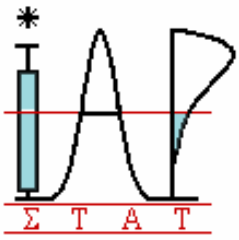

Interuniversity Attraction Pole P5/24

Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data

Workshop 4

HOW TO DEAL WITH HETEROGENEITY?

Program and Abstracts

Leuven, September 30, 2005

WORKSHOP 4 : PROGRAM

The theme of this workshop is one of the meta-modeling aspects as included in the original application of our network: *how to deal with heterogeneity*. In this regard, we will use a broad concept of heterogeneity that covers, for instance, nonstationarity in time series analysis, random effects as included in mixed and survival (frailty) models, and the presence of different subpopulations as assumed in mixture models. On this theme there will be joint contributions, each from at least two network partners. The joint contributions will be based on collaborative work (in the form of joint theoretical work, of data from one partner that are analyzed by another partner, or of any other form).

morning session (chairman: Paul De Boeck)

09:30 – 10:00 *welcome/coffee*

10:00 – 10:30 Sébastien Van Bellegem, Paul De Boeck and Kristof Vansteelandt:
Analysis of longitudinal multilevel emotion data via locally stationary time series.

10:30 – 11:00 Anestis Antoniadis, Jérémie Bigot and Rainer von Sachs:
A multiscale approach for statistical characterization of spatially and temporally heterogeneous brain response images.

11:00 –11:30 Domenico Giannone and Michele Lenza:
The Feldstein-Horioka fact.

11:30 – 12:00 *coffee break*

12:00 – 12:30 Caroline Beunckens, Geert Molenberghs and Geert Verbeke:
Sensitivity analysis for incomplete longitudinal data based on a heterogeneous pattern-mixture model.

12:30 – 12:50 Léopold Simar:
synthesis

12:50 – 14:00 *lunch*

afternoon session (chairman: Léopold Simar)

14:00 – 14:30 Yuri Goegebeur, Paul De Boeck, Geert Molenberghs and Geert Verbeke:
Local influence analysis of non-response in a speeded IRT model.

14:30 – 15:30 Paul Janssen, Arnost Komarek and Emmanuel Lesaffre:
Heterogeneity in clustered survival data.

15:30 – 16:00 *coffee break*

16:00 – 16:30 Taoufik Bouezmarni, Frank Rijmen and Paul De Boeck:
Smoothed kernels for generalized linear mixed models.

16:30 – 17:00 Hans-Hermann Bock and Iven Van Mechelen:
Bi-partitioning methods for data tables: New criteria and algorithms.

17:00 – 17:20 Ingrid Van Keilegom:
synthesis

WORKSHOP 4 : ABSTRACTS

Analysis of longitudinal multilevel emotion data via locally stationary time series

Sébastien Van Bellegem (UCL Louvain-la-Neuve), Paul De Boeck (KULeuven) and Kristof Vansteelandt (KULeuven)

This talk presents a recent collaboration between UCL and KUL-1 for the analysis of emotion data from a panel of 36 persons observed during two weeks. Two sources of heterogeneity are considered: First, for a given person, the observed time series can show some covariance nonstationary behavior; in this talk we propose to model this nonstationarity using a semiparametric model of local stationarity. Second, the level of nonstationarity (i.e., the degree of deviation from stationarity), may be nonidentical for all persons, and may appear as another source of heterogeneity.

A multiscale approach for statistical characterization of spatially and temporally heterogeneous brain response images

Anestis Antoniadis (UJF Grenoble), Jérémie Bigot (UPS Toulouse) and
Rainer von Sachs (UCL Louvain-la-Neuve)

In this project we use an approach of spatial multiscales for an improved characterization of functional pixel intensities of (medical) images. Examples are numerous, such as temporal dependence of brain response intensities measured by fMRI, or frequency dependence of NMR spectra measured at each pixel. The overall goal is to improve the misclassification rate in (unsupervised) clustering of the functional image content into a finite but unknown number of classes. Hereby we adopt a non-parametric point of view to reduce the functional dimensionality of the observed pixel intensities, modelled to be of a very general functional form, opposed to commonly used parametric feature extraction based on a priori knowledge on the nature of the functional response.

Before clustering is applied via an EM-algorithm for estimating a Gaussian mixture model in the domain of the discrete wavelet transform of the pixel intensities over time (or frequency), linear or non-linear wavelet thresholding is applied to denoise the observed intensity curves. Then a dimension reduction "common to all curves" is applied via a recently developed "aggregation estimator".

Our point of reference for comparisons are *monoscale* statistical models, which are typically being used to clean the map of noise and to help extract structure in the underlying measurements. They work at a pixel level resolution and result in a low degree of aggregation of the information underlying the data, since the appropriate choice of scale usually varies with spatial location. We intend to show improvements with our multiscale method, based on *Recursive Dyadic Partitioning* of the image to adaptively choose the locally best scale for clustering, by means of simulated and real data examples, and sketch some ideas for the theoretical treatment of encountered problems.

The Feldstein-Horioka fact

Domenico Giannone (ULB Bruxelles) and Michele Lenza (ULB Bruxelles)

This paper shows that general equilibrium effects can partly realize the high correlation between saving and investment observed in OECD countries. We introduce a novel factor augmented panel regression to control for general equilibrium effects where global shocks are allowed to affect each country with specific magnitude and lag structure. We show that the homogeneity restriction on the propagation of global shocks across countries is rejected by the data and biases the saving-retention coefficient estimated in previous studies. By relaxing this assumption, the saving-retention coefficient remains high in the 70s, but decreases considerably over time, becoming very small in the last two decades. This finding is explained by the increased capital mobility in OECD countries.

Sensitivity analysis for incomplete longitudinal data based on a heterogenous pattern-mixture model

Caroline Beunckens (UHasselt), Geert Molenberghs (UHasselt) and Geert Verbeke (KULeuven)

Standard methodology used to analyze incomplete longitudinal data is mostly based on methods such as last observation carried forward (LOCF), and complete case analysis (CC). Since they are based on extremely strong assumptions (even the strong MCAR assumption does not suffice to guarantee an LOCF analysis is valid) and their validity can be questioned, it is unfortunate that there is such a strong emphasis on these methods. In the selection model framework, under MAR, valid inference can be obtained through a likelihood-based analysis, including the linear mixed model and generalized linear mixed models, without the need for modeling the dropout process. In addition, weighted generalized estimating equations (WGEE) can be used and is valid under MAR. As a consequence, incomplete longitudinal data, both of a Gaussian and of a non-Gaussian nature, can easily be analyzed under the MAR assumption, using standard statistical software tools.

However, MNAR can never be entirely excluded, and one should therefore ideally supplement an ignorable analysis with a suitable chosen set of sensitivity analyses. One such route for sensitivity analysis is to consider pattern-mixture models or shared-parameter models. The latter can also be extended to a latent-class mixture model. Further, a local influence analysis was developed to detect subjects that strongly influence the conclusions.

Local influence analysis of non-response in a speeded IRT model

Yuri Goegebeur (KULeuven), Paul De Boeck (KULeuven), Geert Molenberghs (UHasselt) and Geert Verbeke (KULeuven)

When students are given a test, some of them do not respond to all items in the test, even when they were requested in an explicit way to do so. The non-response typically depends on latent traits such as examinee ability, and, hence, the resulting models are, using the terminology of Rubin (1976) and Little and Rubin (1987), missing not at random (MNAR). Unlike biostatistics, where models for dropout, in particular the selection and pattern-mixture models, are quite popular, psychometric models for omitted items are less widely used and are even relatively new. An early attempt to model non-response in IRT can be found in Lord (1983), some recent contributions are Mislevy and Wu (1996) and Pimentel et al. (2005). We propose a multinomial regression model for non-response in test data. The model is derived from a decision tree that describes the student's possible states and actions when he/she encounters an item. Under the proposed model, the probability of omission depends on item difficulty, examinee ability, an initial propensity to omit and a test speededness effect, where the latter three parameters are considered random in order to model examinee differences in those respects. Test speededness refers to testing situations in which some examinees do not have ample time to answer all questions. In the proposed model, test speededness increases the probability of omission of end-of-test items. The properties of the model are studied, in particular its relationship with classical IRT models such as 2PL and 3PL. The proposed model fits within the MNAR class, and, hence, as such models typically depend on rather strong assumptions and relatively little evidence from the data themselves, caution is in order when drawing conclusions about parameters and/or the non-response mechanism. The local influence measures proposed by Cook (1986) are used to assess the sensitivity of parameter estimates with respect to small perturbations of the model; in particular, we focus on the effect of including test speededness. The model and the related local influence analysis are illustrated with the SIMCE mathematics test data.

References:

Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society Series B*, *48*, 133-169.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Lord, F.M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, *48*, 477-482.

Mislevy, R.J., & Wu, P.K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Technical Report). Princeton, NJ: Educational Testing Service.

Pimentel, J.L., Glas, C.A.W., & Béguin, A.A. (2005, July). *Modeling nonignorable missing data in speeded tests*. Paper presented at the 70th Annual Meeting of the Psychometric Society, Tilburg, The Netherlands.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, *63*, 581-592.

Heterogeneity in clustered survival data

Paul Janssen (UHasselt), Arnost Komarek (KULeuven) and
Emmanuel Lesaffre (KULeuven)

When multicenter clinical trial data are analyzed, it has become more and more important to look for possible heterogeneity in outcome between centers. Beyond the investigation of such heterogeneity, it is also interesting to consider heterogeneity in treatment effect over centers. In this presentation we propose and compare two different approaches:

In the first approach, we propose an extension of the Cox model obtained by including a random center effect and a random treatment by center interaction, and we give a Bayesian approach to fit the proposed model.

In the second approach, we model multicenter survival data using an accelerated failure time (AFT) model where not only heterogeneity between centers is assumed but also heterogeneity with respect to the effect of covariates (e.g., different treatment effect for different centers). For the error distribution of the AFT model, we suggest a flexible expression, using either a classical normal mixture with unknown number of components and unknown locations and scales, or a penalized normal mixture with a prespecified (high) number of components.

The methods will be illustrated on data from an early breast cancer clinical trial, and on data from a bladder cancer clinical trial, both obtained from the European Organization for Research and Treatment of Cancer.

References:

Legrand, C., Ducrocq, V., Janssen, P., Sylvester, R., & Duchateau, L. (in press). A Bayesian approach to jontly estimate center and treatment by center heterogeneity in a proportional hazards model. *Statistics in Medicine*.

Komarek, A., & Lesaffre, E. (2005). *Bayesian accelerated failure time model for correlated interval-censored data with a normal mixture as error distribution.* Manuscript submitted for publication.

Komarek, A., & Lesaffre, E. (2005). *Bayesian accelerated failure time model with multivariate doubly-interval-censored data and flexible distributional assumptions*. Manuscript submitted for publication.

Smoothed kernels for generalized linear mixed models

Taoufik Bouezmarni (KULeuven), Frank Rijmen (KULeuven) and
Paul De Boeck (KULeuven)

Generalized linear mixed models have become a powerful parametric tool in psychometrics. Usually, the estimation of the fixed parameters is based on the normality assumption of the random effects. To avoid this restrictive assumption, we propose a new method to estimate the distribution function of the random effects in a generalized linear mixed model of the logistic type. The method is the smoothed kernel method.

The distribution estimate based on a nonparametric maximum likelihood approach is discrete, and the kernel method is used to smooth this density estimate. Simulation results are shown for the Rasch model. Also, a comparison of the new approach with a nonparametric maximum likelihood approach, with methods based on a normality assumption, and with methods based on a mixture of normal densities, are presented.

Bi-partitioning methods for data tables: New criteria and algorithms

Hans-Hermann Bock (RWTH Aachen) and Iven Van Mechelen (KULeuven)

The paper deals with the simultaneous partitioning of the rows and columns of a two-dimensional data table $X = (x)_{n \times p}$ into m row clusters $A_1, ..., A_m$ and l column clusters $B_1, ..., B_l$ (bi-partitioning). More specifically, we are looking for partitions that are optimal in the sense that they model certain aspects (types) of between-cluster heterogeneity. Beginning with an empirical example, we discuss various possible specifications of the heterogeneity concept and their formalizations in mathematical or statistical terms. Different data types necessitate different instances of heterogeneity and optimality. So we present a schematic overview of several types of data, and a range of associated 'meaningful' clustering criteria. Particular examples we will focus on include the following cases:

(a) real-valued case by variable data where we propose extensions of the classical SSQ clustering criterion
(b) real-valued data of the categorical prediction type (where rows and columns are looked at as values of two qualitative variables) with a 'maximum interaction' criterion
(c) contingency table data where we propose bipartitions with a maximum dependence between row and column clusters, or, alternatively, with an optimum approximation of the original data table by a bi-structured one.

Finally, we will briefly comment on numerical algorithms for calculating or approximating an optimal bi-partition.

References:

Van Mechelen, I., Bock, H.-H., & De Boeck, P. (2004). Two-mode clustering methods: A structured review. *Statistical Methods in Medical Research, 13*, 363-394.

Van Mechelen, I., Bock, H.-H., & De Boeck, P. (2005). Two-mode clustering. In B.S. Everitt & D.C. Howell (Eds.), *Encyclopedia of behavioral statistics* (pp. 2081-2086). Chichester: Wiley.