorta, caused by degeneration of the aortic wall. Abdominal aortic aneurysms occur most commonly in individuals between 65 and 75 years old and are more common among men and smokers.

As abdominal aortic aneurysms expand, they may become painful and lead to pulsating sensations in the abdomen or pain in the chest, lower back, or scrotum. The risk of rupture is high in a symptomatic aneurysm. Rupture of the artery can be life-threatening as large amounts of blood spill into the abdominal cavity. The mortality of AAA rupture is up to 90%. 65 to 75% of the patients die before they arrive at the hospital and up to 90% die before they reach the operating room.

Therefore, symptomatic and large aneurysms are considered for repair by surgical methods. An intervention is often recommended if the aneurysm grows more than 1 cm per year or when it is bigger than 5.5 cm.

In 2006, the academic hospital of Maastricht (the Netherlands) started a follow-up study in patients with an abdominal aorta aneurysm. Between January 2006 en January 2009, all patients with AAA admitted to the department of Vascular Surgery of the academic hospital were invited to participate in the study. Two hundred and eighty-seven AAA patients provided written informed consent. Patients that had a large aneurysm (≥55 mm) or (symptoms of imminent) AAA rupture and patients with either an inflammatory or a mycotic aneurysm were excluded from the follow-up study. Patients with an aneurysm diameter between 30 and 55 mm (n=110) were invited to participate in an imaging surveillance program. A total of 100 patients formally entered the follow-up program. These patients were seen every 6 months. During these follow-up visits the diameter of the artery, a number of patient characteristics and blood measurements were collected. Figure 3.5 shows the diameter curves for the patients in the study. The objective of the study was two-fold. Is it possible to predict the expected diameter of the artery at the next follow-up visit? And secondly, is it possible to detect subgroups (clusters) of patients whose arteries grow in a similar way? To study the evolution of the diameter over time and to find clusters of patients with similar growth, the study researchers applied conventional growth models and growth-mixture models (Hellenthal et al., 2014). The growth-mixture model is presented in Section 7.4.2. The analysis revealed two subgroups of patients. But, for a number of patients, the diameter of the artery is not available at all follow-up visits. Only 7 patients had complete data for all 7 visits, 10 dropped out after the first visit (at

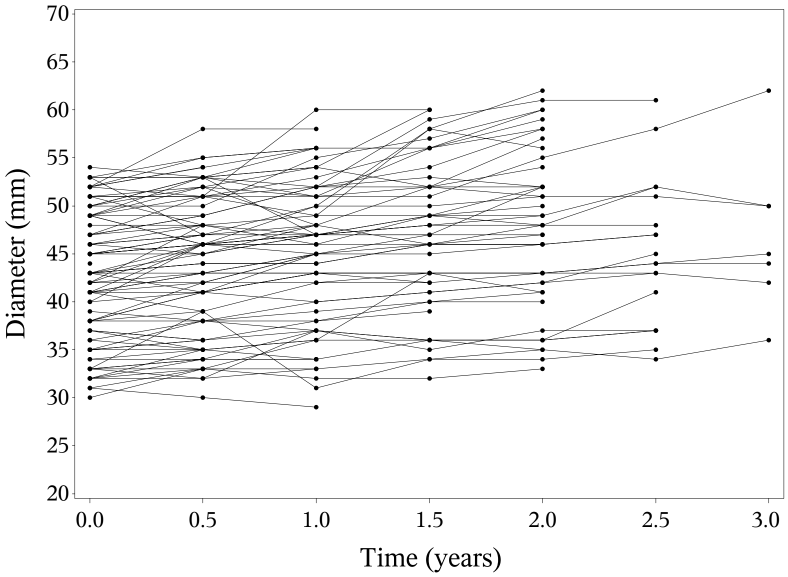


Figure 3.5: *Evolution of the patients' diameter – AAA study.*

6 months of follow-up), 15 after the second visit, 21 after the third visit, 7 after the fourth visit, 27 after visit five and finally 12 patients dropped out after visit 6 (see Table 7.1). The diameter of the artery at visit j can be missing because the patient was not yet $j \times 6$ months in the study, or because the patient did not show up at the visit, or because the patient underwent repair of the artery by surgery, etc. The reason for missing a follow-up visit can depend on the diameter of the artery. The dropout rate is very comparable among the patients that had surgery of the artery and those who did not (92% versus 93%).

In Chapter 7, we study the vulnerability of the cluster analysis results, i.e., the estimated trajectories, and the posterior membership probabilities, by applying different missing-data growth-mixture models for non-ignorable dropout to the evolution of the diameter of the artery. The MNAR-models will be extended by including the information whether or not the patient needed repair of the artery by means of a surgical method.

3.4 Telemonitoring in the Management of Heart Failure Study (HF)

Chronic heart failure (CHF) is characterized by recurrent hospitalizations due to fluid overload and/or worsening of renal function. To reduce morbidity, mortality and healthcare cost, regular adjustment of the treatment of CHF patients is needed. In the TElemonitoring in the MAnagement of Heart Failure study (TEMA-HF1), 160 CHF patients, hospitalized in 7 Belgian hospitals, were included between April 2008 and June 2010. Patients were randomly assigned to receive usual care (UC) after discharge, or to be intensively followed for up to 6 months by telemonitoring (TM). The primary objective of the TEMA-HF1 study, was to investigate whether intensive follow-up of patients through telemonitoring-facilitated collaboration between general practitioners (GPs) and a heart failure clinic could reduce mortality and rehospitalization rate. Details regarding the design and results of TEMA-HF1 are reported elsewhere (Dendale et al., 2012).

We will focus on the 80 patients in the TM group. For these patients, a telemonitoring device daily transferred data on body weight, blood pressure (systolic and diastolic), and heart rate to a web-site, for a period of 6 months. This web-site triggered e-mail alerts to care providers if data were out of limits, or if data had not been received on two consecutive days. At baseline, additional patient characteristics were collected: sex, age, heart rhythm, cardiac muscle fibre stretch as measured through NTprobBNP, a fitness indicator (according NYHA class indication) and the left ventricle ejection fraction (LVEF), which is a measure of heart performance. Profiles of the biomarkers during the first month of the study, for 10 patients in the TM group are displayed in Figure 3.6.

Four TM patients left the study prematurely for motivational reasons, 4 died during the course of the 6 month study, and 16 were hospitalized at least once for heart failure related reasons.

Although alerts were sent out when the longitudinal measurement were missing for two consecutive days, quite some missingness is present in the data. Twenty-eight percent of the patients did receive an alert concerning missing information for the heart rate, 64% concerning the blood pressure measurements and 84% concerning body weight (Dendale et al., 2012). Information about the extent of missingness in the heart failure data is presented in Tables 3.1 and 3.2. Baseline characteristics are fairly complete. About one out of four patients does not have information for the six minute walking test (WALK). On average, 76% of the patients' daily measurements for the biomakers were recorded. Meaning that on average for 137 days out of 180, heart rate, diastolic and systolic blood pressure were communicated to the heart failure clinic by means of the telemonitoring device. The heart failure data has particular features. Heart rate and blood pressure are recorded by the same device and thus simultaneously missing or present. The periods lacking telemonitoring data, are, in general, not too long (average duration is 6 days, median duration is 1 day). However, some patients are featured by longer periods of lacking data. About 5% of the periods, with missing info on consecutive days, lasts longer than 2 weeks. The mean follow-up time is 163 days.

Clustering of the patients in this dataset is challenging for two reasons. First of all, because of the high dimensionality of the data in terms of the number of measurements available per patient. Secondly, because of the presence of missing data. Chapter 6 discusses clustering of high-dimensional data and sketches how multiple imputation can be used in combination with a cluster analysis.

Table 3.1: *Number of patients with missing information at baseline – HF study.*

Characteristic	# of patients	Characteristic	# of patients
Age	0	LVEF	2
Gender	0	NTPROBNP	4
Diastolic Blood Pressure	0	REG-AF	0
Systolic Blood Pressure	0	NYHA	0
Heart Rate	0	WALK	26
Weight	0		

Table 3.2: *Percentage of days with missing information – HF study.*

Biomarker	mean	median
Diastolic Blood Pressure	24	14
Systolic Blood Pressure	24	14
Heart Rate	24	14
Weight	20	7

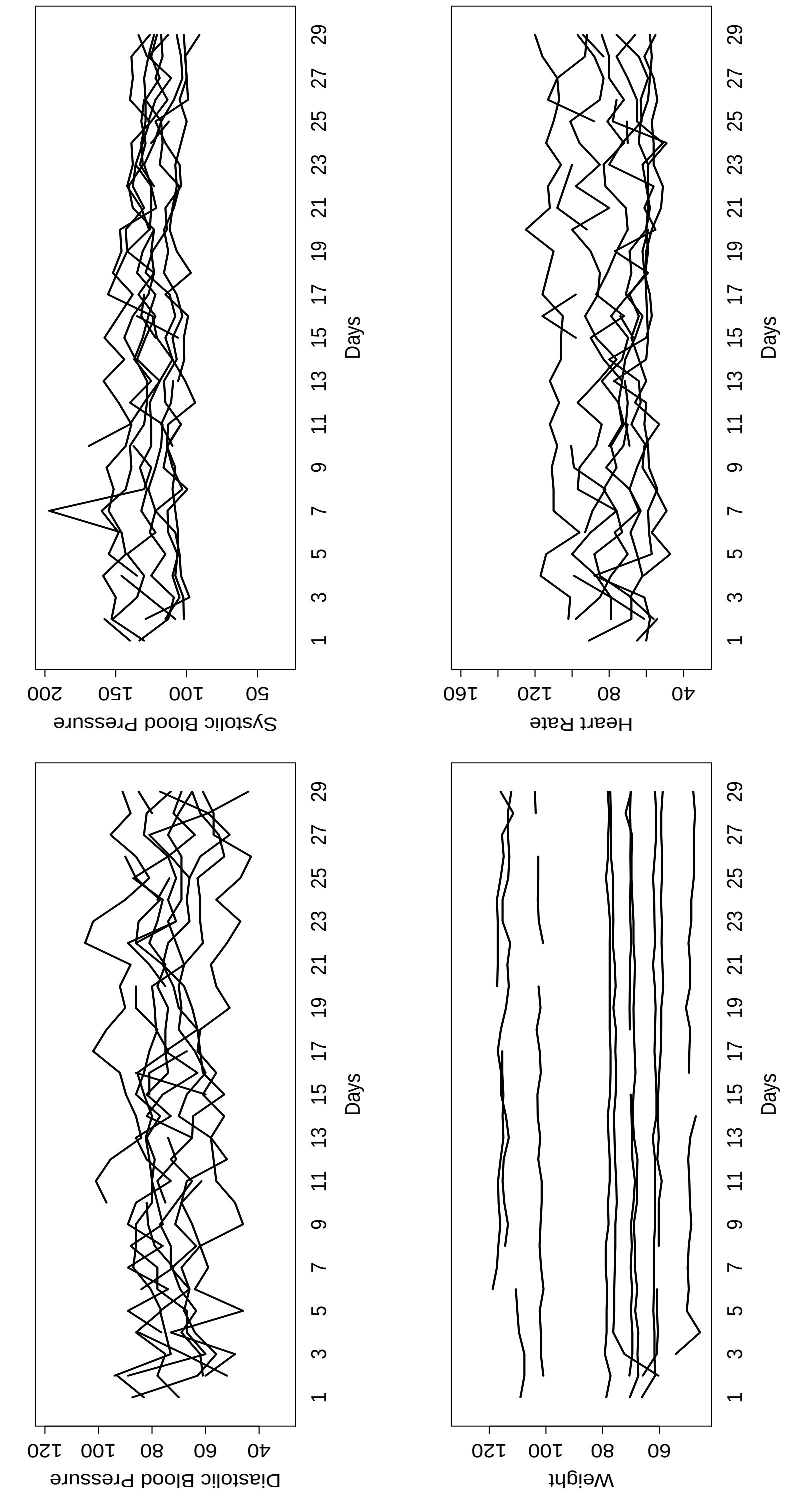


Figure 3.6: *Measurements for the four biomarkers, for 10 patients of the TM group during the first month of the study – HF study.*

Chapter 4

Model-based Clustering for Univariate Longitudinal Data

Repeated measures and multivariate outcomes are very common in social, behavioral, and educational sciences, as well as in clinical trials. A lot of methodological work has been done to extend cluster analysis to repeated measures. When analyzing repeated measurements data, individual differences in evolution are generally captured by random effects, often via linear mixed models for continuous longitudinal data (Laird and Ware, 1982; Verbeke and Molenberghs, 2000), while for other data types generalized linear mixed models can be used (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Molenberghs and Verbeke, 2005). Individual differences can also be described by latent trajectory classes (Land and Nagin, 1996; Nagin and Land, 1993, Nagin, 1999; Nagin and Tremblay, 2001) or by extended growth-mixture models (Muthén and Shedden, 1999; Muthén and Muthén, 1998—2012). These statistical techniques are briefly reviewed in this chapter. The random-effects methodology is discussed in Section 4.1. Models allowing the data to consist of unlabeled subpopulations are reviewed in Section 4.2 and Section 4.3. The methodology is illustrated on the PDB data in Section 4.4.

4.1 Mixed Models for Longitudinal Data

Measurements gathered on the same subject tend to be correlated and this correlation must be taken into account during the statistical analysis to obtain valid inference.

Such correlation is present in longitudinal studies where information for an outcome of interest is obtained repeatedly in time. Observations on a subject tend to be more alike than observations from different subjects.

Methods for continuous longitudinal data are well developed and implemented in standard statistical software packages, due to the elegant properties of the multivariate normal distribution. The most popular model for normally distributed longitudinal data is the linear mixed model (Laird and Ware, 1982; Verbeke and Molenberghs, 2000). The influence of explanatory variables on the mean structure of the data is modelled via fixed-effects parameters. To capture the variance-covariance structure of the data, three types of parameters are used: (1) random (subject-specific)-effect parameters arising from characteristics of individual subjects, (2) serial correlation allowing measurements taken close in time to be more strongly correlated than measurements taken further apart in time, and (3) measurement error.

The general form of the linear mixed model is:

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{4.1}$$

where $\boldsymbol{Y}_i$ denotes the n_i-dimensional response vector for subject $i = 1, \dots, N$. N is the number of subjects, X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ matrices of known covariates, $\boldsymbol{\beta}$ is the p-dimensional vector of population-average regression coefficients called fixed effects, $\boldsymbol{b}_i$ is the q-dimensional vector of random effects for subject i describing how the evolution of the i^{th} subject deviates from the population-average evolution, and $\boldsymbol{\varepsilon}_i$ is a n_i-dimensional vector of measurement error components. It is assumed that $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_i$ are independent with distributions $N(\mathbf{0}, D)$ and $N(\mathbf{0}, \Sigma_i)$, respectively. Σ_i depends on i only through the number of measurements available for subject i. Often, Σ_i is chosen to be equal to $\sigma^2 I_{n_i}$, where I_{n_i} is the identity matrix of dimension n_i. More general forms for the residual covariance structure Σ_i were proposed by Diggle et al. (2002). They assume that $\boldsymbol{\varepsilon}_i$ has constant variance and can be decomposed in a component $\boldsymbol{\varepsilon}_{1i}$, inducing of serial correlation and an independent component $\boldsymbol{\varepsilon}_{2i}$ of measurement errors. The first component results in correlation between serial measurements, and this serial correlation is generally specified as a decreasing function of the time lag between the measurements. Of course, the constant variance assumption can be relaxed for each of the processes.

Conditional on the random effects $\boldsymbol{b}_i$, the distribution of $\boldsymbol{Y}_i$ is

$$\boldsymbol{Y}_i|\boldsymbol{b}_i \sim N(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i, \Sigma_i). \tag{4.2}$$

Inference is based on maximizing the likelihood function of the marginal response $\boldsymbol{Y}_i$. The marginal distribution is obtained by integrating over the random effects

$f(\boldsymbol{y}_i) = \int f(\boldsymbol{y}_i|\boldsymbol{b}_i)f(\boldsymbol{b}_i)d\boldsymbol{b}_i$, with $f(\boldsymbol{y}_i|\boldsymbol{b}_i)$ the density function of $\boldsymbol{Y}_i$ conditional on $\boldsymbol{b}_i$, and $f(\boldsymbol{b}_i)$ the density function of $\boldsymbol{b}_i$. As a result the marginal distribution of $\boldsymbol{Y}_i$ is given by the density of the n_i-dimensional normal distribution $N(X_i\boldsymbol{\beta}, Z_iDZ_i^{'} + \Sigma_i)$.

When $\Sigma_i = \sigma^2 I_{n_i}$ the model specified by (4.2) is called the conditional independence model, since it implies that conditional on $\boldsymbol{b}_i$ the measurements of individual i are independent. The parameters found in $V_i = Z_iDZ_i^{'} + \Sigma_i$ are often grouped in a vector $\boldsymbol{\alpha}$ of variance-covariance parameters, and let $\boldsymbol{\theta}^{'} = (\boldsymbol{\beta}^{'}, \boldsymbol{\alpha}^{'})$ be the vector of all parameters in the marginal model for $\boldsymbol{Y}_i$. The following marginal likelihood function then needs to be maximized with respect to $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N}(2\pi)^{-\frac{n_i}{2}}|V_i(\boldsymbol{\alpha})|^{-\frac{1}{2}} \times \exp[-\frac{1}{2}(\boldsymbol{Y}_i - X_i\boldsymbol{\beta})^{'}V_i^{-1}(\boldsymbol{\alpha})(\boldsymbol{Y}_i - X_i\boldsymbol{\beta})].$$

The population-average evolution of the response can be described in terms of the $\boldsymbol{\beta}$ parameters, but when interest is in discovering groups of individuals evolving differently over time the estimates for the random effects $\boldsymbol{b}_i$ are very useful. The random effects reflect between-subject variability. The marginal distribution of $\boldsymbol{b}_i$ is often assumed to be the multivariate normal distribution $N(\boldsymbol{0}, D)$. Using Bayesian terminology, this distribution is called the prior distribution since it does not depend on the data $\boldsymbol{Y}_i$. The posterior distribution of $\boldsymbol{b}_i$, given the data $\boldsymbol{Y}_i$ is given by

$$f_i(\boldsymbol{b}_i|\boldsymbol{y}_i) = \frac{f_i(\boldsymbol{y}_i|\boldsymbol{b}_i)f(\boldsymbol{b}_i)}{\int f_i(\boldsymbol{y}_i|\boldsymbol{b}_i)f(\boldsymbol{b}_i)d\boldsymbol{b}_i}.$$

A point estimator for $\boldsymbol{b}_i$ is given by the posterior mode of this posterior density function, called the empirical Bayes (EB) estimates.

Mixed-effects models are an attractive tool for analyzing longitudinal data, because of their flexibility to handle data with missing observations, gathered from unbalanced designs, and their ability to explicitly acknowledge that variability in repeated measures is often not constant.

The model assumes that the random effects are sampled from a normal distribution. To address non-normality, as for example seen in data with outliers, mixed-effects models with a multivariate t-distribution (Pinheiro et al., 2001), or a skewed normal distribution (Arellano-Valle et al., 2005) for the random effects have been used to obtain robust estimates.

However, these models still assume that subjects come from a single population, which is often an unrealistic assumption.

4.2 Mixture Models for Longitudinal Data

The assumption of a single-component multivariate normal distribution for the random effects in (4.1) implies that the subjects belong to a homogeneous population that is described by a single mean trajectory and variance-covariance matrix. This assumption may however be unrealistic when subpopulations of subjects exist, each with its own trajectory. Different evolutions can for example be seen for treatment responders and non-responders. In psychiatric studies, different diagnostic classes could be characterized by different mean trajectories. Ignoring this heterogeneity can produce biased estimates of the random-effect parameters and their associated variance terms (Verbeke and Lesaffre, 1996).

This type of non-normality of the random effects can be dealt with by assuming a finite-mixture distribution. This section introduces the concepts of finite-mixture models and their application to repeated-measurements data. Finite-mixture models (McLachlan and Peel, 2000) are latent-variable models that express the distribution of a variable as a mixture of a finite number of component distributions. These models have been used in a wide range of applications in marketing, social and psychosocial sciences, where the data could be seen as arising from two or more populations. A mixture model allows investigation of the performance of estimators in non-normal situations and to develop robust estimators, but it also provides a framework for clustering. Finite-mixture modeling addresses the population heterogeneity in the observed data by means of categorical latent classes, that represent homogeneous subpopulations. Class-membership is latent (not observed) and thus needs to be inferred from the data. In its most general form the finite-mixture model for N p-dimensional observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ is written as: $f(\boldsymbol{y}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y})$. Here π_k is the k^{th} mixing proportion or the probability that an observation belongs to the k^{th} subpopulation (class, component) and $f_k(\boldsymbol{y})$ its corresponding density. K represents the total number of subpopulations and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, with $0 \leq \pi_k \leq 1$, for all $k = 1, \ldots, K$ and $\sum_{k=1}^{K} \pi_k = 1$. The mixture components are often members of the same parametric family, but members of different families are also possible.

Mixture models have been applied to longitudinal data in different settings. Latent-class growth analysis uses a categorical latent variable to represent unobserved heterogeneity in growth trajectories. The assumption is made that conditional on class and covariates the repeated measurements are independent (Land and Nagin, 1996; Nagin and Land, 1993, Nagin, 1999; Nagin and Tremblay, 2001).

Verbeke and Lesaffre (1996) allow for heterogeneity by extending the normality assumption of the random effects $\boldsymbol{b}_i$ in (4.1) to incorporate mixtures of normal components. This model is referred to as the heterogeneity model.

Muthén and Shedden (1999) proposed an extended mixture modeling framework. Their approach allows the joint estimation of (1) a conventional finite-mixture growth model where different curve shapes are captured by class varying random-coefficient means and (2) a logistic regression of a set of binary disease indicators on the classes. The model is thus a combination of latent class modeling (for the disease indicators) and conventional mixture modeling for the repeated measurements.

We first introduce the heterogeneity model proposed by Verbeke and Lesaffre (1996) and then describe the extended finite-mixture model of Muthén and Shedden (1999).

Verbeke and Lesaffre (1996) and Spiessens et al. (2002) allow for model heterogeneity in repeated-measurements data by extending the normality assumption of the random effects $\boldsymbol{b_i}$ to mixtures of normal components,

$$\boldsymbol{b_i} \sim \sum_{k=1}^{K} \pi_k N(\boldsymbol{\mu_k}, R),$$

where, as before, π_k is the proportion of subjects belonging to subpopulation k, described by the multivariate normal distribution $N(\boldsymbol{\mu_k}, R)$. For identifiability we require that, $\boldsymbol{\mu_i} \neq \boldsymbol{\mu_j}$, $\pi_1 \geq \pi_2 \geq \ldots \geq \pi_K$, and $E(\boldsymbol{b}_i) = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu_k} = 0$. Under this assumption, the density function of $\boldsymbol{Y}_i$ is a mixture of densities $f_k(\boldsymbol{y_i})$ with mixture probabilities $\pi_1, \ldots, \pi_K$:

$$f(\boldsymbol{y_i}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y_i}) = \sum_{k=1}^{K} \pi_k \int f(\boldsymbol{y_i}|\boldsymbol{b_i}) \phi_{\boldsymbol{\mu_k},R}(\boldsymbol{b_i}) d\boldsymbol{b_i}, \tag{4.3}$$

with $\phi_{\boldsymbol{\mu_k},R}(.)$ the density of a multivariate normal distribution with mean $\boldsymbol{\mu_k}$ and covariance matrix R.

Estimates for all parameters in the model are obtained by maximizing the log-likelihood for (4.3) by means of the Expectation-Maximisation algorithm (see Section 4.3).

When the goal of the statistical analysis is not only to obtain parameter estimates but also assignment of the subjects to the subpopulation they belong to, we term this model-based clustering. A subject's posterior probabilities, $\pi_k f_k(\boldsymbol{y}_i)/f(\boldsymbol{y}_i)$, are used to classify its longitudinal profile into one of the K components. Spiessens, Verbeke, and Komàrek (2002) have developed a SAS macro, based on the SAS procedure NLMIXED, that implements the EM algorithm for fitting nonlinear and generalised linear models with finite normal mixtures as the random-effect distribution. The macro also classifies the longitudinal profiles into the different components.

Muthén and Shedden (1999) proposed a extended latent-variable modeling framework. Their extended growth-mixture model incorporates the ideas of random-effects mixture models and latent-class models.

Their model is formulated as a structural equation model. Note that the linear mixed model presented in (4.1), with $X_i = Z_i$, can be expressed as a multilevel model,

$$\boldsymbol{Y}_i = \Lambda_i \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i, \tag{4.4}$$

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{\xi}_i, \tag{4.5}$$

where $\Lambda_i = X_i = Z_i$, $\boldsymbol{\alpha} = \boldsymbol{\beta}$, $\boldsymbol{\xi}_i = \boldsymbol{b}_i$. This model is known as a latent growth curve model. Muthén, Asparouhov et al. (2009) and Curran (2003) have demonstrated the isomorphism between these two models analytically and empirically.

To outline the generalized framework of Muthén and Shedden for finite mixtures of latent growth curve models we use the following notation. As before, consider an n_i-dimensional vector $\boldsymbol{Y}_i$ of continuous variables and an r-dimensional vector $\boldsymbol{u}_i$ of binary outcomes, which are related to each other in the model via latent variables. X_i is the vector of covariates, $\boldsymbol{\eta}_i$ is a vector of continuous latent variables (growth factors) and $\boldsymbol{c}$ is a vector of latent categorical variables. Also, let $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{iK})$ have a multinomial distribution, where $c_{ik} = 1$ when subject i belongs to the k^{th} latent class (subpopulation) and is zero otherwise. The random-effect mixture part of the model can be written as:

$$\boldsymbol{Y}_i = \Lambda \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i, \tag{4.6}$$

$$\boldsymbol{\eta}_i = A\boldsymbol{c}_i + \Gamma X_i + \boldsymbol{\xi}_i, \tag{4.7}$$

where Λ is a $n_i \times p$ matrix of constants depending on the data. $\boldsymbol{\varepsilon}_i$ is an error term, distributed as $N(\mathbf{0}, \Sigma_i)$, with Σ_i e.g., diagonal. A and Γ are parameter matrices, relating the classes and covariates to the outcome. $\boldsymbol{\xi}_i$ is a residual vector distributed as $N(\mathbf{0}, \Psi)$.

A logistic regression model is used to link the vector $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK-1})^{'}$, with $\pi_{ik} = p(c_{ik} = 1)$, to covariates:

$$\text{logit}(\boldsymbol{\pi}_i) = \boldsymbol{\gamma}_k X_i, \tag{4.8}$$

where γ_k $(k = 1, \ldots, K-1)$ are parameters associated with the covariates.

For the binary variables $\boldsymbol{u}_i$ it is assumed that they are independent given class-membership of subject i. So,

$$p(u_{i1}, \ldots, u_{ir}|\boldsymbol{c}_i) = p(u_{i1}|\boldsymbol{c}_i) \times \ldots \times p(u_{ir}|\boldsymbol{c}_i). \tag{4.9}$$

Equations (4.6) — (4.9) represent the extended latent mixture model in a hierarchical structure. So, in this framework an individual's response is related to continuous latent variables, i.e., growth factors. The growth factors are modelled as a function of covariates and the latent categorical class. Class-membership is predicted in function of a set of observed covariates. Simultaneously a latent-class model for a clinical event measured via a set of binary indicators is estimated.

4.3 Estimation of the Extended Growth Mixture Model

The likelihood function for the growth mixture model is typically very complex and characterized by local maxima. Closed form solutions are not available and numerical optimization are often not successful.

However, maximum likelihood (ML) estimates can be obtained by means of the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977). A detailed description of the EM algorithm for a extended growth-mixture model can be found in Muthén and Shedden (1999). We briefly sketch the algorithm. To implement the EM-algorithm one considers the complete-data likelihood. The complete data consists of the observed data $\boldsymbol{y}$, $\boldsymbol{u}$, $\boldsymbol{x}$ and the latent categorical variable $\boldsymbol{c}$. The complete-data log-likelihood for the extended growth-mixture model is then given by:

$$\sum_{i=1}^{N}\sum_{k=1}^{K} c_{ik}\left[\log(f(\boldsymbol{u}_i|\boldsymbol{c}_i, \boldsymbol{x}_i)) + \log(f(\boldsymbol{y}_i|\boldsymbol{c}_i, \boldsymbol{x}_i)) + \log(f(\boldsymbol{c}_i|\boldsymbol{x}_i))\right]. \tag{4.10}$$

For ease of notation, we use $f(.)$ to refer to either a probability or a density. This log-likelihood function is easier to maximize since the sum over the latent classes and logarithmic function are swapped pertaining to the ordinary log-likelihood function $l = \sum_{i=1}^{N} \log[\sum_{k=1}^{K} \boldsymbol{\pi}_k f_k(\boldsymbol{y}_i)]$. The EM algorithm involves two steps. In the E-step the expected value of the complete-data log-likelihood (4.10) under the posterior distribution of the latent variable $\boldsymbol{c}_i$ is obtained. The conditional probability of individual i to belong to class k, given the observed data, is obtained as

$$\pi_{ik} = p(c_{ik} = 1|\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{x}_i) = p(c_{ik} = 1)\frac{f(\boldsymbol{y}_i|\boldsymbol{c}_i, \boldsymbol{x}_i)f(\boldsymbol{u}_i|\boldsymbol{c}_i, \boldsymbol{x}_i)}{f(\boldsymbol{y}_i, \boldsymbol{u}_i|\boldsymbol{x}_i)}. \tag{4.11}$$

To start the algorithm, random starting values are chosen for all parameters in the model defined by equations (4.6) – (4.9) and for the prior mixing probabilities $p(c_{ik} = 1)$. Let $\boldsymbol{\Omega}^0$ group all starting values. Using (4.11) this results in values $\pi_{ik}(\boldsymbol{\Omega}^0)$ and allows to obtain the expected value of the complete-data log-likelihood (4.10).

In the M step, this expected log-likelihood is maximized for all parameters in $\boldsymbol{\Omega}$, resulting in updated values $\boldsymbol{\Omega}^1$. The maximization is done separately for the $(\boldsymbol{y}, \boldsymbol{x})$ part of the model and the $(\boldsymbol{u}, \boldsymbol{x})$ part of the model. The $(\boldsymbol{y}, \boldsymbol{x})$ part corresponds to simultaneous estimation of the K groups with posterior-probability weighted sample mean vectors and covariance matrix:

$$E\left[\sum_{i=1}^{N} \mathrm{log} f(\boldsymbol{y}_i|\boldsymbol{c}_i, \boldsymbol{x}_i)|\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{x}_i\right] = \sum_{i=1}^{N}\sum_{k=1}^{K} \pi_{ik}(\boldsymbol{\Omega}^0)\mathrm{log} f_k(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Omega}). \tag{4.12}$$

To maximize the $(\boldsymbol{u}, \boldsymbol{x})$ part of the model, a multinomial regression relating $\boldsymbol{c}$ to covariates $\boldsymbol{x}$, is optimized:

$$\sum_{i=1}^{N}\sum_{k=1}^{K} \pi_{ik}(\boldsymbol{\Omega}^0)\mathrm{log}[p(c_{ik} = 1|\boldsymbol{x}_i, \boldsymbol{\Omega})].$$

The parameters describing the relation between $\boldsymbol{u}$ and $\boldsymbol{c}$ and $\boldsymbol{x}$ are estimated via a logistic regression model:

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{r} \pi_{ik}(\boldsymbol{\Omega}^0)\mathrm{log}\ [p(u_{ij})|\boldsymbol{c}_i, \boldsymbol{x}_i, \boldsymbol{\Omega})].$$

The ML estimates for $\boldsymbol{\Omega}$, i.e., $\boldsymbol{\Omega}^1$, are fed to the E-step, and iteration between the E-step and M-step goes on until convergence in the log-likelihood value or in the parameter estimates is obtained.

This EM algorithm for the extended growth-mixture model is implemented in Mplus (Muthén and Muthén, 1998—2012).

4.4 Growth Mixture Modeling for the PDB Data

The Persistent Disturbing Behavior (PDB) data was introduced in Section 3.1. It was argued that it is extremely important to be able to break up the group of PDB patients into natural, smaller groups. The information and description of these refined subgroups comprise relevant information for policy makers, institution managers and fieldworkers. To this end, conventional cluster analysis methods, starting from cross-sectional data, was executed (Serroyen et al., 2010). To study intra-individual change and inter-individual differences in the evolution of the PDB scores these authors also applied a conventional growth model to the data. Classification of the PDB patients into latent subgroups is not possible with this model, since it treats the data as if collected from a single population. This assumption is relaxed in the subsequent analyses, where growth-mixture models are applied to reveal latent PDB trajectory classes.

Longitudinal PDB-scores were obtained using data from the pilot study of 1998 and the repeated measurements available at patient level in the minimal psychiatric registry data (Section 3.1.4). The minimal psychiatric registry was put in place only at the second semester of 1996 in psychiatric hospitals and at the first semester of 1998 in psychiatric nursing homes. Given that 1996 was the year that the registration system started, it is prudent not to put too much trust in the data for this semester. This semester will be ignored in subsequent analysis. Furthermore, for the first semester of 1997, no data are available, owing to the start of the registration system. End 2000, the legal registration framework changed. Therefore attention was restricted to the period 1997.2–2000.1 for the psychiatric hospitals and 1998.1–2000.1 for the psychiatric nursing homes. Data of the 169 patients (126 in psychiatric hospital, 43 in psychiatric nursing home) that according to the interdisciplinary team are PDB patients, are used in this chapter. Individual profiles of the PDB scores for 20 randomly selected PDB patients are displayed in Figure 3.1.

4.4.1 Modeling Strategy

The following model fitting strategy was employed for the PDB data, in line with the proposal of Muthén (2004). First, a conventional, one-class growth model was fitted to obtain some initial insight into the growth factor variation. The repeated PDB scores, were assumed to change linearly over time. A patient's evolution was allowed to deviate from the population average by incorporating a patient-specific intercept and slope. If Y_{ij} denotes the PDB score for patient i at occasion j, t_{ij} a time-related variable then a multilevel formulation of a growth model would specify that,

$$Y_{ij} = \eta_{0i} + \eta_{1i} t_{ij} + \varepsilon_{ij}, \tag{4.13}$$

$$\eta_{0i} = \alpha_0 + \xi_{0i}, \tag{4.14}$$

$$\eta_{1i} = \alpha_1 + \xi_{1i}. \tag{4.15}$$

The residuals $\boldsymbol{\varepsilon}$, $\boldsymbol{\xi}_0$ and $\boldsymbol{\xi}_1$ are assumed to be zero-mean normally distributed, level 2 residuals $\boldsymbol{\xi}_0$ and $\boldsymbol{\xi}_1$ are possibly correlated but uncorrelated with level 1 residuals

$\boldsymbol{\varepsilon}$. The time-specific residuals $\boldsymbol{\varepsilon}$ have a covariance matrix Σ, the $\boldsymbol{\xi}$-residuals have covariance matrix Ψ.

The time-related variable t was defined such that 1998.2 corresponds to $t = 0$. The time variable increases with 1 for each semester. Resulting in a time variable ranging from -2—3 for the psychiatric hospitals and from -1—3 for the psychiatric nursing homes.

Growth-mixture models were used to identify latent classes. A growth-mixture model is obtained by freeing parameters in (4.14) and (4.15), and allowing different classes (k =1, . . . , K) to vary around different mean growth curves, and by specifying class-specific covariance matrices Ψ_k for the $\boldsymbol{\xi}$-terms and Σ_k for $\boldsymbol{\varepsilon}$. A series of unconditional growth mixture models, i.e., not including covariate information, with two to four classes were considered. Various constraints for the growth factor variances and residual variances of the PDB scores were investigated. The variances of the growth factors were set equal to zero in Model I, which corresponds to the approach of Nagin and Tremblay (2001). They were constrained to be equal across classes in Model II, and they were allowed to range freely in Model III. The residual variances ε_{ij} were constrained to be constant across classes but allowed to change over time, Model A, to be constant over time but with a different variance per class, Model B, and finally the residual variances were left unconstrained in Model C.

Selection of a final model was based on the Bayesian Information Criterion value (Schwarz, 1978) and the sample-size adjusted version of it (Sclove, 1987),

$$BIC = -2\text{log}(L) + p\text{log}(N),$$

and

$$aBIC = -2\text{log}(L) + p\text{log}(\frac{N+2}{24}),$$

where p is the number of model parameters and N is the sample size. Additionally, and by way of sensitivity analysis, we also consider Akaike's Information Criterion (Akaike, 1974), $AIC = -2\text{log}(L) + 2p$, supplemented with the likelihood-ratio test proposed by Lo, Mendell, and Rubin (2001), designed to compare a given model with a model containing one class less. The associated p-value represents the evidence in favour of the simpler model. A parametric bootstrapped likelihood-ratio test (McLachlan and Peel, 2000) could also be considered, but this method is rather time consuming.

To evaluate the certainty in classification the entropy (Ramaswamy et al. 1993) was used. The relative entropy is defined by the model-based probabilities, i.e.,:

$$E = 1 - \frac{-\sum_{i=1}^{N}\sum_{k=1}^{K}\pi_{ik}\text{log}(\pi_{ik})}{N\text{log}(K)}, \tag{4.16}$$

with π_{ik} the posterior probability that subject i belongs to class k. The values of the relative entropy range from 0 to 1, with higher values indicating high certainty in classification. Clark (2010) suggests using a cut-off point of 0.80 for good classification and 0.6 for a medium classification.

The data for the psychiatric hospitals and psychiatric nursing homes are analysed separately. All models were fitted in Mplus (Muthén and Muthén, 1998-2012).

4.4.2 Psychiatric Hospitals

Model fitting procedures for the single-class model resulted in a log-likelihood value of -1040.5 and a BIC of 2099.4. The estimated slope is -0.085 (p=0.012) and indicates that PDB-score decreased over the 6 assessments. The estimated latent intercept variance of 2.365 (p=0.0114) for the PDB-score, and the variance of the latent slope score of 0.050 (p <0.0001), indicate that substantial variation exists among the PDB patients, at time zero, but also in the rate of change over time. The significance of the random effects were investigated by means of likelihood-ratio statistics with the asymptotic null distributions given a mixture of two chi-squared distributions. In graphical displays, the calendar time (semester of the year) will be used, instead of the time variable used in the statistical models.

Table 4.1 displays model fit results of unconditional growth-mixture models with two to four classes. Models I, II and III, defining the growth factor variances and models A, B, and C specifying the residual variances, are defined in Section 4.4.1. We show the value of the log-likelihood, the sample size adjusted BIC, the AIc, the number of parameters in the model, and the entropy. Based on the information criteria, the three-class linear Model IIIB was selected as the optimal model. This choice is a compromise between goodness-of-fit on the one hand and the desire to select a model that is not overly complex on the other, bearing in mind that more elaborate models might be less than optimal for prediction and classification purposes. This three-class mixture model resulted in a log-likelihood value of -975.8, a BIC of 1985.0 and an entropy estimate of 0.71. Parameter estimates and standard errors for all fixed effects and all variance components are reported in Table 4.2. Five percent (n=6) of the patients were allocated to the first class, 33% (n=41) to the second class and 63% (n=79) to the third class. Based on the intercept and slope factor, these classes were labeled: (1) ‘Low group’, (2) ‘High but improving’, and (3) ‘Middle group’.

Figure 4.1 displays the predicted trajectories for the three classes. The linear trends appear to describe the data well.

Table 4.3 shows the posterior class-membership probabilities for the three-class

Table 4.1: *Models for the psychiatric hospitals: log likelihood ℓ, sample size adjusted BIC and AIC, the entropy, the Lo-Mendell-Rubin likelihood-ratio test (p value), and the number of parameters in the model are reported. Growth factor variances are: Model I, equal to zero; II, equal across classes; III, unconstrained. Residual variances are: A, constant across classes; B, constant over time; C, unconstrained – PDB study.*

Classes		ℓ	BIC	AIC	entropy	LMR-LRT (p)	#par
Model I							
2	A	-1162.75	2343.92	2347.51	0.810	173.54 (0.1594)	11
	B	-1161.81	2335.34	2337.62	0.794	190.04 (0.4205)	7
	C	-1150.11	2328.67	2334.22	0.801	206.06 (0.3584)	17
3	A	-1103.13	2229.70	2234.26	0.879	111.56 (0.4513)	14
	B	-1082.05	2182.50	2186.09	0.868	151.68 (0.1280)	11
	C	-1073.26	2190.05	2198.53	0.870	150.24 (0.2031)	26
4	A	-1063.46	2155.38	2160.92	0.893	74.23 (0.0647)	17
	B	-1051.25	2127.62	2132.51	0.901	58.56 (0.0020)	15
	C	-1038.43	2135.44	2146.85	0.882	68.11 (0.1745)	35
Model II							
2	A	-1032.62	2088.38	2092.94	0.933	15.03 (0.2139)	14
	B	-992.31	2001.35	2004.61	0.687	102.13 (0.0000)	10
	C	-983.00	1999.48	2006.00	0.719	112.42 (0.0001)	20
3	A	-1026.95	2082.36	2087.91	0.950	10.33 (0.3051)	17
	B	-986.95	1997.34	2001.91	0.758	10.18 (0.0050)	14
	C	-970.01	1988.57	1998.02	0.702	25.40 (0.3117)	29
4	A	-1023.33	2080.13	2086.65	0.951	6.96 (0.0800)	20
	B	-982.37	1994.88	2000.76	0.736	8.71 (0.3831)	18
	C	-954.10	1971.81	1984.20	0.788	31.11 (0.2241)	38
Model III							
2	A	-1025.28	2079.02	2084.57	0.478	29.43 (0.0198)	17
	B	-988.37	1998.50	2002.74	0.664	111.97 (0.0000)	13
	C	-981.07	2000.64	2008.14	0.688	116.86 (0.0809)	23
3	A	-1018.21	2074.91	2082.41	0.395	13.68 (0.3559)	23
	B	-975.78	1985.03	1991.55	0.708	24.46 (0.0903)	20
	C	-964.87	1988.33	1999.74	0.694	31.85 (0.3329)	35
4	A	-1014.57	2077.69	2087.15	0.528	7.81 (0.7622)	29
	B	-971.96	1989.11	1985.91	0.806	19.08 (0.0846)	27
	C	-946.01	1970.70	1986.02	0.720	28.73 (0.1646)	47

Table 4.2: *Summary of the three-class model, selected for the psychiatric hospitals. Parameter estimates, standard errors and t value for the class-specific intercepts and slopes are shown – PDB study.*

Effect	**Estimate**	**Standard Error**	***t*-value**
Fixed effects			
Intercepts			
Class 1	−0.913	0.758	−1.025
Class 2	1.558	0.296	5.256
Class 3	0.372	0.160	2.325
Slopes			
Class 1	0.064	0.009	6.929
Class 2	−0.271	0.092	−2.944
Class 3	0.021	0.030	0.705
Random effects			
Variance of intercepts			
Class 1	3.507	1.857	
Class 2	2.552	0.698	
Class 3	1.155	0.279	
Variance of slopes			
Class 1	−0.006	0.002	
Class 2	0.132	0.064	
Class 3	0.014	0.010	
Covariance of intercept, slope			
Class 1	0.038	0.018	
Class 2	−0.120	0.164	
Class 3	0.007	0.040	
Residual Variance			
Class 1	0.110	0.045	
Class 2	1.849	0.234	
Class 3	0.376	0.052	

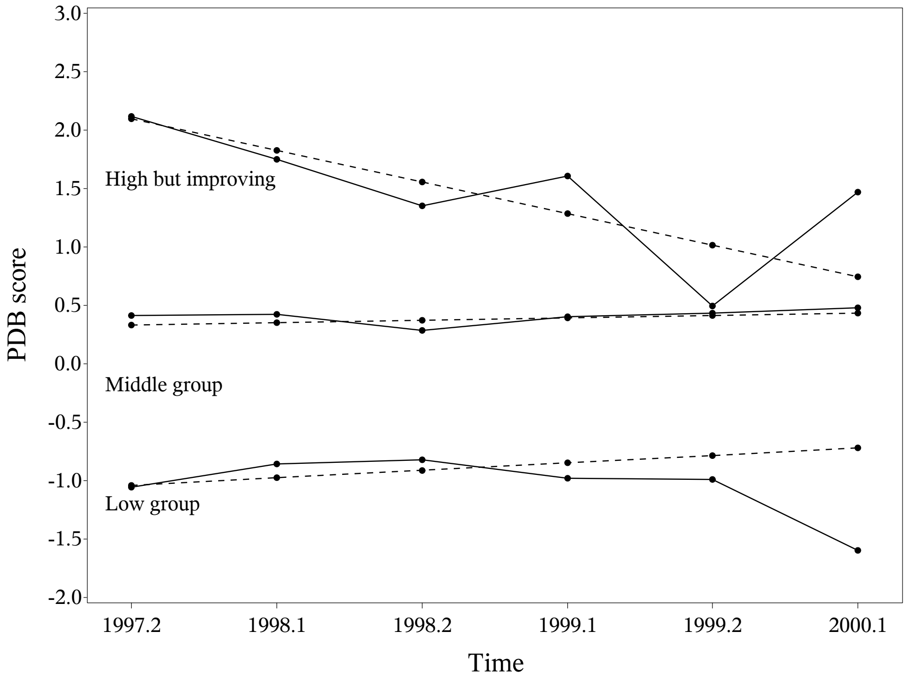


Figure 4.1: *Observed (full lines) and predicted (dashed lines) trajectories for the 3-class model (psychiatric hospitals) – PDB study.*

model. High diagonal and low off-diagonal values indicate good classification. The agreement between the latent-class membership and status in terms of average class probability is highest for class 2, the patients with high PDB-scores. Patients of this class had 91% chance to be assigned to class 2. Classes 1 and 3 are most difficult to distinguish. Patients classified in class 3, i.e., the stable patients, have 15% chance to belong to class 1.

More than 50% of the PDB patients belong to class 3, with an average PDB score of 0.37. About one in three patients has high PDB scores, averaging around 1.56 at time 0 (1998.2). It is sensible to conclude that the behaviour of this group is more disturbing than the behaviour of the other groups. That there is no evolution in the PDB scores over time for class 3 does not mean that the behaviour itself is constant. The only conclusion that can be drawn, is that the behaviour remains disturbing in the same degree. The type of disturbing behaviour can however change over time.

Table 4.3: *Agreement between the classification probabilities based on the average class probabilities and latent class membership, for the three-class model displayed in Table 4.2 – PDB study.*

	Latent class		
Average class prob.	**1**	**2**	**3**
1	0.840	0.011	0.149
2	0.000	0.914	0.086
3	0.052	0.105	0.842

The variances of the intercepts are relatively large, indicating that even within a class there is still heterogeneity. The significance of the random effects was investigated with a likelihood-ratio statistic, with as asymptotic null distributions a mixture of chi-squared distributions. The variance of the intercept is significant for all three classes, the variance of the slope is highly significant for classes 1 and 2, but not for class 3.

When studying the sample variances, weighted by the estimated class probabilities we find that the variances are not constant over time. For classes 1 and 2, the variance is smallest around times 0 and 1. This is when the pilot study was performed. The repeated PDB scores are constructed based on a discriminant function that was built using the data of the pilot study in 1998 (see Section 3.1). It is therefore important that the analysis can accommodate non-constant variances, as fortunately is the case.

4.4.3 Psychiatric Nursing Homes

Model fitting procedures for the single-class model resulted in a log-likelihood value of -586.8 and a BIC of 1183.7. PDB patients residing in a psychiatric nursing home have constant PDB scores over time. The estimated slope of 0.085 is not significant ($p = 0.1905$). The large intercept variance of 4.0334 ($p < 0.0001$) for the PDB-score shows that variation exists among PDB patients in psychiatric nursing homes in terms of their (average) PDB-score. The estimate of the slope variance equals zero, indicating that, under a hierarchical interpretation of the model, the random slope can be removed.

Table 4.4 displays the results of fitting unconditional growth-mixture models with two to four classes. Models I, II, and III, defining the growth factor variances and models A, B, and C specifying the residual variances, are as in Section 4.4.1. The two-class Model IIIC is the preferred choice, using logic similar to the one employed in the case of psychiatric hospitals. This two-class mixture model resulted in a log-likelihood value of -541.7, a BIC of 1104.6, and an entropy estimate of 0.84. Figure 4.2 displays the observed and predicted trajectories of the classes. The estimate of the mean PDB-score in 1998.2 equals -0.10 in class 1 and 1.00 in class 2. Two thirds of patients belong to class 2. The variances of the intercepts show that within a class patients still differ. The heterogeneity is largest in class 2. The average class 1 probability equals 0.91 for class 1, and the average class 2 probability equals 0.98 for class 2. This indicates that the groups are well separated.

4.5 Conclusion

Persistent disturbing behaviour (PDB) is a highly disruptive condition. Proper treatment and organization of care pose important challenges. So far, it had not been properly defined, let alone circumscribed and characterized. Previous analyses, based on a pilot study, indicated that the group of PDB patients is likely larger than generally believed, complicating fieldwork organization because, additionally, it is desirable that PDB wards are small. The growth mixture analyses provide some basis for grouping patients into organizational units based on the degree and evolution of their condition. This does not mean that they might be able to socially function together, but rather that they will be receiving similar types and intensities of care. This will be advantageous for the care givers involved.

The analyses, based on growth mixture modeling, lead to two important conclusions. First, meaningful, plausible groups may well exist, in spite of previous findings that were less optimistic (Serroyen et al., 2010). While previous analyses indicated, at best, the presence of two groups, we reached plausible evidence for three groups, categorized as high, medium, and low, in terms of PDB-score profiles. Second, the analyses clearly show that there is a lot of variability, even within a group of patients whose behaviour is experienced as disturbing by the care team. The group with extremly disturbing behaviour is about one third, 35%, of the entire group. Setting up specialized wards for this group could also lead to better living circumstances for the remaining patients at the wards.

For the entire PDB group in psychiatric hospitals, a linear decrease in the average PDB scores was reported previously. A repeated-measurements analyses indicated that patient-specific characteristics are important and that some patients have intrin-

Table 4.4: *Models for the psychiatric nursing homes: log likelihood ℓ, sample size adjusted BIC and AIC, the entropy, the Lo-Mendell-Rubin likelihood-ratio test (p value), and the number of parameters in the model are reported. Growth factor variances are: Model I, equal to zero; II, equal across classes; III, unconstrained. Residual variances are: A, constant across classes; B, constant over time; C, unconstrained. Empty entries result from convergence failure – PDB study.*

Classes		ℓ	**BIC**	**AIC**	**entropy**	**LMR-LRT** (p)	**#par**
			Model I				
2	A	-622.76	1255.64	1265.53	0.839	99.37 (0.5228)	10
	B	-619.02	1245.13	1252.05	0.796	110.89 (0.0779)	7
	C	-615.02	1245.22	1260.04	0.825	119.19 (0.0912)	15
3	A	-599.77	1212.68	1225.53	0.876	42.59 (0.0959)	13
	B	-584.05	1179.22	1190.09	0.900	65.99 (0.0531)	11
	C	-571.72	1166.71	1189.45	0.945	84.07 (0.0460)	23
4	A	-585.62	1186.91	1202.73	0.899	26.67 (0.1752)	16
	B	-542.17	1099.51	1114.33	0.926	79.01 (0.0639)	15
	C	-544.11	1119.58	1150.22	0.911	53.62 (0.0432)	31
			Model II				
2	A	-576.45	1166.05	1178.90	0.976	19.11 (0.0141)	13
	B	-558.30	1126.71	1136.36	0.741	75.05 (0.0003)	10
	C	-551.18	1120.56	1138.36	0.717	69.10 (0.0023)	18
3	A	-575.24	1166.65	1182.47	0.979	2.25 (0.1838)	16
	B	-553.16	1120.47	1134.31	0.835	9.70 (0.0858)	14
	C	-541.65	1109.61	1135.31	0.700	18.49 (0.3939)	26
4	A	-573.61	1166.44	1185.22	0.895	3.01 (0.7151)	19
	B	-551.67	1121.54	1139.34	0.818	2.81 (0.3170)	18
	C	-	-	-	-	-	-
			Model III				
2	A	-574.88	1165.95	1181.76	0.494	22.86 (0.0485)	16
	B	-557.53	1128.22	1141.07	0.718	78.40 (0.0009)	13
	C	-541.66	1104.55	1125.31	0.835	88.30 (0.0900)	21
3	A	-565.08	1152.42	1174.17	0.578	10.90 (0.2472)	22
	B	-545.58	1111.39	1131.16	0.846	23.11 (0.0403)	20
	C	-516.92	1066.22	1097.85	0.899	48.40 (0.0295)	32
4	A	-	-	-	-	-	-
	B	-536.71	1100.74	1127.43	0.905	20.17 (0.5365)	27
	C	-499.18	1041.85	1084.36	0.816		43

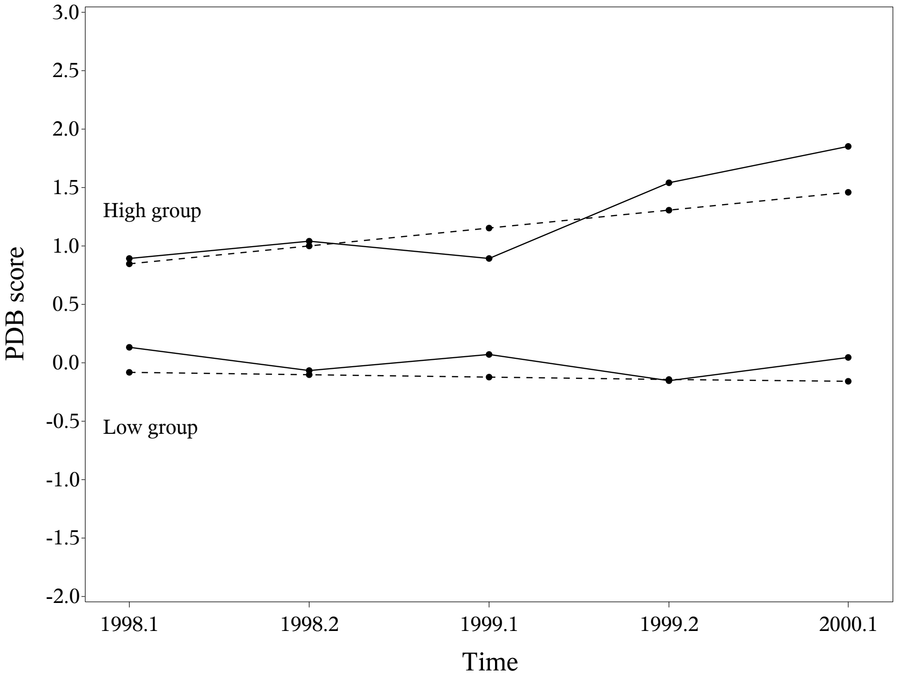


Figure 4.2: *Observed (full lines) and predicted (dashed lines) trajectories for the 2-class model (psychiatric nursing homes) – PDB study.*

sically high values, while others have low values (Serroyen et al., 2010). These findings stemmed from a conventional growth model. Juxtaposing the results of the conventional growth model and the results of the growth mixture model, we discern that part of the heterogeneity in the PDB population is explainable by it being a mixture of classes, which differ not only in their mean values but also in their evolution. Most patients have moderately stable PDB-scores.

For the PDB patients in psychiatric nursing homes, the growth mixture analyses reveal two classes. The distinction between the groups is essentially the average level of the score. Thus, the condition does not worsen or improve. This is not surprising, as we are dealing with a chronic, therapy-resistant group of patients. With current knowledge of therapy and medication, the behaviour of these patients cannot be improved. At the same time, the absence of worsening underscores the chronic nature of the group, which reaches and gets trapped in its worst condition.

Table 4.5: *Summary of the two-class model, selected for the psychiatric nursing homes. Parameter estimates, standard errors and t value for the class-specific intercepts and slopes are shown – PDB study.*

Effect	**Estimate**	**Standard Error**	***t*-value**
Fixed effects *Intercepts*			
Class 1	−0.103	0.446	−0.231
Class 2	0.999	0.372	2.689
Slopes			
Class 1	−0.019	0.011	−1.638
Class 2	0.153	0.086	1.774
Random effects			
Variance of intercepts			
Class 1	2.794	0.608	
Class 2	4.375	1.053	
Variance of slopes			
Class 1	-0.029	0.020	
Class 2	-0.068	0.107	
Covariance of intercept, slope			
Class 1	-0.005	0.012	
Class 2	-0.089	0.188	
Residual Variance			
Class 1			
time -1	0.438	0.198	
time 0	−0.042	0.025	
time 1	0.582	0.340	
time 2	0.467	0.271	
time 3	0.561	0.399	
Class 2			
time -1	2.961	1.093	
time 0	1.167	0.428	
time 1	2.687	0.853	
time 2	5.751	2.022	
time 3	4.368	1.947	

Chapter 5

Model-based Clustering for Multivariate Longitudinal Data

5.1 Introduction

Latent growth modeling approaches, such as growth-mixture models, were introduced in Chapter 4. These methods identify, for a heterogeneous population, subgroups of individuals that are similar in terms of their evolution of a longitudinal response. The methodology was applied to the PDB data (Section 4.4) and did reveal a meaningful subdivision of the patients.

When, for each individual, more than a single outcome is measured over time, a multivariate set of longitudinal profiles is obtained. Interest could be in finding subgroups of individuals that are similar in their evolution over time for the various repeated sequences. Thus, the goal is to find clusters that are unique in the evolutions over time of the different outcomes, as well as in the correlation structure over time and between these outcomes.

However, when fitting growth-mixture models to a multivariate repeated measurements setting, computational problems are likely to occur. In mixed-effects models, the correlation between the multivariate longitudinal profiles is dealt with via specification of a joint distribution of the random effects. When the dimension of this joint distribution becomes too high computations will stall.

Section 5.2 describes some existing techniques to find clusters when information about multiple longitudinally measured responses is available. A clustering algorithm for multivariate repeated data, using pseudo-likelihood and ideas based on K-means clustering, is proposed in Section 5.3. The algorithm is demonstrated on the EEG data in Section 5.4. Section 5.5 investigates the performance of the algorithm by means of a simulation study. Finally, Section 5.6 contains a discussion.

5.2 Review of Relevant Existing Work

The dual trajectory model, to analyze the evolution of two related outcomes, was presented by Nagin and Tremblay (2001) and by Nagin (2005). Let $\boldsymbol{Y_1} = (Y_{11}, Y_{12}, \ldots, Y_{1T_1})$ and $\boldsymbol{Y_2} = (Y_{21}, Y_{22}, \ldots, Y_{2T_2})$ denote two longitudinal profiles to be modeled together. The dual trajectory model assumes that the J trajectories groups of $\boldsymbol{Y_1}$ are probabilistically linked to the K groups for $\boldsymbol{Y_2}$. In addition, the model makes the assumption of conditional independence given group membership (as is the case in the single trajectory group-based approach). So, conditional on j and k, $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$ are independently distributed, $f_{jk}(\boldsymbol{y_1}, \boldsymbol{y_2}) = g_j(\boldsymbol{y_1})h_k(\boldsymbol{y_2})$, where $f(\cdot)$, $g(\cdot)$ and $h(\cdot)$ are suitable probability distributions. If π_{jk} is the joint probability of membership to trajectory j for $\boldsymbol{Y_1}$ and trajectory k for $\boldsymbol{Y_2}$, then $f(\boldsymbol{y_1}, \boldsymbol{y_2}) = \sum_j \sum_k \pi_{jk} g_j(\boldsymbol{y_1}) h_k(\boldsymbol{y_2}) = \sum_j \sum_k \pi_{k|j} \pi_j g_j(\boldsymbol{y_1}) h_k(\boldsymbol{y_2})$. The dual trajectory model results in estimates for the trajectories for both longitudinal responses, probabilities of group membership for each trajectory and conditional probabilities linking membership across the trajectory groups of the two profiles. The model can be used for outcomes that evolve contemporaneously (e.g., depression and alcohol use) or that evolve over different time periods. In their work, Nagin and Tremblay (2001) present the Montreal prospective longitudinal study, where children from kindergarten classes in low socio-economic Montreal neighborhoods were assessed on a wide range of factors, including hyperactivity and physical aggression. These factors were evaluated at age 6 and annually from ages 10 to 15. Much research has documented the overlap of physical aggression and hyperactivity in children. Nagin and Tremblay show how jointly estimating the developmental trajectories can illuminate the nature of this overlap.

Conceptually, the extension of the dual model to more than two outcomes is straightforward. From a practical point, the addition of outcomes results in an unmanageable proliferation of probability matrices linking the trajectories for the various outcomes. Still, there are many circumstances where it is valuable to link trajectories of three or more outcomes of interest. Applications of the multitrajectory modeling approach can be found in Nagin (2005) and Piquero et al. (2002).

An alternative approach for the joint analysis of more than one series of longitudinal measurements is described by Putter et al. (2008). First, a latent class joint model for the longitudinal outcomes is used in order to reduce the dimensionality of the problem. The parameters in this model are estimated in two-stages. In the first stage, the latent classes, their probabilities and the mean and covariance structure are estimated based on the longitudinal data of the first outcome. In the second stage, the relation between the latent classes, patient characteristics, and the other outcome(s) is modelled. This approach is particularly attractive when main interest is in the relation between the longitudinal outcomes. Putter et al. demonstrated the usefulness of the method for data from 195 lung cancer patients in two outpatient clinics of lung diseases in The Hague. The relation between denial and longitudinal health measures was of interest. The analysis revealed an interesting phenomenon: although no difference between classes could be detected for objective measures of health, patients in classes representing higher levels of denial consistently scored significantly higher in subjective measures of health.

Roy and Lin (2000) propose a latent-variable model for repeated measures for different outcomes that are assumed to measure an underlying quantity of main interest. They relate the observed outcomes to a latent variable by means of random (intercept) effects regression models. The random intercepts are independent, i.e., conditional independence of the outcomes given the latent variable applies. The latent variable is modeled as a function of covariates by a separate linear mixed model. The method was illustrated using data from a panel study on changes in methadone treatment practices.

Extensions of the group-based approach for multivariate longitudinal data to the growth-mixture modeling setting are problematic, stemming from the high dimension of the joint distribution of the random effects (see Section 5.3.1).

In the next section, we propose a clustering algorithm for multivariate longitudinal data, based on a bivariate joint-modeling approach.

5.3 Proposed Clustering Algorithm

The clustering algorithm for multivariate longitudinal data that we are proposing resembles a K-means iterative cluster procedure. The idea is to divide the observations in K clusters such that the full likelihood for the m jointly measured repeated outcomes becomes maximized. Therefore, a joint-modeling approach using mixed models is implemented. Fieuws and Verbeke (2008) use this pairwise modeling strategy in a discriminant analysis. They predict renal graft failure by fitting bivariate mixed models to 4 repeatedly measured markers. The obtained estimates were used in Bayes rule to obtain the prognosis for long-term success of the transplant, at each point in time.

Before specifying the different steps of the clustering algorithm, some background on fitting a joint model for multivariate longitudinal data is given.

5.3.1 Joint Model for Multivariate Longitudinal Data

When m longitudinally measured outcomes are available for N subjects, a joint-modeling approach using mixed models can be followed.

Let $\boldsymbol{Y}_i = (\boldsymbol{Y}_{1i}, \dots, \boldsymbol{Y}_{mi})$ denote the vector of all measurements obtained for subject i, where $\boldsymbol{Y}_{mi}$ contains the longitudinal observations of response m. The full log-likelihood

$$\sum_{i=1}^{N} l_i(\boldsymbol{y}_{1i}, \boldsymbol{y}_{2i}, \dots, \boldsymbol{y}_{mi} | \boldsymbol{\Theta}^*) \tag{5.1}$$

then has to be maximized with respect to $\boldsymbol{\Theta}^*$. In this expression, l_i is the log-likelihood contribution of subject i to the full joint mixed model. This full joint model can be specified as a series of random-effects models, one for each outcome process, and the processes are linked by imposing a joint multivariate distribution on the random effects. So, for each outcome, we specify $\boldsymbol{Y}_{mi}|\boldsymbol{b}_{mi} \sim N(X_{mi}\boldsymbol{\beta}_m + Z_{mi}\boldsymbol{b}_{mi}, \Sigma_{mi})$ with $\boldsymbol{b}_{mi}$ the q-dimensional vector containing the random effects for response m for subject i. If we assume the same number of random effects for each outcome, $\boldsymbol{b}_i = (\boldsymbol{b}_{1i}, \boldsymbol{b}_{2i}, \dots, \boldsymbol{b}_{mi})$ is a $q \times m$-dimensional vector containing all random effects with a general $(q \times m) \times (q \times m)$ covariance matrix.

Although this approach has many advantages (it allows for unbalanced designs and response variables can be of different nature), its usability is limited by the dimension (m) of the data. In case the number of repeated outcomes becomes large, computational problems are likely in the estimation process due to the high dimension of the joint distribution of the random effects.

Instead of maximizing the likelihood of the full joint model, a pairwise approach can be used to obtain estimates for all parameters in $\boldsymbol{\Theta}^*$. Pseudo-likelihood estimation (Besag, 1975) replaces the joint full likelihood by a suitable product of marginal or conditional densities, where this product is easier to evaluate than the original likelihood. Fitting all possible pairwise models is equivalent to maximizing a pseudo-likelihood function of the following form:

$$p\ell(\boldsymbol{\Theta}, \boldsymbol{Y}) = \sum_{i=1}^{N} \sum_{(r,s)} l_{rs,i}(\boldsymbol{y}_{r,i}, \boldsymbol{y}_{s,i} | \boldsymbol{\Theta}_{rs}), \tag{5.2}$$

with $r = 1,\ldots, m-1$; $s = r+1,\ldots, m$, $l_{rs,i}$ the log-likelihood contribution of subject i in the bivariate model for outcomes r and s, and N the total number of subjects. $\boldsymbol{\Theta}_{rs}$ represents the vector of all parameters in the bivariate joint mixed model corresponding to the specific pair (r, s) of repeated outcomes. Let $\boldsymbol{\Theta}$ be the stacked vector combining all pair-specific parameter vectors $\boldsymbol{\Theta}_{rs}$. There are $m(m-1)/2$ bivariate joint models to be considered. Some parameters in $\boldsymbol{\Theta^*}$ have a single counterpart in $\boldsymbol{\Theta}$, e.g., the covariance between random effects from two different outcomes. Other elements in $\boldsymbol{\Theta}^*$ have multiple counterparts in $\boldsymbol{\Theta}$, e.g., the covariance between random effects from the same outcome. Given that the pairwise approach fits within the pseudo-likelihood framework, an asymptotic multivariate normal distribution for $\boldsymbol{\Theta}$ can be derived. Asymptotic normality of the pseudo-likelihood estimator in the single parameter case and in the vector-valued parameter case is shown in Arnold and Strauss (1991) and in Geys et al. (1999). Finally, estimates for the parameters in $\boldsymbol{\Theta}^*$ in (5.1) can be calculated by taking averages over all pairs. This is obtained by

$$\widehat{\boldsymbol{\Theta}}^* = A\widehat{\boldsymbol{\Theta}} \sim N(\boldsymbol{\Theta^*}, A\Sigma(\widehat{\boldsymbol{\Theta}})A'),$$

with A a matrix containing the appropriate coefficients to calculate the averages and $\Sigma(\widehat{\boldsymbol{\Theta}})$ equals the covariance matrix for $\widehat{\boldsymbol{\Theta}}$ obtained by an expression shown in Arnold and Strauss (1999). A mean estimate is simply obtained by averaging all corresponding pair-specific estimates in $\widehat{\boldsymbol{\Theta}}$. Standard errors of these estimates take into account the variability amongst the pair-specific estimates. Furthermore estimates corresponding to two pairwise models with a common outcome are based on overlapping information and hence are correlated. This correlation is also accounted for in the sampling variability of the combined estimates in $\widehat{\boldsymbol{\Theta}}^*$.

The idea of replacing the full likelihood by pairwise marginal likelihoods is used in Step 3 of the proposed clustering algorithm.

5.3.2 Clustering Algorithm for Multivariate Longitudinal Data

In this section, we propose an algorithm to reveal latent subgroups in multivariate repeated outcomes. The idea is that the data are not coming from one multivariate distribution, but rather that the generation process behind the data is a mixture of a number of multivariate normal distributions, each described by a density $N(\boldsymbol{\mu}^k, V^k)$. We use superscripts to indicate the k^{th} cluster. In the subsequent expressions, subscripts will be used to refer to the pair (r, s) of responses. The algorithm aims at allocating the N observations in a number of clusters, such that the full likelihood is maximal. The algorithm is iterative in nature and resembles a partition cluster method. The cluster criterion being used is an individual's pseudo-log-likelihood contribution. An illustration of the algorithm is given in Section 5.4.

The algorithm consists of the following steps:

1. Choose the number of clusters, K.

2. Randomly divide the N observations into K initial clusters.

3. Iterate the following steps until no observation switches cluster anymore:

   (a) Fit all bivariate joint models with the K clusters as 'known' groups (see Section 5.3.1). For each bivariate joint model, based on outcomes r and s, this results in K mean profiles over time, $\boldsymbol{\mu}^k_{rs}$ as well as K covariance matrices V^k_{rs} $(k = 1, \dots, K)$.

   (b) For each pair $p = (r, s)$ of longitudinal outcomes $(p = 1, \dots, P)$ the following K likelihoods for observation i are then calculated $(k = 1, \dots, K)$:

   $$L^k_{p,i} = L^k_{rs,i} = (2\pi)^{-n_{p,i}/2}|V^k_p|^{-1/2}e^{\frac{1}{2}(\boldsymbol{y}_{p,i}-\boldsymbol{\mu}^k_p)'(V^k_p)^{-1}(\boldsymbol{y}_{p,i}-\boldsymbol{\mu}^k_p)}.$$

   (c) The full joint likelihood function of all m responses will reach its maximum value when the pseudo-likelihood is maximal. Therefore, the sum of natural logarithms of these pseudo-likelihoods over all P pairs is a natural choice as a cluster criterion:

   $$pl^k_i = \sum_{p=1}^{P} \log(L^k_{p,i}). \tag{5.3}$$

   For each observation i this results in K individual pseudo-likelihood values. Each observations is (re-)classified into the group having the highest individual pseudo-likelihood. Steps (a) through (c) are repeated until no observations change cluster.

We will dwell on Steps 1. and 2. in turn.

Random Division in K Initial Clusters

It is well known that the results of the partition cluster method can depend on the starting cluster seeds, both in the number of clusters found and in their centroids. An unfortunate choice of the division to start the procedure in the first step could lead to a poor final division of the data into K groups. To minimize this risk, it is advisable to repeat the proposed clustering algorithm for a number of times, with different randomly chosen K initial groups. Out of these runs, the replicate giving the highest value for the pseudo-likelihood value (5.2), has to be selected as the final solution. This strategy reduces the possibility of accepting a poor solution due to an inappropriate starting seed. In addition, reproducibility of the pseudo-likelihood value is an indication of how well a particular number of groups fits the natural structure of the data.

Number of Clusters

To determine the optimal number of clusters, it is advisable to perform the clustering algorithm for a range of K-values (e.g., $K = 2, 3, 4, \ldots$). To evaluate the need to include an additional cluster in the model the Akaike and Bayesian information criterion could be used.

Counterparts for the AIC and the BIC information criteria for model selection were derived for the framework of pseudo-likelihood. A pseudo-likelihood is in fact a misspecified likelihood and consequently the second Bartlett identity does not hold, meaning that the Hessian and the variance of the score function are not equal. We refer to Varin and Vidoni (2005) for a derivation of the pseudo AIC criterion, and to Gao and Song (2010) for the pseudo BIC criterion. The criteria have the usual form, but the effective number of parameters is to be estimated from the Hessian matrix, $H(\boldsymbol{\Theta}) = E_{\Theta}\left(-\nabla_{\Theta} u(\boldsymbol{\Theta}, \boldsymbol{y})\right)$, and the variability matrix of the pseudo score functions $J(\boldsymbol{\Theta}) = \mathrm{Var}_{\Theta}\left(u(\boldsymbol{\Theta}, \boldsymbol{y})\right)$ with the score functions $u(\boldsymbol{\Theta}, \boldsymbol{y}) = \nabla_{\theta} p\ell(\boldsymbol{\Theta}, \boldsymbol{y})$. The effective number of parameters is $\dim(\boldsymbol{\Theta}) = \mathrm{tr}(J(\boldsymbol{\Theta})H^{-1}(\boldsymbol{\Theta}))$ and

$$\begin{aligned} AIC &= -2p\ell(\widehat{\boldsymbol{\Theta}}, \boldsymbol{y}) + 2\dim(\widehat{\boldsymbol{\Theta}}), \\ BIC &= -2p\ell(\widehat{\boldsymbol{\Theta}}, \boldsymbol{y}) + \log(N)\dim(\widehat{\boldsymbol{\Theta}}), \end{aligned}$$

where $p\ell$ is the pseudo-likelihood function (5.2) evaluated at $\widehat{\boldsymbol{\Theta}}$.

5.4 Application to EEG Data

To illustrate the proposed clustering algorithm, the method was applied to an Electro-Encephalogram (EEG) study conducted at Janssen Pharmaceutica (Belgium). The data was introduced in Section 3.2. The aim of study is to characterize the effects of psychotropic drugs on cortical brain activity, on the basis of spectral electro-encephalograms. For each rat, 9 activity profiles are obtained, at 6 different positions in the brains. To illustrate the clustering algorithm, we focus on the frequencies obtained at the left prefrontal cortex. So we are facing 9-variate longitudinal profiles. To be able to compare the results with analyses done in the past, we only include the rats, with follow-up data, on the placebo and the highest dose level. This reduces the data set to 139 rats. That said, very comparable results were obtained when including all four dose levels in the analyses (data not shown).

The data for the 9 frequency measurements are visually presented in Figures 3.2 to 3.4. The response of interest is the percentage change with respect to the measurement at baseline Y_{ib} (administration of the drug): $Y_{ij} = 100(Y_{ij} - Y_{ib})/Y_{ib}$. At baseline all percentage changes are by definition equal to zero. Graphical displays therefore exclude the baseline data. In graphical displays and in the statistical models, time zero refers to the first measurement obtained after administering the drug (i.e., after 45 minutes). Heterogeneity is seen in all waves, some rats have a decrease in the frequency while for others an increase is obtained as an effect of the drug. For some waves extreme profiles are seen, such as for the α_1 wave. This heterogeneity is of course largely caused by administrating 10 different drugs at different dose levels. When applying the clustering algorithm, this information was not taken into account. The goal of the analyses is to see if it is possible to identify subpopulations within the set of 139 rats. Subpopulations that are homogeneous in the growth parameters for the 9 waves and in the correlation structure. The information about the compounds and doses will later be used to assess whether the identified groups are meaningful. All analyses were performed in SAS/STAT software, Version 9.3 of the SAS System for Windows.

The cluster analysis method was applied for the 9-variate response profile by fitting 36 bivariate joint models to the follow-up visits, as explained in Section 5.3.2. Generally, one names the frequency measurement ranges, obtained for each rat, as alpha, beta, delta, and gamma activity. We will reserve these Greek letters to refer to the fixed or random effects in our statistical model and use $Y_{m,il}$ for the percentage change of the m^{th} frequency measurement, obtained for rat i at time t_{il}, with t_{i1}=0 referring to the first follow-up measurement taken 45 minutes after administration of the drug.

For each longitudinal profile ($m = 1, \ldots, 9$) the following random-effects model was specified ($k = 1, \ldots, K$; $t_{i1} = 0, \ldots, t_{i8} = 7$):

$$Y_{m,il} = \alpha_m^k + a_{mi}^k + \beta_m^k t_{il} + \gamma_m^k t_{il}^2 + \varepsilon_{m,il}^k.$$

The parameters α_m^k, β_m^k, and γ_m^k describe the average quadratic evolution of outcome m over time. For each of the K clusters a separate trajectory is fitted. In this model, the cluster is incorporated as a known group effect.

The random intercept a_{mi}^k takes into account heterogeneity within cluster k and introduces correlation between the measurements of response m over time. Associations between the nine longitudinal profiles were imposed by assuming that the random intercepts a_{mi}^k and $a_{m'i}^k$ are distributed as a bivariate normal distribution with mean zero and a 2×2 covariance matrix. The errors $\varepsilon_{m,il}^k$ are zero mean normally distributed with variance $\sigma^2_{\varepsilon_m^k}$. More specifically,

$$\begin{pmatrix} a_{mi}^k \\ a_{m'i}^k \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} \;,\; \begin{pmatrix} \sigma^2_{a_m^k} & \sigma_{a_{mm'}^k} \\ \sigma_{a_{mm'}^k} & \sigma^2_{a_{m'}^k} \end{pmatrix} \right].$$

Note that the random-effect and error-component distributions can be specified to be cluster specific, allowing the associations between the nine longitudinal response profiles to differ from cluster to cluster. At maximum, 8 rats are randomized to each dose-compound combination. Cluster-specific covariance matrices for the multivariate normal distribution, estimated via the pseudo-likelihood estimation method, turned out to be singular. This may signal a perfect dependency among some variables, induced by an overspecified model fitted to a small set of data. In the application, we therefore assume the random effects and error distributions to be common to all clusters. This model results in constant correlation over time between measurements of the same response, and between measurements of different responses.

The AIC and BIC value for pseudo-likelihood estimation were obtained for models imposing $K = 1, 2, 3, 4, 5, 6, 8, 10$ clusters. To start the algorithm, the 139 rats were randomly divided into K groups. To minimize the risk of choosing unfortunate starting values the algorithm was executed 35 times for each value of K. The run resulting in the highest pseudo-likelihood value is reported here. Table 5.1 shows the pseudo-likelihood value, the AIC and BIC value, and the effective number of parameters for the models fitted to the set of 139 rats. A sample size of 139 was used to obtain the BIC. These criteria balance the increase in pseudo-likelihood value with the increase

Table 5.1: *Minus twice pseudo-log-likelihood values and information criteria for the final models resulting from the clustering algorithm for multivariate longitudinal data – EEG study.*

# clusters	-2pl	effective # of parameters	AIC	BIC
1	651260	366	651992	653065
2	640466	755	636317	639923
3	633859	1229	636317	639923
4	627028	1581	630190	634829
5	622189	2108	626406	632592
6	617623	2550	622723	630205
8	612800	3179	619157	628485
10	609677	3808	617293	628467

in model complexity. A graphical display is presented in Figure 5.1. It is seen that the BIC value gradually decreases from one to five/six components, from where the BIC appears to level off. The AIC value still decreases when adding components to the model. The AIC is known for overestimating the number of clusters in data (Hu and Xu, 2003). In what follows, we will discuss the results for the 5-component model. The pseudo-likelihood values could be duplicated for the setting with 2–5 clusters (the maximum value is obtained for respectively 20, 4, 4, and 6 runs out of the 35). For the models specifying more than five clusters the maximum pseudo-likelihood value could not be duplicated. For the setting with six clusters a number of runs result in comparable pseudo-likelihood values (minus twice pseudo-log-likelihood values: 617.64; 618.14; 618.19; and 618.25).

Figure 5.2 graphically displays the composition of the clusters. It is seen that the algorithm results in a natural grouping of the different doses and compounds included in the study. The clustering algorithm grouped all rats on the placebo dose levels of the psychoactive compounds and the majority of the rats on the active dose of Buproprion in one cluster (cluster 3: $n = 81$). It is not unexpected that the placebo dose levels of the compounds are found in one group. But it is interesting to see that the effect of Buproprion on the brain activity, as quantified by the 9 frequency measurements, cannot be distinguished from the effect of a placebo level. In the model specifying 8 clusters, the rats on Buproprion are separated from this cluster. The second to largest cluster (cluster 2: $n = 22$) contains the rats that were randomized to the highest dose level of the cholinesterase inhibitors (Donepezil, Galantamine, Tacrine) or to Nicotine. In the model specifying 8 clusters, the rats on Nicotine are separated from this cluster. The rats randomized to the active dose of Memantine or to PCP are grouped together, both psychoactive compounds are NMDA receptor antagonists (cluster 5: $n = 18$). Cluster 1 ($n = 11$) groups the rats randomized to the active dose level of the anti-epileptics (Lamotrigine, Valproate). The clustering algorithm did separate the rats on the highest dose of Amphetamine from the rest (cluster 4: $n = 7$). Indicating that the two psycho-stimulants included in the study (Amphetamine and Nicotine) exhibit different effects on the brains.

The estimated mean growth trajectories for each cluster and for each frequency measurement are displayed in Figures 5.3–5.4. For the placebo dose level cluster, i.e., cluster 3, flat wave profiles at a mean percentage change value close to zero are observed. No effect of the psychoactive compounds is noticed on the rats brain activity. Cluster 4, i.e., the highest dose level of Amphetamine, is characterized by its profiles for the δ, θ_1, and α_1 waves. Amphetamine is the only psychoactive compound in the study that results in a reduction in the δ and θ_1 frequencies and an increase in α_1 frequencies. The induced change for α_1 vanishes by the end of the study. The effect of administering the highest dose of Memantine or PCP (cluster 5) is best seen in the $\beta_1, \beta_2, \alpha_2$, and γ_2 waves. For this cluster, the reduction in frequency measurement for the β_1, β_2, and α_2 waves is larger than observed for the other clusters. The γ_2 frequencies increased. This increase is larger than seen for the other clusters and the effect is still present by the end of the study. Cholinesteras inhibitors and Nicotine at the highest dose (cluster 2) result in distinct profiles for θ_1, α_1, α_2, and γ_1. A positive percentage change is observed for θ_1 and γ_1, this increase seems to level off around time point 5. The percentage change of the α waves are negative during the study. The anti-epileptic compounds (cluster 1) behave different for the δ, β_2, θ_2, and γ_2 frequencies. The percentage change for the δ frequencies is positive and more or less constant during the study, the β_2 percentage change is also positive but keeps increasing during the study period.

These results show that rats randomized to the same compounds were nicely clustered together. The compounds constituting a cluster are known to give rise to similar effects on cortical brain activity, as measured by EEG. Thus, the results could be interpreted from a clinical point of view. The proposed algorithm is an exploratory tool that has potential value to divide a heterogeneous population in homogeneous subpopulations.

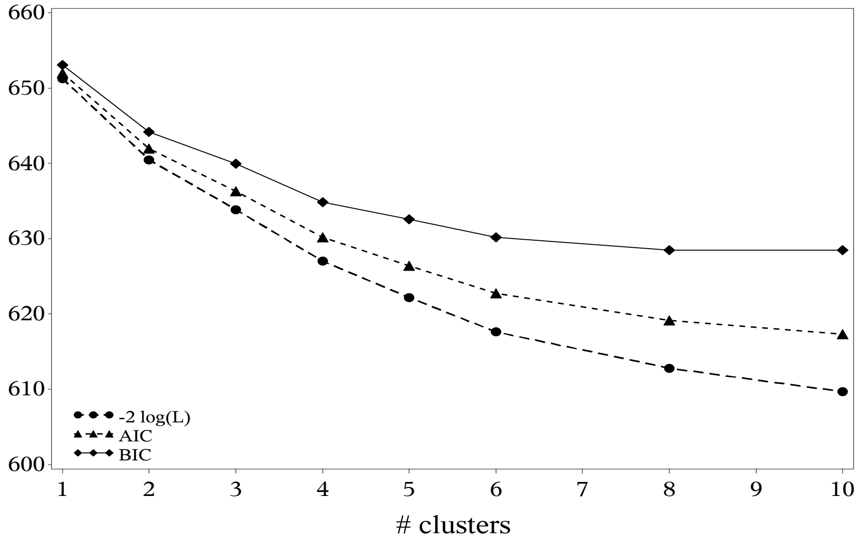


Figure 5.1: *Minus twice pseudo-log-likelihood, AIC, and BIC for the different models.*

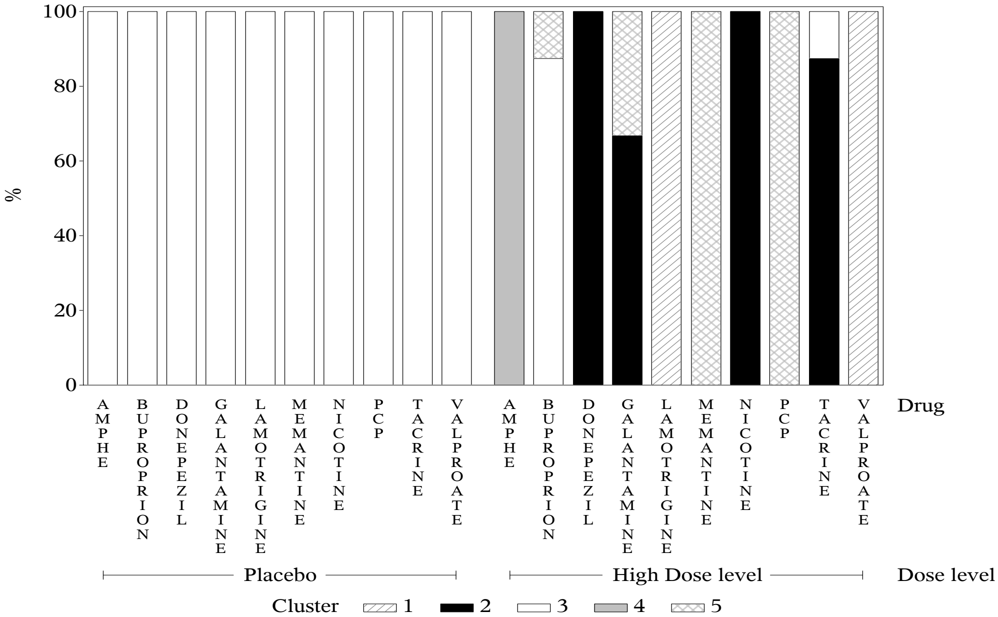


Figure 5.2: *Composition of the clusters in terms of psycho-active compound and dose – EEG study.*

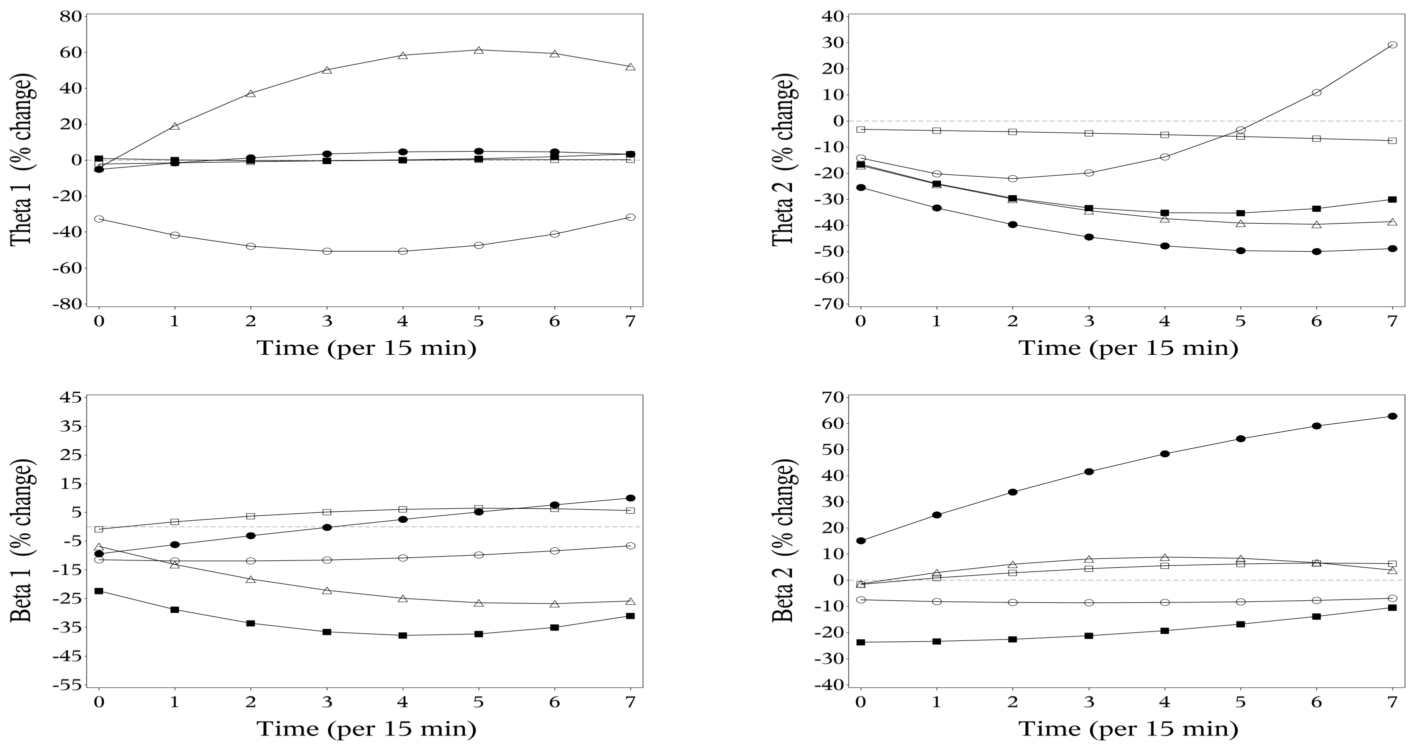


Figure 5.3: *Estimated cluster mean profiles for the 9 repeatedly measured frequencies (time = 0 is 45 min after baseline). The legend is given in Figure 5.4 – EEG study.*

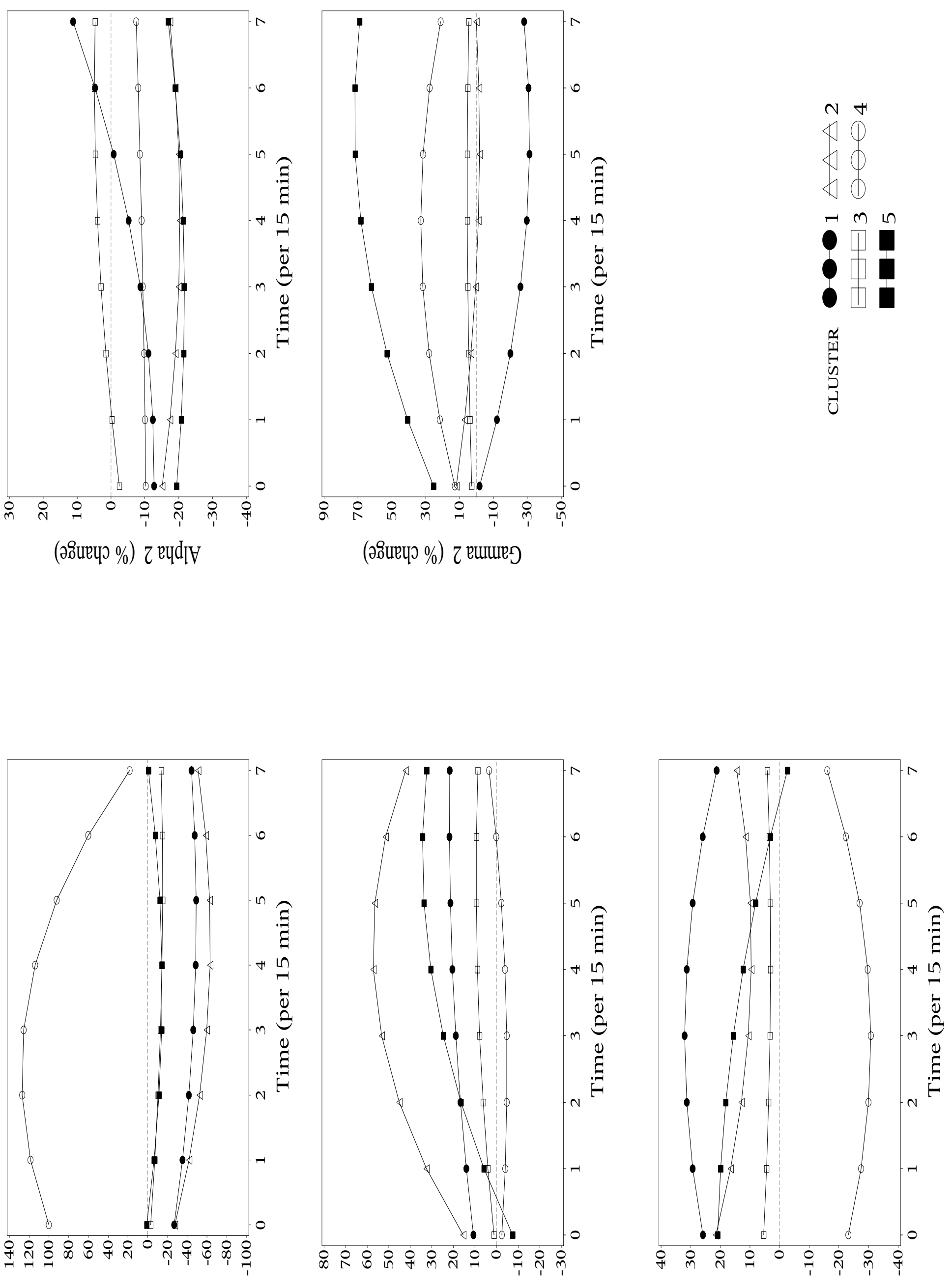


Figure 5.4: *Estimated cluster mean profiles for the 9 repeatedly measured frequencies (time = 0 is 45 min after baseline) – EEG study.*

5.5 Performance of the Algorithm: Simulations

In this section, the performance of the proposed clustering algorithm is investigated through simulations. We explore the performance by means of simulated labelled data, covering settings for separated and overlapping clusters of equal size.

5.5.1 Settings for the Simulations

Multivariate longitudinal data was generated, for four clusters (indexed by k) and 6 profiles (indexed by m), according to the following model:

$$Y_{m,il} = \alpha_m^k + a_{mi}^k + \beta_m^k t_{il} + \varepsilon_{m,il}^k, \tag{5.4}$$

with $m = 1, \ldots, 6$; $l = 0, \ldots, 6$; and $k = 1, \ldots, 4$. Values for the cluster and profile specific fixed intercepts and slopes are given in Table 5.2. The random intercepts a_{mi}^k and random error terms $\varepsilon_{m,il}^k$ were generated from multivariate normal distributions, with variance-covariance matrices common to all clusters. The random intercepts, of

Table 5.2: *Values for the fixed intercept and slope effects in Model (5.4).*

Cluster (k)	Response (m)	α_m^k	β_m^k	Cluster (k)	Response (m)	α_m^k	β_m^k
1	1	-3	3	3	1	-3	4
1	2	3	0	3	2	4	0
1	3	5	6	3	3	5	4
1	4	17	7	3	4	17	8
1	5	74	8	3	5	74	10
1	6	13	6	3	6	20	5
2	1	2	3	4	1	-3	3
2	2	3	3	4	2	-3	3
2	3	6	5	4	3	5	3
2	4	17	7	4	4	16	8.5
2	5	74	9.5	4	5	75	9
2	6	15	5	4	6	19	4

the 6 profiles, were allowed to co-vary. Their correlation matrix was specified as:

$$\begin{pmatrix} 1.00 & 0.25 & 0.10 & 0.30 & 0.20 & 0.00 \\ & 1.00 & 0.20 & 0.10 & 0.10 & 0.10 \\ & & 1.00 & 0.10 & 0.10 & 0.10 \\ & & & 1.00 & 0.20 & 0.10 \\ & & & & 1.00 & 0.10 \\ & & & & & 1.00 \end{pmatrix}$$

Data representing different degrees of overlap among the clusters, was obtained by specifying different values for the random intercept variances and residual variances, see Table 5.3. Setting 1 assumes the following values for the variance of the six random intercepts, (2.0, 0.5, 3.0, 2.0, 2.0, 1.0) and for the variance of the six residual variances the following values were specified (1.0, 0.3, 1.0, 1.0, 2.5, 1.0). Settings 2 through 9 are obtained by multiplying these variance by a factor. Table 5.3 displays the considered settings for the variance terms and the corresponding average Mahalanobis distance between the (centers of the) four clusters. The Mahalanobis distance (Mahalanobis, 1936) is a similarity measure that accounts for the variance of each variable and the covariance between variables. The distance between two observations $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, from two multivariate distributions, with covariance matrix V, is given by $(\boldsymbol{x}_1 - \boldsymbol{x}_2)V^{-1}(\boldsymbol{x}_1 - \boldsymbol{x}_2)'$. For setting 1 and setting 9, Table 5.4 displays the distances between the four clusters. The simulated data, consists of four clusters whereby cluster (1,3) and (2,4) are closer to each other and thus harder to separate, as compared to the other clusters.

For each setting, 50 data sets were generated. Equal cluster sizes were assumed. The sample size per cluster were set equal to 7, 10, 15 and 20. Figure 5.5 shows the profiles for the 6 responses, generated under setting 1 and assuming $n_k = 10$. The four clusters can not be discerned with the naked eye.

5.5.2 Results

The clustering algorithm was entertained for $k = 4$, with 15 random initial divisions and a maximum of 45 iterations. Each iteration involves 15 bi-variate mixed models. For each response, the model given in (5.4) was applied. The two random intercepts were allowed to co-vary. The variance-covariance matrices for the two random intercepts and residual errors were specified to be common to the four clusters.

The algorithm allocates each observation into one of four groups. To circumnavigate the label degeneracy all 4! permutations of the labels, assigned by the algorithm,

Table 5.3: *Settings considered in the simulations: Multiplying factors for the random intercept variances (2.0, 0.5, 3.0, 2.0, 2.0, 1.0) and residual error variances (1.0, 0.3, 1.0, 1.0, 2.5, 1.0) in Model (5.4) and average Mahalanobis distance between the four clusters.*

	Random intercept variances	Residual error variances	Average Mahalanobis distance
Setting 1	1	1	20.3
Setting 2	1	1.5	17.1
Setting 3	1	2	15.2
Setting 4	1.5	1	19.4
Setting 5	1.5	1.5	16.2
Setting 6	1.5	2	14.2
Setting 7	2	1	19.0
Setting 8	2	1.5	15.7
Setting 9	2	2	13.8

Table 5.4: *Mahalanobis distance between the centers of the 4 clusters, for settings 1 and 9, specified in Table 5.2 and Table 5.3.*

Cluster Cluster		1	2	3	4		1	2	3	4
1	Setting 1	0	24.7	11.7	25.8	Setting 9	0	16.5	8.0	18.0
2			0	24.4	13.0			0	16.4	8.3
3				0	22.2				0	15.5
4					0					0

were compared with the true label. Table 5.5 presents the results for the permutation maximally recovering the true labels.

For each data set the proportion of correctly classified observations is obtained. Table 5.5 displays the mean (and standard deviation), minimum and maximum value of the distribution of correctly classified observations. For example, for data generated under setting 2 with clusters of size 7, we conclude that on average the clustering algorithm is able to assign 84% of the observations to the proper cluster.

As is to be expected, we see that the larger the overlap between the clusters, as measured by the Mahalanobis distances between their centers, the harder it is to reveal the grouping in the data (Table 5.5). The four sub-populations are well recovered by the algorithm for the settings 1, 2, 4 and 7. Also for settings 3, 5 and 8 the proportion of correctly classified observations is reasonably good.

One would expect the performance of the algorithm to become better with increasing sample size. This is not seen in this simulation exercise. Possibly the range of sample sizes under consideration is too narrow.

In this simulation exercise the clustering algorithm was entertained for k=4 only, i.e. the correct number of clusters. Simulations to see how the clustering algorithm performs in choosing the correct number of clusters when k is ranged from 3—5 are ongoing; but preliminary results are promising.

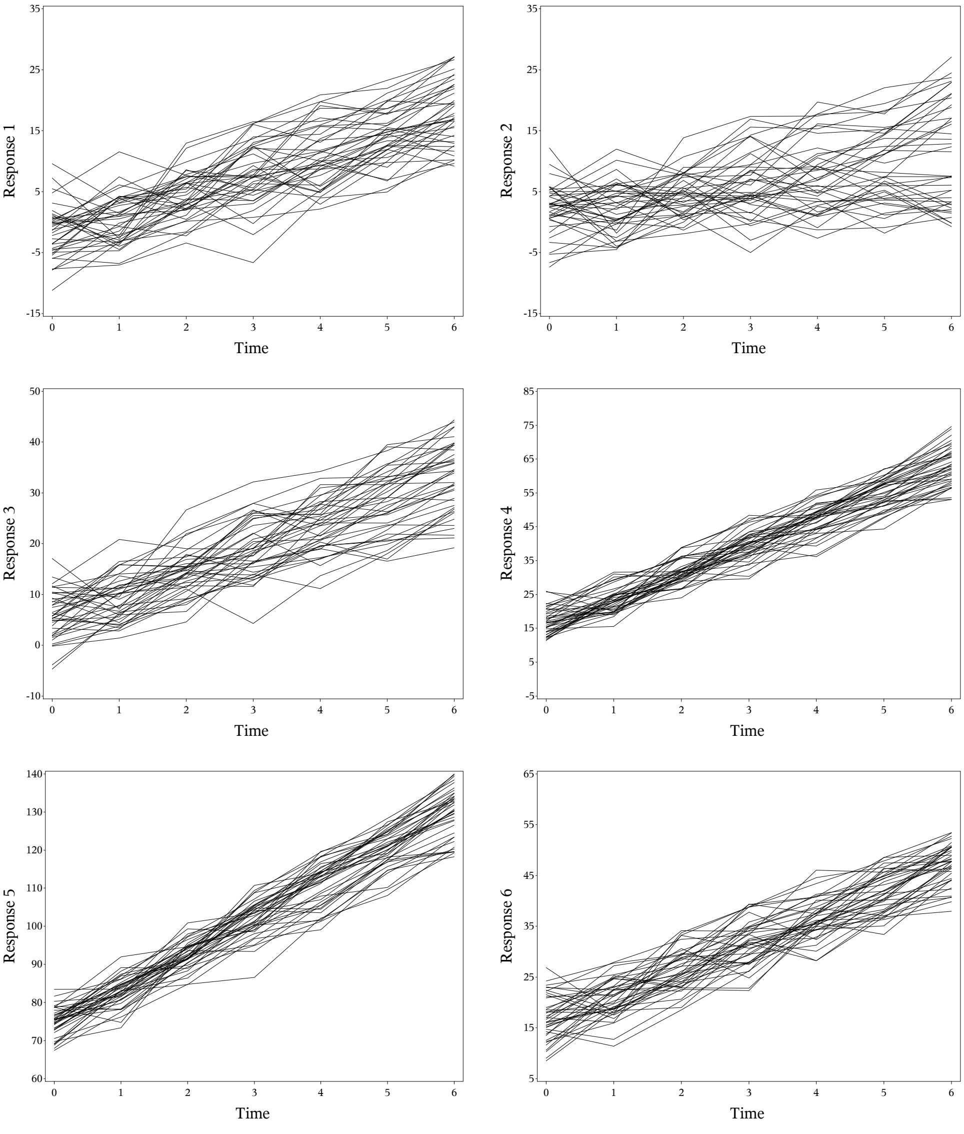


Figure 5.5: *Multivariate longitudinal data for four clusters, generated under setting 1 (Tables 5.2 and 5.3) and equal cluster size (n_k=10).*

Table 5.5: *Proportion of correctly classified observations (average over 50 simulations (standard deviation), minimum (min) and maximum value (max)) under the settings specified in Table 5.2 and Table 5.3– Simulated Data.*

Setting	Average Mahalanobis	$n_k = 7$			$n_k = 10$			$n_k = 15$		
	Distance	mean (std)	min	max	mean (std)	min	max	mean (std)	min	max
1	20.3	99 (4)	79	100	100 (1)	93	100	100 (1)	98	100
2	17.1	84 (15)	46	100	81 (14)	55	100	89 (14)	47	100
3	15.2	65 (14)	43	93	61 (15)	35	98	56 (12)	30	82
4	19.4	97 (6)	75	100	99 (4)	75	100	100 (1)	95	100
5	16.2	75 (14)	46	100	75 (16)	43	100	76 (16)	47	100
6	14.2	59 (11)	39	86	53 (12)	40	93	50 (9)	32	70
7	19.0	96 (8)	71	100	98 (5)	78	100	100 (1)	95	100
8	15.7	71 (3)	43	96	70 (14)	48	100	68 (14)	45	100
9	13.8	54 (9)	39	75	54 (9)	35	85	48 (9)	32	77

Table 5.6: *Average computing time (minutes) per simulated data set with n_k=10 (Dell latitude with 16 GB of RAM and 2.70GHz CPU).*

	Time
Setting 1	39
Setting 3	70
Setting 4	60
Setting 5	36

5.6 Discussion

In this chapter, we presented an algorithm to reveal clusters in the setting of multivariate repeated data. The algorithm mimics a K-means algorithm. The means of the K clusters are obtained via bivariate joint models for the repeated responses. An individuals' pseudo-likelihood contribution is used as the criterion to classify an individual into a cluster.

The longitudinally measured wave responses in the EEG study were all continuous and recorded at the same time points. The proposed algorithm is, however, not confined to this type of data. The clustering algorithm breaks down to repeatedly fitting bivariate mixed models. As such, it benefits from the attractive feature of a multivariate mixed model that different responses (e.g., binary and continuous) can be combined, and can be applied to sequential responses. Study design features, such as, for example, blocking, can be incorporated in the clustering algorithm, by introducing another random effect in the bivariate model.

The primary practical limitation of the algorithm is computation time. Given that the algorithm is iterative in nature, computation time increases with increasing complexity of the data set; i.e., increasing number of individuals, number of repeated outcome measures and number of specified clusters. Table 5.6 presents average computation time per simulated data set for some of the settings presented in Table 5.3.

The BIC criterion was used to choose the optimal number of clusters. The bootstrap likelihood-ratio test is an alternative (McLachlan, 1987), but was not implemented because of computing time.

In general, cluster analyses are sensitive to starting values and to outliers. This is not different for the proposed algorithm. Running the algorithm for a number of times, each time starting from a different random division in K initial groups, allows to evaluate sensitivity to starting values. The maximum pseudo-likelihood value is harder to duplicate when more clusters are specified, again increasing computing time.

The issue of outliers is harder to investigate. Outliers induce clusters with a few units, centered around the outlier. Bivariate mixed models, as many other statistical tools, easily run into problems when applied to small and sparse data sets. Chapter 8 discusses local influence as a tool to investigate the effect of one data point on the results from a model-based clustering.

In this contribution, we did not investigate the effect of model mis-specification — such as misspecified average time evolutions, error distribution, random terms in the model, etc. — on the number and the constitution of the discovered clusters.

Chapter 6

Clustering Multiply Imputed Multivariate High-Dimensional Longitudinal Profiles

6.1 Introduction

A lot of methodological work has been done to extend cluster analysis to repeated and multivariate data structures. When analyzing repeated measurements data, individual differences in evolution are generally captured by random effects, often via linear mixed models (Laird and Ware, 1982; Verbeke and Molenberghs, 2000). Individual differences can also be described by latent trajectory classes (Land and Nagin, 1996; Nagin and Land, 1993, Nagin, 1999; Nagin and Tremblay, 2001) or by growth-mixture models (Muthén and Shedden, 1999; Muthén and Muthén, 1998–2012).

When, for each patient, more than a single outcome is measured over time, a multivariate set of longitudinal profiles is obtained. Interest could be in finding subgroups of patients that are similar in their evolution over time for the various repeated sequences. An application of a clustering procedure for this type of data can be found in Chapter 5 of this dissertation.

Nowadays, data complexity and dimensionality are enhanced by novel data collection techniques. These techniques permit observations to be densely sampled over a continuum, usually time as in the TEMA-HF1. The data then reflect the influence of a (set of) smooth function(s) underlying and generating the observations. Often, the evolutions are not easily described by a mathematical formula. The dependencies between these so-called functional data curves can be analyzed by methods from the functional data analyses framework (Ramsay and Silverman, 2002, 2005; Rice and Silverman 1991; Rice 2004; Ferraty and Vieu, 2006). As usual, observed heterogeneity can be corrected for via explanatory variables. Unobserved sources of population heterogeneity can be investigated via cluster analyses, where the main objective is to classify patients into homogenous groups. However, clustering functional data in general requires first a reduction of the high dimensionality of the data (Abraham et al., 2003; Tarpey and Kinateder, 2003).

Data reduction techniques are hampered by missing values—an issue often intertwined with longitudinal data. In the regression context, a multiple imputation procedure (Rubin, 1987; Schafer, 1997; Carpenter and Kenward, 2013) can be applied to quantify the extra uncertainty in estimators of population parameters due to the missing values. Applying a cluster algorithm on the imputed data sets of the TEMA-HF1 results in multiple partitionings of the patients. It is however not so clear how uncertainty due to the imputation process needs to be reflected in the final result. Basagaña et al. (2013) present a framework for multiple imputation in cluster analysis. They suggest ways to report how the final number of clusters, the result of a variable selection procedure and the assignment of individuals to clusters is affected by the missing values. Their final decision on a patient's cluster membership is based on a majority vote. We propose to approach the problem as a combinatorial optimization problem to summarize the cluster ensemble into a single consolidated clustering and at the same time measure the missing data influence in the cluster analyses.

In this chapter, we assemble techniques from functional data analysis, missing data analysis and ensemble clustering to reveal groups of similar patients when facing high-dimensional multivariate data with missing observations. The final data analysis brings together a number of statistical techniques that are briefly introduced: the idea of multiple imputation is given in Section 6.2, the concept of functional data is briefly introduced in Section 6.3, multivariate functional data and functional principal component analyses, as a data reduction technique, are described in Section 6.4. Section 6.5 describes how the results of a principal component analysis are used to obtain a probability density for the functional data. A model-based cluster method for functional data is given in Section 6.6. The ensemble method for clustering is the topic of Section 6.7. The various steps of the proposed procedure are graphically displayed in Figure 6.1.

The ability of the procedure, to reveal latent structures in the data in the presence of missingness, is investigated by means of a small simulation exercise in Section 6.8. The potential of the method was brought out on the heart failure data, introduced in Section 3.4. In the TEMA-HF1 study, 80 chronic heart failure patients were followed up intensively for 6 months. A telemonitoring device daily transferred data on body weight, blood pressure (systolic and diastolic), and heart rate. Although alerts were sent out when the longitudinal measurements were missing for two consecutive days, quite some missingness is present in the data. Section 6.9 applies the outlined procedure and presents the results.

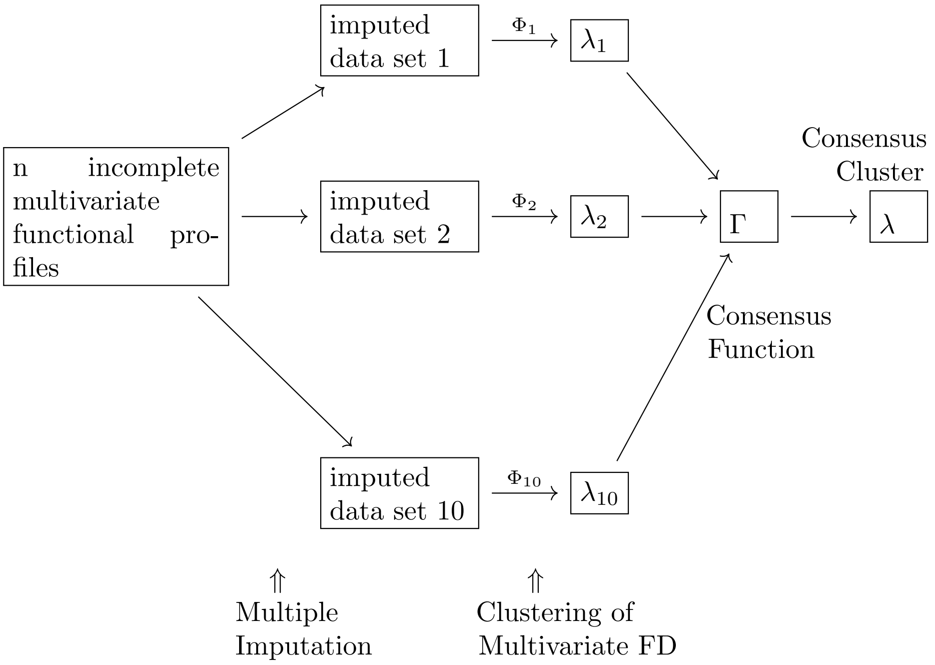


Figure 6.1: *Steps of the proposed procedure to cluster multiply imputed multivariate high-dimensional longitudinal profiles.*

6.2 Multiple Imputation

Data reduction techniques, like principal component analysis, require rectangular data structures. Often, records with missing values are discarded in the analyses. To circumnavigate this problem, multiple imputation was used to create a set of complete/rectangular data sets.

Multiple imputation is a popular tool for dealing with data when they are only partially observed (Rubin, 1987; Schafer, 1997; Carpenter and Kenward, 2013; Molenberghs and Kenward, 2007). The idea is to use the observed information to impute sensible values for the missing ones. To reflect the uncertainty in this prediction, missing values are imputed multiple times. Multiple imputation is appealing because it results in complete data sets, that can be analyzed with standard statistical techniques. Two routes can be followed: multivariate or fully conditionally specified imputation (Schafer, 1997; Little and Rubin, 2002; van Buuren et al., 1999; Raghunathan et al., 2001). The multivariate approach assumes that all variables in the imputation model jointly follow a multivariate normal distribution. Fully conditional imputation or chained equations, specify for each variable in turn a conditional distribution, conditional on all other variables in the model. This approach does not postulate multivariate normality. Both approaches assume the missing data to be missing at random (MAR, Little and Rubin, 2002). Under the MAR assumption, the probability that an observation is missing, is driven only by the observed data, implying that no extra information is contained in the missing part of the data.

Standard imputation models applied to longitudinal data can lead to absurd results (Honaker and King, 2010), e.g., imputations falling far from previous and subsequent observations, or imputations that are very implausible on the basis of common sense. Honaker and King (2010) developed the software package AMELIA that facilitates imputation of (among others) smooth time-series patterns. Before executing the imputations, the data is supplemented with smooth basis functions. To allow subjects to exhibit different evolutions over time, the imputation model includes interactions of the basis functions and the subject indicator. On top of that, lead and lag variables can be specified for the imputation model. Clearly the number of parameters in the imputation model rapidly increases, resulting in computational difficulties in the EM algorithm used to obtain the posterior density of the data and taking random draws from it (i.e., draws of μ and Σ). AMELIA implements a so-called EMB algorithm (Basford, 1994). This algorithm combines the classical EM procedure with a bootstrap approach to take draws from the posterior. For each draw, the data is bootstrapped to simulate estimation uncertainty. Next, the EM algorithm is used to find the mode of posterior for the bootstrapped data.

6.3 Functional Data

Functional data analysis (FDA) can be seen as an extension of classical multivariate methods where data are not vectors but rather functions or curves (Ramsay and Silverman, 2002; Ramsay and Silverman, 2005; Rice and Silverman 1991; Rice 2004; Ferraty and Vieu, 2006). Functional data describe a process that changes smoothly and continuously over a domain. Often, this domain is time, resulting in repeated measurement data, but it can be anything, such as, for example space or energy. Data in many fields result from a process that is functional. Ramsay and Silverman (2002, 2005) provide many examples: daily weather data about temperature and precipitation; human growth data for height and weight, yearly non-durable good index, hip and knee angles observed in a gait cycle, etc.

In functional data analysis, the existence of a smooth function x is assumed. This function gives rise to data y_j, superimposed by measurement error ε_j, usually observed at discrete time points t_j, such that $y_j = x(t_j) + \varepsilon_j$. Although the curves are sampled for a finite set of time-points, the observations are supposed to belong to an infinite-dimensional space. The functional form of the data is often reconstructed from the discrete observations by assuming that the finite-dimensional space is spanned by a basis of functions. Consider a basis $\phi = \{\phi_1, \ldots, \phi_G\}$ and represent the functional data $x_i(t)$, for patient i, by a linear combination of the G basis functions:

$$x_i(t) = \sum_{g=1}^{G} a_{ig}\phi_g(t).$$

The basis coefficients are estimated so that the constructed curve optimally fits the data for a certain degree of smoothing. The number of basis functions can be chosen in terms of a bias-variance trade-off.

6.4 Principal Component Analysis of Functional Data

Principal component analysis for functional data is similar to principal component analysis for multivariate data (Hotelling, 1933). We therefore precede the introduction of PCA for functional data with a brief resume of PCA for multivariate data.

6.4.1 PCA for Multivariate Data

For high-dimensional multivariate data, dimension reduction is usually performed prior to applying a statistical procedure in order to avoid the effects of the curse of dimensionality. Principal component analysis reduces the dimensionality by linear mapping of the data (Hotelling, 1933). The linear combinations are chosen to highlight types of variation strongly represented in the data. The principal components, in the multivariate situation when data for N subjects is obtained for p variables, are defined as:

$$f_{im} = \sum_{j=1}^{p} \beta_{jm} x_{ij}, \qquad i = 1, \ldots, N, \tag{6.1}$$

with β_{jm} a set of orthogonal weights that maximize the variation in the f_{im}. The solutions to this maximization problem are given by the eigenvectors of the eigenequation $\boldsymbol{V\beta} = \lambda\boldsymbol{\beta}$, with $\boldsymbol{V}$ the $p \times p$ sample variance-covariance matrix. A sequence of eigenvalue-eigenvector pairs $(\lambda_m, \boldsymbol{\beta_m})$ satisfy this eigenequation, with $\boldsymbol{\beta_m}$ orthogonal.

6.4.2 PCA for Functional Data

6.4.2.1 Univariate Functional Data

The dimensionality for multivariate data is given by the discrete index j in (6.1), for functional data a continuous index s is taking over this role. The principal component scores, for univariate functional data, are obtained as the inner product of two functions, the weight function and the data function (Ramsay and Silverman, 2005; Besse and Ramsay, 1986; Castro et al., 1986; Dauxois et al., 1982; Cardot, 2000; Hall and Hosseini-Nasab, 2006; Jones and Rice, 1992):

$$f_i = \int \beta(s) x_i(s) ds, \qquad i = 1, \ldots, N.$$

A sequence of weight functions $\beta_m(s)$ is chosen such that they define the most important modes of variation in the curves, conditional on the weights to be orthonormal. So,

1. $\frac{1}{N}(\int \beta_m x_i)^2$ is maximal,
2. $\|\beta_m^2\| = \int (\beta_m)^2 = 1$,
3. $\int \beta_m \beta_n = 0, \;\; n \neq m$.

Functional principal component analysis is also tantamount to solving an eigenequation. Define the sample variance-covariance function as $v(s,t) = \frac{1}{N}\sum_{i=1}^{N} x_i(s) x_i(t)$.

Then V, in the functional version of PCA, is a variance operator and transforms a function β as $V\beta(.) = \int v(.,t)\beta(t)dt$. The eigenequation can then be expressed as:

$$V\beta(s) = \int v(s,t)\beta(t)dt = \lambda\beta(s), \tag{6.2}$$

where β are eigenfunctions now instead of vectors. Solutions to this continuous functional eigenanalysis problem (6.2) can be obtained by approximating the problem as a matrix eigenanalysis task. Theretofore the data (a set of N curves $x_i(t)$) and eigenfunctions are presented as linear combinations of the basis $\boldsymbol{\phi}$: $x_i(t) = \sum_{g=1}^{G} a_{ig}\phi_g(t)$ and $\beta(s) = \sum_{g=1}^{G} b_g\phi_g(s)$.

In compact matrix notation, the expansion of the N curves can be expressed as:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\phi},$$

where $\boldsymbol{A}$ is the $N \times K$ coefficient matrix; $\boldsymbol{x}$ and $\boldsymbol{\phi}$ are vector-valued functions with components respectively $(x_1, \ldots, x_N)$ and $(\phi_1, \ldots, \phi_G)$. In a similar way the eigenfunctions and variance covariance function can be expressed in matrix notation:

$$\boldsymbol{\beta} = \boldsymbol{b}^{'}\boldsymbol{\phi},$$

$$v(s,t) = N^{-1}\boldsymbol{\phi}(s)^{'}\boldsymbol{A}^{'}\boldsymbol{A}\boldsymbol{\phi}(t).$$

with $\boldsymbol{\beta}$ a vector-valued function with components $(\beta_1, \ldots, \beta_M)$.

The eigenequation (6.2) can then be written down as:

$$\begin{array}{lcl} \lambda\beta(s) & = & \int v(s,t)\beta(t)dt \\ \lambda\boldsymbol{\phi}(s)^{'}\boldsymbol{b} & = & \int N^{-1}\boldsymbol{\phi}(s)^{'}\boldsymbol{A}^{'}\boldsymbol{A}\boldsymbol{\phi}(t)\boldsymbol{\phi}(t)^{'}\boldsymbol{b}dt \\ & = & \boldsymbol{\phi}(s)^{'}N^{-1}\boldsymbol{A}^{'}\boldsymbol{A}\boldsymbol{W}\boldsymbol{b} \end{array} \tag{6.3}$$

with $\boldsymbol{W}$ the G symmetric matrix containing the innerproducts of the basisfunctions, i.e., $w_{g_1,g_2} = \int \phi_{g_1}\phi_{g_2}$. Equation (6.3) will hold for all s only if

$$N^{-1}\boldsymbol{A}^{'}\boldsymbol{A}\boldsymbol{W}\boldsymbol{b} = \lambda\boldsymbol{b}. \tag{6.4}$$

The condition of orthogonality of the eigenfunctions β implies that (6.4) is to be solved with the following constraints: $\boldsymbol{b}\boldsymbol{W}\boldsymbol{b} = 1$ and $\boldsymbol{b_1}\boldsymbol{W}\boldsymbol{b_2} = 0$. For an orthonormal basis $\boldsymbol{W} = \boldsymbol{I}$, the functional PCA problems reduces to a standard multivariate PCA of the matrix $N^{-1}\boldsymbol{A}^{'}\boldsymbol{A}$.

6.4.2.2 Multivariate Functional Data

When extending functional PCA to M-variate functional data, the weight functions become M-vector functions $\beta = (\beta^1, \dots \beta^M)'$, with β^l depicting the variation in the l^{th} dimension (Berrendero et al., 2011; Ramsay and Silverman, 2005). The principal component scores are again linear combinations of the data:

$$f_i = \sum_{m=1}^{M} \int \boldsymbol{\beta}^m(s) x_i^m(s) ds,$$

where the weight functions $\boldsymbol{\beta}^m$ are solutions of an eigenequation system $\boldsymbol{V\beta} = \lambda\boldsymbol{\beta}$. $\boldsymbol{V}$ is the covariance operator as defined before, $v_{ii}(s,t)$ is the covariance operator for the i^{th} functional data dimension and $v_{ij}(s,t)$ the cross-covariance operator between dimensions i and j. The eigenequation translates to a system of equations:

$$\begin{cases} v_{11}\boldsymbol{\beta^1} + v_{12}\boldsymbol{\beta^2} + \ldots + v_{1m}\boldsymbol{\beta^m} = \lambda\boldsymbol{\beta^1}, \\ v_{21}\boldsymbol{\beta^1} + v_{22}\boldsymbol{\beta^2} + \ldots + v_{2m}\boldsymbol{\beta^m} = \lambda\boldsymbol{\beta^2}, \\ \vdots \\ v_{m1}\boldsymbol{\beta^1} + v_{m2}\boldsymbol{\beta^2} + \ldots + v_{mm}\boldsymbol{\beta^m} = \lambda\boldsymbol{\beta^m}. \end{cases}$$

In practice, a standard principal component analysis is carried out on a vector Z_i concatenating all data functions of patient i.

6.5 Density for Functional Data

Model-based clustering identifies homogenous subgroups of patients using a mixture model for the density function of the data. Delaigle and Hall (2010) use the Karhunen-Loève expansion to introduce the notion of a probability density for functional data.

The basis, yielding a minimum value for the total mean squared error when decomposing a stochastic process $\boldsymbol{X}(t)$ as an infinite linear combination, is the set of orthogonal eigenfunctions of the process itself:

$$\boldsymbol{X}(t) = \boldsymbol{\mu}(t) + \sum_{j=1}^{\infty} f_j \beta_j(t).$$

If $\boldsymbol{\mu}(t) = 0$, i.e., for a centered process, the composition is referred to as the Karhunen-Loève expansion (Karhunen, 1947; Loève, 1978). The basis coefficients are random variables, in contrast to the coefficients resulting from, for example, a polynomial basis. The random variables f_j are uncorrelated, have zero mean and variance λ_j.

We denote the distribution of f_j by $\mathfrak{f}_j$. The variables f_j follow a Gaussian distribution and are stochastically independent for a Gaussian process.

Let $p(\boldsymbol{x}|h) = P(\|\boldsymbol{X} - \boldsymbol{x}\| \leq h)$ for $h > 0$ and $\|\boldsymbol{X} - \boldsymbol{x}\|$ the L_2-distance between $\boldsymbol{X}$ and $\boldsymbol{x}$. Then, $p(\boldsymbol{x}|h)$ is the probability that $\boldsymbol{X}$ belongs to a ball of radius h centered at $\boldsymbol{x}$. Delaigle and Hall (2010) show that this probability can be written as a product of the densities $\mathfrak{f}_j$, corresponding to the largest eigenvalues:

$$\log[p(\boldsymbol{x}|h)] = C_1(r, \theta) + \sum_{j=1}^{r} \log \mathfrak{f}_j(f_j) + O(r), \tag{6.5}$$

where $\mathfrak{f}_j(f_j) = \mathfrak{f}_j(f_j(\boldsymbol{x}))$ is the density of the j principal component score evaluated for the j component score for $\boldsymbol{x}$; $r = r(h)$ diverges to infinity as h decreases to zero, and C_1 depends on h and on the infinite eigenvalue sequence, θ. Based on (6.5), a natural surrogate for the log density of functional data is provided by the average of log densities of the r largest principal components. This log-density $l(\boldsymbol{x}|r) = r^{-1} \sum_{j=1}^{r} \log \mathfrak{f}_j\, (f_j)$ captures variation with $\boldsymbol{x}$ up to order r.

6.6 Clustering of Functional Data

An excellent review of approaches to clustering functional data is presented by Jacques and Preda (2014). They classify the approaches into four categories: raw-data clustering, two-stage procedures (Abraham et al., 2003; Peng and Müller, 2008), model-based procedures (Bouveyron and Jacques, 2011; Jacques and Preda, 2012) and nonparametric techniques for clustering functional data (Ferraty and Vieu, 2006; Ieva et al., 2012; Yamamoto, 2012). Techniques not using the functional form of the data, but applying cluster algorithms for high-dimensional vector data, directly on the info observed at the discrete time points, are labeled raw-data clustering. Clearly these techniques do not reckon with time dependencies present in the data. In two-stage procedures, the dimension of the data is first reduced, e.g., by means of a functional principal component analyses, in the second step a cluster algorithm is applied on the newly constructed data summaries. The functional nature of the data is dealt with in the first step. Model-based clustering techniques assume a probability distribution for the data, and perform the dimension reduction and clustering simultaneously. Both the (second step of the) two-stage and model-based procedures can be applied on the basis coefficients when approximating the curves or on the principal component scores. Nonparametric clustering of functional data comes down to executing a classical clustering algorithm for finite-dimensional data, on distances or similarities between curves. We opt for model-based clustering, using principal components. This procedure tackles the functional nature of the data, simultaneously performs a data reduction and cluster exercise, while at the same time allowing for complex covariance structures in the multivariate longitudinal profiles.

Jacques and Preda use the approximation of the probability density for functional random variables to fit a parametric mixture model to univariate functional data (Jacques and Preda, 2012) and to multivariate functional data (Jacques and Preda, 2014). We briefly summarize the different steps of their algorithm.

Assume the existence of a latent group indicator $Z_i = (Z_i^1, \cdots, Z_i^K)$ for K clusters. For subject i, $Z_i^g = 1$ if its curves $\boldsymbol{x}_i$ belong to group g, 0 otherwise. Let Z_i have a multinomial distribution with mixing proportions $\pi_1, \ldots, \pi_K$ ($\sum_{k=1}^{K} \pi_k = 1$). Under these assumptions, the unconditional approximated density of $\boldsymbol{X}$ is equal to

$$\mathfrak{f}_{\mathbf{X}}^{(q)}(\boldsymbol{x};\theta) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{q_k} \mathfrak{f}_{j,k}(f_{j,k}(\boldsymbol{x});\lambda_{j,k}).$$

When $\boldsymbol{X}$ is a Gaussian process, the $\mathfrak{f}_{j,}$ are Gaussian. The parameters $\theta = \{(\pi_k, \lambda_{1,k}, \cdots, \lambda_{q_k,k})_{1 \leq k \leq K}\}$ and $q = (q_1, \cdots, q_K)$ are estimated by maximizing the pseudo completed log-likelihood via an iterative EM algorithm:

$$\begin{aligned} & L^{(q)}(\theta; \{X_1, \cdots, X_N\}, \{Z_1, \cdots, Z_N\}) \\ = & \sum_{i=1}^{N} \sum_{k=1}^{K} Z_i^k \left(\log(\pi_k) + \sum_{j=1}^{q_k} \log(\mathfrak{f}_{j,k}(f_{i,j,k}(\boldsymbol{x}_i)) \right), \end{aligned}$$

where $f_{i,j,k}$ is the j^{th} principal component of curves $\boldsymbol{x}_i$ belonging to group k.

At iteration h, the E-step of the EM-algorithm evaluates the conditional expectation of the pseudo completed log-likelihood, with respect to unknown Z_i^k, given the observed data and current parameter estimates, $\Theta(\theta, \theta^{(h)})$.

$$\begin{aligned} \Theta(\theta, \theta^{(h)}) &= E_{\theta^{(h)}}[L^{(q)}(\theta; \boldsymbol{X}, \boldsymbol{Z}) | \boldsymbol{X} = \boldsymbol{x}] \\ &= \textstyle\sum_{i=1}^{N} \sum_{k=1}^{K} E_{\theta^{(h)}}[Z_i^k | \boldsymbol{X} = \boldsymbol{x}] \\ &\quad \times \left[\log(\pi_k) + \textstyle\sum_{j=1}^{q_k} \log(\mathfrak{f}_{j,k}(f_{i,j,k}(\boldsymbol{x_i}; \lambda_{j,k})))\right] \\ &\simeq \textstyle\sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\pi_k \prod_{j=1}^{q_k} \mathfrak{f}_{j,k}(f_{i,j,k}(\boldsymbol{x_i});\lambda_{j,k})}{\sum_{k=1}^{K} \pi_k \prod_{j=1}^{q_k} \mathfrak{f}_{j,k}(f_{i,j,k}(\boldsymbol{x_i});\lambda_{j,k})} \\ &\quad \times \left[\log(\pi_k) + \textstyle\sum_{j=1}^{q_k} \log(\mathfrak{f}_{j,k}(f_{i,j,k}(\boldsymbol{x_i}; \lambda_{j,k})))\right] \end{aligned} \tag{6.6}$$

where $\mathfrak{f}_{j,k}(f_{i,j,k}(\boldsymbol{x_i}); \lambda_{j,k})$ is the value of $\mathfrak{f}_{j,k}$ for $\boldsymbol{X}_i = \boldsymbol{x}_i$.

The M-step maximizes this conditional expectation with respect to θ. Before executing the M-step, Jacques and Preda (2014) update the group-specific principal components $f_{j,k}$. For this purpose, a weighted principal component analyses is fitted, with weights $E_{\theta^{(h)}}[Z_i^k|\boldsymbol{X} = \boldsymbol{x}]$. Furthermore, the class-specific dimensions q_k are selected by means of the scree-test of Cattell (Cattell, 1966). After these intermediate steps, the M-step maximizes $\Theta(\theta, \theta^{(h)})$ with respect to θ.

Jacques and Preda note that this procedure does not guarantee an increase in the pseudo-likelihood between two iterations. The reason for this is that an approximation to the density of functional data is used. They advise to pre-run the algorithm a couple of times with different (random) starting values, using a small number of iterations. The best solution among these is then to be used as the starting point for the algorithm with a large number of iterations (Biernacki, 2004). This empirical strategy increases the chance of convergence to a local maximum.

6.7 Consensus Clustering

Cluster ensembles are collections of individual solutions to a given clustering problem (Strehl and Ghosh, 2002). Let $\mathfrak{X} = \{x_1, x_2, \ldots, x_N\}$ denote a set of objects/samples, where each x_i is some p-dimensional data vector. A partitioning of the N objects into K clusters can be represented as a set of K sets of objects $(\mathcal{C}_k|k = 1, \ldots K)$ or as a label vector $\delta \in \mathbb{N}^N$. The clustering algorithm (function) to obtain this label vector is called a clusterer Φ.

The resulting partition can be a soft (fuzzy) or a hard (crisp) partition. If all non-negative numbers μ_{ik}, quantifying the probability that object i belongs to class k, with $\sum_{k=1}^{K} \mu_{ik} = 1$, are in {0,1} the obtained partition is referred to as a hard partition. Otherwise the partition is soft. The matrix M containing the μ_{ik}'s is called the membership matrix, with rows corresponding to objects and columns to the classes. The co-membership matrix $C(M) = MM^{'}$ has entries $c_{ij} = 1$ if objects i and j are in the same class in a hard partition, and 0 otherwise.

The label vector δ containing the class identification numbers is not unique. The class labels can be permuted arbitrarily without changing the underlying partition. For a partition in K clusters there are $K!$ equivalent representations. The canonical form is the representation that satisfies the following two constraints (i) λ_1 =1 (the first object's label is cluster 1); (ii) for all $i = 1, \ldots, N-1 : \lambda_{i+1} \leq \max_{j=1,\ldots,i}(\lambda_j) + 1$ (the cluster label λ_{i+1} of any successive object has a label that occurred before or a label that is one greater than the highest label so far). In terms of the membership matrix M, permuting class IDs means replacing M by $M\Pi$, where Π is a suitable permutation.

Results obtained from applying different clusterers Φ on a dataset can be quite different but all equally plausible. The problem of combining multiple partitionings into a single clustering is referred to as cluster ensembles (Strehl and Ghosh, 2002).

Intuitively, the final consensus is the partition of the N objects that shares most information with the original clusterings. It is assumed that the consensus cluster is less likely to be biased towards the models (Φ) used in the separate analyses and more likely to reflect the underlying structure of the data. Day (1986) and Leclerc (1998) studied the consensus of hard partitions; fuzzy consensus clustering has been investigated by Gordon and Vichi (2001).

Consensus clustering synthesizes the information in the elements of a cluster ensemble into a single clustering, often by minimizing a criterion function measuring how (dis)similar consensus candidates are from the ensemble (the so-called optimization approach to consensus clustering). Since there is no relation between the labels assigned to object i by a clusterer (Φ_1) and another clusterer (Φ_2) the cluster ensemble problem is more difficult than a classifier ensemble problem. This label correspondence issue is the main problem that has to be dealt with when clustering ensembles. The problem can be solved via the Hungarian method (Kuhn, 1955). An additional issue is that the number and shape of the input clusters may be different and that the optimal final number of clusters is often not known in advance.

To state the cluster ensemble as a problem of mapping a set of r labelings, $\delta^{(1,\cdots,r)}$, to a single consensus clustering, δ, a consensus function Δ, $\mathbb{N}^{N\times r} \to \mathbb{N}^{N}$ is needed $\Delta : \{\delta^{(q)}|q \in \{1, \cdots, r\}\} \to \delta$. An estimate $\widehat{\delta}$ is often obtained by maximizing (minimizing) a criterion/objective function measuring how (dis)similar consensus candidates are from the ensemble. Measures for dissimilarity and similarity are key ingredient to clustering (ensembles). Let d be a suitable dissimilarity measure; most popular criterion functions are of the form

$$L(\delta) = \sum w_b d(\delta^b, \delta)^p, \tag{6.7}$$

where w_b is a weight given to element δ^b of the ensemble, and $p \geq 1$. If $p = 1$ the consensus solution is called a median of the ensemble, while $p = 2$ gives least squares consensus partitions (Gordon, 1999). A variety of methods are available to minimize criteria of this form; fixed-point algorithms for soft Euclidean and Manhattan consensus partitions, greedy algorithms, SUMT algorithms, and exact solvers (Hornik, 2005). A multiplicity of (dis)similarity measures are described in the literature. Among the ones commonly used are the Euclidean and Manhattan dissimilarity of the memberships (Dimitriadou, Weingessel and Hornik 2002), the Rand index (Rand 1971, Gordon 1999), Normalized Mutual Information (Strehl and Ghosh 2002), the Katz-

Powell index (Katz and Powell 1953), the Jaccard index, etc. The maximization in (6.7) ranges over all possible K-partitions (Strehl and Ghosh, 2002). An exhaustive search over all possible clusterings with K labels for the one with the maximum criterion is in general not possible. Dimitriadon, Weingessel and Hornik (2002) have shown that optimal matching can be determined very efficiently when agreement is expressed as Euclidean partition dissimilarity.

To evaluate the reliability of a partition of a data set, the fuzziness in the partitioning can be investigated. In fuzzy clustering, a data point does not completely belong to just one cluster but has a probability of belonging to each cluster. Denote by μ_{ik} the probability that data point i belongs to cluster k. The uncertainty of a fuzzy partition can be quantified via the the Partition Coefficient, $\sum_{i,k} \mu_{i,k}^2$, and the Partition Entropy, $\sum_{i,k} H(\mu_{i,k})$, where $H(u) = u\text{log}(u) - (1-u)\text{log}(1-u)$ (Bezdek 1981).

6.8 Simulated Data for Multivariate Functional Data with Missing Observations

The described approach relies on the approximation of the notion of a probability density for functional random variables by means of the Karhunen-Loève expansion. Jacques and Preda (2012, 2014) have illustrated the efficiency of their model-based clustering algorithm using this functional random variable density approximation. The data they used, both simulated and real data sets, were complete data matrices. The obtained correct classification rates compared well with rates obtained using competitors for clustering functional data and methods to cluster traditional multivariate data.

To illustrate numerically the validity of the proposed procedure, we simulated bivariate functional data according to the following model, for two clusters.

$$\begin{aligned} \text{Cluster 1}: \quad & X_1(t) = -5 + t/2 + U_2 h_3(t) + U_3 h_2(t) + \sqrt{0.1}\varepsilon(t), \\ & X_2(t) = -5 + t/2 + U_1 h_1(t) + U_2 h_2(t) + U_3 h_3(t) + \sqrt{0.5}\varepsilon(t), \\ \text{Cluster 2}: \quad & X_1(t) = U_3 h_2(t) + \sqrt{10}\varepsilon(t), \\ & X_2(t) = U_1 h_1(t) + U_3 h_3(t) + \sqrt{0.5}\varepsilon(t), \end{aligned}$$

with $U_1 \sim N(0.5, 1/12)$, $U_2 \sim N(0, 1/12)$, $U_3 \sim N(0, 2/3)$ and $\varepsilon(t) \sim N(0,1)$ independent normally distributed variables. The functions h_1, h_2, and h_3 are defined for $t \in [1, 21]$, as $h_1(t) = (6 - |t-11|)_+$, $h_2(t) = (6 - |t-7|)_+$, $h_3(t) = (6 - |t-15|)_+$, where $()_+$ indicates the positive part. A sample of 50 curves was simulated, with equal mixing proportions and the curves are observed in 41 equidistant points (t= 1, 1.5, . . . , 21). Figure 6.2 displays the profiles for the first 100 simulated data sets.

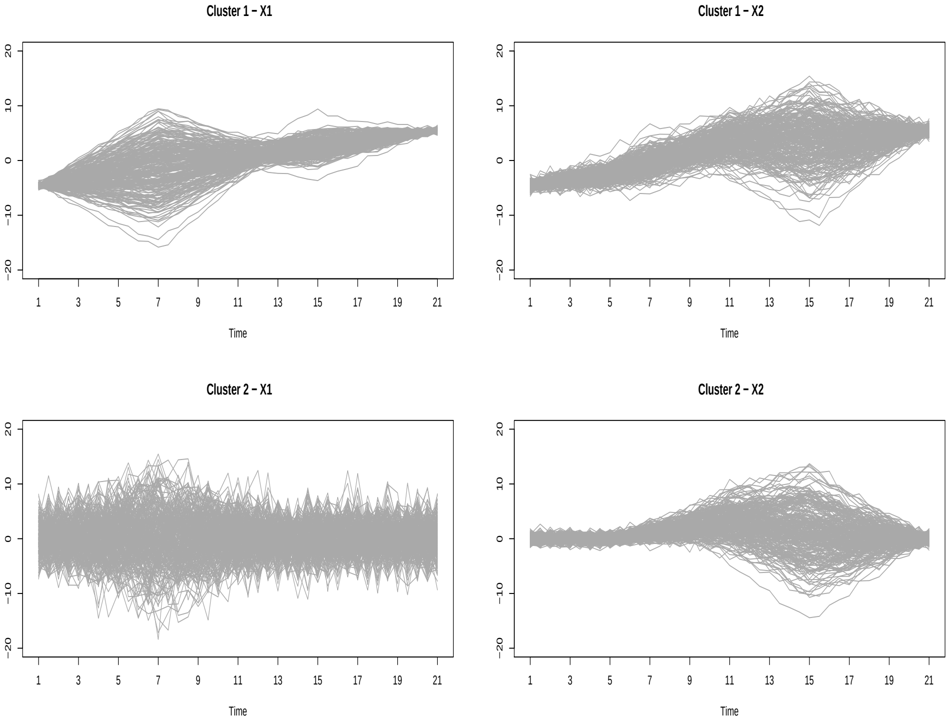


Figure 6.2: *Bi-variate simulated profiles for two clusters, data for first 100 simulated sets.*

This is the model used by Jacques and Preda (2014) to simulate bivariate functional data, for the complete data setting. We used the same model but have put a random proportion of the data to be unobserved. The percentage of missing data was set equal to 10, 20, and 30%. For each setting, 250 incomplete data sets were simulated. Ten imputed data sets were created. The completed profiles were smoothed by means of a cubic spline basis with 28 basis functions. Each imputed data set was then clustered via the outlined model-based procedure using the surrogate densities. The algorithm used 15 initializations for 40 iterations and 200 in the final run with a stopping criteria of 1e-5, the Cattell scree test threshold was fixed to 0.05. Ensemble clustering, with a Euclidean distance measure, resulted in the final partition. For each simulated data set the proportion of correctly classified observations was calculated. Figure 6.3 shows the distribution of the proportion of correctly classified observations over the

250 simulated data sets. The correct classification rate obtained for the complete data set is also presented (of course no imputation was done in this case).

It can be seen that the method performs well, for the considered amounts of missing data. On average 72%–80% of the observations are classified into the correct cluster. The performance on the incomplete data settings is comparable to the correct classification rate obtained for the complete data. So, although the procedure faces a number of sources introducing uncertainty — noise/errors, incomplete observations and thus uncertainty in the estimated principal component scores and eigenfunctions — it is well able to recover the latent cluster structure in the data.

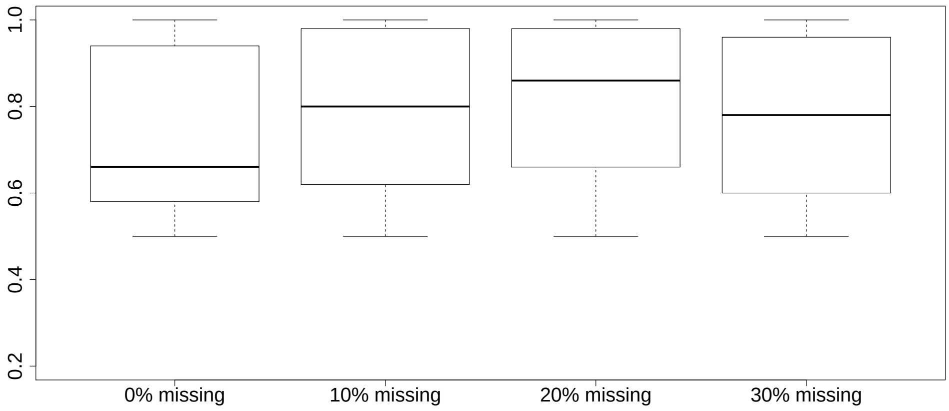


Figure 6.3: *Correct classification rate over 250 simulations, for the complete data sets and after introducing a proportion of missing observations.*

6.9 Cluster Analysis for the Telemonitoring in the Management of Heart Failure Study

The Heart Failure Study is introduced in Section 3.4. To illustrate the methodology outlined in this article, only data from the TM group was used. For this group the telemonitoring device daily transferred data on body weight, blood pressure (systolic and diastolic), and heart rate. Missing information on two consecutive days provoked an alert, patients were contacted to motivate them to make the measurements. At baseline additional patient characteristics were collected. Although alerts were send out when the longitudinal measurement were missing for two consecutive days, quite some missingness is present in the data.

The ability, of the 4 daily-measured biomarkers, to discriminate between patients needing re-hospitalization in the near future and patients not needing to be hospitalized, has been investigate by Njagi et al. (2013). They fitted a joint model for the time to re-hospitalization and the longitudinal biomarker. The model results in a dynamic prediction, i.e., a patient-specific probability for re-hospitalization. This probability is estimated based on information in the longitudinal biomarker (the level of the biomarker and changes in the biomarker), and can (theoretically) be updated daily with every new value of the biomarker being collected.

Information about the extent of missingness in the heart failure data is presented in Tables 3.1 and 3.2. Baseline characteristics are fairly complete. About one out of four patients does not have information for the six minute walking test (WALK). On average, 76% of the patients' daily measurements for the biomakers were recorded, meaning that on average for 137 days out of 180, heart rate, diastolic and systolic blood pressure were communicated to the heart failure clinic by means of the telemonitoring device. The heart failure data has particular features. Heart rate and blood pressure are recorded by the same device and thus simultaneously missing or present. The periods lacking telemonitoring data, are, in general, not too long (average duration is 6 days, median duration is 1 day). However, some patients are featured by longer periods of lacking data. About 5% of the periods, with missing info on consecutive days, lasts longer than 2 weeks.

The EMB algorithm implemented in AMELIA (Honaker and King, 2009) was used to obtain ten complete data matrices. The imputation model included all patients' baseline characteristics and daily-measured biomarker data. A natural logarithm transformation was applied to the longitudinal measurements of heart rate, blood pressure, and body weight in order to normalize the distributions. For these daily-measured biomakers a smooth model over time was imposed, with patient-specific time trends. In particular, a cubic spline model was specified. The EM algorithm can suffer from numerical instability when the number of parameters in the imputation model is high and/or when the degree of missingness is high. Therefore, a ridge prior of 10% was used. Multiple imputation leads to valid results when the imputation model is correctly specified and missingness is missing at random (MAR). MAR cannot be formally tested for. But the accuracy of the imputed values can be judged by over-imputing. Each observed value, in succession, is treated as if it were missing. After a large number of imputations, it can be investigated whether the actual observed value falls within the range of imputed values. Based on this technique it can be concluded that the imputation model is acceptable (graph not shown).

The model-based clustering algorithm for multivariate functional data, described in Section 6.6, was then carried out on each completed data set. Basically the method boils down to applying a parametric mixture model to the surrogate density of the functional data. Multivariate functional principal components analysis is a key building block for as much as the surrogate density function is determined by the PC scores. Since the units of the four biomarkers are different (kg, bpm, and mm Hg), the data were first normalized, $\boldsymbol{Y}(t) = R(t,t)^{-1}\boldsymbol{X}(t)$ with $R(t,t) = \sqrt{(V(t,t))}$, whereupon the contribution of the 4 biomarkers, in defining the principal components, is the same. The response profiles were first smoothed by means of a cubic spline basis with 69 basis functions. A patient's evolution in diastolic and systolic blood pressure, heart rate and weight can be well summarized by the first three principal component scores. Sixty-nine percent of the variability in these biomarkers is explained by three principal components: 28% (range 27–29%) is attributable to the first principal component, 22% (range 21–25%) to the second principal component and finally the third component adds another 19% (range 18–20%). These are percentages averaged over the ten imputed data sets.

The model-based clustering algorithm was applied to the surrogate densities of each of the ten completed data sets separately. For each data set, the algorithm was initialized by running fifty random initializations, for 40 iterations. The random initialization resulting in the best solution (i.e., the highest pseudo-likelihood value), is used as the starting point for a longer algorithm with 500 iterations. The threshold of the Cattell scree test was set to 0.05. An increase in the pseudo-log-likelihood value less than 1e-5 was specified as the stopping criteria. Code for R (package Funclustering) developed by Jacques and Preda (2013) was used.

For the obtained soft two-class solutions, information about the cluster sizes, the estimated orders for the surrogate density functions, and the fuzziness are given in Table 6.1. The Euclidean agreement between the 10 elements of the ensemble ranges from 0.67 (data set 4 and 10) to 0.94 (data set 3 and 6), with a mean Euclidean agreement of 0.80. The agreement among the ten imputed data sets is of particular interest. This measurement quantifies the uncertainty in partitioning the heart failure patients, induced by the presence of missing data. The two-class cluster solution for member 4 of the ensemble, results in a partition of (31,49) patients, for member 6 this is (15,65).

Subsequently, two-class consensus clustering was used to synthesize the information in the 10 partitions—resulting from the model-based clustering— into a single clustering. The Euclidean distance was used as dissimilarity measure, and the con-

Table 6.1: *Heart failure data: number of patients per cluster, number of selected principal components, and fuzziness – for each imputed data set and for the consensus partition.*

	Imputed Dataset										
	1	2	3	4	5	6	7	8	9	10	consensus
	# of patients										
Cluster 1	63	62	63	49	62	65	53	62	65	62	63
Cluster 2	17	18	17	31	18	15	27	18	15	18	17
	# of principal components										
Cluster 1	6	6	6	6	6	6	6	7	6	7	-
Cluster 2	5	5	5	5	5	5	6	6	5	6	-
Fuzziness	0.24	0.22	0.25	0.37	0.28	0.23	0.53	0.27	0.22	0.22	0.36

sensus solution was obtained by maximizing the objective function. A fixed-point algorithm, implemented in the R package CLUE (Hornik, 2005), was used. This algorithm results in a soft consensus partition.

The results are presented in Table 6.1. Partitioning of the 80 patients, based on their profiles for diastolic and systolic blood pressure, heart rate, and weight results in groups of sizes 63 and 17. The average agreement between the consensus clustering and the 10 members of the ensemble equals 0.78 (range 0.65–0.86). The fact that a patient is not necessarily assigned to the same cluster for each of the 10 imputed data sets introduces uncertainty in the consensus cluster assignment. This uncertainty is measurable via a patient's probability of belonging to the cluster. The normalized partition coefficient—measuring the uncertainty in a fuzzy partition — equals 0.36 for the resulting consensus clustering. The fuzziness for the consensus clustering is generally higher than the fuzziness of the 10 members of the ensemble. The fuzziness for the consensus result reflects uncertainty in allocation as present in any cluster procedure, increased by uncertainty due to missing information in a patient's profile. The cluster allocation is clear cut for most patients. For the 63 patients assigned to cluster 1, the average probability of belonging to cluster 1 is 87%. For cluster 2 this probability equals 89%. No relation has been found between the proportion of missingness in a patient's pattern and its cluster membership. Twelve patients (19%) of cluster 1 were re-hospitalized at least once during the study, in cluster 2 four patients (24%) were re-hospitalized at least once. This difference is not statistically significant (χ^2 =0.005, p-value = 0.94). The average evolution (with 95% confidence interval) for the four biomarkers, per cluster, is shown in Figure 6.4. The average evolution for a biomarker, is obtained by averaging the cluster-specific evolution of the 10 imputed data sets. The variability is estimated as a weighted sum of the within and between imputation variability, to reflect the uncertainty in the evolution due to missingness. On average, the patients assigned to cluster 1 have a slightly higher weight and also their diastolic blood pressure is elevated as compared to the patients assigned to cluster 2. Given the small differences in evolutions in the biomarkers, we conclude that there is no substantial evidence for a latent structure in the population of heart failure patients.

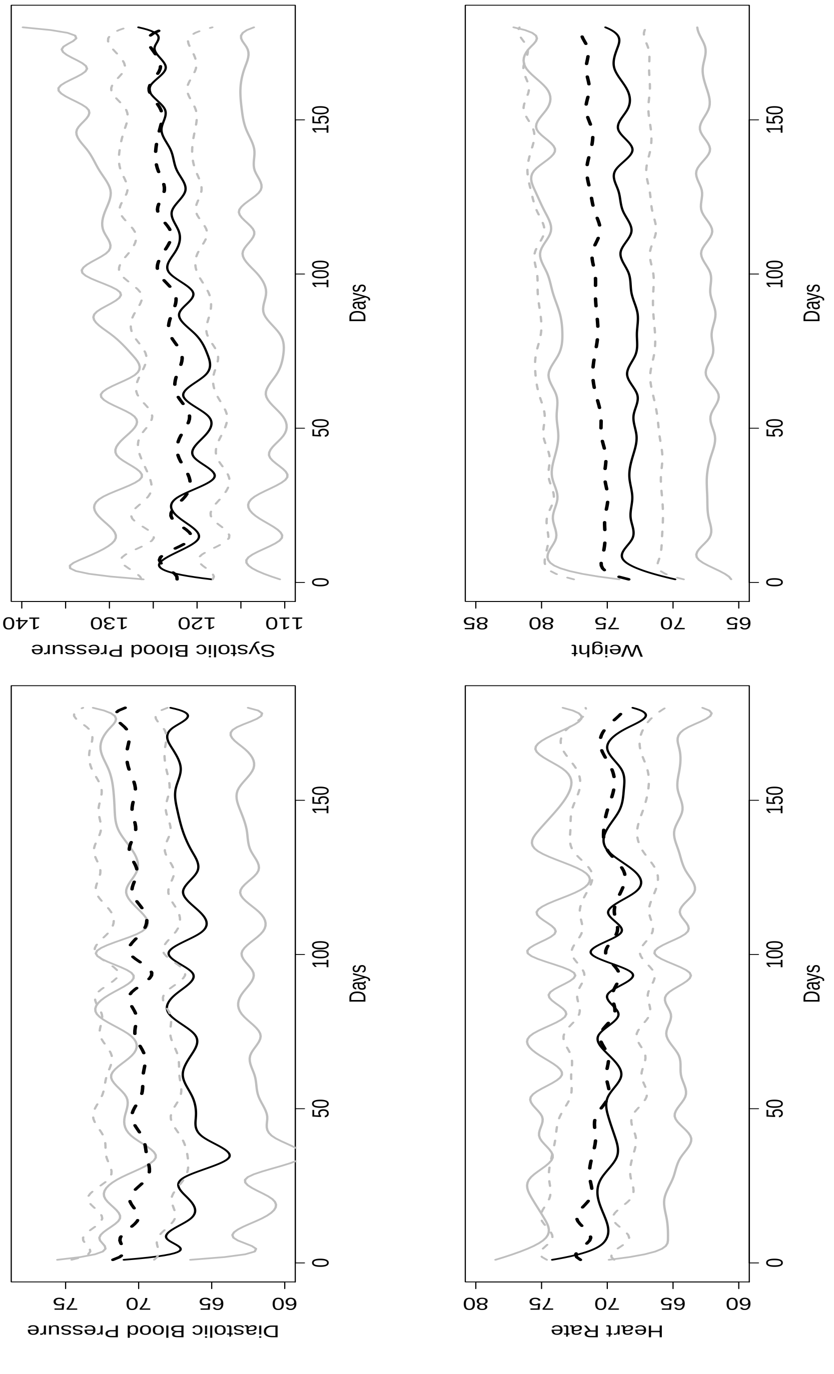


Figure 6.4: *Average evolution and 95% confidence interval for the four biomarkers (average evolution:black; 95% confidence interval: gray; cluster 1: dashed lines; cluster 2: full lines).*

It is well documented (Hajnal and Loosveldt, 2000; Bradley and Fayyad (1998); Pena, Lozano and Larranaga, 1999) that cluster results are sensitive to the preferred algorithm and the randomly selected starting values. Likewise for the proposed approach, alternative options and settings could lead to different partitions of the heart failure data.

The final step in the outlined procedure (Section 6.7), i.e., the consensus clustering, involves a number of choices. The (dis)similarity measure, the objective function, and the optimization algorithm have to be decided. For the heart failure data, Section 6.10 describes the susceptibility of the method in terms of some of these choices. The choice of the distance measure and procedure to optimize the objective function was not very important. The choice of the scree-test threshold, or the number of principal components, to be used in the approximation of the surrogate density, on the other hand did influence the final partition.

6.10 Sensitivity of the Procedure

The algorithm proposed involves a number of choices. The (dis)similarity measure, the objective function, and the optimization algorithm have to be decided.

Consensus clustering carried out on the two-class partitions of the ten imputed data sets, using different distance measures and different procedures to optimize the objective function, resulted in identical results in terms of the hard assignment to the two consensus clusters. The following four methods were compared: fixed-point algorithm for soft least squares consensus partitioning with a Euclidean (1) and GV1 (2) (Gordon and Vichi, 2001) dissimilarity, fixed-point algorithm for soft median consensus clustering with a Manhattan dissimilarity (3), and finally the objective function was minimized with a SUMT algorithm and a co-membership dissimilarity (4). All methods resulted in exactly the same consensus cluster composition, except for one patient.

The cluster results obtained for each of the imputed data sets are more susceptible to the settings specified. The pseudo-density function of the multivariate functional data is specified in terms of a finite number of principal components. The class-specific orders (q_k) are estimated by means of the scree-test of Cattell (1966). The scree-test relies on a threshold to be specified. The number of principal components, selected by the scree-test, increases with decreasing thresholds. Figure 6.5 illustrates the sensitivity of the procedure to the choice of the scree-test threshold. The results are obtained for the first imputed data set. The number of principal components, to be used in the approximation of the surrogate density, was forced to be equal for the two clusters and to change from 1–10. For each setting, 50 random initializations were performed (with 40 iterations) and the best solution among these was subjected to an additional 500 iterations. The stopping criteria for the EM-like algorithm was set equal to 1e-5.

The evolution of the pseudo-likelihood in the last 250 iterations is shown in Figure 6.5. The optimization converges only when four principal components are used to approximate the surrogate density. Non-convergence is seen when less or more information (principal components) are used. For the non-converging pseudo-likelihood profiles, the pseudo-likelihood gets trapped between two values or the evolution is cyclic. Adding more iterations will not unlock the algorithm. The jumps in pseudo-likelihood are large, indicating that the cluster sizes can change substantially from one iteration to the next. This is visualized in Figure 6.6. When five principal components are selected, a group of 14 patients changes cluster, causing the pseudo-likelihood to bounce between two values. The number of patients, with unstable group allocation increases with the number of principal components diverging from 4. The group membership probabilities of patients with unstable cluster allocations are closer to 0.5, as compared to these probabilities for patients with stable cluster allocation. It is obvious that the partitions, corresponding to the maximum value in the pseudo-likelihood profiles, differ. The size of the largest cluster ranges from 46 to 77; and the percentage of patients classified in the same cluster, when using 4 or 10 principal components, was as low as 68%. Clearly, the choice of the scree-test threshold, or the number of principal components, influences the cluster result.

The phenomena, of the likelihood function not converging, does not necessarily indicate problems, but is a possible by-product of the method. The EM-like algorithm iterates between the E-step, the weighted principal component analyses and scree-test, and the M-step. The algorithm stops when the change in pseudo-log-likelihood becomes negligible. It is important to stress that convergence to a local maxima is not insured. The pseudo-likelihood is not based on the entire information contained in the profiles of the four biomarkers, but only on the part of information encompassed in the finite number (q_k) of principal components used to approximate the density function. These dimensions, are updated at each iteration and therefore the data used at different iterations can be different. Consequently, the pseudo-likelihood can increase or decrease during the effectuation of the optimization algorithm.

The maximum number of iterations specified for the EM-like algorithm is not very critical. Forty iterations executed for each of 50 random divisions seem to offer some warranty that the algorithm stabilizes, whether or not reaching a local maximum or iterating between solutions. From figure 6.5 it is clear that when the pseudo-likelihood is trapped between two values the difference in pseudo-likelihood is much bigger than the specified stopping criteria.

6.11 Discussion

In this chapter, an approach for clustering high-dimensional multivariate data with missing observations was proposed. Functional data analysis often utilizes dimension reduction techniques such as principal component analysis. Dimension reduction techniques require complete data matrices. To overcome this problem, the data were first completed by means of multiple imputation. Each imputed data set was subjected to a cluster procedure for multivariate functional data. Consensus clustering was subsequently applied to summarize the ensemble of partitions into the final cluster result.

When the time points vary widely across subjects and are sparse, the functional principal components scores obtained through the Karhunen-Loève expansion are not well approximated by the usual integration method. Yao, Müller and Wang (2005) proposed a nonparametric method to perform functional principal component analysis for the case of irregularly spaced longitudinal data where the number of repeated measurements is small. Although a sufficient number of observations per patient is available in the Heart Failure Study, this procedure could have been applied to the data. Given the large number of time points and patients, it could take a long time to execute the analysis.

We have however chosen for an alternative route, i.e., to impute the missing data and construct a number of complete data matrices. We see two reasons for doing so. The imputation process can benefit from all available information, e.g., information in baseline covariates, and associations between the biomarkers. Further, the uncertainty due to incomplete data is also reflected in and quantified during the clustering process.

The uncertainty in cluster membership, due to missing data, was characterized by means of the agreement between the members of the ensemble and the fuzziness of the consensus clustering. The usefulness of the method was illustrated on a simulated data set and on the heart failure data. However, a number of topics are still open for further investigation.

The functional representation of raw data in general involves some smoothing. In this work the data was smoothed by a cubic spline basis with 69 basis functions. But alternative smoothing methods—including other basis function, local weighting methods and roughness penalty approaches—could have been used. They all have in common that smoothing parameters (e.g., the number of basis functions, bandwidth of kernel function or penalty parameters) have to be optimally chosen.

The class-specific orders, used to describe the pseudo-likelihood, are chosen through the threshold of the Cattle scree test. This is a heuristic method. Other heuristic and statistical procedures could be used to determine the number of components to be retained (Jackson, 1993).

Information criteria like AIC and BIC are generally used to determine the optimal number of clusters. These criteria can be obtained from the pseudo-likelihood, but are not very useful. Only relative comparisons between a set of models attempting to fit a given dataset can be done with these. The amount of data used in the algorithm, depends on the class-specific orders resulting from the Cattle Scree test. Thus it is not guaranteed that the data used in different models is identical, which hampers the determination of the number of clusters.

Breaban and Luchian (2011) have defined a new information criterion, CritCF. This criterion takes into account the number of clusters and the number of variables for ranking partitions. This criterion could be valuable in addressing two issues at once, the issue of selecting the class-specific orders and the issue of determining the optimal number of clusters.

The proposed algorithm was applied on ten completed data sets, but the choice of the number of imputed data sets is still an open topic.

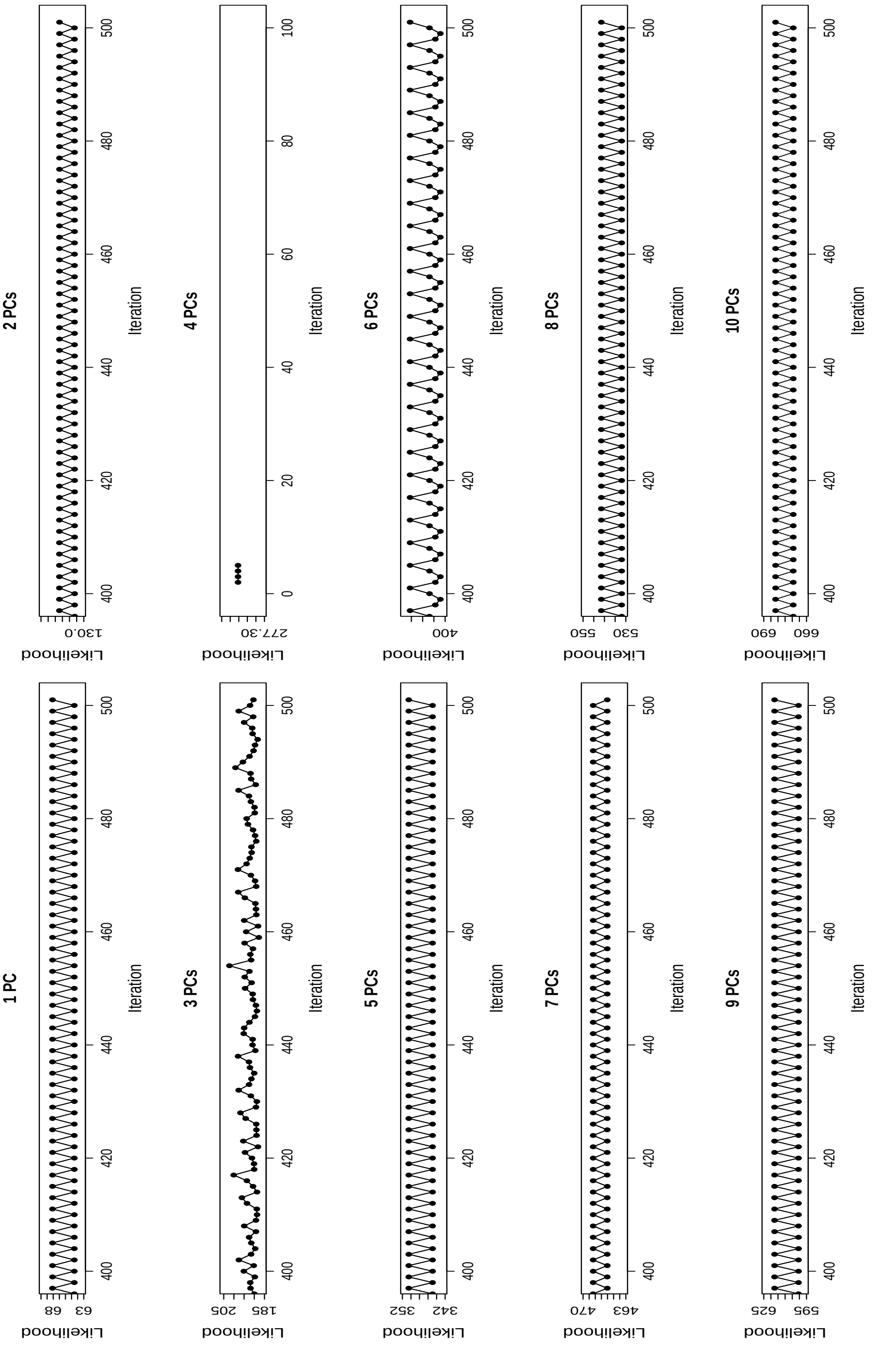


Figure 6.5: *Evolution of the pseudo-likelihood for the first imputed data set, for different dimensions of principal components*

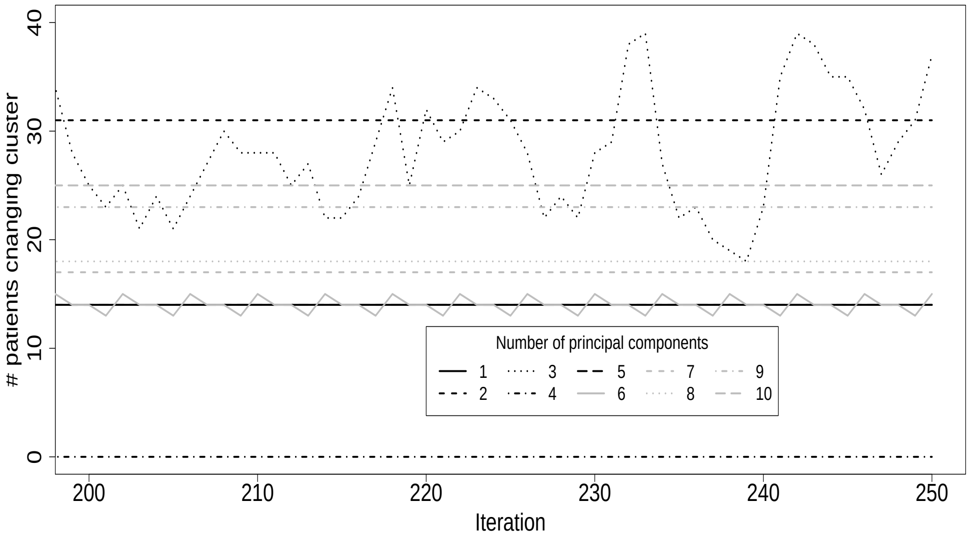


Figure 6.6: *Number of patients with unstable cluster allocation (the lines for 1 and 5 principal components coincide).*

Chapter 7

Sensitivity Analysis for a Growth-Mixture Model

Missing data are inevitable when collecting information about patients, especially in studies collecting data repeatedly over time. The pattern and cause for missingness can vary. Intermittent missing observations are often referred to as non-monotone missingness. Patients with information for all planned visits until a given point in time, and no information thereafter are referred to as dropout. The reasons for not obtaining data at all planned visits can be diverse. Adverse events, no response to study treatment, protocol violations, and loss to follow-up are frequently reported in clinical studies as reasons for dropout.

The occurrence of incomplete records introduces specific challenges in statistical analyses (Little and Rubin, 2002). Not only has the loss of information a negative impact on the precision and the power, but furthermore a complete case analysis will often result in biased estimators. Quite a number of methods are available to handle missing data. Likelihood-based analyses are valid when the missing-data mechanism is missing at random (MAR), in the sense that the mechanism describing missingness is allowed to depend on covariates and observed outcomes but, given these, not further on unobserved outcomes or unobserved covariates. However, the possibility of the more general missing not at random (MNAR) mechanism, where there is further dependence of missingness on unobserved outcomes, can never be definitively excluded. Both MAR and MNAR rely on untestable assumptions.

In this chapter, we introduce MNAR models when clustering of longitudinal profiles with missing information, by means of a growth-mixture model, is of interest. Muthén et al. (2011) have provided an excellent overview of these recent innovations. We demonstrate the use of these MNAR models as a tool to study the vulnerability of the cluster analysis results. Where Muthén et al. focussed on the comparison of the results of the different models in terms of the optimal number of clusters and the estimated trajectories; we will complement the model comparison with an evaluation of the sensitiveness of the posterior membership probabilities under the different missing-data models for non-ignorable dropout.

Deciding which of these missing-data models is most suitable is difficult. The log-likelihoods of MAR under ignorability, pattern-mixture and selection models cannot be compared. Only comparison of log-likelihoods and BIC within the same family is meaningful. However, comparison of the models in terms of their predictive value for an event occurring in the (near) future is possible and could assist in choosing between the models. Such an event is called a distal event.

This chapter is organised as follows. Section 7.1 introduces terminology, missing-data mechanisms, and models for analyzing longitudinal data in the presence of missingness. Section 7.2 positions the non-verifiable assumptions, underlying these models, in the framework of enriched data. Section 7.3 introduces a number of incomplete-data growth-mixture models, some incorporating a distal event, that were applied to the abdominal aorta aneurysm (AAA) data (Section 3.3) in the context of a sensitivity analysis. Results are presented in Section 7.4. Conclusions are formulated in Section 7.5.

7.1 Incomplete Longitudinal Data

We will use the terminology of Rubin (1976) and Little and Rubin (2002) to distinguish between the different missing-data mechanisms. A process is said to be missing completely at random (MCAR) if the missingness is independent of both the unobserved and observed data. The process is missing at random (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. If the process is neither MCAR nor MAR it is missing not at random (MNAR). The process then depends on the unobserved measurements.

In certain circumstances, the missing-data process is ignored and simple methods such as complete case analyses are used. A complete case analysis takes into account only those patients for which all planned measurements are obtained. Although very simple, this method has severe drawbacks. Due to the loss of information there is a negative impact on the precision and the power. The estimators, in a complete case analysis, will be unbiased only under MCAR.

The method of last observation carried forward (LOCF) replaces every missing value by the last observed value from the same patient. Many authors describe the problems with an LOCF analysis (Gibbons et al., 1993; Molenberghs and Verbeke, 2005; Carpenter et al., 2004; Beunckens et al., 2005). Even under the very strong assumption of MCAR, LOCF can be biased. The direction of the bias depends on the true but unknown treatment effect, as well as on other aspects of the measurement and missingness process.

These traditional techniques for analyzing incomplete data work well only in a limited set of circumstances and are generally prone to bias. Even when the techniques are unbiased, they tend to be less efficient than modern missing-data methods, such as direct maximum likelihood (DML) and multiple imputation (MI). DML and MI make weaker, and probably more realistic, assumptions about the missing data. Consequently, they should produce efficient and broadly unbiased parameter estimates. An application of multiple imputation in the context of cluster analysis was given in Chapter 6. In this chapter we use a direct maximum likelihood approach.

7.1.1 Missing Data Modeling Frameworks

In case of missing data in the repeated measurements $\boldsymbol{y_i}$, one needs to specify the full-data likelihood. The full-data likelihood considers not only the repeated measurements, observed and unobserved, as data but also the missing-data indicators for each time point. The missing-data indicator for subject i, $\boldsymbol{m_i}$, are binary variables with values for m_{it} equal to 1 if the repeated measurement at time t, y_{it}, is missing and 0 if y_{it} is observed. In case of pure dropout, the vector $\boldsymbol{m}_i$ can be replaced by a scalar D_i, denoting the time at which the patient drops out from the study.

The observed repeated measurements are grouped in $\boldsymbol{y_i}^{\text{obs}}$ and the missing measurements in $\boldsymbol{y_i}^{\text{mis}}$ such that $\boldsymbol{y_i} = (\boldsymbol{y_i}^{\text{obs}}, \boldsymbol{y_i}^{\text{mis}})$. The full-data density then becomes:

$$f(\boldsymbol{y_i}, \boldsymbol{m_i} | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}), \tag{7.1}$$

with $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ the parameter vectors associated with the measurement process and the missingness process, and X_i and W_i the associated design matrices.

Different modeling frameworks for incomplete longitudinal data are obtained by factorizing this full-data density in different ways. This results in selection, pattern-mixture and shared-parameter models. The interpretation of the results depends on the chosen framework and different results and conclusions can be obtained. When interest is also in unobserved latent constructs (such as latent-class membership) one should realise that the different models imply different assumptions for the unobserved data, but that the observed data alone does not allow to choose between these models. Auxiliary information can be insightful to judge the plausibility of the different models (see Section 7.1.3).

A selection model (Diggle and Kenward, 1994) factorizes the full-data density as the product of the marginal density of the measurements and the conditional density of the missingness process, conditioned on the measurements.

$$f(\boldsymbol{y_i}, \boldsymbol{m_i}|X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y_i}|X_i, \boldsymbol{\theta}) f(\boldsymbol{m_i}|\boldsymbol{y_i}, W_i, \boldsymbol{\psi}). \tag{7.2}$$

Whenever $m_{it} = 1$ the measurement is missing, thus the regression of $\boldsymbol{m}_i$ on $\boldsymbol{y}_i$ is inherently inestimable. Only by imposing distributional assumptions for the repeated measurements, often multivariate normality, the model can be estimated. Departures from these assumptions can produce biased results.

The pattern-mixture model (Little, 1995) specifies a different measurement model for each pattern of missing values, and the full-data density is obtained as the mixture of the models weighted by the probability of each missing value pattern.

$$f(\boldsymbol{y_i}, \boldsymbol{m_i}|X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y_i}|\boldsymbol{m_i}, X_i, \boldsymbol{\theta}) f(\boldsymbol{m_i}|W_i, \boldsymbol{\psi}). \tag{7.3}$$

The above specified model is also inestimable without imposing assumptions. A quadratic growth model with only two observed data points is not identified. Therefore, values are assumed for the inestimable parameters. Information from across patterns can be shared (Hedeker and Gibbons, 1997, 2006) or identifying restrictions can be used (Little, 1995; Thijs et al., 2002; Kenward et al., 2003). Implementing different constraints can produce different results and unfortunately there is no way to gauge the credibility of the assumptions. The marginal estimates, obtained via a pattern-mixture model, are function of the model parameters and therefore standard errors are not automatically produced. Approximate standard errors are routinely obtained via the delta method (Hedeker and Gibbons, 1997).

Shared-parameter models (Wu and Caroll, 1988) assume the existence of latent variables, that are shared between both factors in the full-data density. Often conditional independence is assumed, i.e., the measurement process and missingness process are independent conditional on the latent variable. The latent variable can be a random effect $\boldsymbol{b}_i$, resulting in the following model:

$$f(\boldsymbol{y_i}, \boldsymbol{m_i}|X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{b}_i) = f(\boldsymbol{y_i}, |X_i, \boldsymbol{\theta}, \boldsymbol{b}_i) f(\boldsymbol{m_i}|W_i, \boldsymbol{\psi}, \boldsymbol{b}_i). \tag{7.4}$$

By regressing the missing-data indicators on the random effects (e.g., intercept and slope), the probability of missingness depends on the entire set of repeated measurements, including the unobserved ones. This model is identified by making assumptions for the distribution of the random effects, and further by assuming that the measurement process and the missingness process are independent conditional on the random effects.

Alternatively, a categorical latent variable can be used (Roy, 2003). The missing-data patterns are assumed to be related to unobserved latent-class membership, where the number of classes is less than the number of observed missing-data patterns. The likelihood for the response is a mixture of latent dropout classes, as opposed to the observed dropout patterns themselves. This approach allows observed patterns to be sparse without needing to specify additional identifying restrictions.

The full-data density is never available and inference is made on the observed data. Depending on the postulated missing-data mechanism different implications are ensued. The implications of the postulated missing-data mechanism are easiest seen in a selection-modeling framework as defined by Rubin (1976). The conditional distribution of the missingness process, conditional on the measurements, plays a central role in the selection model. This conditional distribution simplifies under the assumption of MCAR and MAR, and this has implications for the joint density of the observed data and the missing-data indicators. The selection probability $f(\boldsymbol{m_i}|\boldsymbol{y_i}, W_i, \boldsymbol{\psi})$ under the three missing-data mechanisms is given by:

$$\begin{aligned}
\text{MCAR} &: \; f(\boldsymbol{m_i}|W_i, \boldsymbol{\psi}),\\
\text{MAR} &: \; f(\boldsymbol{m_i}|\boldsymbol{y_i}^{\text{obs}}, W_i, \boldsymbol{\psi}),\\
\text{MNAR} &: \; f(\boldsymbol{m_i}|\boldsymbol{y_i}^{\text{obs}}, \boldsymbol{y_i}^{\text{mis}}, W_i, \boldsymbol{\psi}).
\end{aligned}$$

The joint density of the observed data and missing-data indicators thus simplifies as:

$$\begin{aligned}
\text{MCAR} &: \; f(\boldsymbol{y_i}^{\text{obs}}|X_i, \boldsymbol{\theta}) f(\boldsymbol{m_i}|W_i, \boldsymbol{\psi}),\\
\text{MAR} &: \; f(\boldsymbol{y_i}^{\text{obs}}|X_i, \boldsymbol{\theta}) f(\boldsymbol{m_i}|\boldsymbol{y_i}^{\text{obs}}, W_i, \boldsymbol{\psi}),\\
\text{MNAR} &: \; \int f(\boldsymbol{y_i}^{\text{obs}}, \boldsymbol{y_i}^{\text{mis}}|X_i, \boldsymbol{\theta}) f(\boldsymbol{m_i}|\boldsymbol{y_i}^{\text{obs}}, \boldsymbol{y_i}^{\text{mis}}, W_i, \boldsymbol{\psi}) d\boldsymbol{y_i}^{\text{mis}}.
\end{aligned}$$

Under MAR the likelihood factorizes into two components. When the parameter space of $(\boldsymbol{\theta},\ \boldsymbol{\psi})$ is given by the product of the individual parameter spaces (separability condition), inference for $\boldsymbol{\theta}$ can be obtained on the marginal observed data density alone, ignoring the missingness process. So, in the likelihood framework and when the separability condition is satisfied, ignorability is equivalent to MAR ∪ MCAR. Under MNAR the joint distribution of $\boldsymbol{y_i}$ and $\boldsymbol{m}_i$ cannot be simplified. Inference is only possible by making unverifiable assumptions. Whenever $m_{it} = 1$ the measurement is missing, thus the regression of $\boldsymbol{m_i}$ on $\boldsymbol{y_i}$ is inherently inestimable. Only by imposing distributional assumptions, often multivariate normality, the model can be estimated. Departures from these assumptions can produce biased results.

7.1.2 Missing Data in Growth-Mixture Models

Chapter 4 introduced finite-mixture models as a tool to identify homogeneous subgroups in a heterogeneous population. A finite-mixture model specifies the density of the measurements as a mixture of K density functions: $f(\boldsymbol{y}_i) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i)$. In the presence of missing observations, the full-data density for a mixture model is given as (where for simplicity of notation dependence of covariates is suppressed):

$$f(\boldsymbol{y_i}, \boldsymbol{m_i}) = \sum_{k=1}^{K} f(\boldsymbol{y}_i, \boldsymbol{m}_i | c_{ik} = 1) f(c_{ik} = 1), \tag{7.5}$$

with $c_{ik} = 1$ when subject i belongs to the k^{th} mixture and zero otherwise. Factorizing this full-data density according to the selection model and assuming MAR for the missing-data mechanism conditional on class, implies the following observed-data likelihood for a growth-mixture model:

$$f(\boldsymbol{y_i}^{\text{obs}}, \boldsymbol{m}_i) = \sum_{k=1}^{K} f(\boldsymbol{y_i}^{\text{obs}} | c_{ik} = 1) f(\boldsymbol{m}_i | \boldsymbol{y_i}^{\text{obs}}, c_{ik} = 1) f(c_{ik} = 1). \tag{7.6}$$

The standard missing-data procedure of using all available data is only valid when,

(1) MAR holds conditional on class (resulting in decomposition (7.6)), and
(2) the missing-data mechanism remains the same for the different classes given the observed data, i.e., $f(\boldsymbol{m}_i | \boldsymbol{y_i}^{\text{obs}}, c_{ik} = 1) = f(\boldsymbol{m}_i | \boldsymbol{y_i}^{\text{obs}})$, and
(3) the first two terms in (7.6) do not share parameters.

Under these assumptions, maximization of the first term of the likelihood is sufficient and the missing-data mechanism can be ignored. When assumption (2) is violated ignorability does not hold for the mixture distribution of $f(\boldsymbol{y})$. The term $f(\boldsymbol{m_i} | \boldsymbol{c_i}, *)$ can not be ignored in the EM algorithm. Missingness depends on a latent variable (in the above situation on $\boldsymbol{c_i}$) leading to a non-ignorable missing-data situation.

The assumption of MAR leads to convenient simplification of the likelihood, but in general one cannot justify this assumption. Often one is forced to make assumptions which cannot be underpinned with the observed data alone.

7.1.3 Sensitivity Analysis Using a Distal Event

Choosing between the MAR, selection models, and pattern-mixture models is difficult since the models use different sets of dependent variables, resulting in different likelihoods and information criteria metrics. When auxiliary information about the reasons for dropout is available, model comparison becomes possible. Given that such a distal or ultimate outcome contains information about $\boldsymbol{y_i}^{\mathrm{obs}}$ but on top of that also about side-effects and subjects preferences it is useful for MNAR modeling. A sensitivity analysis incorporating this information in the models can help in choosing between different models fitted to the data. The relationship between the latent-class membership and the distal event provides a predictive validity check of the latent-class formulation. Furthermore, the congruence between the latent class formation when not including versus including the distal event in the model is a measure for the ability of the original model to capture non-ignorable missingness.

Muthén and Shedden (1999) describe an extended finite-mixture model that allows joint estimation of (1) a conventional finite growth-mixture model where different curve shapes are captured by class-varying random-coefficient means and (2) a logistic regression of the distal event on the classes. Let $\boldsymbol{u_i}$ be a categorical distal outcome with R categories. A multinomial regression model is used to link the latent class $\boldsymbol{c_i}$ to the distal outcome $\boldsymbol{u_i}$ $(r = 1, \ldots, R)$:

$$p(\boldsymbol{u_i} = \boldsymbol{r} | \boldsymbol{c_i} = k) = \frac{e^{\alpha_{rk}}}{\sum_{r=1}^{R} e^{\alpha_{rk}}}.$$

The complete-data likelihood is given by:

$$f(\boldsymbol{y_i}, \boldsymbol{u_i}) = \sum_{c} f(\boldsymbol{y_i}|\boldsymbol{c_i}) f(\boldsymbol{u_i}|\boldsymbol{c_i}) f(\boldsymbol{c_i}).$$

Before sketching the MNAR growth-mixture models, that will be applied to the AAA data (Section 3.3), we place the issues of MNAR and unverifiable assumptions in the framework of enriched data.

7.2 Latency and Unidentifiability

Using the terminology of Verbeke et al. (2010) a growth-mixture model encompasses three types of data. Of course, the observed data (outcomes $\boldsymbol{y_i}^{\text{obs}}$, $\boldsymbol{m_i}$ and covariates $\boldsymbol{x_i}$) are one type of data. The latent variables, continuous ones ($\boldsymbol{b_i}$) and categorical ones ($\boldsymbol{c_i}$) are so-called augmented data, referring to the addition of constructs to the observed data. The augmented data are always fully unobserved. The reason for the augmentation of the data is that in general the model development simplifies. The term coarsened data is used to refer to the fact that the observed data are coarser than hypothetically conceived. Ideally, the data are observed fully but in practice this is not the case. In the growth-mixture model, formulated above, the missing observations ($\boldsymbol{y_i}^{\text{mis}}$) are coarsened data. The augmented and coarsened data taken together are called enriched data.

Verbeke et al. (2010) show that always a part of the model for enriched data is unidentifiable from the observed data. By replacing these unidentified parts, which rest completely on assumptions, an entire class of models can be obtained. All of these models produce the same fit to the observed data. The authors assume that data $\boldsymbol{z_i}$ for an independent unit $i = 1, \ldots, N$ are augmented with $\boldsymbol{a_i}$. The $\boldsymbol{a}_i$ can take any enriched data form and is thus broader than the latent classes used before.

The joint model $f(\boldsymbol{z_i}, \boldsymbol{a_i}|\boldsymbol{\theta}, \boldsymbol{\psi})$ can be factorized in the following ways:

$$f(\boldsymbol{z_i}, \boldsymbol{a_i}|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{z_i}|\boldsymbol{a_i}, \boldsymbol{\theta}, \boldsymbol{\psi}) f(\boldsymbol{a_i}|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{z_i}|\boldsymbol{\theta}, \boldsymbol{\psi}) f(\boldsymbol{a_i}|\boldsymbol{z_i}, \boldsymbol{\theta}, \boldsymbol{\psi}). \tag{7.7}$$

Replacing the posterior density, of the enriched data given the observed data $f(\boldsymbol{a_i}|\boldsymbol{z_i}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}})$, where $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}$ are the estimates obtained from the original model, by an arbitrary conditional density $f(\boldsymbol{d_i}|\boldsymbol{z_i}, \boldsymbol{\gamma})$ retains the fit to the original model. Here, $\boldsymbol{d_i}$ is used to indicate that the original and substituted enriched data can be different.

Using the notation of the previous section, in a growth-mixture model, it is the posterior density $f(\boldsymbol{b_i}, \boldsymbol{c_i}, \boldsymbol{y_i}^{\text{mis}}|\boldsymbol{y_i}^{\text{obs}}, \boldsymbol{m_i}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}})$ that could be replaced without changing the fit to the data. This posterior function could further be factorized in the following three terms :

$$\begin{aligned} f(\boldsymbol{b_i}, \boldsymbol{c_i}, \boldsymbol{y_i}^{\text{mis}}|\boldsymbol{y_i}^{\text{obs}}, \boldsymbol{m_i}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}) &= f(\boldsymbol{y_i}^{\text{mis}}|\boldsymbol{y_i}^{\text{obs}}, \boldsymbol{m_i}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}) \\ &\quad \times f(\boldsymbol{b_i}|\boldsymbol{c_i}, \boldsymbol{y_i}^{\text{mis}}, \boldsymbol{y_i}^{\text{obs}}, \boldsymbol{m_i}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}) \\ &\quad \times f(\boldsymbol{c_i}|\boldsymbol{y_i}^{\text{mis}}, \boldsymbol{y_i}^{\text{obs}}, \boldsymbol{m_i}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}). \end{aligned} \tag{7.8}$$

A simplification, resulting in the conventional growth-mixture model, assumes that the random effects $\boldsymbol{b_i}$ are influenced by the class, and that prediction of the class is done solely by covariates. In this case (7.8) becomes:

$$\begin{aligned} f(\boldsymbol{b_i}, \boldsymbol{c_i}, \boldsymbol{y_i}^{\text{mis}} | \boldsymbol{y_i}^{\text{obs}}, \boldsymbol{m_i}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}) &= f(\boldsymbol{y_i}^{\text{mis}} | \boldsymbol{y_i}^{\text{obs}}, \boldsymbol{m_i}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}) \\ &\quad \times f(\boldsymbol{b_i} | \boldsymbol{c_i}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}) \\ &\quad \times f(\boldsymbol{c_i} | \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}). \end{aligned} \tag{7.9}$$

Replacing one of these densities will not change the value of the observed-data log-likelihood. The first term of (7.9) could, for example, be replaced by its MAR counterpart $f(\boldsymbol{y_i}^{\text{mis}} | \boldsymbol{y_i}^{\text{obs}})$. The posterior density of the random effects given the class, is generally assumed to be the normal distribution. These densities could be replaced by other distributions, e.g., gamma random effects. Finally, the last term in (7.9) specifies the number of latent classes. A model specifying L classes instead of K could be considered, nevertheless resulting in the same fit.

None of these densities is identifiable from the observed data, they are solely determined by modeling assumptions. To investigate how sensitive the results are to the assumptions, a sensitivity analysis should be conducted. In a broad sense, a sensitivity analysis is one in which several statistical models are considered simultaneously and/or a statistical model is further scrutinized using specialised tools (such as diagnostic measures). In terms of a sensitivity analysis for the missing-data model, the simplest procedure is to fit a selected number of (MNAR) models which seem plausible, or one in which a primary analysis is supplemented with a number of modifications. The degree to which conclusions are stable across these models provides an indication of the confidence that is placed in them. Comparison of the models is also possible in terms of their predictive value for an event occurring in the (near) future. By studying the congruence between the latent class formation when including and not including a distal event in the model, the ability of the model to capture non-ignorable missingness can be studied (see Section 7.1.3).

7.3 Incomplete-data Models for the AAA Study

In this section we describe the MAR and MNAR models that were applied to the AAA data. The purpose of the analysis was to group patients with similar growth profiles for the diameter of the artery, therefore a finite-mixture model was used.

7.3.1 Measurement Model

The measurement model was kept common for all MAR and MNAR models. Within the time range of the follow-up, the diameter of the artery exhibits a linear growth

(Figure 3.5). Thus, a linear growth model will be assumed. The likelihood functions are easily maximized with the latent-variable modeling program Mplus. Hence, we formulate the growth model as a latent-variable model (see Section 4.2). For class k:

$$y_{it}|_{c_{ik}=1} = \eta_{0i} + \eta_{1i}t + \varepsilon_{it}, \tag{7.10}$$

where $\boldsymbol{c_i}$ is the latent-class variable. The random-effects distributions vary as a function of the class k:

$$\begin{array}{rcl} \eta_{0i}|_{c_{ik}=1} & = & \alpha_{0k} + \zeta_{0i}, \\ \eta_{1i}|_{c_{ik}=1} & = & \alpha_{1k} + \zeta_{1i}. \end{array} \tag{7.11}$$

The residuals are assumed to be normally distributed with zero mean. The $\boldsymbol{\zeta_i}$ have a 2×2 covariance matrix D. The $\boldsymbol{\varepsilon_i}$ a 7×7 diagonal matrix Σ_k, the error variances are assumed to be constant over time.

7.3.2 Missing at Random Model

Model (7.10) reflects the first term in (7.6). This model provides unbiased estimates when in addition to ignorability, MAR conditional on the latent class holds and missingness is not influenced by the class.

7.3.3 Missing Not at Random Models

Different modeling frameworks have been proposed to jointly model the outcomes and the missing-data process. We will in turn discuss selection and pattern-mixture models. The conventional models and the models incorporating a latent-class variable $\boldsymbol{c_i}$ are presented. Some of these models carry features of shared-parameter models.

7.3.3.1 Selection Modeling

Selection models factorize the full-data likelihood $f(\boldsymbol{y}, \boldsymbol{m})$ as $f(\boldsymbol{y})f(\boldsymbol{m}|\boldsymbol{y})$. The first factor is the marginal density of the measurement process and the second one the density of the missingness process, conditional on the outcomes. One often assumes the following form for the conditional density, proposed by Diggle and Kenward (1994):

$$\text{logit}(m_{it}|y_{it}, y_{i,t-1}) = \beta_{0t} + \beta_1 y_{it} + \beta_2 y_{i,t-1}. \tag{7.12}$$

The probability of missing data at time t depends directly on the repeated measurements at time t as on the preceding measurement. MAR holds if $\beta_1 = 0$ and $\beta_1 \neq 0$ is an indication of MNAR. A logistic regression could be used to make inference about the missing-data mechanism. However, the association between $\boldsymbol{m}_i$ and $\boldsymbol{y}_i$ is estimated under distributional assumptions.

This Diggle-Kenward model can be extended to a mixture model by specifying the following model for the missing-data indicators, conditional on the class:

$$\text{logit}(m_{it}|y_{it}, y_{i,t-1}, c_{ik} = 1) = \beta_{0tk} + \beta_{1k}y_{it} + \beta_{2k}y_{i,t-1}, \tag{7.13}$$

The logistic regression coefficients are allowed to vary across the latent classes. In the AAA study, this can be important when patients in high-trajectory classes tend to drop out because of large diameters, and patients in the low class tend to drop out because of small diameters of the artery. When the influence of y_{it} and $y_{i,t-1}$ on the dropout probability varies across time, the model can further be generalized:

$$\text{logit}(m_{it}|y_{it}, y_{i,t-1}, c_{ik} = 1) = \beta_{0tk} + \beta_{1tk}y_{it} + \beta_{2tk}y_{i,t-1}. \tag{7.14}$$

To lower the number of parameters in the model, the parameters β_{*tk} could be specified to vary as a linear function of time.

An alternative model was proposed by Beunckens et al. (2008). Their model combines features from a selection model and a shared-parameter model. Dropout is influenced by the latent-class variable and the random effects. For the AAA study, we specified the dependence of dropout on latent class as a linear function of time t and as a function of the random intercept η_{0i}.

$$\text{logit}(m_{it}|\eta_{0i}, c_{ik} = 1) = \beta_{0k} + \beta_{1k}t + \beta_{2k}\eta_{0i}. \tag{7.15}$$

The selection-model features of these models are given by the (random effects of the) outcome process influencing the dropout indicators. Given that the dropout indicators and the diameters outcomes are both influenced by the latent-class variable (and the random intercept), these models also are of the shared-parameter type.

7.3.3.2 Pattern-mixture Modeling

Pattern-mixture models (Little 1993, 1994) factorize the full-data likelihood $f(\boldsymbol{y}, \boldsymbol{m})$ as $f(\boldsymbol{m})f(\boldsymbol{y}|\boldsymbol{m})$. This is a mixture density over different populations, each defined by the observed pattern of missingness. A simple version of $f(\boldsymbol{y}|\boldsymbol{m})$ was specified by allowing the random effects of the growth model to vary as a function of the missing-data indicators. For each dropout pattern, a linear growth model was specified.

$$y_{it}|m_{i1}, \ldots, m_{iT} = \eta_{0i} + \eta_{1i}t + \varepsilon_{it}, \tag{7.16}$$

and

$$\begin{array}{r}\eta_{0i}|m_{i1},\ldots,m_{iT} = \alpha_0 + \sum_{t=1}^{T}\beta_{0t}m_{it} + \zeta_{0i},\\ \eta_{1i}|m_{i1},\ldots,m_{iT} = \alpha_1 + \sum_{t=1}^{T}\beta_{1t}m_{it} + \zeta_{1i}.\end{array} \tag{7.17}$$

The slope of the time effect for patients dropping out after the first visit was set equal to the slope of patients dropping out after the second visit. The slope is not identified for patients dropping out after the baseline visit.

An alternative model has been proposed by Roy (2003). Where in the conventional pattern-mixture models the distributions for the different dropout patterns are mixed, Roy's model considers a latent class defined by the missing-data indicators, so that $f(\boldsymbol{y}|\boldsymbol{m}) = \sum_c f(\boldsymbol{y},\boldsymbol{c}|\boldsymbol{m}) = \sum_c f(\boldsymbol{y}|\boldsymbol{c})f(\boldsymbol{c}|\boldsymbol{m})$. Latent-class membership is specified as a multinomial logistic regression model, with the missing-data indicators as explanatory variables. The random effects of the measurement outcomes are influenced by this latent-class variable. So, (7.10) and (7.11) describe the measurement process, where the class variable $\boldsymbol{c_i}$ is determined by the missing-data indicators:

$$\text{logit}(c_{ik} = 1|m_{i1},\ldots,m_{iT}) = \gamma_0 + \sum_{t=1}^{T}\gamma_t m_{it}. \tag{7.18}$$

A generalization to this model was proposed by Muthén (2011). Two latent categorical variables are used. One latent class provides information about the outcome trajectory classes, the second latent class is based on the dropout pattern. We will refer to this model as the Roy-Muthén model. Define a latent-class variable $\boldsymbol{cd}$ for the L dropout groups and a latent-class variable $\boldsymbol{cy}$ for K trajectory classes for the outcome $\boldsymbol{y}$. The model is specified by the following equations.

$$\begin{array}{rl} y_{it}|_{cd_{il}=1,cy_{ik}=1} & = \eta_{0i} + \eta_{1i}t + \varepsilon_{it},\\ \eta_{0i}|_{cd_{il}=1,cy_{ik}=1} & = \alpha_{0lk} + \zeta_{0i},\\ \eta_{1i}|_{cd_{il}=1,cy_{ik}=1} & = \alpha_{1lk} + \zeta_{1i}.\end{array} \tag{7.19}$$

The two latent-class memberships are given by a bivariate loglinear model of the form:

$$\text{logit}(cd_{il} = 1, cy_{ik} = 1|m_{i1},\ldots,m_{iT}) = \gamma_{0,dl} + \gamma_{0,yk} + \gamma_{0,ydkl} + \sum_{t=1}^{T}\gamma_{tl}m_{it}. \tag{7.20}$$

Here, $\gamma_{0,ydkl}$ capture the correlation between $\boldsymbol{cd}$ and $\boldsymbol{cy}$. Both latent classes influence the random effects. However, $\boldsymbol{cy}$ is not influenced by the missing-data indicators. So, $\boldsymbol{cy}$ represent different trajectories which are moderated by $\boldsymbol{cd}$ within each class of $\boldsymbol{cy}$. Hence,

$$\begin{array}{rcl} f(\boldsymbol{y}|\boldsymbol{m}) & = & \sum_{cy,cd} f(\boldsymbol{y},\boldsymbol{cy},\boldsymbol{cd}|\boldsymbol{m})\\ & = & \sum_{cy,cd} f(\boldsymbol{y}|\boldsymbol{cy},\boldsymbol{cd})f(\boldsymbol{cy},\boldsymbol{cd}|\boldsymbol{m})\\ & = & \sum_{cy,cd} f(\boldsymbol{y}|\boldsymbol{cy},\boldsymbol{cd})f(\boldsymbol{cd}|\boldsymbol{m})f(\boldsymbol{cy}).\end{array} \tag{7.21}$$

7.3.3.3 Extended Models

The AAA study contains useful auxiliary information. For each patient, it is known whether or not s/he underwent surgery. Symptomatic and large aneurysms are considered for repair by surgical methods when the aneurysm grows more than 1 cm per year or when it is bigger than 5.5 cm. When patients have an operation they are removed from the follow-up study. The models discussed in Sections 7.3.2 and 7.3.3 were extended incorporating this information. For this, a logistic regression model for the operational status was used. So, $u_i = 1$ for patients that underwent surgery and $u_i = 0$ for patients that did not have an operation. The probability to undergo surgery was allowed to be different for the classes. This sensitivity analysis, incorporating the auxiliary information in the models can help in choosing between different models fitted to the data.

7.4 Results for the AAA study

7.4.1 Missing Data in the AAA Study

Of the 100 patients included in the AAA study only 7 patients had complete data for all 7 visits, 10 dropped out after the first visit (at 6 months of follow-up), 16 after the second visit, 21 after the third visit, 7 after the fourth visit, 27 after visit five and finally 12 patients dropped out after visit 6 (Table 7.1). The diameter of the artery at visit j can be missing for a multitude of reasons: the patient was not yet $j \times 6$ months in the study, the patient did not show up at the visit for a reason not related to his/her condition, the patient underwent repair of the artery by surgery, etc. But it is clear that missingness can depend on the diameter of the artery. The percentage of dropout is very comparable among the patients that had surgery of the artery and those who did not (92% versus 93%). Figure 7.1 shows the mean diameter curve for the patients operated and not operated. The figure also includes the mean diameter among those who did and did not drop out at the next visit. Among the patients that had surgery, patients leaving the study early have smaller diameters (although the number of patients is small) and patients dropping out later in the study are the ones with larger artery diameter. For the group of patients not having surgery the mean artery diameter of patients dropping out varies with time in a non systematic way. The difference in sample mean at visit t between patients dropping out at visit $t+1$ and patients not dropping out is relatively small.

Figure 7.2 shows, per dropout pattern, the mean diameter curve. Patients not

Table 7.1: *Proportion of obtained measurements per visit — AAA Data.*

Visit	Time (years)	Proportion obtained measurements
1	0	1
2	0.5	0.90
3	1	0.74
4	1.5	0.53
5	2	0.46
6	2.5	0.19
7	3	0.07

dropping out or only at the last visit (visit 7) have smaller diameters than patients dropping out around visits 5 and 6. The results when applying the models discussed in Section 7.3 to the AAA data are presented in the next section. For all models, 1 to 4 classes are studied. For some models convergence was not obtained. Some Mplus sample code and information on computing time can be found in Section 7.7.

7.4.2 MAR model

For the missing at random model, the results are presented in Table 7.2. It is clear that the 1-class model is performing worse than the multi-class models. Based on the BIC the two-class model is preferable. The parameter estimates will be unbiased if MAR conditional on class holds, $\boldsymbol{c}$ does not influence the missingness, and ignorability holds.

Based on the intercept and slope factors (Table 7.6), these classes could be labeled: (1) ‘small diameter and slowly growing’, and (2) ‘large diameter and faster growing’. The variances of the intercepts are relatively large, indicating that even within a class there still is heterogeneity. This heterogeneity seems to be more pronounced in class 2. The same is seen in terms of the random slope. So, patients of class 2 differ more in their growth than patients of class 1. For both classes, the covariance between the random intercept and slope is not statistically significant, indicating that, conditionally on class-membership, the value of the diameter at the start of the study and the growth are independent. The two-class growth-mixture model solution estimated under MAR is shown in Figure 7.5. The entropy is 0.64. Forty percent of the patients are in the low class showing slow growth of the artery. Sixty percent are in the high class with faster growing arteries.

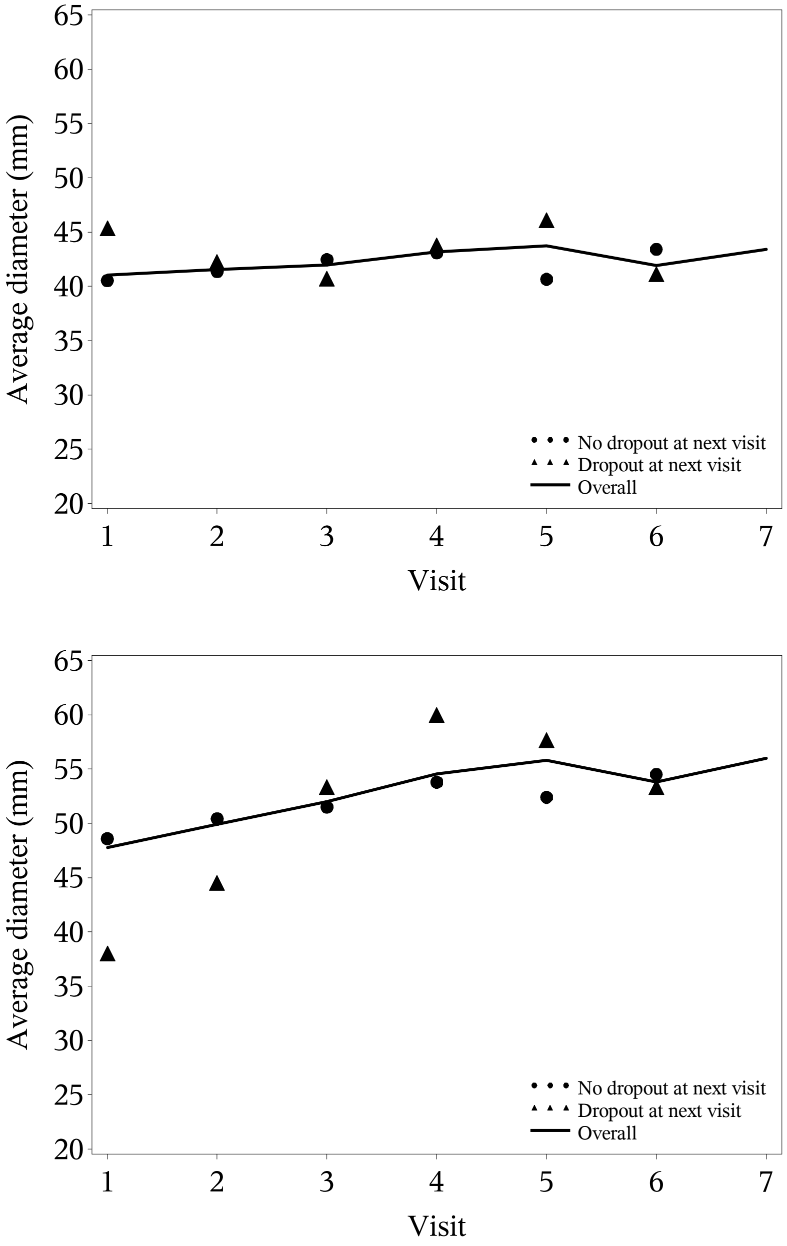


Figure 7.1: *Evolution of the (sample) average diameter for patients that were not operated (upper panel) and for patients that under went surgery (lower panel). At each visit the mean diameter is also presented according to dropout status at the next visit — AAA Data.*

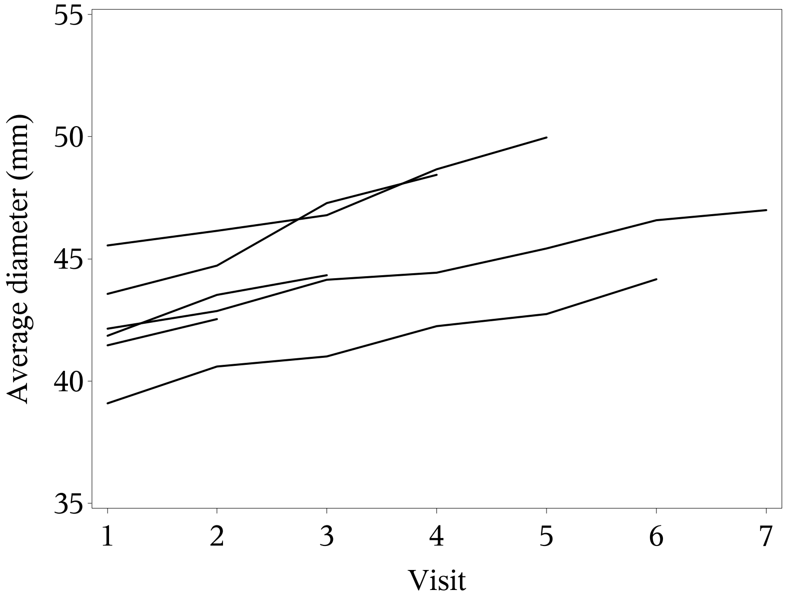


Figure 7.2: *Evolution of the (sample) average diameter per dropout pattern — AAA Data.*

7.4.3 MNAR Models

Results of the selection models are also presented in Table 7.2. Multi-class models were not always converging. The number of parameters in the model becomes large in comparison with the number of patients in the study. Based on the BIC statistic the 1-class Diggle-Kenward (DK) model is best. The two-class Beunckens model is comparable. The estimated mean trajectories for the two-class models are shown in Figure 7.5. The class of patients with small diameters is estimated to be 70% under the Diggle-Kenward Model (7.13), 43% under the Diggle-Kenward Model (7.14) and 66% in the Beunckens model.

Results of the pattern-mixture models are presented in Tables 7.2 and 7.3. In terms of BIC, the two and three class Roy and Roy-Muthén models outperform the traditional pattern-mixture model. The two-class Roy model is preferable. The estimated mean trajectories for the two-class Roy model are shown in Figure 7.5. The class of patients with small diameters is estimated to be 44% under the Roy Model. Table 7.6 presents for the different two-class models the parameter estimates for the

Table 7.2: *Summary of the MAR and MNAR models: number of classes, log-likelihood, number of parameters and BIC — AAA Data.*

Model	# classes	log(L)	# par.	BIC
MAR	1	-952.040	6	1932.039
MAR	2	-925.211	13	1910.289
MAR	3	-911.006	20	1914.116
MAR	4	-901.437	27	1927.214
Selection Models:				
Diggle-Kenward	1	-1132.386	14	2329.244
Diggle-Kenward Model (7.13)	2	-1128.814	29	2391.178
Diggle-Kenward Model (7.13)	3	-1090.499	44	2383.625
Diggle-Kenward Model (7.13)	4	-1069.043	59	2409.791
Diggle-Kenward Model (7.14)	2	-1105.308	41	2399.428
Diggle-Kenward Model (7.14)	3	-1087.660	62	2460.841
Beunckens Model (7.15)	2	-1122.674	19	2332.846
Beunckens Model (7.15)	3	-1107.441	29	2348.432
Pattern-Mixture Models:				
Pattern-mixture		-964.797	16	1967.277
Roy	2	-920.584	19	1928.666
Roy	3	-900.032	32	1947.430
Roy	4	-884.476	45	1976.184

Table 7.3: *Summary of the Roy-Muthén models: number of classes, log likelihood, number of parameters and BIC — AAA Data.*

Model	# $\boldsymbol{c}$y	# $\boldsymbol{c}$d	log(L)	# par.	BIC
Roy-Muthén	2	2	-917.823	22	1936.960
Roy-Muthén	3	2	-897.564	30	1933.282

measurement model.

7.4.4 Model Comparision

7.4.5 Posterior Probabilities

Figure 7.3 shows, for the 2-class models presented in Sections 7.4.2 and 7.4.3, the posterior probabilities to be classified into the group with the largest and fastest increasing diameters. The graph clearly shows that for a number of patients this probability can change substantially, depending on the dropout model that was chosen.

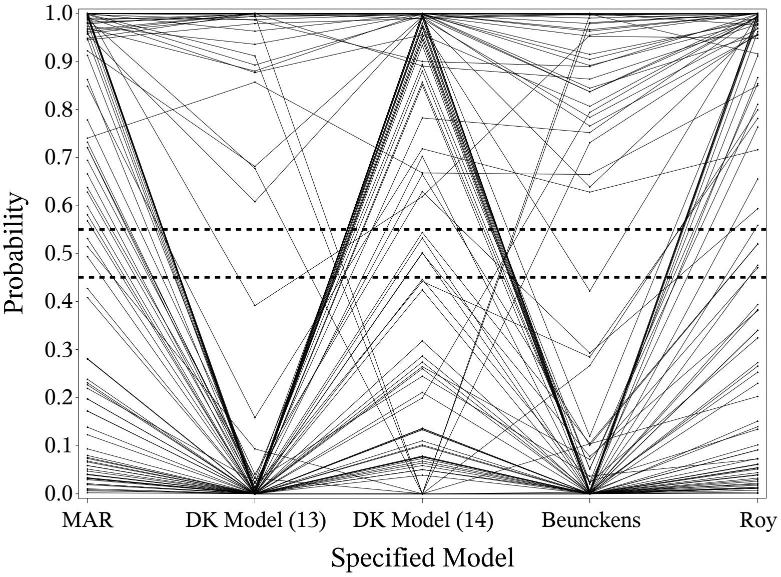


Figure 7.3: *Posterior probability to be classified into the group with the largest and fastest increasing diameters, for the two-class models presented in Table 7.6. Posterior probabilities between the dashed horizontal lines (at 0.45 and 0.55) indicate large uncertainty in classification — AAA Data.*

Table 7.4 shows the agreement in classification in low versus high class for the different models. The table presents the Kappa statistic and the percentage of patients classified in the same class (high or low). Cohen's Kappa coefficient is a statistical measure of inter-rater agreement for categorical items. This measure takes into account the agreement occurring by chance. According to Fleiss (1981), Kappa over .75 characterize excellent agreement, values between .40 to .75 indicate fair to good agreement, and Kappa below .40 point to poor agreement. These guidelines are, however,

Table 7.4: *Kappa statistic (κ) and percentage of patients in the same class for the two-class models presented in Table 7.6 — AAA Data.*

	Kappa statistic					% patients in same class				
	MAR	DK (7.13)	DK (7.14)	Beunckens	Roy	MAR	DK (7.13)	DK (7.14)	Beunckens	Roy
MAR	1.00	0.44	0.69	0.51	0.88	100	70	85	74	94
DK (7.13)		1.00	0.38	0.86	0.50		100	67	91	74
DK (7.14)			1.00	0.41	0.70			100	69	85
Beunckens				1.00	0.58				100	78
Roy					1.00					100

arbitrary. Based on these measures, we conclude that the classification according to MAR and Roy's model are very similar, also that Diggle-Kenward Model (7.13) and Beunckens' model result in very similar classifications of the patients. Figure 7.3 also reveals how certain the classification in the low or high group is. Posterior probabilities between 0.45 and 0.55 indicate that it is uncertain to which group the patient should be classified. Some patients have posterior probabilities in this region, and therefore it is expected that they can switch group when changing dropout model. However, for a number of patients, the classification according to MAR, Diggle-Kenward Model (7.14) and Roy's model is clearly into the high group (posterior probability > 0.6) where, according Diggle-Kenward Model (7.13) and Beunckens model, it is clearly in the low group (posterior probability < 0.6).

As a result of the susceptibility of the posterior probability to the chosen dropout model, it is seen that the number of patients classified in the low or high class, based on their most likely latent-class membership, strongly depends on the assumptions made. Under the MAR assumption 40% of the patients are classified in the lowest latent class, under the Diggle-Kenward Model (7.13) this increases to 70% (Table 7.5). Sixty patients out of a total of 100 are always classified into the same latent class, irrespective of the assumptions made for the missing-data mechanism. Using the classification resulting from the MAR model, we see that the effect of the chosen missing-data mechanism on the posterior probability is smaller for the patients in the low group as compared to the ones in the high group. Of the 60 patients belonging to the group with larger diameter, only 26 (43%) are always classified into the high group for the selection and pattern-mixture models under consideration. Patients classified in the low group almost always (34 out of 40, 85%) end up in the low group.

Table 7.5: *Class sizes, entropy, and average posterior probability for the two-class MAR and MNAR models — AAA Data.*

Model	Low Class	High Class	Entropy	Average post. prob.	
				Low Class	High Class
MAR	40	60	0.640	0.88	0.90
DK(7.13)	70	30	0.890	0.99	0.94
DK(7.14)	43	57	0.665	0.89	0.91
Beunckens (7.15)	66	34	0.787	0.97	0.88
Roy	44	56	0.662	0.87	0.92

Figure 7.4 graphically shows the relationship between the random slope and random intercept. Empirical Bayes estimates for the random effects were obtained for the two-class MAR model. The figure shows this relationship by dividing the patients into two groups. A group of patients that, irrespective of the chosen dropout model, is always classified into the high or low class and a second group of patients whose classification does depend on the assumptions made for the missing-data mechanism. For the patients always classified in the same latent class, i.e., always in the low or always in the high class, the Spearman correlation coefficient between the random slope and intercept equals 0.73 ($p<$ 0.001). For the patients whose classification depends on the chosen dropout model the correlation equals -0.17 (p=0.285). It is thus seen that the relation between a patients' intercept and slope goes in the reverse direction for the two groups. The second group consists of patients with small and fast growing arteries and patients with large but slowly growing arteries.

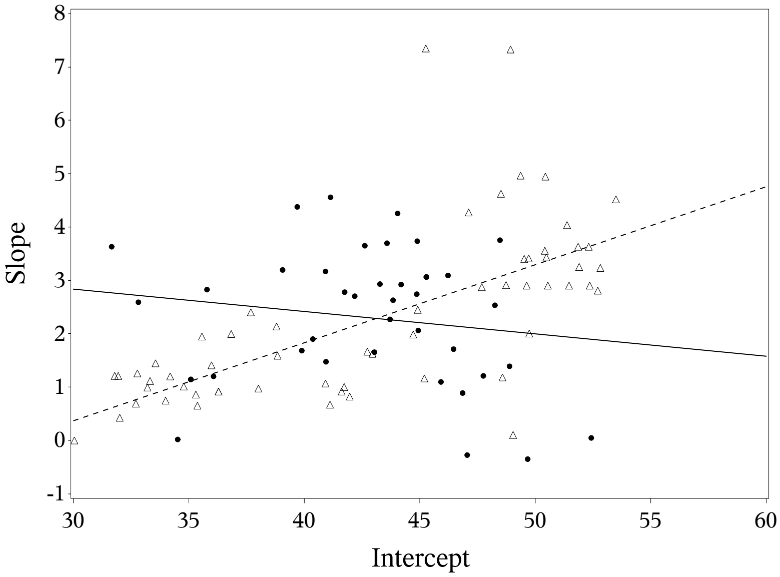


Figure 7.4: *Empirical Bayes estimates for random intercept and random slope, under the two-class MAR model. Patients always classified in same latent class are represented by triangles and dashed line, patients whose classification depends on the chosen dropout model are represented by dots and full line — AAA Data.*

Table 7.6: *Summary of the measurement part of the two-class MAR and MNAR models: class-specific parameter estimates (standard errors) are shown — AAA Data.*

			DK	**DK**			
		MAR	**(7.13)**	**(7.14)**	**Beunckens**	**Roy**	
E	ffect	Par.	Est. (s.e.)	Est. (s.e.)	Est. (s.e.)	Est. (s.e.)	Est. (s.e.)
					Class I: Low Class		
Measurement Model							
Fixed Effects:	Intercept	α_0	38.294(1.337)	39.819(0.701)	39.686(1.078)	39.512(0.865)	38.494(1.259)
	Slope	α_1	1.328(0.195)	2.113(0.250)	1.171(0.254)	1.859(0.236)	1.377(0.188)
Covariances	var(η_{0i})	d_{11}	23.138(6.601)	27.818(3.405)	35.669(6.263)	25.622(3.348)	23.294(5.518)
	var(η_{1i})	d_{22}	0.554(0.255)	1.383(0.451)	1.175(0.354)	1.173(0.527)	0.556(0.241)
	cov(η_{0i}, η_{1i})	d_{12}	1.301(1.093)	-1.120(1.171)	-1.761(1.752)	1.753(1.342)	1.548(1.040)
Residual variance	var(ε_{ij})	σ^2	0.601(0.082)	1.725(0.235)	1.204(0.256)	1.397(0.329)	0.629(0.098)
					Class II: High Class		
Measurement Model							
Fixed Effects	Intercept	α_0	46.002(0.974)	49.986(0.473)	45.299(1.060)	49.781(0.761)	46.057(1.081)
	Slope	α_1	2.917(0.395)	3.867(0.705)	2.427(0.378)	2.950(0.624)	2.931(0.400)
Covariances	var(η_{0i})	d_{11}	30.089(8.136)	3.232(2.048)	33.199(7.077)	4.538(2.184)	30.827(9.031)
	var(η_{1i})	d_{22}	4.388(1.393)	4.162(1.275)	3.046(1.288)	6.440(2.049)	4.541(1.410)
	cov(η_{0i}, η_{1i})	d_{12}	-0.091(1.986)	-0.097(1.410)	2.001(2.140)	-0.860(1.942)	-0.267(2.092)
Residual variance	var(ε_{ij})	σ^2	3.307(0.452)	3.691(0.710)	3.863(0.616)	3.195(0.552)	3.097(0.447)

7.4.6 Parameter Estimates

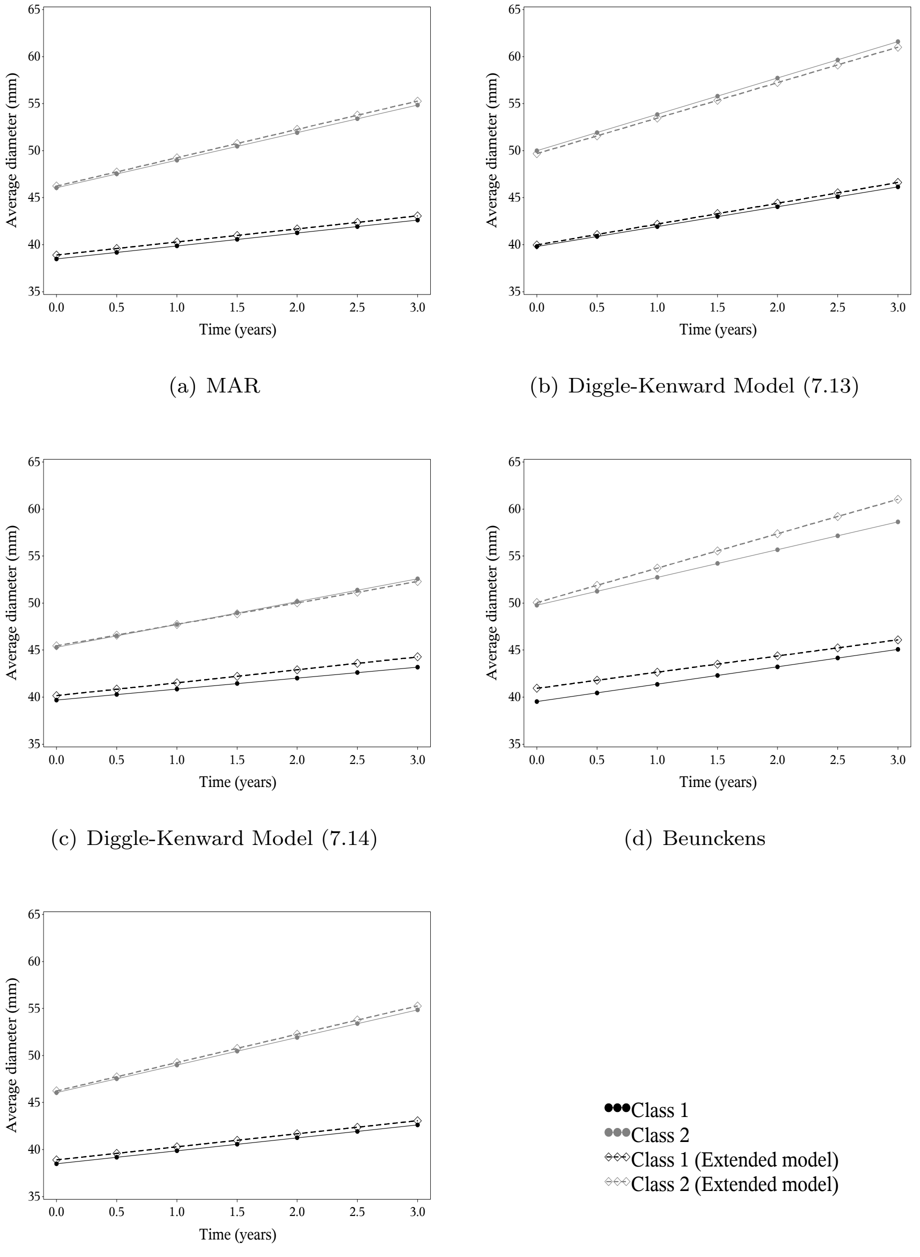
The choice of the dropout model also influences the parameter estimates for the fixed effects. Compared with the model assuming MAR, the largest effect on the intercept and slope is seen in Diggle-Kenward Model (7.13) and the Beunckens Model (Table 7.6). The differences in the fixed effects between the models is most pronounced for the latent class with largest and fastest increasing diameters. The class-specific mean evolution for the two latent classes, under the different missing-data mechanisms, are graphically presented in Figure 7.5. The variance of the random intercept for the high class in the Diggle-Kenward Model (7.13) and in the Beunckens model is much smaller then this variance component in the other models, indicating that these models result in more homogeneous latent classes in terms of the arteries diameter at the start of the study. The other variance components (variance of the random slope, covariance of random intercept and slope, and the residual variance) under the different models did not change substantially.

Choosing between these models is difficult. All these models contain unidentifiable parts, which rely completely on assumptions that are not testable from the observed data. The log-likelihoods of MAR, pattern-mixture and selection models cannot be compared. Only comparison of log-likelihoods and BIC within the same family is informative. The models can, however, be compared in terms of their predictive value for a distal event. This is studied in Section 7.4.7.

7.4.7 Extended Models

The two-class models for the AAA study discussed in Sections 7.4.2 and 7.4.3 were extended with a logistic regression model for the operational status. So, u_i=1 for patients that underwent surgery and u_i=0 for patients that did not have an operation. Overall, 26% of the patients underwent surgery. This probability was allowed to be different for the 2 classes by adding one additional parameter to the models.

Figure 7.5 shows the mean curves for the two-class models as discussed in Sections 7.4.2 and 7.4.3. The trajectories estimated under the original model and the extended versions, including the distal event, are presented in the same graph. In general, the extended versions retain the trajectory shapes. The Beunckens model performs worse. For this model, it is also noted that the number of patients classified in the highest class decreases from 34% under the original model to 20% under the extended version (Table 7.9). With the exception of the Diggle-Kenward Model (7.13), the proportion of patients classified in the high class is lower for the extended model



(a) MAR

(b) Diggle-Kenward Model (7.13)

(c) Diggle-Kenward Model (7.14)

(d) Beunckens

(e) Roy

(f) Legend

Figure 7.5: *Estimated average trajectories under the MAR and MNAR two-class models (original model and the extended version to include a distal event) — AAA Data.*

Table 7.7: *Class sizes, entropy and average posterior probabilities for the extended two-class MAR and MNAR models — AAA Data.*

Model	Low Class	High Class	Entropy	Average post. prob.	
				Low Class	High Class
MAR	44	56	0.665	0.87	0.91
Diggle-Kenward Model (7.13)	70	30	0.889	0.99	0.92
Diggle-Kenward Model (7.14)	53	47	0.733	0.89	0.96
Beunckens Model (7.15)	80	20	0.884	0.98	0.96
Roy	48	52	0.718	0.90	0.94

as compared with the original model. Table 7.8 shows the transitions between low and high classes for the original and extended models. The percentage of patients that underwent surgery is also given. As noted before, transition from the low to the high class is not as frequent as the opposite transition. The few transitions from the low class in the original model to the high class in the extended model occur for patients that underwent surgery. None of the patients being classified in the high class in the original model and in the low class in the extended model, except one, underwent an operation.

Table 7.8: *Number of patients with transitions; percentage of operated patients; κ statistic and percentage of patient classified in the same class measuring the agreement in classification for the original (Table 7.6) and extended model — AAA Data.*

		Extended Model				κ statistic	% patients in
		Low class		High class		[95% CI])	same class
		N	% operated	N	% operated		
Original Model							
MAR	Low	36	2.8	4	100.0	0.754 [0.624; 0.884]	88
	High	8	0.0	52	42.3		
DK Model (7.13)	Low	68	7.4	2	100.0	0.905 [0.813; 0.996]	96
	High	2	0.0	28	71.4		
DK Model (7.14)	Low	38	0.0	5	100.0	0.603 [0.451; 0.756]	80
	High	15	0.0	42	52.4		
Beunkens	Low	66	12.1	0	-	0.654 [0.496; 0.811]	86
	High	14	7.1	20	90.0		
Roy	Low	43	7.0	1	100.0	0.879 [0.786; 0.973]	94
	High	5	0.0	51	45.10		

Table 7.8 also displays Kappa statistics and the percentage of patients classified in the same class, measuring the agreement between the classification according to the original and extended versions of the MAR and MNAR models. Diggle-Kenward Model (7.13) and Roy's model have κ values of about 0.9 and more than 90% of patients are classified in the same class under the original and extended model, pointing to very good agreement. The κ values for the other models are not too bad, but clearly Diggle-Kenward Model (7.13) and Roy's model stand out. This shows that the original Diggle-Kenward Model (7.13) and Roy's model are best capturing the potential for non-random missingness.

The predictive power of these extended models is presented in Table 7.9. For all models, the lowest class is clearly a group of patients not needing surgery. In terms of the probability to undergo surgery, for the patients being classified in the high class, Diggle-Kenward Model (7.13) and Beunckens' models perform best. When classified in the high class according to the Beunckens' extended model, the probability the patient needs surgery is 86%. In summary, for Diggle-Kenward Model (7.13) we see that the shape of the latent trajectories is kept, that the proportion of patients classified in the latent classes stays the same for the original and extended model and that this model has reasonably good predictive power for the distal event (Table 7.9). Almost three out of four patients classified in the high class underwent surgery, whereas only 8 percent of the patients from the low class underwent surgery. The entropy of this models equals 0.89, indicating that the two latent classes are well separated. Therefore we conclude that the Diggle-Kenward Model (7.13) likely best captured the missingness mechanism.

Table 7.9: *Estimated probability to undergo surgery given latent-class membership under two class MAR and MNAR models — AAA data.*

	Low Class	High Class
Extended Model	$p(u=1)$	$p(u=1)$
MAR	0.05	0.44
Diggle-Kenward Model (7.13)	0.08	0.74
Diggle-Kenward Model (7.14)	<0.01	0.52
Beunckens Model (7.15)	0.11	0.86
Roy	0.06	0.45

7.5 Discussion and Conclusion

In this chapter, a number of models for non-random missingness were fitted to the abdominal aorta aneurysm data. Similarities in terms of the trajectory shapes are seen, irrespective of the model chosen for dropout. The division of patients in two latent classes, however, changed substantially. This phenomenon has been studied by a number of authors in specific data and model settings.

Verbeke and Molenberghs (2000) applied selection models of the Diggle-Kenward type to the so-called toenail dermatophyte onychomycosis (TDO) data. They considered a measurement model including a serial component. This model was expanded to also include measurement error. Both selection models and a MAR model result in very comparable estimated mean profiles and non-significant treatment effects. The sensitivity of the selection model is, however, seen in the estimated correlation matrix. Inclusion of measurement error results in higher estimated correlations, as compared to only including serial correlation. This directly has an effect on the dropout model. The current observation is less needed for predicting dropout, once the previous observation is known. This analysis clearly shows that finding an appropriate covariance model is important when dropout is present.

Kenward (1998) presented maximum likelihood estimates of random and non-random dropout models, deleting a different set of influential observations from the Mastitis in Dairy Cattle data. He noted that the influence of the deletion scheme on the measurement model parameters is small. The estimates for the dropout model, however, changed substantially.

A unifying framework was presented by Verbeke and Molenberghs (2010), bringing together the frameworks of coarse data (e.g., missing data, censoring, grouping) and augmented data (e.g., latent classes, latent variables, random effects, mixture model membership). The term coarse data is reserved for settings where the observed data are less refined than what ideally might have been observed. The authors refer to augmented data when convenience or interpretation-enhancing structures are added to the data, without being observable. These authors term this unification as data enrichment and showed that every model for enriched-data settings can be factored as a product of two components. The first one, the marginal model, is fully identifiable from the observed data. The second one, the predictive model, i.e., the conditional distribution of the enriched data given the observed data cannot be identified from the data without additional, non-verifiable model assumptions. To a large extent, inferences purely in terms of the marginal models (e.g., treatment effects), are unaffected by enrichment, whereas others, such as empirical Bayes estimates or conclusions based on latent-class membership, strongly depend on assumptions made.

This stresses the need for sensitivity analysis. Rather than fitting a single model and putting belief into it, it is advisable to consider a set of alternative models and asses how vulnerable the results are to the choices made. This has been done for the AAA study in this chapter. However when results turn out to be sensitive to the assumptions made (e.g., with respect to the missing-data mechanism) it is hard to choose between these models since they all rely on non-verifiable assumptions.

In this chapter, model validation was done by incorporating a so-called distal event in the model. A distal event is an ultimate outcome that relates to the growth curves studied. The relationship between the latent-class membership and the distal event provides a predictive validity check of the latent classes. Furthermore, the congruence between the latent class formation when not including versus including the distal event in the model is a measure for the ability of the original model to capture non-ignorable missingness.

For the AAA study, the fact whether or not a patient had surgery of the artery was used as auxiliary information in the model. We conclude that the Diggle-Kenward Model (7.13) is preferable. The model has good predictive power for a patients' need to undergo surgery. The trajectories described by the original and extended model Diggle-Kenward Model (7.13) are very similar; and they result in the same classification of the patients. Given these results we conclude that the Diggle-Kenward Model (7.13) picked up best the missingness mechanism.

7.6 Addendum 1: Performance of Information Criteria in the Presence of Missing Data

In this chapter, the decision on the optimal number of components in the mixture model was based on the BIC. However, deciding on the number of components in a mixture model is a difficult task, which has not yet been completely resolved. Often, the problem of determining the optimal value for K is separated from estimating the parameters of the component distributions for a fixed K. Two main approaches are being used to decide on the order of a mixture model. A likelihood-ratio statistic can be used to test the null hypothesis $H_0 : K = k_0$ versus $H_1 : K = k_1$ for some $k_1 > k_0$. Unfortunately, in the case of mixture models, regularity conditions do not hold for the likelihood-ratio test (LRT) statistic, and the asymptotic null distribution is no longer a χ^2-distribution. Lo, Mendell and Rubin (2001) approximated the LRT-distribution and McLachlan and Peel (2000) proposed a re-sampling approach to the assessment of the p-value of the LRT.

The second approach - and most often used in applied research - to estimate the order of a mixture model is based on penalized information criteria. The likelihood increases with the addition of components to the mixture model and therefore needs to be corrected by a term to penalize for model complexity (the number of estimated parameters in the model). A wide variety of such criteria exists, but the Akaike information criterion (AIC, Akaike, 1974), the Bayesian information criterion (BIC, Schwarz, 1978) and the sample-size adjusted BIC (aBIC, Sclove, 1987) are among the ones most commenly used. Although AIC and (a)BIC both are penalized goodness-of-fit criteria, the motivation behind them is different.

Akiake's Information Criterion selects the model that minimizes

$$-2\text{log}L(\hat{\boldsymbol{\theta}}) + 2p$$

where p is equal to the total number of parameters in the model. This criterion is based on the Kullback-Leibler (1951) discrepancy between the true density $f(\boldsymbol{y})$ and its modelled approximation $f(\boldsymbol{y};\hat{\boldsymbol{\theta}})$.

The Bayesian Information Criterion selects the model that minimizes

$$-2\text{log}L(\hat{\boldsymbol{\theta}}) + p\text{log}(N).$$

The Bayesian information criterion is derived within a Bayesian framework for model selection. The penalty for model complexity increases as the sample size increases. Sclove (1987) proposed to replace N by $(N+2)/24$. Adding the sample size correction reduces the sample size penalty, and should lead to better performance in case of either a large number of parameters or small sample size (Yang, 2006). Thus the sample size adjusted BIC, aBIC, equals:

$$-2\text{log}L(\hat{\boldsymbol{\theta}}) + p\text{log}\left(\frac{N+2}{24}\right).$$

BIC is consistent (Haughton, 1988), so it tends to select the correct model more frequently when sample size increases. AIC on the other hand is not consistent (Woodruffe, 1982). AIC is an asymptotically efficient model selection criterion. This means that for $N \to \infty$, with probability approaching one, the model with the minimum AIC score will also possess the smallest Kullback-Leibler divergence.

The performance of the AIC and BIC criterion have been investigated in the mixture context. AIC tends to overestimate the correct number of components (Soromenho, 1993; Celeux and Sormenho, 1996). Fraley and Raftery (1998) note that there is considerable support for the use of BIC in the mixture situation.

Although these model selection criteria have been evaluated by many researchers, less is known about the performance in a missing-data setting. Model selection criteria typically depend on the likelihood function based on the observed data. For missing-data problems this observed data likelihood function can be complicated and without a closed form. Hence, it is challenging to get an accurate approximation of the observed data likelihood and to compute AIC and BIC. Ibrahim et al. (2008) present a general class of information criteria for missing-data problems, which yields the Akaike and Bayesian information criteria as special cases. Cavanaugh and Shumway (1998) introduce an Akaike information criterion for model selection in the presence of incomplete data.

Ibrahim et al. use the fact that the observed data log-likelihood can be expressed as the difference between two functions, the Q-function of the EM algorithm and a quantity called the H-function. The H-function can be analytically approximated, and can then be computed as part of the EM output. The resulting criteria $IC_{\tilde{H}(k),Q}$ depend solely on the EM output. Cavanaugh and Shumway develop an AIC criterion based on the complete data Kullback-Leibler discrepancy, instead of the discrepancy based on the incomplete data. They show that the complete-data discrepancy is potentially more sensitive than the incomplete-data discrepancy to deviations from the true $\boldsymbol{\theta}$.

In this section we will describe the performance of the traditional information criteria AIC, aBIC, and BIC in a modest simulation exercise.

7.6.1 Settings for the Simulations

Data were generated for 4 different settings, each consisting of a mixture of two latent classes with equal class probability. Repeated measurements at four time points were generated assuming linear individual trajectories. Residual variances were normally distributed, homoscedastic, uncorrelated and class invariant. Also the residual variances of the growth factors (random intercept and slope) were specified to be normally distributed, invariant across the latent classes, and with a zero covariance between the growth factors. The values specified for the growth parameters and their variances are presented in Table 7.10. The trajectory for class 1 represents an increasing trend over time, class 2 represents a zero growth class. The standardized between-class differences for the intercept and slope is largest for setting 1. The standardized between-class differences are the same for settings 3 and 4, but the percentage of explained variance by the growth factor is largest for setting 3 (80% versus 60%). The sample size, for each simulated data set, was fixed to be either 100 or 500. For each setting, 500 data sets were generated.

Table 7.10: *Growth Mixture Model Specifications: Growth Factors.*

	Class 1		Class 2		
	Intercept	Slope	Intercept	Slope	Residual Variance
Setting	(variance)	(variance)	(variance)	(variance)	
1	100	4.7	89	0	10
	(20)	(5)	(20)	(50)	
2	100	4.7	89	0	15
	(50)	(10)	(50)	(10)	
3	100	3.16	93	0	15
	(50)	(10)	(50)	(10)	
4	100	3.16	93	0	40
	(50)	(10)	(50)	(10)	

For the follow-up visits, missing observations were introduced according to the following models:

1. $\text{logit}(m_{it} = 1) = \gamma$;

2. $\text{logit}(m_{it} = 1) = \gamma + \delta y_{it}$.

Model 1 corresponds to MCAR and Model 2 will result in data to be MNAR. Values used for γ and δ are given in Table 7.11. Figure 7.6 displays the amount of missing data, introduced by the models specified above, for each of the considered settings. The MCAR mechanism results in data such that the amount of missing data in the follow-up measurements is constant and about equal to 27%. The MNAR setting specifies the probability for a missing observation to depend on the unobserved value of y_{it}. Under MNAR-1 the model introducing missing data is common for class 1 and 2. Setting MNAR-2, specifies a different missing-data model for class 1 and 2. For class 1, the probability for a missing observation is larger for large values of y, for class 2 the probability for an observation to be missing decreases as a function of y. For the zero-growth class (i.e., class 2) the proportion of missingness is constant over time. For class 1, the proportion of missingness increases over time (given the positive growth and positive slope parameter in the MNAR model). Figure 7.7 displays the

Table 7.11: *Growth Mixture Model Specifications: Missing Data Model.*

	Class 1		Class 2	
Setting	γ	δ	γ	δ
MCAR	-1	0	-1	0
MNAR - 1	-12	0.1	-12	0.1
MNAR - 2	-12	0.1	7	-0.1

observed average profiles and a ± two standard deviations interval under the different settings.

7.6.2 Performance of the IC

Data generation and model estimation were done with the Monte Carlo facilities in Mplus. A maximum of 5000 iterations was used for convergence of the mixture models. Nonconvergence occurred because of a singular information matrix. Starting values were chosen to be close to the values specified to generate the data. Convergence rates are given in Table 7.12.

We looked at the performance of the information criteria for one- through three-class models and identified the model were the lowest value was observed. The settings with non-ignorable missing data (MNAR-1 and MNAR-2) were analysed twice. Once ignoring the missing-data process, i.e., assuming MAR under ignorability, and once entertaining the selection model that was used to generate the data. The summary is displayed in Tables 7.13 and 7.14. For example, for the complete data sets, generated under setting 3 with a sample size of 500, we note that BIC identifies the correct number of classes 40% of the times; aBIC and AIC identify the correct number of classes about 80% of the times.

Although this simulation study is too limited to overly generalize, a few results are worth mentioning.

We first consider the situation with a sample size of N=500. Not surprisingly, for the complete data settings the information criteria behave as expected based on theory and previous simulation studies. For settings 1 and 2, i.e., the settings considering reasonably well separated classes, all information criteria are able to identify the correct number of classes (i.e., 2). More overlap between the classes is present in settings 3 and 4, and as a result the rate at which the correct number of classes is identified decreases. In general the penalty term in BIC is larger than in AIC (for $N > 8$), therefore BIC tends to underestimate the number of classes and AIC tends to overestimate.

For settings 1 and 2, the performance of the IC, when 20% of the follow-up data is missing completely at random, is comparable with the performance under the complete data setting. When the overlap between the clusters is larger (settings 3 and 4), the loss of information is reflected in a slight reduction in the performance of the IC (as seen by the more frequent underestimation of the number of clusters).

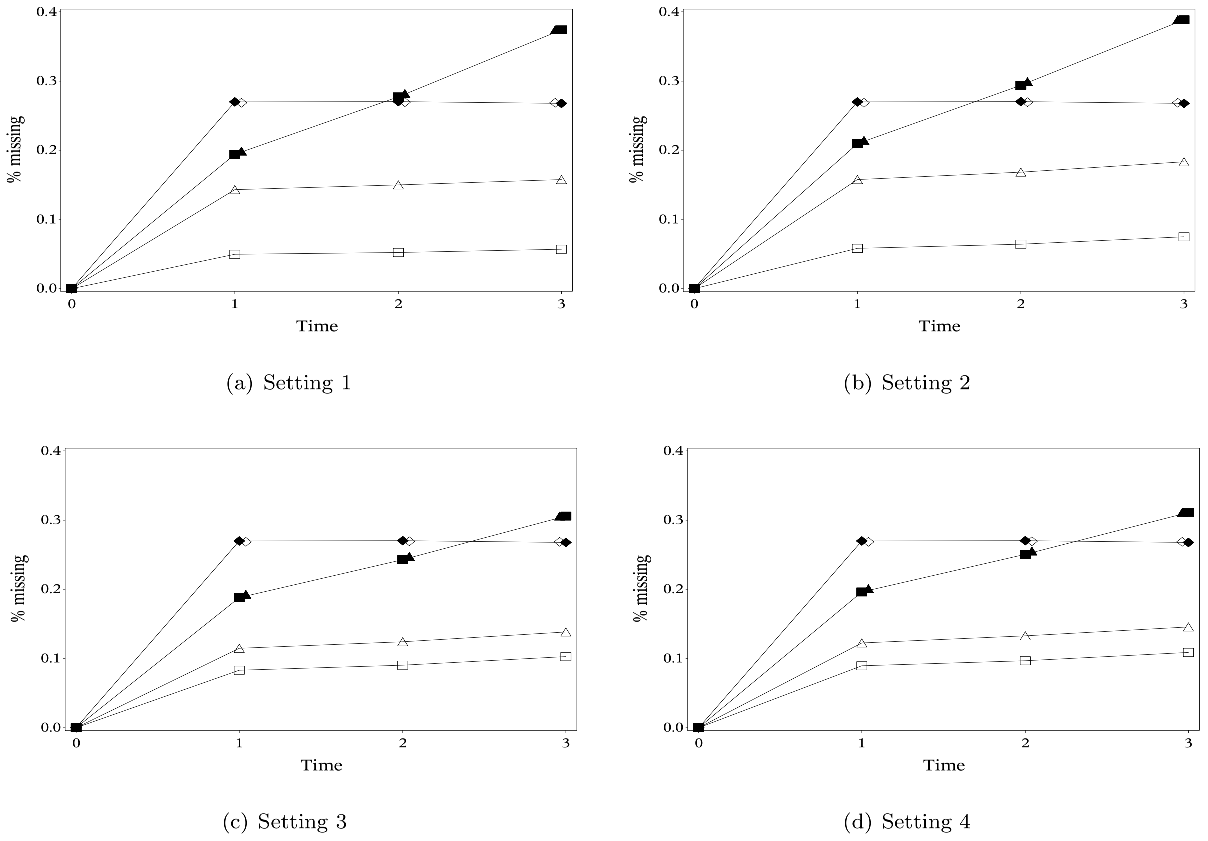
When considering the settings with non-ignorable missing data (MNAR-1 and MNAR-2), analysed by means of a selection model, we see that BIC is performing best at identifying the true number of latent classes, but also aBIC and AIC perform reasonably well.

For well separated data settings (settings 1 and 2), the MAR analysis under ignorability for the MNAR generated data seems to identify the correct number of classes. But since missingness was generated to be non-ignorable, the obtained estimates for the cluster specific growth factors are biased. Table 7.15 diplays the coverage rates for all parameters under the different settings. Data generated under setting 3, MNAR-1, and $N = 500$ results in a coverage rate of 41% for the slope parameter of class 1 when missingeness is assumed to be ignorable. When a selection model is assumed during clustering, the coverage rate turns out to be 93.4%.

Similar conclusions can be drawn for a sample size of 100. When the classes are well separated, BIC performs best, while for settings 3 and 4 AIC and aBIC more frequently select the correct number of classes. Missing data that is missing completely at random results in a reduced performance of the information criteria for settings 3 and 4 (i.e., more overlap between the classes). When data is MNAR, all information criteria perform well in identifying the correct number of classes by means of a selection model. Results from a MAR analyses results in biased estimates for the class-specific slopes.

In this inelaborate simulation study, simplifying assumptions were made and a limited range of settings was considered. Linear profiles over times were assumed for two equally sized latent classes. Residual variances and the variance-covariance of the growth factors were specified to be class-invariant. The missing-data model considered is very simple and more elaborate models can easily be imagined. In fact, this simulation study investigates the performance of the information criteria for incomplete data settings, to correctly identify the number of latent classes, given that the correct data and missing-data models are used.

The results should thus be interpreted very cautiously.



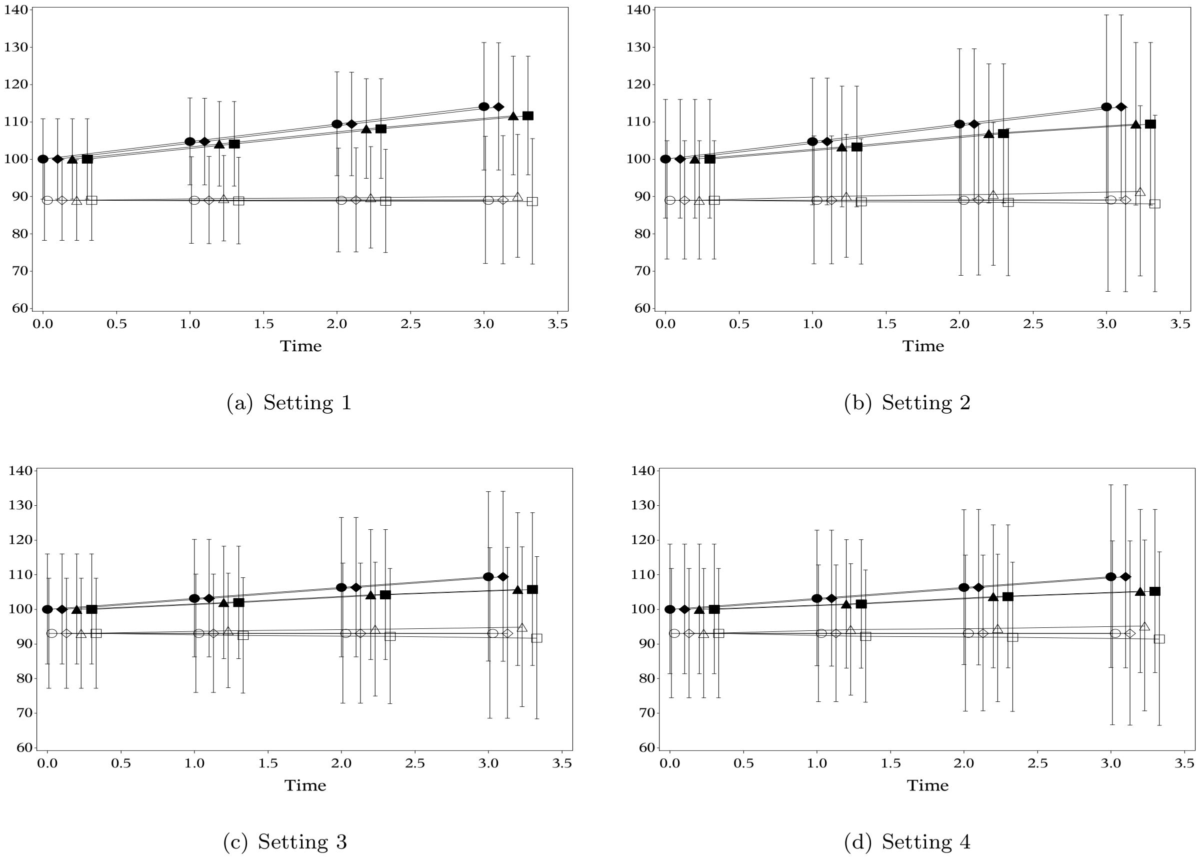
(a) Setting 1

(b) Setting 2

(c) Setting 3

(d) Setting 4

Figure 7.6: *Evolution of cluster specific proportions of missing data, for the different settings: ◇ MCAR, □ MNAR-1, △ MNAR-2. Filled symbols are used for class 1 and open symbols for class 2.*



(a) Setting 1

(b) Setting 2

(c) Setting 3

(d) Setting 4

Figure 7.7: *Plot of cluster specific average profiles and ± 2 standard deviations: ○ complete data, ◊ MCAR, □ MNAR-1, △ MNAR-2. Filled symbols are used for class 1 and open symbols for class 2.*

Table 7.12: Convergence rates (number out of 500 runs).

	Data	**Complete**	**MCAR**	**MNAR - 1**		**MNAR - 2**	
# classes	**Model**			**MAR**	**Selection**	**MAR**	Selection
			Setting 1 - N=500				
1		500	500	500	500	500	500
2		500	500	50	500	500	500
3		489	495	488	497	493	476
			Setting 2 - N=500				
1		500	500	500	500	500	500
2		500	500	500	500	500	500
3		495	498	499	500	496	474
			Setting 3 - N=500				
1		500	500	500	500	500	500
2		500	500	500	500	500	500
3		490	485	490	483	481	476
			Setting 4 - N=500				
1		500	500	500	500	500	500
2		500	500	500	500	500	500
3		493	483	485	487	482	473
			Setting 1 - N=100				
1		500	500	500	500	500	500
2		500	500	50	500	500	500
3		496	497	497	500	499	472
			Setting 2 - N=100				
1		500	500	500	500	500	500
2		500	500	500	500	500	500
3		500	499	500	500	496	478
			Setting 3 - N=100				
1		500	500	500	500	500	500
2		500	499	500	500	500	500
3		490	481	480	495	479	466
			Setting 4 - N=100				
1		500	500	500	500	500	500
2		500	500	500	500	500	498
3		489	485	481	487	474	469

Table 7.13: *Percentage of Times that the Lowest Value Occurred in each Class Model for BIC, aBIC and AIC, N= 500*

Data	**Complete**			**MCAR**			**MNAR - 1**						**MNAR - 2**					
Model							**MAR**			**Selection**			**MAR**			**Selection**		
# Classes	1	**2**	3	1	**2**	3	1	**2**	3	1	**2**	3	1	**2**	3	1	**2**	3
Setting 1																		
BIC	0	**100**	0	0	**100**	0	0	**100**	0	0	**100**	0	0	**100**	0	0	**74**	26
aBIC	0	**96**	4	0	**95**	5	0	**95**	5	0	**93**	7	0	**96**	4	0	**67**	33
AIC	0	**84**	16	0	**81**	19	0	**82**	18	0	**79**	21	0	**83**	17	0	**59**	41
Setting 2																		
BIC	0	**100**	0	0	**100**	0	1	**99**	0	0	**100**	0	6	**94**	0	0	**100**	0
aBIC	0	**94**	6	0	**94**	6	0	**93**	7	0	**90**	10	0	**91**	9	0	**93**	7
AIC	0	**84**	16	0	**81**	19	0	**82**	18	0	**76**	24	0	**79**	21	0	**77**	23
Setting 3																		
BIC	60	**40**	0	71	**29**	0	89	**11**	0	27	**73**	0	95	**5**	0	0	**100**	0
aBIC	12	**83**	5	18	**76**	6	43	**54**	3	4	**89**	7	60	**37**	3	0	**93**	7
AIC	4	**80**	16	6	**75**	19	20	**64**	16	1	**84**	15	34	**53**	13	0	**78**	22
Setting 4																		
BIC	89	**11**	0	93	**7**	0	99	**1**	0	75	**25**	0	100	**0**	0	1	**99**	0
aBIC	39	**56**	5	50	**45**	5	78	**20**	2	23	**71**	6	90	**9**	1	0	**95**	5
AIC	19	**66**	16	23	**60**	17	48	**37**	15	9	**75**	16	68	**22**	10	0	**78**	22

Table 7.14: *Percentage of Times that the Lowest Value Occurred in each Class Model for BIC, aBIC and AIC, N= 100*

Data	**Complete**			**MCAR**			**MNAR - 1**						**MNAR - 2**					
Model							**MAR**			**Selection**			**MAR**			**Selection**		
# Classes	1	**2**	3	1	**2**	3	1	**2**	3	1	**2**	3	1	**2**	3	1	**2**	3
Setting 1																		
BIC	0	**99**	1	0	**99**	1	0	**98**	2	0	**59**	41	2	**97**	1	0	**100**	0
aBIC	0	**65**	35	0	**64**	36	0	**64**	36	0	**51**	49	0	**65**	35	0	**62**	38
AIC	0	**81**	19	0	**80**	20	0	**79**	21	0	**52**	48	0	**79**	21	0	**80**	20
Setting 2																		
BIC	45	**55**	0	55	**44**	1	69	**31**	0	35	**64**	1	78	**22**	0	10	**88**	2
aBIC	2	**58**	40	3	**60**	37	7	**57**	36	3	**61**	36	10	**54**	36	0	**61**	39
AIC	6	**71**	23	7	**69**	24	15	**64**	21	5	**73**	22	24	**58**	17	0	**78**	22
Setting 3																		
BIC	95	**5**	0	95	**5**	0	97	**3**	0	60	**10**	30	98	**2**	0	46	**54**	0
aBIC	28	**41**	31	29	**38**	33	37	**32**	31	14	**45**	41	42	**29**	29	2	**59**	39
AIC	48	**37**	15	53	**32**	15	63	**25**	12	23	**39**	39	67	**22**	12	4	**73**	23
Setting 4																		
BIC	99	**1**	0	98	**2**	0	99	**1**	0	69	**3**	28	99	**1**	0	75	**25**	0
aBIC	38	**34**	28	42	**28**	30	48	**26**	26	27	**31**	42	46	**28**	26	4	**56**	40
AIC	60	**27**	13	60	**25**	15	67	**21**	11	40	**25**	35	70	**20**	10	12	**69**	19

Table 7.15: *Coverage Rate for the Intercept and Slope in the Two-class Models, N=500*

Data	**Compleet**		**MCAR**		**MNAR - 1**				**MNAR - 2**			
Model					**MAR**		**Selection**		**MAR**		**Selection**	
	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
	Setting 1											
intercept	95.0	94.2	94.6	94.6	95.0	94.4	95.4	94.8	95.4	94.8	94.2	95.4
slope	95.4	95.2	95.0	94.0	51.0	92.6	94.8	93.5	53.6	87.2	93.8	94.4
	Setting 2											
intercept	93.0	92.4	93.2	92.8	90.0	88.6	93.2	92.6	88.6	88.2	94.2	91.6
slope	91.6	92.8	91.6	90.8	41.0	87.2	93.4	92.6	49.4	76.4	91.0	91.0
	Setting 3											
intercept	81.4	82.8	81.0	82.4	72.6	79.2	84.6	87.4	72.6	74.0	88.2	88.0
slope	79.8	84.6	78.8	80.0	49.8	72.4	85.8	84.2	53.8	64.2	86.0	85.8
	Setting 4											
intercept	79.6	81.0	80.0	81.0	70.6	73.4	81.8	82.2	74.0	71.8	87.2	86.2
slope	80.6	77.2	77.4	76.6	37.0	78.6	81.6	79.6	38.8	52.8	85.0	87.2

Table 7.16: *Coverage Rate for the Intercept and Slope in the Two-class Models, N=100*

Data	**Compleet**		**MCAR**		**MNAR - 1**				**MNAR - 2**			
Model					**MAR**		**Selection**		**MAR**		**Selection**	
	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
	Setting 1											
intercept	86.8	86.8	87.2	87.6	83.4	85.8	87.0	87.2	85.6	84.8	85.6	85.6
slope	89.0	89.8	88.2	88.4	72.6	83.6	88.6	86.0	70.8	80.6	86.4	85.8
	Setting 2											
intercept	79.4	77.2	81.8	76.4	78.2	75.8	82.8	79.0	77.2	77.0	80.0	78.6
slope	79.6	82.4	79.8	81.2	68.0	77.2	81.0	78.4	69.6	72.2	78.4	77.4
	Setting 3											
intercept	79.6	77.0	77.8	76.8	78.4	74.6	79.4	79.8	77.0	74.6	76.5	78.7
slope	81.2	82.8	81.4	80.4	57.6	75.8	82.8	80.2	58.0	67.8	82.3	79.1
	Setting 4											
intercept	94.2	92.8	93.0	95.0	92.2	93.6	91.8	93.8	92.0	94.0	92.4	93.4
slope	92.6	95.2	93.4	93.2	83.0	91.4	93.8	92.6	82.6	90.6	92.6	91.8

7.7 Addendum 2: Mplus Sample Code

This appendix presents Mplus data format and Mplus commands to execute the mixture models for the AAA study discussed in see Tables 7.2 and 7.6.

7.7.1 Data Format

Mplus data files are ASCII text files. The example below contains the first 5 patients of the AAA study (AAA.dat). Each row contains the information for a patient, i.e, the diameter at the seven visits and a missing-data indicator. The values for each of the variables are separated by a delimiter (here a blank was used) and the variable names are not on the first line. Missing observations were denoted by a $*$.

```
41    43    43    43    43    44    44     0     0     0     0     0     0
41    41    43    42    43    43    42     0     0     0     0     0     0
33    32    37    36    36    41    *      0     0     0     0     0     1
45    46    48    52    52    *     *      0     0     0     0     1     1
32    32    33    34    35    34    36     0     0     0     0     0     0
```

7.7.2 Mplus Programs

Mplus command input files (.inp) for the two-class models presented in Tables 7.2 and 7.6 are given in the following sections.

7.7.2.1 MAR Model

```
TITLE: AAA DATA : MODEL TABLE 6.2 - MAR WITH 2 CLASSES
  !provide title for analysis

 DATA: FILE = C:\AAASTUDY\AAA.DAT;
  !datafile to be used

 VARIABLE:
     NAMES        ARE  Y1 Y2 Y3 Y4 Y5 Y6 Y7
                          U2 U3 U4 U5 U6 U7;
      !assign names to variables in the data set
     USEVARIABLES ARE  Y1 Y2 Y3 Y4 Y5 Y6 Y7;
      !specifies which variables to use in the analysis
     MISSING= * ;

     CLASSES  = C(2);
      !assigns a name to the latent categorical variable in the model,
      !and the number of components in the mixture

 ANALYSIS:
     TYPE=MIXTURE;
      !specifies a mixture model for the analysis
```

```
    STARTS= 50 20;
     !50 initial stage random sets of starting values and
     !20 final stage optimizations are used
    COVERAGE = 0.05;
     !mimimum acceptable proportion of cases that contribute
     !in calculation of variance/covariance

MODEL:
          %OVERALL%
          !describes part of the model that is common to all components
          I S  | Y1@0 Y2@0.5 Y3@1 Y4@1.5 Y5@2 Y6@2.5 Y7@3;

          %C#1%
          !component specific specifications
          Y1 Y2 Y3 Y4 Y5 Y6 Y7 (VARRES1);
          I (I1); S (S1); I WITH S (IS1);

          %C#2%
          Y1 Y2 Y3 Y4 Y5 Y6 Y7 (VARRES2);
          I (I2); S (S2); I WITH S (IS2);

SAVE:    FILE IS C:\AAASTUDY\MAR2_CP.DAT;
         FORMAT IS FREE;SAVE=CPROBABILITIES;
```

7.7.2.2 DK Model (7.13)

```
TITLE:  AAA DATA : MODEL TABLE 6.2 - DK MODEL (6.13)  WITH 2 CLASSES

DATA:  FILE = C:\AAASTUDY\AAA.DAT;

VARIABLE:
    NAMES ARE  Y1 Y2 Y3 Y4 Y5 Y6 Y7
                  U2 U3 U4 U5 U6 U7 ;
    USEVARIABLES ARE Y1 Y2 Y3 Y4 Y5 Y6 Y7
                        D2 D3 D4 D5 D6 D7 ;
    MISSING = *;
    CATEGORICAL = D2 D3 D4 D5 D6 D7;

    CLASSES = C(2);

  DEFINE:
    D7=0; D6=0; D5=0; D4=0; D3=0; D2=0; D1=0;

    IF(U7 EQ 1) THEN D7=1;

    IF(U7 EQ 1 AND U6 EQ 1) THEN D6=1;
    IF(U7 EQ 1 AND U6 EQ 1) THEN D7=_MISSING;

    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1) THEN D5=1;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1) THEN D6=_MISSING;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1) THEN D7=_MISSING;

    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1)  THEN D4=1;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1)  THEN D5=_MISSING;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1)  THEN D6=_MISSING;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1)  THEN D7=_MISSING;


    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1) THEN D3=1;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1) THEN D4=_MISSING;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1) THEN D5=_MISSING;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1) THEN D6=_MISSING;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1) THEN D7=_MISSING;

    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1 AND U2 EQ 1) THEN D2=1;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1 AND U2 EQ 1) THEN D3=_MISSING;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1 AND U2 EQ 1) THEN D4=_MISSING;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1 AND U2 EQ 1) THEN D5=_MISSING;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1 AND U2 EQ 1) THEN D6=_MISSING;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1 AND U2 EQ 1) THEN D7=_MISSING;

   ANALYSIS:
    TYPE=MIXTURE;
    STARTS= 50 20;
    COVERAGE =0.05;
    ALGO=INT;
    INTEGRATION=MONTECARLO;

  MODEL:
            %OVERALL%
            I S  | Y1@0 Y2@0.5 Y3@1 Y4@1.5 Y5@2 Y6@2.5 Y7@3;

            D2 ON Y1 Y2 ; D3 ON Y2 Y3 ; D4 ON Y3 Y4 ;
            D5 ON Y4 Y5 ; D6 ON Y5 Y6 ; D7 ON Y6 Y7 ;

            %C#1%
            Y1 Y2 Y3 Y4 Y5 Y6 Y7 (VARRES1);
            I (I1); S (S1); I WITH S (IS1);

              D2 ON Y1 (BETA21);  D2 ON Y2 (BETA11);
              D3 ON Y2 (BETA21);  D3 ON Y3 (BETA11);
              D4 ON Y3 (BETA21);  D4 ON Y4 (BETA11);
              D5 ON Y4 (BETA21);  D5 ON Y5 (BETA11);
              D6 ON Y5 (BETA21);  D6 ON Y6 (BETA11);
              D7 ON Y6 (BETA21);  D7 ON Y7 (BETA11);

            %C#2%
            Y1 Y2 Y3 Y4 Y5 Y6 Y7 (VARRES2);
            I (I2); S (S2); I WITH S (IS2);

              D2 ON Y1 (BETA22);  D2 ON Y2 (BETA12);
              D3 ON Y2 (BETA22);  D3 ON Y3 (BETA12);
              D4 ON Y3 (BETA22);  D4 ON Y4 (BETA12);
              D5 ON Y4 (BETA22);  D5 ON Y5 (BETA12);
              D6 ON Y5 (BETA22);  D6 ON Y6 (BETA12);
              D7 ON Y6 (BETA22);  D7 ON Y7 (BETA12);
```

7.7.2.3 DK Model (7.14)

This model can be obtained by making the following changes to the model for DK (7.13):

- allow the coefficients in the logistic regression of d_i on y_i and y_{i-1} to be class and time specific,
- specify the intercept in the logistic regression model of d_i as a linear function of time.

The class specific description, for example for class $C\#1$, in the Mplus program has to be changed as follows:

```
MODEL:
        %C#1%
        Y1 Y2 Y3 Y4 Y5 Y6 Y7 (VARRES1);
        I (I1); S (S1);I WITH S (IS1);

          D2 ON Y1 Y2 ; D3 ON Y2 Y3 ; D4 ON Y3 Y4 ;
          D5 ON Y4 Y5 ; D6 ON Y5 Y6 ; D7 ON Y6 Y7 ;
          [D2$1] (P2); [D3$1] (P3); [D4$1] (P4);
          [D5$1] (P5); [D6$1] (P6); [D7$1] (P7);

MODEL CONSTRAINT:
          NEW(G1);
          P3 = P2+G1*.5;
          P4 = P2+G1*1;
          P5 = P2+G1*1.5;
          P6 = P2+G1*2;
          P7 = P2+G1*2.5;
```

7.7.2.4 Beunckens Model

To obtain model (7.15) the model command in the Mplus program should be specified as follows:

```
 MODEL:
      %OVERALL%
      I S  | Y1@0 Y2@0.5 Y3@1 Y4@1.5 Y5@2 Y6@2.5 Y7@3;

      D2 ON I; D3 ON I; D4 ON I;
      D5 ON I; D6 ON I; D7 ON I;

      %C#1%
      Y1 Y2 Y3 Y4 Y5 Y6 Y7 (VARRES1);
      I (I1); S (S1); I WITH S (IS1);

      [D2$1] (P2); [D3$1] (P3); [D4$1] (P4);
      [D5$1] (P5); [D6$1] (P6); [D7$1] (P7);

       D2 ON I(1); D3 ON I(1); D4 ON I(1);
       D5 ON I(1); D6 ON I(1); D7 ON I(1);

      %C#2%
      Y1 Y2 Y3 Y4 Y5 Y6 Y7 (VARRES2);
      I (I2); S (S2); I WITH S (IS2);

      [D2$1] (P12); [D3$1] (P13); [D4$1] (P14);
      [D5$1] (P15); [D6$1] (P16); [D7$1] (P17);

      D2 ON I(2); D3 ON I(2); D4 ON I(2);
      D5 ON I(2); D6 ON I(2); D7 ON I(2);

 MODEL CONSTRAINT:
      NEW(G1 G2);
      P3 = P2+G1*.5;
      P4 = P2+G1*1;
      P5 = P2+G1*1.5;
      P6 = P2+G1*2;
      P7 = P2+G1*2.5;

      P13 = P12+G2*.5;
      P14 = P12+G2*1;
      P15 = P12+G2*1.5;
      P16 = P12+G2*2;
      P17 = P12+G2*2.5;
```

7.7.2.5 Roy Model

```
TITLE:  AAA DATA : MODEL TABLE 6.2 - ROY MODEL WITH 2 CLASSES

 DATA:  FILE = C:\AAASTUDY\AAA.DAT;

 VARIABLE:
     NAMES ARE  Y1 Y2 Y3 Y4 Y5 Y6 Y7
                   U2 U3 U4 U5 U6 U7 ;
     USEVARIABLES ARE Y1 Y2 Y3 Y4 Y5 Y6 Y7
                         D2 D3 D4 D5 D6 D7 ;
     MISSING = *;

     CLASSES = C(2);

   DEFINE:
     D7=0; D6=0; D5=0; D4=0; D3=0; D2=0; D1=0;

     IF(U7 EQ 1) THEN D7=1;

     IF(U7 EQ 1 AND U6 EQ 1) THEN D6=1;
     IF(U7 EQ 1 AND U6 EQ 1) THEN D7=0;


     IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1) THEN D5=1;
     IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1) THEN D6=0;
     IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1) THEN D7=0;
```

```
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1)  THEN D4=1;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1)  THEN D5=0;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1)  THEN D6=0;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1)  THEN D7=0;


    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1)   THEN D3=1;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1)   THEN D4=0;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1)   THEN D5=0;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1)   THEN D6=0;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1)   THEN D7=0;

    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1 AND U2 EQ 1)   THEN D2=1;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1 AND U2 EQ 1)   THEN D3=0;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1 AND U2 EQ 1)   THEN D4=0;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1 AND U2 EQ 1)   THEN D5=0;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1 AND U2 EQ 1)   THEN D6=0;
    IF(U7 EQ 1 AND U6 EQ 1 AND U5 EQ 1 AND U4 EQ 1 AND U3 EQ 1 AND U2 EQ 1)   THEN D7=0;


  ANALYSIS:
    TYPE=MIXTURE;
    STARTS= 50 20;
    COVERAGE =0.05;

  MODEL:
    %OVERALL%
    I S  | Y1@0 Y2@0.5 Y3@1 Y4@1.5 Y5@2 Y6@2.5 Y7@3;
    C ON D2-D7;

   %C#1%
      Y1 Y2 Y3 Y4 Y5 Y6 Y7 (VARRES1);
      I (I1); S (S1); I WITH S (IS1);

   %C#2%
      Y1 Y2 Y3 Y4 Y5 Y6 Y7 (VARRES2);
      I (I2); S (S2); I WITH S (IS2);
```

7.7.3 Computing Time

The time needed to execute the MAR and MNAR models specified in this appendix are given in Table 7.17. Mplus version 5 was used, on a Dell latitude with 16 GB of RAM and 2.70GHz CPU.

Table 7.17: *Computing time (min:sec) for the MAR and MNAR models - AAA Data.*

Model	# classes	Elapsed Time
MAR	1	00:00
MAR	2	00:00
MAR	3	00:02
MAR	4	00:04
Selection Models:		
Diggle-Kenward	1	01:03
Diggle-Kenward Model (7.13)	2	03:02
Diggle-Kenward Model (7.13)	3	06:41
Diggle-Kenward Model (7.13)	4	18:30
Diggle-Kenward Model (7.14)	2	08:42
Diggle-Kenward Model (7.14)	3	16:12
Beunckens Model (7.15)	2	01:24
Beunckens Model (7.15)	3	02:55
Pattern-Mixture Models:		
Pattern-mixture		00:07
Roy	2	00:03
Roy	3	00:09
Roy	4	00:17

Chapter 8

Local Influence Diagnostics for a Growth-Mixture Model

8.1 Introduction

Outlying and influential observations impact the performance of many clustering algorithms and also the performance of model-based clustering. Methods for identifying outliers in a finite-mixture model have been described in the literature. Approaches to identify influential observations are less common.

For hierarchical cluster analysis of multivariate data, Jolliffe et al. (1995) propose measures that allow quantification of the influence of a single observation on the clustering process. Prediction of observations with no effect and those with the greatest effect are done by the minimum spanning tree and the number of neighbours of each observation. Starting from a single-link dendogram, a subgraph completely linking all observations in the dendogram, without any closed loops, and minimizing the sum of the dissimilarities corresponding to the linked observations is constructed. This subgraph is called a minimum spanning tree. It turns out that observations with little effect are at the terminal points of the minimum spanning tree, and observations with the most edges are the most influential ones. Kim et al. (2000) discuss interactively visualizing hierarchical clustering using multidimensional scaling and the minimal spanning tree to detect influential observations. Cheng and Milligan (1996) consider a single data point to be influential when different cluster partitions result from the removal of the observation from the data set. They provide measures to quantify the influence of a data point and investigate the nature of the influence, whether beneficial or detrimental to the clustering.

Methods for detecting influential observations in nonhierarchical cluster analysis have been studied by Cerioli (1998) and Cuesta-Albertos et al. (1997). Cerioli (1998) compares the reference partition obtained for the complete data with the partition obtained after deleting a single case, using the same clustering algorithm. Via a fast forward search algorithm the multivariate observations are ordered from those most in agreement with a specified clustering structure to those least in agreement with it. Identification of multiple outliers and influential observations is then possible via simple graphical displays. Their procedure is not affected by masking and swamping problems. An outlier masks a second outlier, if the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier. An outlier swamps a second observation, if the latter can be considered as an outlier only in the presence of the first one. Cuesta-Albertos et al. (1997) describe a procedure, called trimmed K-means, to robustify the K-means clustering algorithm. The idea of the trimmed K-means is that the outliers should be discarded or trimmed in the calculation of cluster centers. For model-based clustering algorithms, outlier detection is studied by McLachlan and Peel (2000) and by Wang et al. (1997). McLachlan and Peel (2000) used the Mahalanobis distance to decide if a suspicious observation is genuinely outlying for all groups in the finite-mixture model. Wang et al. (1997) proposed a modified likelihood-ratio test, comparing a model built with all observations included, and a model based on all observations excluding the tested observation.

Identification of influential observations in model-based clustering algorithms has, to our knowledge, not yet been described. In this chapter, we will apply local-influence diagnostics, as introduced by Cook (1986), to a finite-mixture model assuming an a priori given number of components. The chapter is organized as follows. Section 8.2 briefly summarizes the measures introduced by Cheng and Milligan (1996) to quantify the influence of an observation and the outlier detection method proposed by Wang (1997). The concept of local-influence diagnostics is sketched in Section 8.3 and applied to real-life data in Section 8.4.

8.2 Outlier Detection and Influence on Cluster Partition

8.2.1 Outlier Detection for a Finite-Mixture Model

Sain et al. (1999), based on Wang et al. (1997), developed an outlier-detection procedure that applies to mixture populations where no labelled training data is available. Specifically, the authors assume that the training data of size N is a sample from a mixture of K distributions, $f(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}; \boldsymbol{\mu_k}, \Sigma_k)$. To investigate if a new observation, $\boldsymbol{y_{N+1}}$, is obtained from the mixture population or from an outlier population they use a likelihood-ratio test statistic that does not require the distribution of the outlier. The classical likelihood-ratio test statistic is the ratio of the maximized likelihood functions $L_0(\boldsymbol{\theta_0}) = [\prod_{i=1}^{N} f(\boldsymbol{y_i}; \boldsymbol{\theta_0})] f(\boldsymbol{y_{N+1}}; \boldsymbol{\theta_0})$ under H_0, and $L_1(\boldsymbol{\theta_0}, \boldsymbol{\theta_1}) = [\prod_{i=1}^{N} f(\boldsymbol{y_i}; \boldsymbol{\theta_0})] h(\boldsymbol{y_{N+1}}; \boldsymbol{\theta_1})$ under H_1, with $h(\boldsymbol{y}; \boldsymbol{\theta_1})$ the density associated with the outlier population. When there is only a single observation from the outlier population, maximizing L_1 is hard. Wang et al. (1997) noted that a viable test statistic could be based on the ratio $L_0/\tilde{L}_1$, with $\tilde{L}_1(\boldsymbol{\theta_0}) = \prod_{i=1}^{N} f(\boldsymbol{y_i}; \boldsymbol{\theta_0})$, eliminating the need for $h(\boldsymbol{y}; \boldsymbol{\theta_1})$. Thus, in essence, the principle of equi-ignorance is employed, and the distribution of the outlier is not needed. The resulting modified likelihood-ratio test statistic

$$W(\boldsymbol{y_{N+1}}; \boldsymbol{y_1}, \ldots, \boldsymbol{y_N}) = \frac{\sup\limits_{\theta_0 \in \Theta} L_0(\boldsymbol{\theta_0})}{\sup\limits_{\theta_0 \in \Theta} \tilde{L}_1(\boldsymbol{\theta_0})}$$

will take small values when $\boldsymbol{y_{N+1}}$ departs from f. The null distribution of W is obtained through nonparametric bootstrap (Efron and Tibshirani, 1993). The authors examined the power of the outlier test based on W, via simulations.

8.2.2 Influence of Individual Data Points on the Cluster Partition

To address the problem of measuring the impact of an individual data point in a cluster analysis, Cheng and Milligan (1996) consider the difference in cluster partitions resulting from removing an element from the data set as a measure of influence. The similarity between the partition sets obtained when clustering based on all N observations and $N-1$ observations is quantified by means of the Rand Index (1971). For the influential cases, the nature of the influence, facilitating or inhibiting, was quantified by two internal criteria. The authors use the γ and the point-biserial internal criterion to measure the agreement between the classification resulting from the cluster exercise and the true class-membership. The sign difference of these statistics computed on the full data set and on the reduced data set with $N-1$ observations, suggests whether or not clustering may have been improved by the introduction of the data point in question. A positive values for $\gamma_{\text{full}} - \gamma_{\text{reduced}}$ indicates that the selected observation is a facilitator, a negative value points to an inhibiting observation.

Similar ideas can be applied to partitions resulting from a model-based cluster exercise. Finite-mixture models are used to classify subjects using (1) all N observations and (2) $N-1$ observations into the K components. For each subject, two weighted γ criteria are obtained, with a subject's posterior probabilities to belong to each of the K components as weights. γ_{full} is based on a clustering method using all N observations, but quantifies the correct classification of the remaining $N-1$ observations. γ_{reduced} on the other hand quantifies the correct classification of $N-1$ observations when running the clustering algorithm on these $N-1$ observations.

8.3 Review of General Theory for Local Influence

Developments in this review follow Lessafre and Verbeke (1998).

Local influence was presented by Cook and Weisberg (1982) and Cook (1986) and used by several authors since. The impact of individuals and measurements on the analysis is assessed by comparing standard maximum likelihood estimates with those resulting from slightly perturbing the contribution of an individual or a measurement. The method is to be contrasted with global influence, or case deletion, where impact is assessed by simply deleting an individual or measurement. While local influence comes with a certain amount of technicality, it is easy and fast to calculate in practice, and in many cases leads to interpretable components of influence. Lesaffre and Verbeke (1998) introduced an influence assessment paradigm for the linear mixed model. A review of several diagnostic procedures for the linear mixed model is given in Mun and Lindstrom (2013). Verbeke et al. (2001) used local influence for longitudinal Gaussian data with dropout, while incomplete binary data were studied by Jansen et al. (2003). Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005) study the method and provide ample references. Ouwens, Tan, and Berger (2001) applied local influence to the generalized linear mixed model for count data, i.e., the Poisson-normal model.

Let the log-likelihood for the chosen model take the form

$$\ell(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{\mathrm{N}} \ell_i(\boldsymbol{\theta}), \tag{8.1}$$

in which $\ell_i(\boldsymbol{\theta})$ is the contribution of the i^{th} individual to the log-likelihood. Let

$$\ell(\boldsymbol{\theta}|\boldsymbol{\omega}) \;=\; \sum_{i=1}^{\mathrm{N}} w_i \ell_i(\boldsymbol{\theta}), \tag{8.2}$$

now denote the perturbed version of $\ell(\boldsymbol{\theta})$, depending on an N-dimensional vector $\boldsymbol{\omega}$ of weights, assumed to belong to an open subset Ω of $I\!R^N$. The original log-likelihood (8.1) follows for $\boldsymbol{\omega} = \boldsymbol{\omega_0} = (1, 1, \ldots, 1)'$. Here, the perturbed log-likelihood gives more or less weight to log-likelihood contributions of single subjects, but other perturbation schemes are possible (Verbeke and Molenberghs, 2000).

Let $\widehat{\boldsymbol{\theta}}$ be the maximum likelihood estimator for $\boldsymbol{\theta}$, obtained by maximizing $\ell(\boldsymbol{\theta})$, and let $\widehat{\boldsymbol{\theta}}_\omega$ denote the estimator for $\boldsymbol{\theta}$ under $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$. Cook (1986) proposed to measure the distance between $\widehat{\boldsymbol{\theta}}_\omega$ and $\widehat{\boldsymbol{\theta}}$ by the so-called likelihood displacement, defined by

$$\mathrm{LD}(\boldsymbol{\omega}) \;=\; 2\left(\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})\right).$$

This measure for the difference in estimates, takes into account the variability in $\widehat{\boldsymbol{\theta}}$. $\mathrm{LD}(\boldsymbol{\omega})$ will be large if $\ell(\boldsymbol{\theta})$ is strongly curved at $\widehat{\boldsymbol{\theta}}$ (which means that $\boldsymbol{\theta}$ is estimated with high precision) and small otherwise. A graph of $\mathrm{LD}(\boldsymbol{\omega})$ versus $\boldsymbol{\omega}$ brings out information on the influence of the case-weight perturbations. The graph is the geometric surface formed by the values of the $(N+1)$-dimensional vector

$$\boldsymbol{\xi}(\boldsymbol{\omega}) \;=\; \begin{pmatrix} \boldsymbol{\omega} \\ \mathrm{LD}(\boldsymbol{\omega}) \end{pmatrix}$$

as $\boldsymbol{\omega}$ varies throughout Ω. Following Cook (1986) and Verbeke and Molenberghs (2000), we will refer to $\boldsymbol{\xi}(\boldsymbol{\omega})$ as an influence graph. It is unfeasible to evaluate $\mathrm{LD}(\boldsymbol{\omega})$ for all $\boldsymbol{\omega}$. Cook (1986) describes the sensitivity of $\ell(\hat{\boldsymbol{\theta}})$ by looking at small perturbations for case weights around $\boldsymbol{\omega_0}$, i.e., the local behaviour of $\mathrm{LD}(\boldsymbol{\omega})$ around $\boldsymbol{\omega_0}$. This was done using the normal curvature C_h of $\mathrm{LD}(\boldsymbol{\omega})$ at $\boldsymbol{\omega_0}$, in the direction of a unit vector $\boldsymbol{h}$ in Ω. Cook (1986) derived a convenient computational scheme. Let $\boldsymbol{\Delta}_i$ be the s-dimensional vector of second-order derivatives of $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$, with respect to ω_i and all components of $\boldsymbol{\theta}$, and evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ and at $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. Also, write Δ for the $s \times N$ matrix with $\boldsymbol{\Delta}_i$ in the i^{th} column. Let $\ddot{L}$ denote the $s \times s$ matrix of second-order derivatives of $\ell(\boldsymbol{\theta})$ w.r.t the components of $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. For any unit vector

$\boldsymbol{h}$ in Ω, it follows that:

$$C_h \quad = \quad 2\left|\ \boldsymbol{h}'\Delta'\ddot{L}^{-1}\Delta\boldsymbol{h}\ \right|. \tag{8.3}$$

Various choices for $\boldsymbol{h}$ have received specific attention. First, one can focus on a single subject i only, by choosing $\boldsymbol{h} = \boldsymbol{h_i}$, the zero vector with a sole value 1 in the i^{th} position. The normal curvature is then called the total local influence and is given by

$$C_i \ \equiv \ C_{h_i} \quad = \quad 2\left|\ \boldsymbol{\Delta}_i'\ddot{L}^{-1}\boldsymbol{\Delta}_i\ \right|. \tag{8.4}$$

Large values of C_i are obtained for subjects for which small perturbations in case weight result locally in a large log-likelihood displacement.

Second, $\boldsymbol{h} = \boldsymbol{h}_{\max}$ can be chosen as the direction of maximal normal curvature $C_{\max}$. It was shown that $\boldsymbol{h}_{\max}$ is the eigenvector of $-\Delta'\ddot{L}^{-1}\Delta$ corresponding to the largest eigenvalue (Beckman, 1987 ; Verbeke, 1997 , 2000; Seber, 1984). $\boldsymbol{h}_{\max}$ permits detection of individuals that are simultaneously influential.

The total local influence of individual i can be expressed in terms of the nonzero eigenvalues, $\lambda_1 \geq \ldots \geq \lambda_s > 0$ and normalized orthogonal eigenvectors $\boldsymbol{\nu}_1 \equiv \boldsymbol{h}_{\max}, \ldots, \boldsymbol{\nu}_s$ of $-\Delta'\ddot{L}^{-1}\Delta$:

$$C_i = 2\sum_{j=1}^{s}\lambda_j\nu_{ji}^2,$$

with ν_{ji} the i^{th} component of $\boldsymbol{\nu}_j$. $C_{\max}$ is twice the largest eigenvalue, $C_{\max} = 2 \cdot \lambda_1$. This holds a warning: it is possible for C_i to be large without the same holding for the i^{th} component in $\boldsymbol{h}_{\max}$, provided the corresponding components are large for some of the secondary eigenvectors. It is thus recommended to examine both the C_i and $\boldsymbol{h}_{\max}$.

Lesaffre and Verbeke (1998) proposed a threshold for C_i above which an individual is defined as "remarkable". They state that the i^{th} subject is influential if C_i is larger than the cutoff value $2\sum_{i=1}^{\mathrm{N}} C_i/N$.

The methodology still applies when interest is in a subset $\boldsymbol{\theta}_1$ of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$. It follows that (Verbeke and Molenberghs, 2000) the influence on the estimation of the subset $\boldsymbol{\theta}_1$ is given by:

$$C_h(\boldsymbol{\theta}_1) = C_h \ - \ 2\left|\boldsymbol{h}'\Delta'\begin{pmatrix} 0 & 0 \\ 0 & \ddot{L}_{22}^{-1}\end{pmatrix}\Delta\boldsymbol{h}\right| \leq C_h, \tag{8.5}$$

$\ddot{L}_{22}$ is defined by the partition of $\ddot{L} = \begin{pmatrix} \ddot{L}_{11} & \ddot{L}_{12} \\ \ddot{L}_{21} & \ddot{L}_{22}\end{pmatrix}$ according to the dimensions of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Should $\ddot{L}_{12} = 0$, then $C_h = C_h(\boldsymbol{\theta}_1) \ + \ C_h(\boldsymbol{\theta}_2)$. For weakly correlated sub-vectors, this decomposition holds approximately.

To obtain local-influence diagnostics for a finite-mixture model with K components (see Section 4.2), second derivatives of $l(\boldsymbol{\theta}|\boldsymbol{\omega})$ with respect to ω_i and all components of $\boldsymbol{\theta}$, have to be obtained. But using case-weights perturbations this simplifies to obtaining first-order derivatives of $l_i(\boldsymbol{\theta})$ with respect to the components of $\boldsymbol{\theta}$. Second-order derivatives of $l(\boldsymbol{\theta})$ are also needed. Under the finite-mixture model, $\boldsymbol{\theta}$ contains the fixed- and random-effects parameters describing the profiles for the K components, and the mixture probabilities. Further, the contribution $l_i(\boldsymbol{\theta})$ of the i^{th} subject to the log-likelihood is $l_i(\boldsymbol{\theta}, \boldsymbol{\pi}) = \log \left[\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i|\boldsymbol{\theta})\right]$. Expressions for the first and second derivatives, involved in C_i, with respect to the components of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ are easily obtained as:

$$\begin{aligned}
\frac{\partial l_i}{\partial \pi_k} &= \frac{1}{f(\boldsymbol{y}_i)}\left[f_k(\boldsymbol{y}_i) - f_K(\boldsymbol{y}_i)\right] \\
\frac{\partial l_i}{\partial \theta_k} &= \frac{1}{f(\boldsymbol{y}_i)}\left[\sum_{k=1}^{K} \pi_k \frac{\partial f_k(\boldsymbol{y}_i)}{\partial \theta_k}\right]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l_i}{\partial \pi_k \partial \pi_l} &= \frac{-1}{f^2(\boldsymbol{y}_i)}\left[(f_l(\boldsymbol{y}_i) - f_K(\boldsymbol{y}_i))(f_k(\boldsymbol{y}_i) - f_K(\boldsymbol{y}_i))\right] \\
\frac{\partial^2 l_i}{\partial \theta_k \partial \theta_l} &= \frac{-1}{f^2(\boldsymbol{y}_i)}\left[\sum_{k=1}^{K} \pi_k \frac{\partial f_k(\boldsymbol{y}_i)}{\partial \theta_k}\right] + \frac{1}{f(\boldsymbol{y}_i)}\left[\sum_{k=1}^{K} \pi_k \frac{\partial^2 f_k(\boldsymbol{y}_i)}{\partial \theta_k \theta_l}\right] \\
\frac{\partial^2 l_i}{\partial \pi_k \partial \theta_l} &= \frac{-1}{f^2(\boldsymbol{y}_i)}\left[\sum_{k=1}^{K} \pi_k \frac{\partial f_k(\boldsymbol{y}_i)}{\partial \theta_l}\right]\left[f_k(\boldsymbol{y}_i) - f_K(\boldsymbol{y}_i)\right] \\
&\quad + \frac{1}{f(\boldsymbol{y}_i)}\left[\frac{\partial f_k(\boldsymbol{y}_i)}{\partial \theta_l} - \frac{\partial f_K(\boldsymbol{y}_i)}{\partial \theta_l}\right]
\end{aligned}$$

Often, interest is not only in the stability of the components' mean profiles, but also in the stability of the classification of an individual subject.

The theory of local influence, as described above, allows, in an elegant way, quantification of the influence of subject i on the posterior probability of subject j. To this end, log-likelihood (8.2) has to be parameterized as a function of the posterior probabilities. Given the relation between the posterior probabilities and the mixture probabilities, $\pi_{jk} = \pi_k f_k(\boldsymbol{y}_j)/\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_j)$, it is straightforward to express the contribution of the i^{th} individual to the log-likelihood as a function of the posterior probability of the j^{th} individual. The log-likelihood then takes the form:

$$l(\boldsymbol{\theta}, \boldsymbol{\pi}) \qquad \Rightarrow \qquad l(\boldsymbol{\theta}, \boldsymbol{\pi}_j) = \sum_{i=1}^{N} l_i(\boldsymbol{\theta}, \boldsymbol{\pi}_j),$$

with $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jK})^{'}$ the vector of posterior probabilities for subject j and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^{'}$ the mixing probabilities. The local influence of subject i, on the subset $\boldsymbol{\pi}_j$ of the vector of all parameters can be obtained via (8.5). The first and second derivatives of $l(\boldsymbol{\theta}, \boldsymbol{\pi}_j|\boldsymbol{\omega})$ are obtained via the chain rule for differentiating composite functions:

$$\begin{aligned}\frac{\partial l_i}{\partial(\boldsymbol{\theta}, \boldsymbol{\pi}_j)} &= \frac{\partial l_i}{\partial(\boldsymbol{\theta}, \boldsymbol{\pi})}\frac{\partial(\boldsymbol{\theta}, \boldsymbol{\pi}^{'})}{\partial(\boldsymbol{\theta}, \boldsymbol{\pi}_j)},\\ \frac{\partial^2 l_i}{\partial^2(\boldsymbol{\theta}, \boldsymbol{\pi}_j)} &= \frac{\partial(\boldsymbol{\theta}, \boldsymbol{\pi})}{\partial(\boldsymbol{\theta}, \boldsymbol{\pi}_j)}\frac{\partial^2 l_i}{\partial^2(\boldsymbol{\theta}, \boldsymbol{\pi})}\frac{\partial(\boldsymbol{\theta}, \boldsymbol{\pi})}{\partial(\boldsymbol{\theta}, \boldsymbol{\pi}_j)}.\end{aligned}$$

8.4 Data Applications

8.4.1 Orange Tree Data

To illustrate the local influence approach on a nonlinear model we use the orange tree data of Draper and Smith (1981). The data consists of trunk circumference recordings of five orange trees, at seven occasions.

A logistic nonlinear mixed model for Gaussian data was used before on these repeated measurements by Lindstrom and Bates (1990) and Pinheiro and Bates (1995):

$$Y_{ij} = \frac{\beta_1 + b_i}{1 + \exp[-(t_{ij} - \beta_2)/\beta_3]} + \varepsilon_{ij}, \tag{8.6}$$

where Y_{ij} represents the j^{th} measurement on the i^{th} tree ($i = 1, \ldots, 5; j = 1, \ldots, 7$), t_{ij} is the corresponding day, β_1, β_2, β_3 are the fixed-effects parameters, b_i are the random-effect parameters assumed to be i.i.d. $N(0, \sigma_b^2)$, and ε_{ij} are the residual errors assumed to be i.i.d. $N(0, \sigma_\varepsilon^2)$ and independent of the b_i.

Spiessens and Verbeke (2002) analysed the data assuming that the random effects follow a mixture of two normal distributions $b_i \sim \pi_1 N(\mu_1, \sigma_b^2) + \pi_2 N(\mu_2, \sigma_b^2)$ (see Section 4.2). Based on the posterior probabilities, the five trees were classified into two clusters. The trees with the highest growth rate (trees 1, 3, and 5) were classified in the second component, trees 2 and 4 in the first component. The growth profiles of the five trees are plotted in Figure 8.1. The results of an influence analysis are displayed in Table 8.1. Tree 5, classified in the first component, has a large impact on the parameter estimates as shown by a large value of C_i. Tree 5 is the only influential tree. This tree has a large impact on the fixed parameters μ_1 and β_2 but also on μ_2, and on the variance component σ_b^2 of the model. The maximal normal curvature equals 6.76. The indexplot of $\boldsymbol{h}_{\max}$ shows that tree 5 also has a large contribution

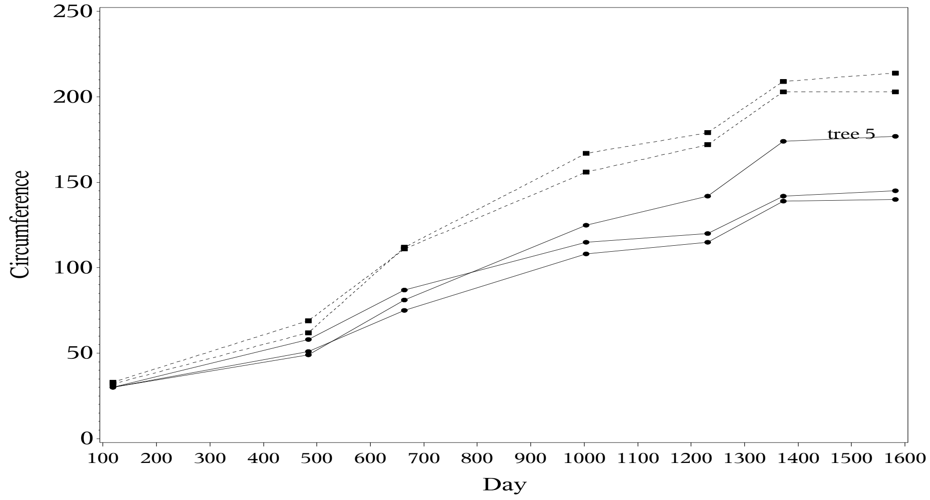


Figure 8.1: *Observed growth profiles for the five orange trees.*

in the direction of maximal curvature. Tree 5 is not considered to be influential for the mixture probability. The effect of perturbing the case-weight of tree 5, on the posterior probabilities to belong to the first component is given in Figure 8.2. The heatmap displays the local influence that tree i has on the posterior probability for tree j, in the crossing of i^{th} column with the j^{th} row. The values in the graph are standardized, a value of 1 corresponds to a $C_i(\pi_{j1})$ equal to the cutoff value above which tree i is considered to be influential for the posterior probability of tree j to belong to component 1. It is seen that the influence of tree 5 on the fixed-effect parameters is reflected in the influence on the posterior probabilities of the remaining trees. The largest influence is observed on the trees not belonging to the cluster containing tree 5 (i.e., trees 2 and 4). On the other hand, the posterior probability of tree 5 is not very sensitive to perturbations in the case-weights of the other trees.

8.4.2 Pharmacokinetic Data

Nonlinear mixed models are also widely used in pharmacokinetics to study how a drug disperses through subjects. Pinheiro and Bates (1995) present data on serum concentrations of the drug theophylline in 12 subjects measured over a 25-hour period after oral administration. They considered a first-order compartment model, allowing for random variability between subjects. Let Y_{ij} denote the observed concentration of the i^{th} subject at time t_{ij}, D the dose of theophylline, k_{ei} the elimination rate

Table 8.1: *Local-influence diagnostics for the parameters in model (8.6) - Orange Tree Data.*

	C_i	$h_{\max,i}$	$C_i(\beta_2)$	$C_i(\beta_3)$	$C_i(\mu_1)$	$C_i(\mu_2)$	$C_i(\sigma_b)$	$C_i(\sigma_\varepsilon)$	$C_i(\pi_1)$
tree 1	1.059	-0.399	0.228	0.023	0.262	0.035	0.222	0.008	0.265
tree 2	0.086	-0.146	0.111	0.052	0.059	0.263	0.227	0.047	0.603
tree 3	1.207	-0.346	0.000	0.111	0.396	0.012	0.023	0.265	0.271
tree 4	0.599	0.058	0.055	0.513	0.117	0.004	0.22	0.166	0.609
tree 5	4.697	0.834	1.090	0.211	2.982	0.418	1.643	0.171	0.272
$\frac{2\sum_i C_i}{N}$	3.059		0.593	0.364	1.527	0.292	0.935	0.263	0.807

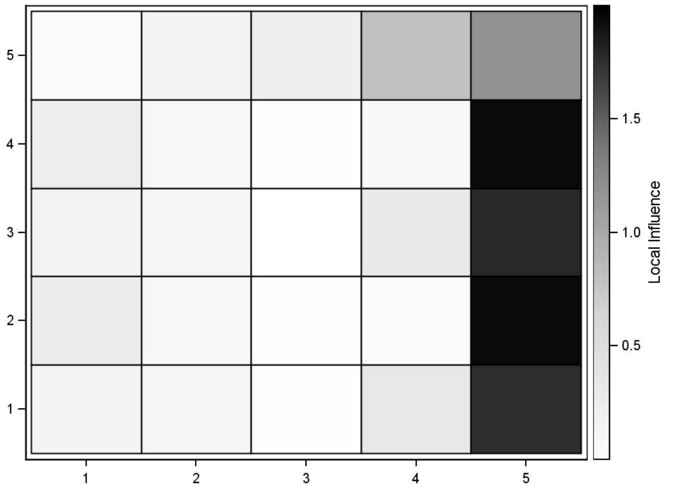


Figure 8.2: *Local-influence diagnostics for the posterior probabilities to belong to the first component of the mixture - Orange Tree Data. The crossing of i^{th} column with the j^{th} row displays the local influence that tree i has on the posterior probability of tree j to belong to the first component of the mixture. Values above 1 are considered to be influential.*

constant for subject i, k_{ai} the absorption rate constant for subject i, Cl_i the clearance for subject i, and ε_{it} normal errors. The model for the observed concentration is specified as:

$$Y_{ij} = \frac{Dk_{ei}k_{ai}}{Cl_i(k_{ai} - k_{ei})}[\exp(-k_{ei}t_{ij}) - \exp(-k_{ai}t_{ij})] + \varepsilon_{ij}. \tag{8.7}$$

The clearance, absorption, and elimination rates for subject i were functions of fixed and random effects:

$$Cl_i = \exp(\beta_1 + b_{i1}), \tag{8.8}$$

$$k_{ai} = \exp(\beta_2 + b_{i2}), \tag{8.9}$$

$$k_{ei} = \exp(\beta_3). \tag{8.10}$$

The random effects allow for heterogeneity between subjects. The $\boldsymbol{b_i} = (b_{i1}, b_{i2})^{'}$ are assumed to follow a multivariate normal distribution with mean zero and an unknown covariance matrix.

The expected concentration level in the body as a function time, for a typical patient (i.e., random effects equal to zero) is displayed in Figure 8.3. The model fit criteria for this homogeneity model are as follows: log-likelihood -178.2, AIC 368.5, and BIC 371.4. When carrying out a two-component heterogeneity model the fit criteria are: log-likelihood -162.9, AIC 343.9, and BIC 348.3; indicating a better fit. The mixing probabilities are .53 and .47. Based on their posterior probabilities, 6 subjects were classified into the first component and 6 in the second. The expected concentration for both components, for a typical subject, can be found in Figure 8.3. The two components distinguish in terms of the maximum concentration level attained

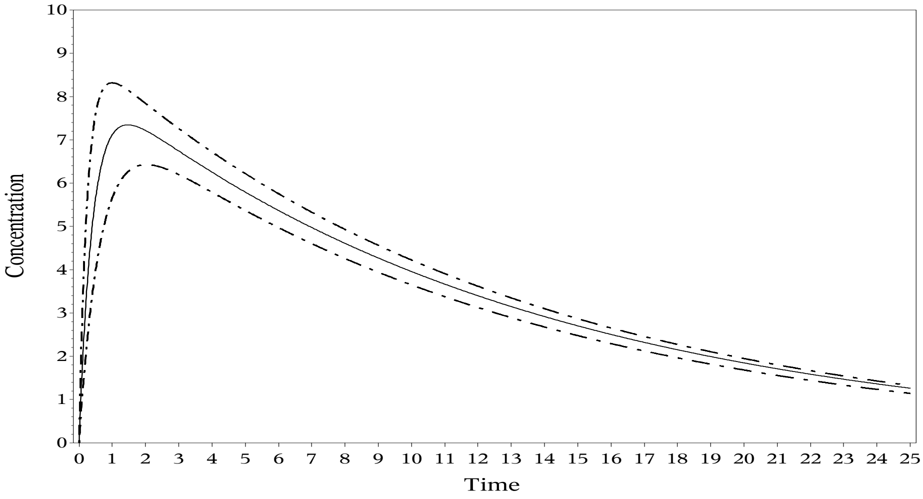


Figure 8.3: *Evolution of the concentrations - Theophylline data. For each component in the mixture distribution, the evolution for a typical patient is displayed (full line: one-component model, −. − . : two-component model).*

and the time after adminstration that the maximum concentration is reached. The first component reaches its maximum concentration level faster, and the maximum level attained is higher than compared to the maximum level of the second component.

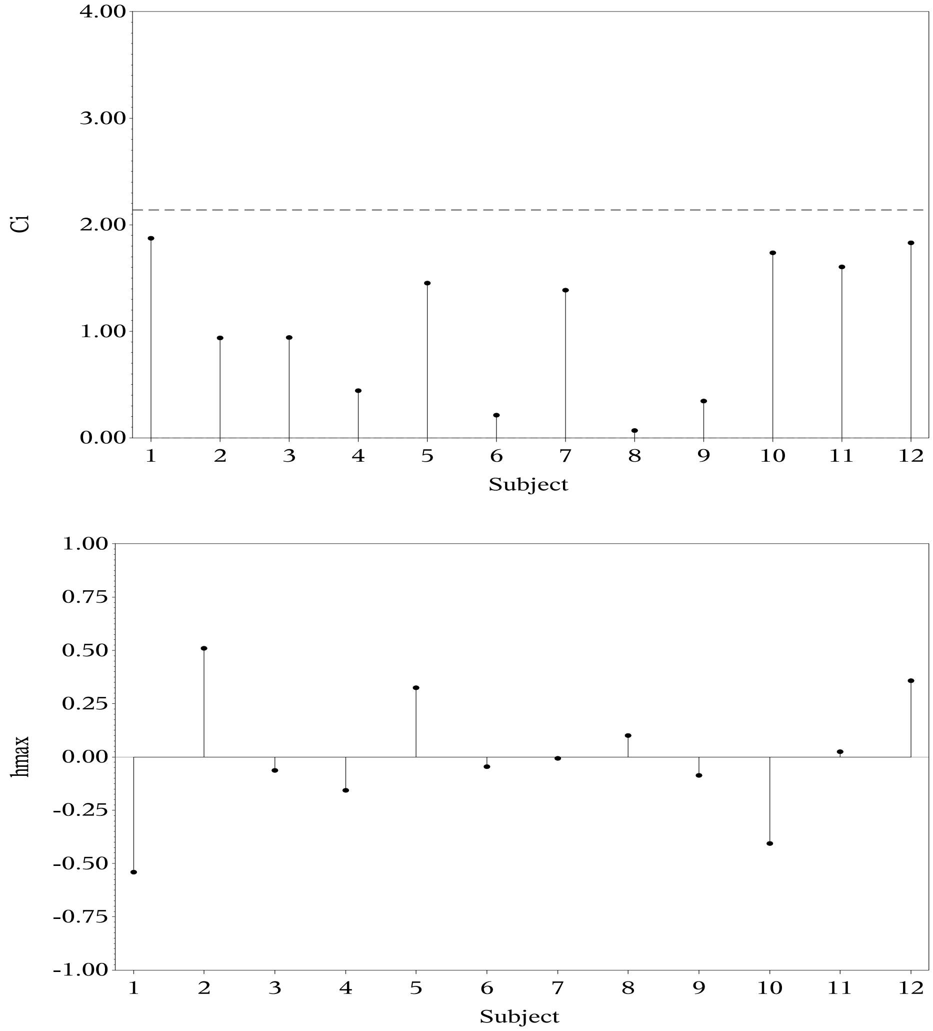


Figure 8.4: *Local-influence analysis - Theophylline data. The upper-panel figure displays the total local influence versus the patient numbers. The horizontal line represents the cut-off value for C_i. The lower-panel figure shows the indexplot of the components of $\boldsymbol{h}_{max}$.*

An influence analysis for the two-component heterogeneity model does not reveal cases being locally influential (see Figure 8.4). All subjects' C_i are below the cutoff value of 2.14. The components of $\boldsymbol{h}_{\max}$ indicate that subjects 1, 2, 5, 10, and 12 exhibit larger contributions in the direction of maximal curvature; C_{max} equals 12.76. When looking into the plots of the total local influence for a specific fixed or random parameter, these subjects have value of C_i exceeding the cut-off (Results not displayed).

8.4.3 EEG Data

The aim of EEG studies is to characterize the effects of psychotropic drugs on cortical brain activity, on the basis of spectral electro-encephalograms. The EEG data is introduced in Section 3.2.

In this chapter, we focus on data of the γ_2 waves in the left prefrontal cortex of two psychoactive agents, PCP and Donepezil, administered at the highest dose. Gamma waves are related to strong mental activity like solving problems, fear, and awareness. PCP in low to moderate doses acts as a stimulant, whilst at higher doses it has a sedative effect. Donepezil is a cholinesterase inhibitor and is used to treat moderate to severe dementia of the Alzheimer's type.

To visualize the data, the individual γ_2 longitudinal profiles are given in Figure 8.5. The response of interest is the γ_2 percentage change as compared to the measurement at baseline (administration of the drug): $Y_{ij} = 100(Y_{ij} - Y_{ib})/Y_{ib}$. In graphical displays and in the statistical models, time zero refers to the first measurement obtained after administering the drug (i.e., after 45 minutes). Heterogeneity is seen in the γ_2 waves, some rats have a decrease in the frequency while for others an increase is obtained as an effect of the drug. This heterogeneity is of course largely caused by administrating 2 different drugs.

The heterogeneity model will be assumed, with a quadratic evolution of the γ_2 percentage change over time and a random intercept that is a mixture of 2 normal distributions. So for component k ($k = 1, 2$) in the mixture we have:

$$Y_{ij}^k = \beta_0^k + \beta_1^k t_{ij} + \beta_2^k t_{ij}^2 + b_i + \varepsilon_{ij}, \tag{8.11}$$

where β_0^k, β_1^k, and β_2^k are component-specific fixed parameters describing the mean γ_2 profiles, b_i are rat-specific intercepts sampled from a 2-component model, and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon)$.

When applying the cluster algorithm, information about the drug a rat was given, was not taken into account. The log-likelihoods (BIC) for the one and two-component model were respectively -571.1 (1156.1) and -531.7 (1088.4). The model hypothesizing a mixture is outperforming the one-component model. Classification of the rats into

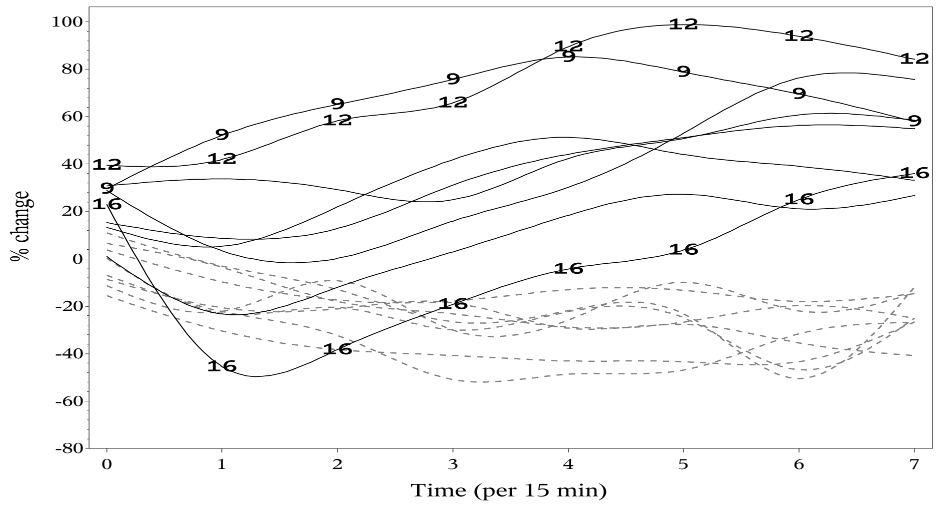


Figure 8.5: *Smoothed observed %change for γ_2 profiles (full lines: PCP, dotted lines: Donepezil) - EEG data. The origin of the time axis is at the first measurement after administration of the drug.*

the two components, based on their posterior probabilities, perfectly coincides with the two drug groups. The 8 rats on PCP (IDs 9–16) are classified together into cluster 1, and the 8 rats on Donepezil (IDs 1–8) into cluster 2.

The results of a local-influence analysis for the two-component model are displayed in Figure 8.6. Three rats (IDs 9, 12, and 16) of the PCP group are locally influential, based on their C_i value. The observed profiles of these rats are highlighted in Figure 8.5. These rats also have a large component in the direction of maximal curvature $\boldsymbol{h}_{\max}$. The maximal curvature equals 10.63. The contribution to $\boldsymbol{h}_{max}$ of rat 14 is also substantial, although its C_i was rated as not exceptionally high. To study the influence on subsets of parameters of model (8.11), expression (8.5) was used. The local-influence diagnostics were obtained for the cluster specific average profiles, the random components, and the mixture probability. The results are presented in Figure 8.7.

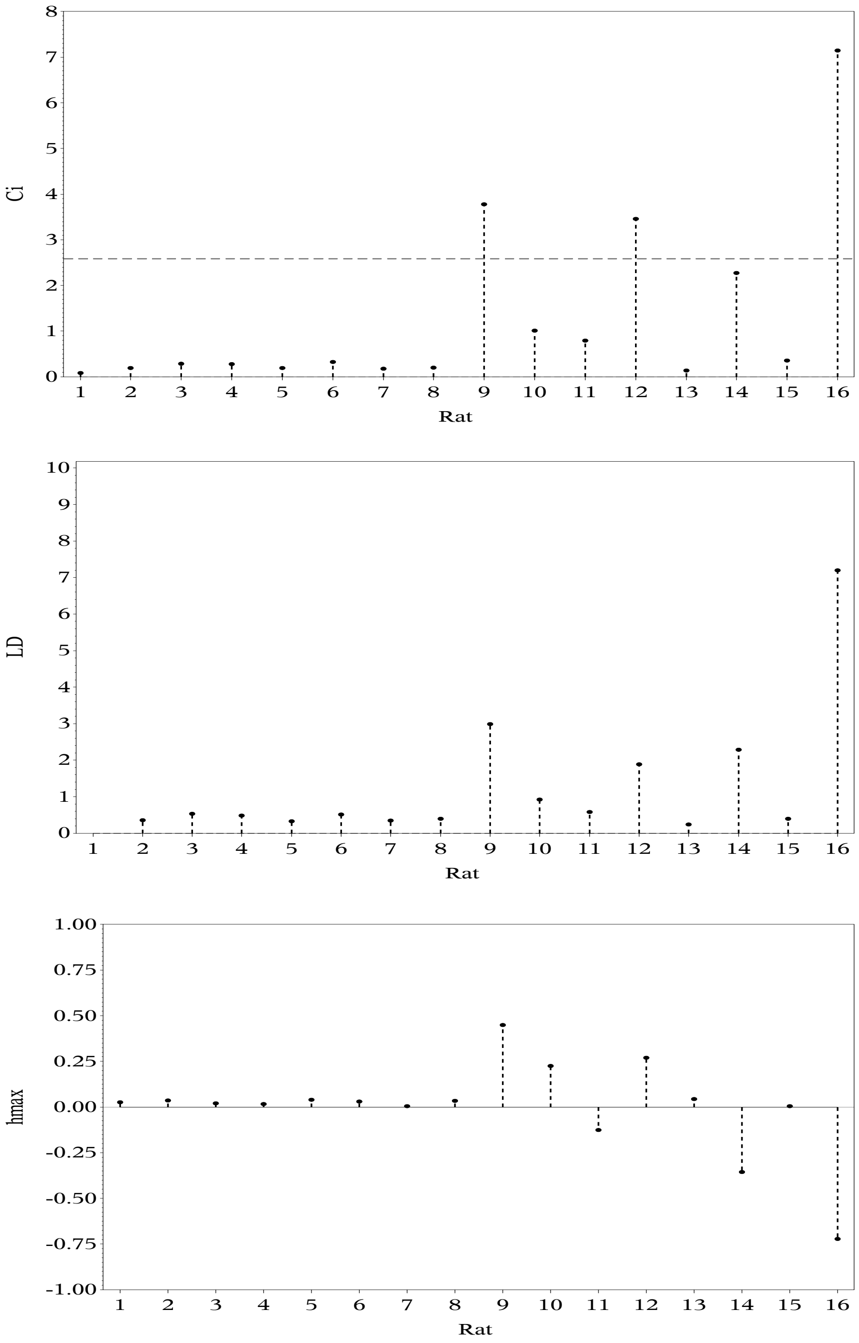


Figure 8.6: *Total local influence, likelihood displacement, and direction of maximal curvature versus patient identification numbers - EEG data. The horizontal line in the total local influence graph represents the cut-off value for C_i.*

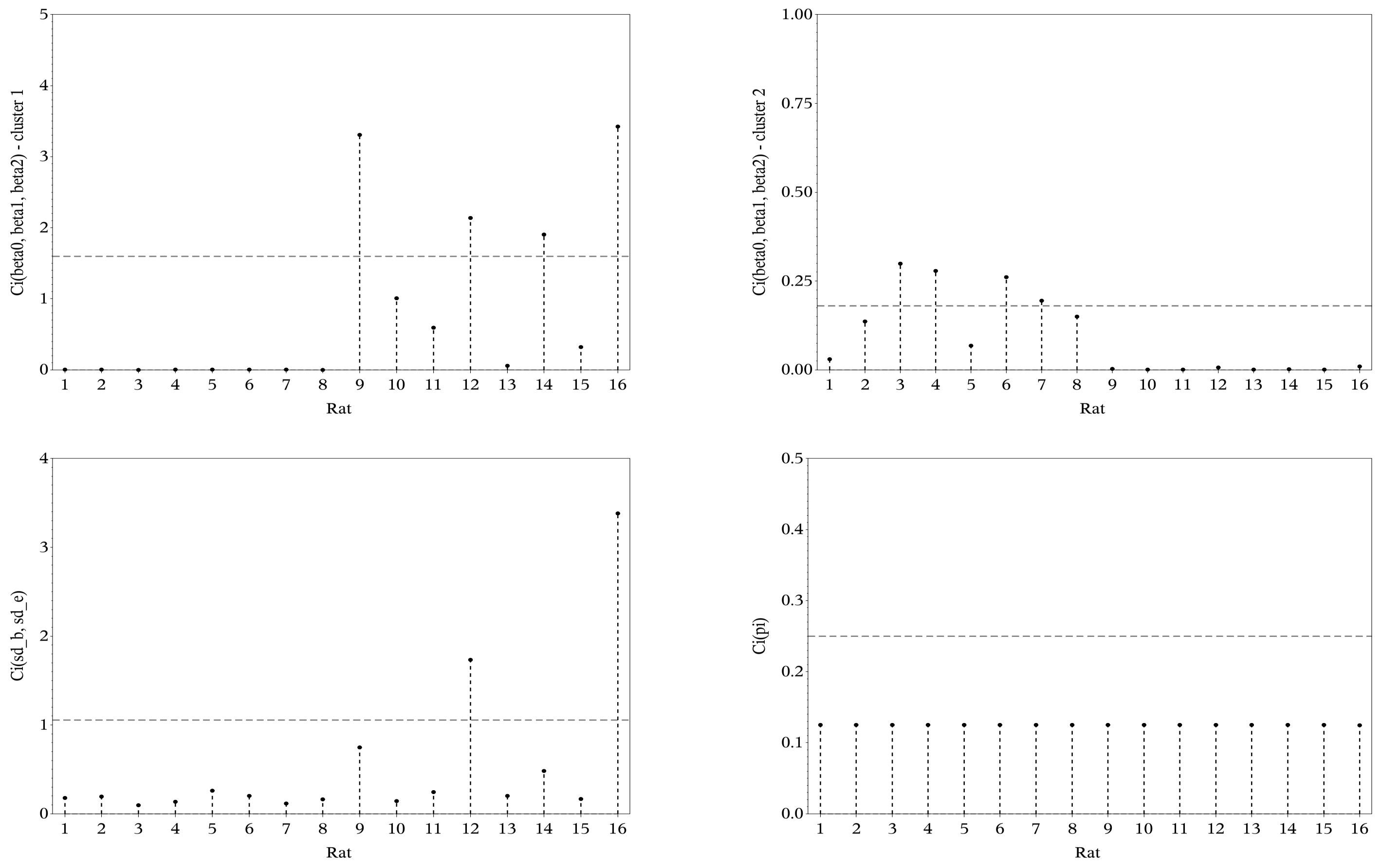


Figure 8.7: *Plot of local-influence diagnostics for the cluster specific average profiles, the random components and the mixture probability - EEG data. The horizontal lines represent the cut-off values for the displayed influences.*

The influence of rats 9, 12, and 16 is visible in the set of fixed parameters characterizing the average evolution of the cluster they belong too (cluster 1), their influence on the average profile of cluster 2 is negligible. Rats 12 and 16 are also influential for the random effects $(\sigma_b, \sigma_\varepsilon)$, their diagnostics exceed the cutoff value. The mixture probability is not subject to small perturbations in the case-weights.

Local-influence diagnostics for the posterior probabilities are presented in a heatmap (Figure 8.8), summarizing the local influence that rat i has on the posterior probability of rat j, in the crossing of i^{th} column with the j^{th} row. As before, the values in the graph are standardized, a value of one corresponds to a $C_i(\pi_{j1})$ equal to the cutoff value above which rat i is considered to be influential for the posterior probability of rat j to belong to component 1. The influence on the posterior probabilities of rats 7, 12, and 14 could not be investigated. The second derivative of the log-likelihood is singular when parameterizing it as a function of the posterior probabilities of these rats. It can be seen that a perturbation in the case-weight of rat 16 influences the posterior probabilities of the other rats. Rat 9 influences the posterior probability of rats 11 and 16, rat 12 influences the posterior probabilities of rats 9, 10, 11, 13 and 15. The other influence diagnostics did not exceed the cutoff values.

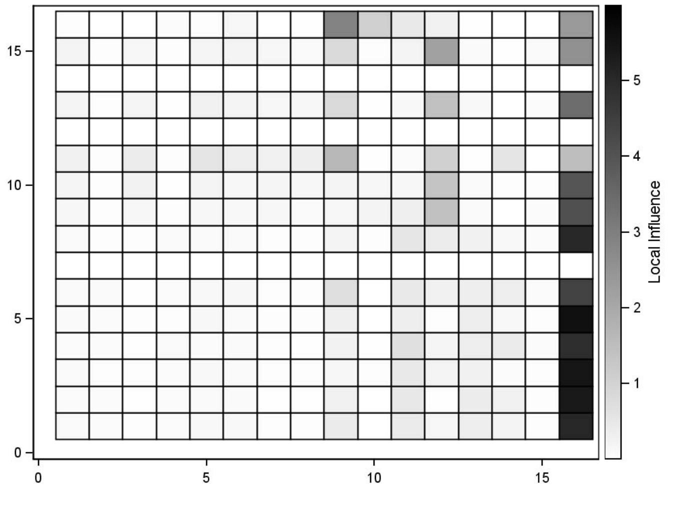


Figure 8.8: *Local-influence diagnostics for the posterior probabilities to belong to the first component of the mixture - EEG data. The crossing of i^{th} column with the j^{th} row displays the local influence that rat i has on the posterior probability of rat j to belong to the first component of the mixture. Values above 1 are considered to be influential.*

Classical diagnostics are generally based on case deletion. The likelihood displacements obtained by deletion of one rat at a time from the analysis is given in Figure 8.6. The largest likelihood displacements are seen for rats 9, 12, 14 and 16. It is reassuring that these are also the rats that stand out in the local-influence analysis. The influence measures do however not agree for the ranking of rats 12 and 14.

For this study, the true cluster membership, i.e., PCP or Donezepil, is in fact know. Therefore the statistics proposed by Cheng and Milligan (1996) (Section 8.2.2) can be obtained. All γ values are extremely small (Figure 8.9), indicating that inclusion/exclusion of an observation hardly has any effect on the correct classification of the other observations. Negative values for the γ criterion are obtained for rats 9, 12 and 16, the values are, respectively -18, -10, and -19×10^{-6}. This indicates that, if any effect, these rats impede correct classification of the other rats. Combining this with the results from the local-influence analysis for the posterior probability, we can conclude that rats 9, 12, and 16 have an influence on the posterior probabilities of the other rats but that the effect is not strong enough to change the cluster allocation.

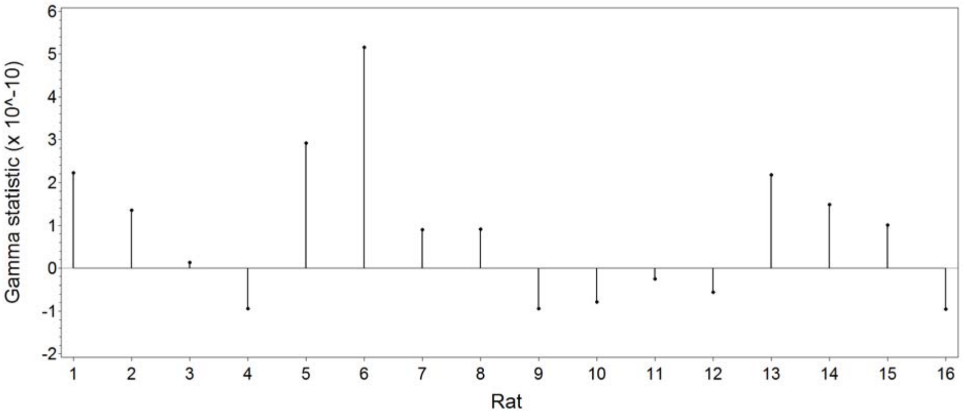


Figure 8.9: *γ-statistics versus the identification numbers of the rats - EEG data. Negative values for rat i indicate that this rat hinders correct classification of the other rats.*

To investigate if rat 16 is to be considered an outlier, the detection procedure described by Sain et al. (1999) was employed. The profiles of the first 15 rats are assumed to be sampled from a two-component mixture population, and the modified likelihood-ratio test is used to see whether the profile of rat 16 belongs to an outlier population or not. The value of the modified likelihood-ratio test statistic W equals 0.46. Applying this procedures for rats 9 and 12, the value of the modified likelihood-ratio statistic equals 0.74 and 0.84, respectively. The null distribution of the test statistic was obtained via 999 nonparametric bootstrap samples. The 1st (5th) percentile of the distribution equals 0.52 (0.68). Thus, the profile of rat 16 is an outlying observation at the 1% level of significance. On the other hand, the profiles of rat 9 and 12, also flagged in the local-influence analysis, are considered to belong to the two-component mixture population, according to this approach.

8.5 Discussion

This chapter elucidates the usefulness of local influence in a model-based cluster analysis.

Local influence quantifies the impact of observations on the analysis. This can, for instance, be done by introducing case-weights in the log-likelihood, such that the contribution of an individual is slightly perturbed. Focus can be put on the effect of individual i only, by choosing the vector of case-weights to be the zero vector with one value of 1 in the i^{th} position. The total local influence is then defined as the normal curvature of the likelihood displacement in the direction of the i^{th} individual.

In this chapter, we demonstrated the usefulness of local-influence diagnostics when clustering longitudinal profiles by means of a finite-mixture model, with an a priori given number of components. The total local influence diagnostics quantify an individual's influence on the vector of all parameters in the model. Generally, this parameter vector contains (1) a number of fixed-effect parameters to describe the average evolution of each component in the mixture, (2) random-effect parameters reflecting heterogeneity in the population, and (3) the mixture probabilities. The influence on a subset of this vector of parameters – for example the influence on the average profile of a specific cluster, or on the mixture probabilities – can also be obtained.

When interest is not only in the stability of the parameters describing the components in the population, but also in the stability of an individual's classification the influence on the posterior probabilities is to be investigated. Local influence is an elegant approach for this. The stability of the posterior probabilities of individual j, can easily be inspected by re-parameterizing the log-likelihood in terms of the fixed effects, random effects and the posterior probabilities of individual j. For the two-component mixtures carried out in this chapter, the $i \times j$ influence measures were displayed in a heatmap. Local-influence diagnostics were obtained for three real-life datasets subjected to a finite mixture model. For the EEG data, the results were compared with an outlier detection procedure for finite-mixture models and a method quantifying the impact of individual data points on the cluster partition when the correct classification is available. Local-influence diagnostics highlighted influential observations, that were not revealed by the traditional case-deletion methods.

Chapter 9

Concluding Remarks and Future Work

In this dissertation, we have addressed clustering for high dimensional data, possibly subject to missingness. The research was inspired by a number of data sets, ranging from data collected in a mental care setting, studies in patients with abdominal aortic aneurysm or heart failure, to an EEG study in rats. The communality in these studies is the believe that the population under investigation is not homogenous, but instead consists of subpopulations. A direct labelling of these subpopulations is not available. But given that these sub-populations are characterized by different structures in the collected data, it is possible to uncover the latent subpopulations. Model-based clustering is a statistical tool that can be entertained for this purpose.

However, for the given data sets, clustering is impeded by the high dimensionality and longitudinal character of the data, and by the fact data is not always fully observed.

Often a set of outcomes is measured over time, resulting in multivariate longitudinal data. Due to the dimension of the joint distribution of the random effects, computational problems are likely to occur when mixture models are applied to a multivariate longitudinal setting. In this dissertation, we have proposed an algorithm to reveal latent subgroups for multivariate repeated outcomes. The approach is inspired by work of Fieuws and Verbeke (2008), the authors perform a discriminant analysis for repeatedly measured data. Instead of maximizing the full joint model a pseudo-likelihood approach, based on bivariate joint models for the repeated outcomes, was utilized. The iterative algorithm mimics a partition cluster method. The performance of the proposed algorithm was looked into by means of a simulation study.

Complexity is enhanced when observations are densely sampled over a continuum, e.g., time. In such a situation, the data are generated by an underlying smooth function or by a set of smooth functions that are not easily described by a mathematical expression. Functional data analysis methods are used to reduce the dimensionality of the data and latent subgroups are then discovered for the reduced data. Such an approach was, e.g., used in Jacques and Preda (2013). The fact that their approach uses a data reduction technique, requiring a complete data structure, limits the practical usefulness of the cluster algorithm. In this dissertation, we combined methods from functional data analysis, missing data and ensemble clustering to discover latent subgroups in high-dimensional data, in terms of the number of responses and the number of repeated measurements, contaminated by missing observations. Data were completed by means of multiple imputation, whereupon the model-based clustering of Jacques and Preda (2013) was used to find latent subgroups in the principal components, and finally ensemble clustering was employed to summarize the set of partitions into a final data partition. The amalgamation of statistical techniques allows to cluster complex data and at the same time to quantify the influence of the missing data on the composed groups. Ensemble clustering has to our knowledge not yet been used in combination with multiple imputation. A small simulation study was designed to explore its utility.

When the missing-data mechanism is believed to be non-random, the joint distribution of the data and the missing-data indicators should be considered. In this work, we have investigated various mixture models for non-random missingness as proposed by Muthén et al. (2011). We assessed the vulnerability of the results not only in terms of the number of clusters, the cluster-specific profiles, but also in terms of the group-membership probabilities. It is however impossible to decide on the best model, since all models rely on non-verifiable assumptions. We have illustrated how an ultimate outcome, related to the growth curves, can be supportive in choosing between the models.

Cluster results are of course also sensitive to outlying and influential observations. We used ideas presented by Lesaffre and Verbeke (1998) for a mixed model, and applied local-influence diagnostics to a mixture model. This allowed quantification of the influence an observation has on the cluster-specific profiles and on the group-membership probabilities of the other observations.

A number of issues were not or partially addressed in this dissertation and could be topic for further deepening.

The local-influence diagnostics, described in Chapter 8, are obtained by introducing weights for the log-likelihood contributions of single subjects, where focuss was on the influence of a single subject. Other perturbation schemes could be worthwhile to consider. The method of local influence could for example be used to study the impact of MNAR mechanisms on the cluster result.

The approach presented in Chapter 6, to cluster sets of smooth but incomplete functions, has a number of flaws. The method is sensitive to the class-specific orders to approximate the pseudo-likelihood for the functional data. A heuristic test is used to determine these orders. More formal procedures could be implemented. Determination of the number of clusters is also difficult. An information criterion similar to the one proposed by Breaban and Luchian (2011) could be developed for functional data. This would address the selection of the class-specific orders and the optimal number of clusters at the same time. The set of partitions is reduced into a final partition by means of consensus clustering. This step could be replaced by other techniques, for example a latent class analysis with the cluster-indicators as variables.

We have focussed on repeated measurements for continuous responses. But the methods presented in this dissertation, can be applied to non-continuous responses/data with other structures. It would be interesting to see how the methods perform in spatial or temporal-spatial settings and for combinations of responses not belonging to the same parametric family.

Bibliography

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

[2] Abraham, C., Cornillon, P.A., Matzner-Lober, E., Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scand J Statist*, **30**, 581–595.

[3] Agresti, A. (2002). *Categorical Data Analysis.* New York: John Wiley & Sons.

[4] Arellano-Valle, R.B., Bolfraine, H., and Lachos V.H. (2005). Skew-normal Linear Mixed Models. *Journal of Data Science*, **3**, 415–438.

[5] Arnold, B.C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhya B.*, **53,** 233–243.

[6] Basagaña, X., Barrera-Gömez, J., Benet, M., Antö, J.M., and Garcia-Aymerich, J. (2013). A framework for multiple imputation in cluster analysis. *American Journal of Epidemiology*, **177,** 718–25.

[7] Beckman, R.J., Nachtsheim, C.J., and Cook, R.D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, **29,** 413–426.

[8] Berrendero, J.R., Justel, A., and Svarc, M. (2011). Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, **55,** 2619–2634.

[9] Besag, J. (1975). Statistical Analysis of Non-Lattice Data. *The Statistician*, **24,** 179–195.

[10] Besse, P. and Ramsay, J.O. (1986). Principal components analysis of sampled functions. *Psychometrika*, **51**, 285–311.

[11] Beunckens, C., Molenberghs, G., and Kenward, M.G. (2005). Tutorial: Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical Trials,* **2**, 379–386.

[12] Beunckens, C., Molenberghs, G., Verbeke, G., and Mallinckrodt, C. (2008). A latent-class mixture model for incomplete longitudinal Gaussian data. *Biometrics* **64**, 96–105.

[13] Bezdek, J.C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms.* Kluwer Academic Publishers Norwell, MA, USA.

[14] Biernacki, C. (2004). Initializing EM Using the Properties of its Trajectories in Gaussian Mixtures. *Statistics and Computing*, **14,** 267–279.

[15] Bonner, R.E. (1964). On some clustering techniques. *Journal of Research and Development*, **8**, 22–32.

[16] Bouveyron, C. , and Jacques, J. (2011). Model-based Clustering of Time Series in Group-specific Functional Subspaces. *Advances in Data Analysis and Classification*, **5**, 281–300.

[17] Bradley, P. and Fayyad, U. (1998). Defining Initial Points for K-Means Clustering. *Proceedings of the 15th International Conference on Machine Learning (ICML98)*, Microsoft Research, May 1998.

[18] Breaban, M. and Luchian, H. (2011). A unifying criterion for unsupervised clustering and feature selection. *Pattern Recognition*, **44,** 854–865.

[19] Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed modles. *Jounral of the American Statistical Association*, **88,** 9–25.

[20] Bruckers, L., Molenberghs, G., Poncelet, J., Brouns, K., Cuypers, W., Slaets, H., and Vanheyst, I. (2000). Identificatie en inschatting van de omvang van de groep patiënten met persisterend storend gedrag. *Acta Hospitalia*, **40**, 21–30.

[21] Bruckers, L., Serroyen, J., Molenberghs, G., Slaets, H., and Goeyvaerts, W. (2010). Latent class analysis of persistent disturbing behaviour patients by using longitudinal profiles. *Journal of Royal Statistical Society, Series C*, **59**, 495-512.

[22] Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park: Sage Publications.

[23] Cardot, H. (2000). Nonparametric estimation of the smoothed principal components analysis of sampled noisy functions. *J. Nonparametr. Stat.*, **12,** 503–538.

[24] Carpenter, J., Kenward, M.G., Evans, S., and White, I. (2004). Last observation carry-forward and last observation analysis. Letter to the Editor. *Statistics in Medicine*, **23**, 3241–3244.

[25] Carpenter, J. and Kenward, M. (2013). *Multiple Imputation and Its Application.* New York: John Wiley & Sons.

[26] Castro, P.E., Lawton, W.H., and Sylvestre, E.A. (1986). Principal modes of variation for process with continuous sample curves. *Technometrics*, **28**, 329–337.

[27] Cattell, R.B. (1966). The Scree Test for the Number of Factors. *Multivariate Behavioral Research*, **1,** 245–276.

[28] Cavanaugh, J. E. and Shumway, R. H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *J. Stat. Plan. Infer.*, **67,** 45–65.

[29] Celeux, G., Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, **13,** 195–212.

[30] Cerioli, A. (1998). A New Method for Detecting Influential Observations in Nonhierarchical Cluster Analysis. Advances in Data Science and Classification. In Rizzi, A. , Vichi, M., and Bock, BOCK H. *Studies in Classification, Data Analysis, and Knowledge Organization.* Springer-Verlag Berlin Heidelberg.

[31] Cheng, R. and Milligan, G.W. (1996). Measuring the influence of individual data points in a cluster analysis. *Journal of Classification*, **13**, 315–335.

[32] Clark, S. and Muthén, B. (2009). Relating latent class analysis results to variables not included in the analysis. *h*ttps://www.statmodel.com/download/relatinglca.pdf.

[33] Cook, R. and Weisberg, S. (1982). *Residuals and influence in regression.* New York, NY: Chapman & Hall.

[34] Cook, R. (1986). Assessment of Local Influence. *Journal of the Royal Statistical Society. Series B*, **48**, 133–169.

[35] Cuesta-Albertos, J.A., Gordaliza, A., and Matran, C. (1997). Trimmed kMeans: an Attempt to Robustify Quantizers. *The Annals of Statistics*, **25**, 553–576.

[36] Curran, P.J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, **38**, 529–569.

[37] Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Multivariate Anal*, **12**, 136–154.

[38] Day, W.H.E. (1986). Foreword: comparison and Consensus of Classifications. *Journal of Classification*, **3,** 183–185.

[39] Delaigle, A. and Hall, P. (2010). Defining probability density for a distribution of random functions. *The Annals of Statistics*, **38,** 1171–1193.

[40] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[41] Dendale, P., De Keulenaer, G., Troisfontaines, P., Weytjens, C., Mullens, W., Elegeert, Y., Ector, B., Houbrechts, M., Willekens, K., and Hansen D. (2012). Effect of a Telemonitoring-facilitated Collaboration Between General Practitioner and Heart Failure Clinic on Mortality and Rehospitalization Rates in Severe Heart Failure The TEMA-HF 1 (Telemonitoring in the Management of Heart Failure) Study. *European Journal of Heart Failure*, **14,** 333–340.

[42] Diggle, P.J. and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *A*pplied Statistics, **43**, 49–93.

[43] Diggle, P.J., Heagerty, P.J., Liang, K.Y., and Zeger, S.L. (2002). *A*nalysis of Longitudinal Data (2n ed.), Oxford Science Publications, Oxford: Clarendon Press.

[44] Dimitriadou, E., Weingessel, A., and Hornik, K. (2002). A combination scheme for fuzzy clustering. *International Journal of Pattern Recognition and Artificial Intelligence*, **16,** 901–912.

[45] Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis, Second Edition.* New York: John Wiley & Sons.

[46] Efron, B. and Tibshirani. R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[47] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Verlag.

[48] Fieuws, S. and Verbeke, G. (2008). Predicting Renal Graft Failure Using Multivariate Longitudinal Profiles. *Biostatistics,* **9,** 419–431.

[49] Fleiss, J.L. (1981). *Statistical methods for rates and proportions. 2nd ed.* New York: John Wiley & Sons.

[50] Fraley, C. and Raftery A.E. (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, **41**, 578–588.

[51] Gao, X. and Song, P.X.K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, **105,** 1531–1540.

[52] Geys, H., Molenberghs, G., and Ryan, L. (1999). Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, **94,** 34–745.

[53] Gibbons, R.D., Hedeker, D., Elkin, I., Waternaux, C., Kraemer, H.C., Greenhouse, J.B., et al. (1993). Some conceptual and statistical issues in analysis of longitudinal psychiatric data. *Arch. Gen. Psychiatry* **50**, 739–750.

[54] Goldstein, H. (1995). *Multilevel Statistical Models.* Kendall's Libary of Statistics 3. London: Arnold.

[55] Goodman, L.A. and Kruskal, W.H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, **49**, 732–764.

[56] Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61,** 215–231.

[57] Gordon, A.D. (1999). *Classification* (second edition). Chapman & Hall/CRC, Boca Raton, Florida.

[58] Gordon, A.D. and Vichi M. (2001). Fuzzy partition models for fitting a set of partitions. *Psychometrika*, **66,** 229–248.

[59] Gower, J.C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, **27,** 857–871.

[60] Hajnal, I. and Loosveldt, G. (2000). *Data Analysis, Classification, and Related Methods Studies in Classification, Data Analysis, and Knowledge Organization.* Springer Berlin Heidelberg.

[61] Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal component analysis. *J.R. Stat. Soc. Ser. B Stat. Methodol.*, **68,** 109–126.

[62] Han, J. and Kamber, M. (2001). *D*ata Mining: Concepts and Techniques. Morgan Kaufmann Publishers, USA.

[63] Haughton, D.M.A. (1988). On the choice of a model to fit data from an exponential family. *Ann. Statist.*, **16**, 342–355.

[64] Hedeker, D. and Gibbons, R.D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, **2,** 64–78.

[65] Hedeker, D. and Gibbons, R.D. (2006). *L*ongitudinal Data Analysis. John Wiley & Sons, Inc., Hoboken, New Jersey.

[66] Honaker, J., King, G., and Blackwell, M. (2009), *Amelia II: A Program for Missing Data.*

[67] Honaker, J. and King, G. (2010). What to Do about Missing Values in Time-Series Cross-Section Data. *American Journal of Political Science*, **54,** 561–581.

[68] Hornik, K. (2005). A CLUE for CLUster Ensembles. *Journal of Statistical Software*, **14**.

[69] Hotelling, H. (1933). Analysis of a Complex of Statistical Variables Into Principal Components. *Journal of Educational Psychology*, **24,** 417–441 and 498–520.

[70] Hu, X. and Xu, L. (2003). A Comparative Study of Several Cluster Number Selection Criteria. *Intelligent Data Engineering and Automated Learning Lecture Notes in Computer Science,* **2690,** 195–202.

[71] Ibrahim, J.G., Zhu, H., and Tang. N. (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association*, **103,** 1648–1658.

[72] Ieva, F., Paganoni, A., Pigoli, D., and Vitelli, V. (2011). ECG signal reconstruction, landmark registration and functional classification. In: 7th Conference on Statistical Computationand Complex System. Padova.

[73] Jackson, D.A. (1993). Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. *Ecology*, **74,** 2201–2014.

[74] Jacques, J. and Preda, C. (2012). Model-based clustering of functional data. *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges*, 459–464.

[75] Jacques, J. and Preda, C. (2013). Funclust: a curves clustering method using functional random variable density approximation. *Neurocomputing*, **112**, 164-171.

[76] Jacques, J. and Preda, C. (2014). Model-based clustering of multivariate functional data. *Computational Statistics and Data Analysis*, **71,** 92–106.

[77] Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification* , DOI 10.1007/s11634-013-0158-y.

[78] Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data.* Prentice Hall.

[79] Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., and Van Steen, K. (2003). A Local influence approach applied to binary data from a psychiatric study. *Biometrics*, **59**, 410–419.

[80] Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, Pearson Prentice Hall.

[81] Jolliffe I.T., Jones B., and Morgan B.J.T. (1995). Identifying Influential Observations in Hierarchical Cluster Analysis. *Journal of Applied Statistics*, **22**, 61–80.

[82] Jones, R. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine*, **30,** 3050–3056.

[83] Jones, M.C. and Rice, J. (1992). Displaying the important features of large collections of similar curves. *Amer. Statist.*, **46,** 140–145.

[84] Jones, B.L., Nagin, D.S., and Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods and Research*, **29**, 374–393.

[85] Karhunen, K. (1947). Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.,* **37,** 1–79.

[86] Katz, L. and Powell, J. H. (1953). A proposed index of the conformity of one sociometric measurement to another. *Psychometrika*, **18,** 249–256.

[87] Kenward, M.G. (1998). Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in Medicine* **17**, 2723–2732.

[88] Kenward M., Molenberghs G., and Thijs H. (2003). Pattern-mixture models with proper time dependence. *Biometrika* **90**, 53–71.

[89] Kim, S., Kwon, S., and Cook, D. (2000). Interactive visualization of hierarchical clusters using MDS and MST. *Metrika*, **51**, 39–51.

[90] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38,** 963–974.

[91] Land, K.C. and Nagin, D.S. (1996). Micromodels of Criminal Careers: A Synthesis of the Criminal Careers and Life Course Approaches via Semiparametric Mixed Poisson Regression Models, with Empirical Models. *Journal of Quantitative Criminology*, **12,** 163–191.

[92] Leclerc, B. (1998). *Consensus of classifications: the case of trees. In: Advances in Data Science and Classification, Studies in Classification, Data Analysis and Knowledge Organization.* Berlin, Springer-Verlag, pp. 81-90.

[93] Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, **54**, 570–582.

[94] Lindstrom, M.J. and Bates, D.M. (1990). Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics*, **46**, 673–687.

[95] Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125–134.

[96] Little, R.J.A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471–483.

[97] Little, R.J.A. (1995). Modeling the Drop-Out Mechanism in Longitudinal Studies. *Journal of the American Statistical Association*, **90**, 1112–1121.

[98] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.

[99] Lo, Y., Mendell, N.R., and Rubin, D.B. (2001). Testing the number of components in a normal mixture. *Biometrika*, **88**, 767–778.

[100] Loève, M. (1978). *Probability theory. Vol. II, 4th ed. Graduate Texts in Mathematics.* Springer-Verlag.

[101] Lora, A., Cosentino, U., Rossini, M.S.and Lanzara, D. (2001). A Cluster Analysis of Patients with Schizofrenia in Community Care. *Psychiatric Services*, **52**, 682–684.

[102] Mahalanobis, P.C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, **2(1)**, 49—55.

[103] McLachlan, G. (1987). On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture. *Applied Statistics*, **36**, 318–324.

[104] McLachlan G. and Basford K. (1988). *Mixture models, Inference and Applications to Clustering.* Marcel Dekker, Inc.

[105] McLachlan G. and Peel D. (2000). *Finite mixture models.* New York: Wiley.

[106] Melnykov, V. (2013). Challenges in model-based clustering. *W*IREs Comput Stat, **5**, 135–148.

[107] Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

[108] Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies.* Chichester: John Wiley & Sons.

[109] Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M.G. (2008). Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of the Royal Statistical Society, Series B*, **70**, 371–388.

[110] Molenberghs, G., Njeru Njagi, E., Kenward, M.G., and Verbeke, G. (2012). Enriched-data problems and essential non-identifiability. *International Journal of Statistics in Medical Research*, **1**, 16–44.

[111] Mun, J. and Lindstrom, M.J. (2013). Diagnostics for repeated measurements in linear mixed effects models. *Statistics in Medicine*, **32**, 1361–1375.

[112] Munson, C.E. (2001). *The Mental Health Diagnostic Desk Reference.* New York: Haworth Press.

[113] Muthén, L. K. and Muthén, B. O. (1998–2012). *Mplus User's Guide. Seventh Edition.* Los Angeles, CA: Muthén and Muthén.

[114] Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM-algorithm. *Biometrics*, **55,** 463–469.

[115] Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. *In D. Kaplan (ed.), Handbook of quantitative methodology for the social sciences.* Newbury Park, CA: Sage Publications.

[116] Muthén, B., Asparouhov, T., Boye, M.E., Hackshaw, M.D., and Naegeli, A.N. (2009). *Applications of continuous-time survival in latent variable models for the analysis of oncology randomized clinical trial data using Mplus.* Technical Report.

[117] Muthén, B., Asparouhov, T., Hunter, A. and Leuchter, A. (2011). Growth modeling with non-ignorable dropout: Alternative analyses of the STAR*D antidepressant trial. *Psychological Methods*, **16**, 17–33.

[118] Nagin, D.S. (1999). Analyzing developmental trajectories: a semiparametric, group-bases approach. *Psychological Methods*, **4,** 139–157.

[119] Nagin, D.S. and Land, K.C. (1993). Age, criminal careers, and population heterogeneity: specification and estimation of a nonparametric, mixed Poisson model. *Criminology*, **31,** 327–362.

[120] Nagin, D.S. and Tremblay, R.E. (2001). Analyzing developmental trajectories of distinct but related behaviours: a group-based method. *Psychological Methods*, **6**, 18–34.

[121] Nagin, D.S. (2005). *Group-based Modeling of Development.* Cambridge, MA.: Harvard University Press.

[122] Nylund, K.L., Asparouhov, T., and Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: An Interdisciplinary Journal*, **14,** 535–569.

[123] Ouwens, M.J.N.M., Tan, F.E.S., and Berger, M.P.F. (2001). Local influence to detect influential data structures for generalized linear mixed models. *Biometrics*, **57**, 1166–1172.

[124] Pena J., Lozano J., and Larranaga P. (1999). An Empirical comparison of four initialization methods for the K-means algorithm. *Pattern Recognition Letters*, **20,** 1027–1040.

[125] Peng, J., and Müller, H.G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Ann. Appl. Stat.*, **2**, 777–1142

[126] Pinheiro, J.C. and Bates, D.M. (1995). Approximations to the Log-likelihood Function in the Nonlinear Mixed-effects Model. *Journal of Computational and Graphical Statistics*, **4**, 12–35.

[127] Pinheiro, J.C., Liu, C., and Wu Y.N. (2001). Efficient Algorithms for Robust Estimation in Linear Mixed-Effects Models Using the Multivariate t Distribution. *Journal of Computational and Graphical Statistics*, **10**, 249–276.

[128] Piquero, A.R., Brame R., Mazzerole P., and Haaplanen R. (2002). Crime in Emerging Adulthood. *Criminology*, **40,** 137–70.

[129] Putter H., Vos T., de Haes H., and van Houwelingen H. (2008). Joint analysis of multiple longitudinal outcomes: Application of a latent class model. *Statistic in Medicine*, **27,** 6228–6249.

[130] Raftery, A.E. and Dean, N. (2006). Variable Selection for Model-Based Clustering. *Journal of American Statistical Association*, **101,** 168–178.

[131] Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, **27,** 85–95.

[132] Ramaswamy, V., DeSarbo, W., Reibstein, D., and Robinson, W. (1993). An empirical pooling approach for estimating marketing elasticities with PIMS data. *Marketing Science*, **12**, 103–124.

[133] Ramsay, J.O. and Silverman, B.W. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* New York: Springer-Verlag.

[134] Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis, 2nd ed.* New York: Springer.

[135] Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66,** 846–850.

[136] Rice, J.A. and Silverman, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, **53**, 233–243.

[137] Rice, J.A. (2004). Functional and longitudinal data analysis: Perspectives on smoothing. *Statist. Sinica*, **14,** 631–647.

[138] Roy, J. and Lin, X. (2000). Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics*, **56,** 1047–1054.

[139] Roy, J. (2003). Modeling Longitudinal Data with Nonignorable Dropouts Using a Latent Dropout Class Model. *Biometrics*, **59**, 829–839.

[140] Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

[141] Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

[142] Sain, S.R., Gray, H.L., Woodward, W.A., and Fisk. M.D. (1999). Outlier Detection from a Mixture Distribution When Training Data Are Unlabeled. *Bulletin of the Seismological Society of America*, **89**, 294–304.

[143] Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London: Chapman and Hall.

[144] Schwarz, G. (1978). Estimating the dimension of a model. *The annals of Statistics*, **6**, 461–464.

[145] Seber, G.A.F. (1984). *Multivariate Observations.* Wiley, New York.

[146] Serroyen, J., Bruckers, L., Rogiers, G., and Molenberghs, G. (2010) Characterizing persistent disturbing behaviour using longitudinal and multivariate techniques. *Journal of Applied Statistics*, **37**, 341–355.

[147] Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, **52**, 333–343.

[148] Soromenho, G. (1993). Comparing approaches for testing the number of components in a finite mixture model. *Computational Statistics*, **9,** 65–78.

[149] Spiessens B., Verbeke G., and Komàrek A. (2002). *A* SAS-macro for the classification of longitudinal profiles using mixtures of normal distributions in nonlinear and generalized linear models. http://www.med.kuleuven.ac.be/biostat/research/software.htm.

[150] Strehl, A. and Ghosh, J. (2002). Cluster ensembles. A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, **3** 583–617.

[151] Tarpey, T. and Kinateder, K. (2003). Clustering functional data. *J Classif*, **20**, 93–114.

[152] Thijs H., Molenberghs G., Michiels B., Verbeke G., and Curran D.(2002). Strategies to fit pattern-mixture models. *Biostatistics.*, **3**, 245–265.

[153] van Buuren S., Boshuizen H.C., and Knook D.L. (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine*, **18,** 681–694.

[154] Varin, C. and Vidoni, P. (2005), A note on composite likelihood inference and model selection. *Biometrika*, **92,** 519–528.

[155] Verbeke G. and Lesaffre E. (1996). A linear mixed effects model with heterogeneity in the random effects population. *Journal of the American Statistical Association*, **91**, 217–221.

[156] Verbeke G. and Lesaffre E. (1997). The linear mixed model. A critical investigation in the context of longitudinal data, In: *P*roceedings of the Nantucket conference on Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions, Gregoire, T. (Ed.), Lecture Notes in Statistics. New York: Springer.

[157] Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.

[158] Verbeke G., Molenberghs G., Thijs H., Lesaffre E., and Kenward M.G. (2001). Sensitivity analysis for non-random dropout: A local influence approach. *Biometrics*, **57**, 7–14.

[159] Verbeke, G. and Molenberghs, G. (2010). Arbitrariness of models for augmented and coarse data, with emphasis on incomplete-data and random-effects models. *Statistical Modelling*, **10**, 391–419.

[160] Wang, S., Woodward, W.A., Gray, H.L., Wiechecki, S., and Sain, S.R. (1997). A new test for outlier detection from a multivariate mixture distribution. *Journal of Computational and Graphical Statistics*, **6**, 285–299.

[161] Wolfinger, R. and O‘Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation adn Simulation*, **48**, 233–243.

[162] Woodroofe, M. (1982). On model selection and the arc-sine laws. *Ann. Statist.*, **10**, 1182–1194.

[163] Wu, M.C. and Bailey, K.R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics* **45**, 939–955.

[164] Wu, M.C. and Caroll, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics*, **44**, 175–188.

[165] Yamamoto, M. (2012). Clustering of functional data in a low-dimensional subspace. *Advances in Data Analysis and Classification,* **6**, 219–247

[166] Yang, C. (2006). Evaluating latent class analyses in qualitative phenotype identification. *Computational Statistics and Data Analysis*, **50**, 1090–1104.

[167] Yao, F., Müller, H.G. and Wang, J.L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc*, **100**, 577–590.

Samenvatting

Het blootleggen van natuurlijke groeperingen binnen een set van multivariate gegevens noemt men clusteren. De term clusteren verscheen voor het eerst in een artikel gepubliceerd in 1954, met antropologische gegevens. Het K-means algoritme, één van de populairste cluster technieken, werd voor het eerst in 1955 gepubliceerd. Het organiseren van gegevens in logische en natuurlijke groeperingen vindt toepassingen in een groot aantal disciplines. Zo werd clusteren succesvol toegepast in o.a. transcriptomics om genen met dezelfde functie te groeperen, in astronomie om sterren te klasseren, in plant- en dierkunde om organismen in gemeenschappen te verdelen, in geneeskunde om patiënten met dezelfde klinische en psychosociale noden te identificeren, ...

De term 'cluster analyse' omvat in feite verschillende algoritmen, alle met als doel gegevens te groeperen zodat gelijkaardige observaties, volgens een bepaalde afstandsmaat, in dezelfde groep zitten en ongelijksoortige observaties in verschillende groepen. In de literatuur wordt een waaier van cluster algoritmen beschreven, deze verschillen in hun definitie van waaruit een cluster bestaat en hoe deze clusters gedetecteerd worden.

Traditionele cluster algoritmen, zoals hiërarchische methoden en (iteratieve) verdelingsmethoden, worden veelvuldig en met succes toegepast. Maar voor complexe data structuren –zoals het geval is bij herhaalde metingen, ruimtelijke gegevens, enz. – blijken deze technieken minder efficiënt in het blootleggen van de natuurlijke groeperingen. Voor dit soort van gegevens kan het wenselijk zijn om een onderliggend model voor de gegevens te veronderstellen. Het model specifieert enerzijds de structuur van de gegevens (bijv. gemiddelde evolutie, correlatie tussen de metingen van eenzelfde individu) en laat anderzijds toe dat de populatie heterogeen is en uit een (eindig) aantal subpopulaties/clusters bestaat. Een dergelijk model is bijvoorbeeld het finite-mixture model. In een finite-mixture benadering wordt elke cluster wiskundig vertegenwoordigd door een parametrische verdeling, bijv. de normale of Poisson verdeling. De volledige set van gegevens wordt beschreven als een gewogen mengeling (mixture) van de cluster specifieke verdelingen. De parameters in dit model zijn enerzijds de parameters uit de cluster specifieke verdelingen en anderzijds de a-priori kansen om tot de verschillende clusters te behoren. Waarschijnlijkheidstheorie wordt gebruikt om de parameters te bepalen zodat de fit tussen het opgegeven model en de gegevens optimaal is. Clusters worden vervolgens gedefinieerd als observaties die, onder dit model, waarschijnlijk tot dezelfde verdeling behoren. (On)zekerheid over toewijzing van een observatie tot een bepaalde cluster wordt gevat in de a-posteriori kans. Hoofdstuk 4 introduceert finite-mixture modellen en illustreert de toepassing voor gegevens uit de geestelijke gezondheidszorg.

Ondanks het feit dat cluster analyse veelvuldig gebruikt wordt in toegepaste wetenschappen en onderwerp is van een groot aantal methodologische artikels, zijn er nog steeds een reeks open en controversiële vragen: de validiteit van de uiteindelijke groepering, de keuze van de te beschouwen response-variabelen in de cluster analyse, de beslissing van het optimaal aantal clusters, het probleem van lokale oplossingen voor de te maximaliseren waarschijnlijkheidsfunctie (voor model-gebaseerde technieken), gevoeligheid van de uitkomst van het algoritme aan de gekozen startwaarden, . . .

In dit proefschrift behandelen we een aantal beperkingen van cluster analyses die tot nu toe minder aandacht kregen in de statistische literatuur. Deze onderwerpen worden het eenvoudigst behandeld in een model-gebaseerde cluster benadering. Specifiek focussen we op het clusteren van hoog-dimensionale gegevens en bestuderen we het effect van enerzijds ontbrekende gegevens en anderzijds invloedrijke gegevens.

Cluster analyse wordt uitdagend wanneer de dimensionaliteit van de gegevens stijgt. De reden hiervoor is het toenemend aantal parameters in de waarschijnlijkheidsfunctie. In deze thesis zijn twee hoofdstukken gewijd aan dit onderwerp. Hoofdstuk 5 beschouwt een setting waar de dimensionaliteit toeneemt doordat gelijktijdig een aantal response-variabelen doorheen de tijd opgemeten worden. Hoofdstuk 6 behandelt de situatie waarbij één of meerdere responsevariabelen bijna continue opgemeten worden, wat resulteert in functies.

Hoofdstuk 5 beschrijft een algoritme om soortgelijke patronen te ontdekken in een multivariate herhaalde metingen setting. Zoals reeds aangegeven kan clusteren van herhaalde metingen vlot door middel van een finite-mixture model. Voorbeelden beschreven in de literatuur bespreken echter enkel situaties waar tot een maximum van drie response-variabelen gelijktijdig beschouwd worden. Meer response-variabelen leidt snel tot computationele problemen. Hoofdstuk 5 stelt een procedure voor om in een set van longitudinale data voor meerdere response variabelen, groepen bloot te leggen. De procedure is gebaseerd op pseudo-waarschijnlijkheid schattingen. Een EEG studie bij ratten, waarbij verschillende hersensignalen (α, β, γ, δ, ... - golven) geregistreerd worden, dient als case-studie.

Tegenwoordig is men in staat om gegevens bijna continue op te meten, we denken bijvoorbeeld aan ruimtelijke metingen op een zeer dicht raster of metingen kort in de tijd. Bloeddruk en hartslag bijvoorbeeld kunnen door middel van een elektronisch apparaat continue gemonitord worden. Zelfs indien men maar in één response-variabele geïnteresseerd is, resulteert dit in hoog dimensionale gegevens. In dergelijke situaties, worden de waarnemingen gezien als een verwezenlijking van een vloeiend onderliggend proces en spreekt men van functionele data. Voorafgaand aan de eigenlijke statistische verwerking moet de dimensionaliteit van de functionele data omzeild worden. Vaak gebeurt dit via een principaal-component analyse. Populatie heterogeniteit, m.a.w. aanwezigheid van natuurlijke groeperingen, in de oorspronkelijke curves wordt ook weerspiegeld in de gereduceerde gegevens. Een cluster analyse voor functionele gegevens, wanneer deze bestaat uit bivariate functies, werd door Jacques en Preda (2013) beschreven. Een essentiële bouwsteen van deze techniek is de principaal-component analyse, dewelke volledige datastructuren veronderstelt. Hierdoor wordt de toepasbaarheid van hun procedure beperkt door records met ontbrekende gegevens. Ontbrekende gegevens zijn echter bijna onvermijdelijk in longitudinale studies.

Hoofdstuk 6 licht toe hoe een combinatie van technieken uit verschillende domeinen gebruikt kan worden om hoog dimensionale onvolledige gegevens te clusteren. Volledige datasets worden eerst bekomen door meervoudige imputatie van de ontbrekende informatie. Elk van deze datasets wordt vervolgens onderworpen aan een cluster algoritme voor functionele data. Dit resulteert in een collectie van partities van de data, ook een ensemble genoemd. Om op basis van de bekomen collectie van groeperingen tot een finale consensus groepering te komen, wordt ensemble clustering aangewend. Een studie over hartfalen wordt als illustratie gebruikt. Gedurende 6 maanden, na ontslag uit het ziekenhuis, wordt dagelijks informatie over gewicht, hartslag, diastole en systole bloeddruk van de patiënten via een telemonitoring apparaat verzameld. Voor het merendeel van de patiënten zijn er periodes dat de metingen ontbreken.

Bij het vervolledigen van de gegevens wordt verondersteld dat de onvolledigheid van de gegevens afhangt van de geobserveerde metingen maar, gegeven deze, niet van de niet geobserveerde metingen (MAR, willekeurig ontbreken). Indien dit niet het geval is, wordt het mechanisme dat aanleiding geeft tot het ontbreken van gegevens MNAR (niet-willekeurig ontbreken) genoemd. MAR is vaak een beperkende veronderstelling en er kan nooit definitief uitgesloten worden dat het mechanisme niet verder afhangt van ontbrekende informatie. Zodra de onvolledigheid MNAR is, is het nodig een expliciet model voor de ontbrekende gegevens te formuleren. De gezamenlijke verdeling van de metingen en het onvolledigheidsproces moet dan beschouwd worden, dit resulteert in o.a. selectie modellen en pattern-mixture modellen. Er is reeds veel gepubliceerd over niet-willekeurig ontbrekende gegevens, maar weinig over MNAR modellen voor cluster-doeleinden. Hoofdstuk 7 presenteert een aantal MNAR-modellen waarbij een finite-mixture model verondersteld wordt voor de metingen. Elk van deze modellen gaat uit van niet te verifiëren veronderstellingen. De modellen worden toegepast op slagader-diameter bepalingen bij patiënten met abdominale aorta aneurysma (AAA-studie). De resultaten van de verschillende modellen worden vergeleken in termen van de gemiddelde cluster specifieke profielen, voor de a-posteriori kansen en de definitieve groepering van de patiënten. Er zijn opmerkelijke verschillen in de resultaten, maar een beste model kiezen is moeilijk. Externe informatie, die samenhangt met de groepering kan helpen bij een verstandige ranking van de verschillende MNAR modellen. Voor de AAA-studie is deze externe informatie vervat in het feit of de patiënt al dan niet een operatie diende te ondergaan. Dergelijke oefening wordt geïmplementeerd en beschreven in Hoofdstuk 7.

Hoofdstuk 8, tenslotte, bestudeert de invloed van individuele observaties op het cluster resultaat. De parameter schatters, die de verschillende verdelingen in een finite-mixture model beschrijven, zijn onderhevig aan invloedrijke observaties en uitschieters. Het identificeren van uitschieters in een cluster analyse is reeds beschreven in de literatuur. Ook het bepalen van de invloed van individuele observaties in een herhaalde metingen setting voor homogene populaties en van het ontbreken van gegevens werd reeds bestudeerd. Hiervoor werd onder andere gebruik gemaakt van een lokale invloedsanalyse. Deze analyse werd tot nu toe echter nog niet toegepast voor een finite-mixtue model. In Hoofdstuk 8 worden de lokale invloedsstatistieken berekend voor een finite-mixture model. Deze aanpak maakt het mogelijk de invloed van een observatie op de resultaten van de cluster analyse te kwantificeren. Enerzijds de invloed op de parameter schatters, m.a.w. op de cluster specifieke gemiddelde profielen en finale partitie van de gegevens. Maar anderzijds is het ook mogelijk om de invloed van observatie i op de a-posterior kans van observatie j te bepalen. De invloed op de a-posterior kans kan aanzienlijk zijn, zelfs als observatie i geen invloed heeft op de samenstelling van de clusters. De techniek wordt geïllustreerd op de EEG data.