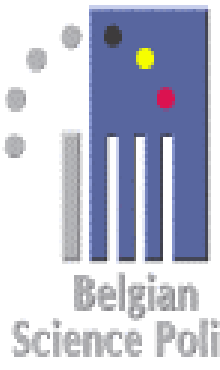
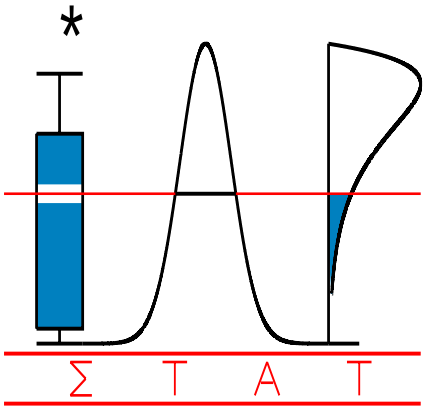


Progress Report 2003

IAP-network in Statistics

Contract P5/24

March 29, 2004



Contents

1 Accomplished Research Projects

1.1 Introduction and overview

1.1.1 Introduction

The research project has been built up around six Work Packages. Table 1 below gives the *main* contributors to each Work Package and indicates per package the partner that is coordinating the work.

Work package	Contributing partners
WP1: Functional estimation	UCL* , ULB, UJF
WP2: Time series	ULB*, UCL
WP3: Survival analysis	LUC*, UCL
WP4: Mixed models	KUL-2*, KUL-1, LUC
WP5: Classification and mixture models	KUL-1*, KUL-2, RWTH
WP6: Incompleteness and latent variables	LUC*, KUL-1, KUL-2

Table 1: *Main contributors per Work Package and coordinating partner per work package (indicated with a *).*

In the subsections we describe the progress that has been made in the various work packages. Within each Work Package we report on the progress that has been achieved on the various *primary objectives* mentioned in the research proposal. For each of the work packages we also indicate **interactions** with research results in other packages: this is done by refering to the other WP as **WP**. The references mentioned in the text can be found in Section 3 which contains a complete list of all publications under the IAP-statistics network.

1.1.2 Overview

The overall achievements for the research project can be summarized as follows:

* Among the most important achievements in frontier estimation are the development of appropriate bootstrap procedures as well as robust procedures. Further, consistent nonparametric estimators of boundary and support in the deconvolution problem have been proposed. The developed fully data-driven procedures for change-point detection have been studied in detail (including estimation and testing issues). Moreover, an appealing direct estimation method of non-smooth curves and surfaces has been worked out.

* In forecasting nonstationary processes, a major step has been made with methods based on locally nonstationary wavelet processes. In the dynamic factor methodology, results have been obtained on consistency rates and forecasting, filling an important gap in the literature. The L1 approach to density estimation, and the proposed local linear methodology for nonparametric regression estimation on random fields are a significant step towards a nonparametric approach in the analysis of spatial processes. The results on the close relation between invariance (typically,

rank-based methods) and semiparametric optimality are providing a very fundamental insight on the nature of semiparametrically optimal inference, and pave the way for the construction of rank-based methods in a number of models where they have never been considered so far. Finally, the papers on multivariate ranks and signs are the first ones in the literature extending optimal rank-based methods to a multivariate setting, both for serial and nonserial semiparametric problems

* For estimation with censored data, new results have been obtained related to survival function and hazard function estimation. Also some important steps have been undertaken in the situation where some kind of dependence is allowed between survival times and censoring times. In the domain of frailty modelling, some important results have been obtained for likelihood ratio testing for the presence of heterogeneity in the data.

* A very important achievement in the modelling of random effects is the development of random effects models in accelerated failure time (AFT) survival models for interval censored data. Allowing a flexible distribution for the error component in combination of a parametric frailty model is a first step towards a general AFT model. As a second major achievement we consider the connection that has been established between the (generalised) linear mixed models which are very popular in biostatistics and the Item Response Models (IRT models). Finally, the classical IRT models have been further developed to incorporate in a general way random effects.

* The major achievements concerning mixture models are extensions of psychometric models (PMD, IRT) with mixture components and transitions between these. Also smoothing approaches are developed to deal with mixtures of random effects in various contexts. Concerning classification methods other than mixture models, we have largely extended the hierarchical classes models, both in the direction of the extant two-way and three-way PCA family and of clustering methods. Finally, a convexity-based method for simultaneous clustering of rows and columns has been developed and a review article has been realized on the general topic of two-way clustering.

* Work has been done towards the assessment of sensitivity to models for incomplete longitudinal and hierarchical data in biometry and social sciences. Specific emphasis has been put onto the use of such methods in the context of clinical trials. Apart from mixed-model applications in meta-analysis, latent structures have been investigated in the setting of multivariate, longitudinal and hierarchical ordinal outcomes. The issue of identifiability has been investigated, especially arising due to incomplete observability of the latent structure, e.g., due to discretization.

1.2 Work package 1: Functional estimation

1.2.1 Nonparametric estimation of a frontier function and deconvolution problems

Frontier estimation

This year several papers were published in the field of frontier estimation: they cover

different approaches and problems involved in this framework.

Most of the work has been devoted to the nonparametric envelopment estimators (DEA and FDH). Simar and Wilson (2003a) show that the double bootstrap is particularly useful for doing inference with DEA estimators. Simar (2003a) proposes a very simple algorithm, based on order-m frontiers, to detect outliers. Beguin and Simar (2004) illustrates the approach in an analysis of the expenses linked to hospital stays. Explaining efficiencies in productivity analysis is quite important. Simar and Wilson (2003b) explain the issues raised by two-stage semiparametric models often used wrongly in the literature; the use of an appropriate bootstrap allows to correct for these errors. Daraio and Simar (2004) propose, in a probabilistic formulation of the problem, a one-stage nonparametric estimator which allows to introduce environmental (or explanatory) variables in the production process. Simar and Zelenyuk (2003) and Steinmann and Simar (2003) investigate the problem of comparing the efficiency of groups of firms.

Kneip, Simar and Wilson (2003) investigate the asymptotic properties of the DEA estimators in a full multivariate context. They provide the limiting distribution of the efficiency scores and propose various versions for bootstrapping in this context (smoothed bootstrap and subsampling), proving the consistency of these bootstrap algorithms. Jeong and Park (2003) also investigate the limit distribution of convex hull estimators for boundaries in the multidimensional case. Bootstrap might be time consuming in terms of the computations. Badin and Simar (2003) propose a very simple way to construct confidence intervals for efficiency scores avoiding to perform the bootstrap, in the homoscedastic case: it is based on the order statistics of the estimated efficiency scores. Daouia and Simar (2003) propose an alternative to the order-m frontiers to get robust estimators of the frontier. It is based on a nonstandard order-α conditional quantile function, adapted to the case of a monotone boundary.

Deterministic frontier models and their nonparametric estimators (DEA and FDH) do not allow for errors or noise in the data. Simar (2003b) suggests a way for improving the performances of these estimators in the presence of noise.

Parametric approaches are also very popular in the field of frontier estimation. Florens and Simar (2004) present a new method based on a parametric approximation of a nonparametric model, which seems to provide a robust parametric estimator, robust to the stochastic hypothesis on the production process but also robust to extreme or outlying observations.

The semiparametric modelling of stochastic frontiers in the presence of a panel of data has been pursued: Park, Sickles and Simar (2003a) adapt their previous works for the case of AR(1) error term and Park, Sickles and Simar (2003b) analyzes dynamic models.

Measuring the imperfection of a market is a challenging issue in economics. Mouchart and Vandresse (2003) propose an empirical method to measure the market imperfection and the bargaining power of the agents by extending the methods of frontier analysis. They propose a nonparametric estimation of the support of a multivariate distribution based on a compact convex closure of a finite set of points, under a restriction of so-called "free disposability". A case study in the field of freight transport illustrates the proposed method.

Deconvolution problems

Delaigle and Gijbels (2003, 2004a,b) deal with the problem of how to estimate nonpara-

metrically the density of a random variable when the measurements on this variable contain errors. The kernel density deconvolution estimator is a consistent estimator in this context. The problem of selecting an appropriate bandwidth is as usual very important in kernel estimation, but is an even more difficult problem in the deconvolution context. In her PhD-thesis (under supervision of Prof. Gijbels and defended in January 2003) Dr Delaigle focused on two problems: (i) how to select an appropriate practical bandwidth in this context? (ii) how to estimate the endpoints of the support of a density. Practical bandwidth selection procedures have been proposed in Delaigle and Gijbels (2004a,b). A finite sample comparison between the discussed procedures, a bootstrap procedure, a plug-in procedure and a cross-validation procedure, has been carried out. Theoretical properties (including consistency) of the bootstrap and plug-in procedures have been established. Delaigle and Gijbels (2003) propose a consistent estimator for the unknown support of a density in case of measurement errors. They establish the bias and variance properties of the estimator, as well as its rate of convergence.

Related topics on semi- and non-parametric inference can be found also under **WP2**. See Sections 1.3.1 and 1.3.4.

1.2.2 Automatic detection of change-points in regression and landmark detection

A first problem studied is the detection of abrupt changes in a univariate regression function. A survey of nonparametric kernel-based methods for this estimation problem can be found in Gijbels (2003). Gijbels and Goderniaux (2004a,b) propose a fully data-driven procedure for detecting jump points in a regression curve or its derivative. They rely on the two-steps estimation method as introduced by Gijbels, Hall and Kneip (1999) and discuss a bootstrap procedure for choosing the smoothing parameters involved. Gijbels and Goderniaux (2004a) also discuss how to estimate the number of jump points in a regression curve, using cross-validation. They also provide a fully data-driven algorithm that should be used when dealing with an identification problem, that typically occurs when a jump point is difficult to distinguish from points with a high derivative (in absolute value). The basic methodology can be adopted to the case of jump points in a derivative of the regression curve. This requires some adjustments such as the choice of an appropriate diagnostic function and a cross-validation criterion for derivative estimation.

Of interest is also the testing problem. A brief discussion on available kernel-based methods for testing for a continuous versus a discontinuous regression function is given in Gijbels (2003). Gijbels and Goderniaux (2004c) dealt with testing the null hypothesis that a regression is continuous versus the alternative hypothesis that it is a discontinuous function, relying on the two-steps data-driven procedure. In addition they introduce a bootstrap procedure for assessing the distribution of the test statistic under the null hypothesis. This bootstrap procedure has also been used in Gijbels, Hall and Kneip (2004) for constructing confidence bands for discontinuous regression curves. The testing procedure proposed by Gijbels and Goderniaux (2004c) has been compared with other testing procedures available in the literature. The latter procedures rely on the use of the asymptotic distribution of the involved test statistics, and do not treat in a satisfactory way the choice of the smoothing parameters involved.

Jeremie Bigot has defended his PhD thesis in 2003 and is currently a post-doctoral researcher of the IAP-statistics network. In Bigot (2002) he developed a scale-space approach with wavelets for landmark detection. Bigot (2003a) proposes a nonparametric approach to estimate the landmarks of a signal observed with noise which yields a new technique to automatically align two sets of landmarks.
The paper by Bigot (2003b) is concerned with the problem of the alignment of multiple sets of curves and their comparison with FANOVA techniques. A fixed-effects FANOVA model combined with a registration step is used to test the significance of main/interaction effects. Some real examples arising from the biomedical area are used to illustrate the methodology.

1.2.3 Modelling of heterogeneous regularities

Delouille, Simoens and von Sachs (2004) have presented a new methodology to construct smooth wavelets which adapt automatically to the stochastic design of a nonparametric curve estimation problem, and circumvent the usual restrictions of classical wavelets (dyadic sample size, boundary treatment, non-equispaced design).
Delouille and von Sachs (2003b) furnish some theoretical results on the optimality of this new estimator which parallel the classical wavelet thresholding for equispaced data. Moreover they develop these in the more difficult time series context of a non-linear autoregressive design and show how their original algorithm can be adapted to various time series situations (including ARCH-type models). This research is linked with research under **WP2**.
Delouille and von Sachs (2003a) and Delouille, Jansen and von Sachs (2003) deal with the two-dimensional situation. Whereas Delouille and von Sachs (2003b) relies on a tensor-product construction (with a sufficiently regular design but possibly different degrees of smoothness in the two directions), Delouille, Jansen and von Sachs (2003) present an approach to treat fully irregularly spaced data in two dimensions. This new method to construct smooth bi-variate estimators is based on lifting on triangulation schemes using Lagrange interpolating polynomials and a Bayesian approach for wavelet thresholding.

1.2.4 Functional estimation for microarray data

The efforts of the Grenoble team have continued with respect to developing nonparametric methods for analyzing and classifying microarray data. Julie Peyre has pursued her work on statistical analysis of microarray data and she is at the final stage of her PhD thesis which will be defended at the end of 2004. In a recent joint work with A. Antoniadis and a team of biologists of the Institute Marie Curie in Paris (see Antoniadis, Peyre et al. (2003)), she has studied the determination of the biological effects of low doses of pollutants by means of a nonparametric ANOVA like DNA microarray analysis designed for the investigation small intracellular changes induced by irradiation at varying low doses.

With respect to developing adapted classification methods for microarray data, following the work of Antoniadis, Lambert-Lacroix and Leblanc (2003), Fort and Lambert-Lacroix (2003) propose a new method combining Partial Least Squares and Ridge penalized logistic regression. After reviewing the existing methods based on PLS and/or

penalized likelihood techniques, they outline their interest in some cases, and explain theoretically their poor behavior. Their procedure is compared with these other classifiers and the predictive performance of the resulting classification rule is illustrated on two well known data sets: the Leukemia data set and the Colon data set. Software that implements the procedures on which their paper focus are freely available at http: //www-lmc.imag.fr/SMS/software/microarrays/. The paper by Antoniadis, Lambert-Lacroix and Leblanc (2003) on "Effective Dimension Reduction Methods for Tumor Classification using Gene Expression Data" has been selected to be reproduced in the 2004 "Year Book of Medical Informatics".

1.2.5 General methodological issues

Nonparametric regression and density estimation.
With respect to developing a general methodology for nonparametric regression via regularization, the joint work by A. Antoniadis (UJF) and I. Gijbels (UCL) and with an external collaborator M. Nikolova (ENST Paris), devoted to penalized likelihood regression for generalized linear models with nonquadratic penalties is at his end.

Einmahl and Van Keilegom (2003) consider the nonparametric regression model $Y = m(X) + \varepsilon$, where the function m is smooth, but unknown, and ε is independent of X. The authors construct omnibus goodness-of-fit tests for the independence of ε and X.

Hall and Van Keilegom (2003) suppose a nonparametric regression model with autoregressive errors, and propose a new method to estimate the autoregressive parameters and the error covariances. A new bandwidth selection method is also proposed. This work relates to **WP2**.

Claeskens and Van Keilegom (2003) study confidence bands for a regression function and its derivatives, using local polynomial estimation. Two types of confidence bands are obtained: those based on asymptotic normality and those based on a smoothed bootstrap approach.

In his doctoral research Taoufik Bouezmarni focuses on estimation of a density function that has a bounded support and/or is unbounded. Bouezmarni and Rolin (2003a) considered the estimation of densities by a Beta kernel. They found the exact asymptotic behavior of the mean integrated absolute error and its upper bound. The L_1-rate of convergence of this estimator for the uniform density on $[0, 1]$ was also established. Uniform weak convergence is shown for the class of continuous densities. Bouezmarni and Scaillet (2003) considered asymmetric kernel estimators and smoothed histograms for the class of densities defined on the positive real line. Weak uniform convergence on each compact for these estimators as well as weak convergence in L_1 were established. They proved the weak convergence to the infinity if the density is unbounded at zero. Bouezmarni and Rolin (2003b) showed the weak and strong consistency of Bernstein estimators of a density function defined on $[0, 1]$, if the underlying density presents an infinite pole at $x = 0$ and/or $x = 1$. Bouezmarni, Mesfioui and Rolin (2003) have established the asymptotic expression and upper and lower bounds of the mean integrated absolute error of the asymmetric kernel estimators and the smoothed histograms.

Nonparametric estimation of a density and a regression in random fields has been a subject of study under **WP2**.

Inference for curves under shape restrictions.
Nonparametric estimation of a curve when assuming some shape constraints on the unknown curve is an interesting problem. If the shape constraint is justified then a constraint nonparametric estimator will perform better than an unconstrained estimator.

An overview of nonparametric methods for estimating a monotone regression function has been provided in Gijbels (2004). Gijbels and Heckman (2004) focus on the hazard function as it is a very important characteristic in survival analysis. They developed a nonparametric testing procedure for testing whether a hazard function is increasing (decreasing) or not. The method is based on normalized spacings and does not involve any smoothing parameter. The inspiration for the test was based on a test used by Proschan and Pyke (1967) for testing for a constant hazard function, and on considering local versions of this test relying on recent developments in nonparametric testing.

Zhang and Gijbels (2003) rely on sieve empirical likelihood methods to deal with estimation in constrained parametric or nonparametric regression models with unspecified error distributions. Asymptotic properties of the sieve empirical likelihood estimators are established, and efficiency properties are discussed. The estimator is shown to be adaptive for inhomogeneity of the conditional distribution of the error with respect to the predictor, especially for heteroscedasticity.

Dimension reduction methods.
A. Antoniadis (Grenoble) has also pursued his work on dimension reduction in functional regression. In collaboration with U. Amato and I. De Feiss (external collaborators from CNR in Naples) he developed some new two-dimensional reduction regression methods to predict a scalar response from a discretized sample path of a continuous time covariate process. The methods take into account the functional nature of the predictor and are both based on appropriate wavelet decompositions. Using such decompositions, they derive prediction methods that are similar to minimum average variance estimation (MAVE) or functional sliced inverse regression (FSIR). The method is succesfully applied for analysis of some real calibration data from near infrared spectroscopy. See Antoniadis, Amato and De Feiss (2003).

G. Geenens in his doctoral research (advisor: L. Simar) analyzes, in collaboration with M. Delecroix (ENSAI, Rennes) the possibility of modelling the probabilities in a multinomial process by single index models (SIM). The idea is to let these probabilities be related to some explanatory variables in a semiparametric way. All the techniques of estimation for SIM have been investigated in this particular setup. Then these methods will be used for doing inference in contingency tables (test of independence, of partial independence,...). A paper is in preparation.

The work on classification trees has been pursued. In De Macq and Simar (2004) an exact algorithm based on hyper-rectangular partitioning trees has been implemented. It has optimal properties but the computing complexity limits its applicability. An approx-

imated (faster) algorithm, based on the quantiles for building the splitting rules has been implemented. Its performances are compared with more classical classification algorithms through some classical data sets and it appears that, for these data sets, this new classifier performs much better. I. De Macq will defend her PhD thesis in May 2004.

Functional estimation and samples of curves.
In collaboration with A. Antoniadis and R. Von Sachs, J. Bigot is actually working on a dense matching approach with wavelets to register sample curves for functional analysis of variance purposes. Indeed, in an ANOVA setup, to compare similar objects, it is generally necessary to find a common referential to represent them. The curve registration problem consists in finding appropriate transformations to synchronize a set of signals. Dense matching methods for curve alignment are based on the definition of an appropriate functional to represent the quality of the registration. In one dimension, dynamic time warping (DTW) is an effective technique to minimize such functionals. The alignment of smooth curves by DTW has been studied from a statistical point of view with kernel estimators. However, there is not much work on developing dense matching approaches to register one-dimensional signals that are not uniformly regular (e.g. with isolated singularities). Wavelet decompositions and the previous work in Bigot (2003b) allow to design new functionals for the alignment of multiple sets of curves, and to derive nonparametric approaches for efficient functional analysis of variance methods.

Model selection issues.
Hjort and Claeskens (2003a,b) and Claeskens and Hjort (2003) deal with model selection and model averaging. Hjort and Claeskens (2003a,b) developed methodology and theory to deal with estimators averaged across different models. As a special case this includes estimators found via a model selection procedure. The traditional use of model selection methods in practice is to proceed as if the final selected model had been chosen a priori. This often leads to underreporting variability and wrong confidence intervals. The results in the paper give a correct way of dealing with these issues. In Claeskens and Hjort (2003) it is explained that the model selector should focus on the parameter singled out for interest; in particular, a model which gives good precision for one estimand may be worse when used for inference for another estimand. This yields a new, focussed information criterion, the FIC. While the papers above concentrated on estimators after model selection, the work by Claeskens and Hjort (2004) focuses on testing hypotheses where the test statistic involves a model selection procedure. In particular, they treat the case of testing whether data come from a certain parametric distribution.

1.2.6 Other topics in functional estimation

Chen, Linton and Van Keilegom (2003) consider M-estimators in non-standard contexts. In particular, they suppose that the criterion function is not necessarily continuous and that it depends on a preliminary nonparametric estimator. They obtain general conditions that guarantee the consistency and asymptotic normality of this type of estimators.

A nonparametric approach based on a generalization of the Wilcoxon test is used in the context of quantitative trait loci methods in Tilquin, Van Keilegom et al. (2003).

Road safety research is traditionally based on the detection of "black spots" supposed to identify accidents-prone loci in the road network. Flahaut, Mouchart, San Martin and Thomas (2003) extends this approach toward the detection of dangerous zones. They propose and compare two different approaches: one by a local decomposition of an autocorrelation coefficient and another one by a nonparametric kernel-type estimation of the intensity of a non-homogenous Poisson process. As such, they obtain a dangerousity index distributed continuously on a given road.

1.3 Work package 2: Time series

1.3.1 Modelling and estimation for high-dimensional non-stationary time series

Fryźlewicz, Van Bellegem and von Sachs (2003) present, both from theoretical and practical point of view, an approach of nonparametrically forecasting a second-order non-stationary time series with the help of the model of "Locally stationary wavelet processes". Applications to prediction of financial data and a variety of comparisons with classical methods are discussed in Van Bellegem and von Sachs (2004). Van Bellegem and von Sachs (2003a) propose a new estimator of the time-varying spectrum of a locally stationary wavelet process (cf. Fryźlewicz et al. (2003)). The estimator is based on the approach of local adaptivity and on a non-asymptotic result controlling the large deviations of a periodogram-like statistic. This new theoretical approach includes possibly very irregular spectra over time (of bounded total variation), thus allowing for a finite number of abrupt changes of the second-order structure over time. Van Bellegem and von Sachs (2003b) present a concrete algorithm and various applications of this estimator. These include a new test on covariance stationarity and a test on local significance of the coefficients of the possibly sparse wavelet representation of the underlying model of locally stationary wavelet processes.

Ombao, von Sachs and Guo (2003) is to be placed into the series of papers the authors have written on modelling and estimation of non-stationary statistical signals (such as EEGs with local transients due to epilepsy) by means of the SLEX ("Smooth Localised complex EXponentials") methodology (cf. previous publications). Here, the paradigm is to find the best-adapted time segmentation (in a truly multivariate sense) of a high-dimensional, long time series by a fast algorithm, which is realised using a new approach of time-varying Principal Components Analysis in the frequency (SLEX) domain.

Donoho, Mallat, von Sachs and Samuelides (2003) present an alternative, though univariate, to finding the best-adapted segmentation of a possibly covariance non-stationary time series by means of local cosine functions. Optimal rates of convergence are derived for spectra which vary with a Sobolev regularity over time by using the concept of "macro-tiles".

Guo, Dai, Ombao and von Sachs (2003), finally, apply the smoothing spline methodology to estimation of a time-varying spectrum of a locally stationary time series. For the price of needing to impose more regularity in order to show the theoretical results, the

advantage of this method compared to the previous ones is that no explicit segmentation in time is needed for constructing the estimator.

The methods described in Guo, Dai, Ombao and von Sachs (2003) and Van Bellegem and von Sachs (2003a,b) are strongly linked with methodologies developed in **WP1**.

Prediction from panel data with a high proportion of missing data raises substantial statistical issues. Mouchart and Rombouts (2003) propose an efficient approach for the case of nowcasting, *i.e.* forecasting present values based on recent past data. A progressive specification strategy is elaborated and illustrated on R&D data for a panel of countries from the European Community.

1.3.2 Analysis of high-dimensional time series data

The papers on dynamic factor models that were submitted last year have made their way to publication, and either appeared or were accepted: see Forni, Hallin, Lippi and Reichlin (2003a, 2004) and Cristadoro, Forni, Reichlin and Veronese (2004).

Several papers were submitted: Forni, Hallin, Lippi and Reichlin (2003b) develops the forecasting aspects of the problem; Forni, Lippi and Reichlin (2003) studies structural identification in these models, while Giannone, Reichlin and Sala (2003a,b) develop empirical applications to monetary economics. See also Reichlin (2003). All these papers also have appeared as working papers at the Center for Economic Policy Research (CEPR), London.

Still in connection with dynamic factor models and their applications, two doctoral theses were successfully defended within the ULB group: L. Sala (Essays in Monetary and Fiscal Policy) and D. Giannone (Essays on Dynamic Factor Models of Large Cross-Sections). L. Sala's thesis is concerned with applications of the model to monetary economics while D. Giannone's develops new econometric results. In particular, an important contribution of the thesis is the asymptotic analysis of dynamic principal components.

New projects are ongoing. L. Reichlin, C. Doz and D. Giannone are working on maximum likelihood estimation of the parameters of a dynamic factor model in large cross-sections and studying misspecification and efficiency issues. See Doz, Giannone and Reichlin (2003).

D. Giannone and L. Reichlin are also working with the Federal Reserve Board in Washington, the European Central Bank in Frankfurt, and the Swiss National Bank in Switzerland to apply the techniques developed by the ULB-team to problems of signal extraction and forecasting. They have a paper in preparation which summarizes this research for *Macroeconomic Annual*, the yearly volume published by the National Bureau of Economic Research which includes the important contributions in macroeconomics during the year.

1.3.3 Random and time-dependent coefficient time-series models

Azrak and Mélard (2003) deal with estimating the parameters of ARMA models with time-dependent coefficients. A generalization to multidimensional processes has been started. Akharif and Hallin (2003) study the detection of random coefficients in autoregressive models.

1.3.4 Multivariate time series

Spatial data, image analysis and inverse problems.
Hallin, Lu and Tran (2003) discuss an L_1 approach to estimation for random fields. Hallin, Lu and Tran (2004) develop local linear methods for regression on random fields.

Christine De Mol has pursued her work on image restoration in Astronomy, which resulted in a joint publication with an Italian team (see Bertero, Boccacci, Custo, De Mol and Robberto (2003)), as well as her collaboration with I. Daubechies and M. Defrise on linear inverse problems with a sparsity constraint. In Daubechies, Defrise and De Mol (2003) it is proved that a weighted L_1-norm penalty on the coefficients of the expansion of the solution on an arbitrary pre-assigned orthonormal basis (such as e.g. a wavelet basis) can regularize a linear ill-posed problem. It is also proved that an iterative algorithm that amounts to Landweber iteration with soft-thresholding applied at each iteration step converges in norm to a regularized solution. These results extend to the case of L_p-penalties with p larger than 1 and apply straightforwardly to nonparametric regression with such sparsity-enforcing penalties ("lasso" or "bridge" regression), as well as to penalized maximum likelihood with a Gaussian noise model and a (generalized) Laplacian prior.

Semiparametric inference based on multivariate signs and ranks.
M. Hallin and B. Werker (University of Tilburg) have established the role of invariants such as signs and ranks in producing semiparametrically efficient inference procedures. See Hallin and Werker (2003). These ideas are applied further in Hallin, Vermandele and Werker (2003a,b) two papers on sign-and rank-based methods for median-restricted models (including a number of time-series models).

M. Hallin and D. Paindaveine developed rank-based methods for a variety of problems involving multivariate elliptical densities, with three papers on elliptical VARMA models. See Hallin and Paindaveine (2004a,b,c). See also Hallin and Paindaveine (2003). A new concept of multivariate ranks based on hyperplane counts is presented in Oja and Paindaveine (2004).

In a similar spirit, Hallin and Paindaveine (2003b) developed rank-based inference procedures for shape, with a distribution-free test of the hypothesis of sphericity that does not require any moment assumptions.

See Paindaveine (2003a,b) and Paindaveine (2004) for work related to the efficiency of the class of methods he developed in his thesis. D. Paindaveine in 2003 received the "Prix du Jeune Statisticien", a prize awarded every third year by the Société française de Statistique to the best dissertation in Statistics defended in a French-speaking university, and was invited to contribute a review paper on his work to the Journal de la Société française de Statistique.

1.3.5 Other topics in time series

A. Saidi has successfully defended his thesis "Tests de non-corrélation entre deux séries chronologiques univariées ou multivariées". See also Hallin and Saidi (2004).

A. Ouakasse comes close to the end of his thesis on a new method for recursive estimation of ARMA and SISO models. During the year he has worked on the SISO part and presented his results.

S. Lotfi is also close to the end of her dissertation, devoted to periodic time series models. See also Hallin and Lotfi (2004).

H. Njimi has presented a first chapter of his thesis on an improvement of the automatic ARMA modelling procedure of Mélard and Pasteels. He has started working on a comparison with Gòmez and Maravall's TRAMO/SEATS method. Azrak, Mélard and Njimi (2003) have promoted that new strategy in a problematic data set with many missing values and outliers.

With A. Klein, T. Zahaf and P. Spreij, G. Mélard has worked on several projects based on the information matrix of time series models. See Klein and Mélard (2004). With A. Saidi and R. Roy, G. Mélard has worked on the computational estimation of the parameters of VARMA models. The paper will be submitted soon. The research will be prolonged with J. Niemczyk in order to use outlier detection in a multivariate context, with an improvement of seasonal adjustment in mind.

Dufour, Fahrat and Hallin (2003) established exact bounds for the tail areas of distributions of autocorrelation coefficients under unspecified innovation density.

Finally, Cohen, Mélard and Ouakasse (2003a,b) delivered communications on the scientific aspects of the course on time series analysis they developed for the National Bank of Belgium.

1.4 Work package 3: Survival Analysis

1.4.1 Nonparametric estimation with censored data

Braekers and Veraverbeke (2003) studied nonparametric estimation of the conditional survival function under dependent censoring. The common (but untestable) assumption that lifetimes and censoring times are independent is indeed not always justifiable. Therefore, a model is considered in which the joint distribution of lifetime and censoring time if described by a known copula function. For the class of Archimedean copulas, an explicit form of the estimator can be obtained. Asymptotic properties have been established.

Members of the LUC team (Paul Janssen and Noël Veraverbeke) in collaboration with Ricardo Cao and Ignacio Lopez de Ullibarri (La Coruna, Spain) further explored presmoothing in survival analysis. See Cao, Lopez de Ullibarri, Janssen and Veraverbeke (2003). Cao, Janssen and Veraverbeke (2003) also obtained new results on relative hazard function estimation with left truncated and right censored data. Paul Janssen and Noël Veraverbeke in collaboration with Jan Swanepoel (Universiteit Potchefstroom, South Africa) studied goodness-of-fit tests derived from a characterization of the uniform distribution. They derived normal and bootstrap approximations in Janssen, Swanepoel and Veraverbeke (2003).

Akritas and Van Keilegom (2003) study nonparametric estimation of the bivariate (and marginal) distribution of two random variables that are subject to censoring. Asymptotic properties of the proposed estimators are established, a bandwidth selection method is proposed and simulations are carried out.

The work by Du, Akritas and Van Keilegom (2003) is situated in the domain of analysis of covariance with censored data. The authors develop a nonparametric method, which is useful in situations where the classical models (like additive risk model, proportional hazards model or proportional odds model) are not adequate. A test statistic for this

model is proposed and its performance is studied for small and large sample sizes.

Van Keilegom (2004) proposes a new estimator of the bivariate and marginal distribution of two random variables subject to censoring. The estimators do not require the common assumption of independence between the vector of survival and censoring times, but allow for a certain type of dependent censoring.

Heuchenne and Van Keilegom (2003) consider a polynomial regression model, in which the response is subject to random right censoring. A new estimation procedure for the parameters in this model is proposed, and the estimators are studied both via asymptotic properties and via finite sample simulations.

In many applications it is reasonable to assume that the hazard function is an increasing (or decreasing) function. In this case nonparametric estimation of the hazard function should be done under this constraint. Gijbels and Heckman (2004) deal with the problem of testing whether the hazard function can indeed be assumed to be increasing (or decreasing). They also illustrate how to apply their methodology to type II censored data. This work is related to work in **WP1**.

Gijbels and Gürler (2003) consider estimation of the hazard function, based on censored data, when the hazard function is assumed to be a step function with a (unknown) jump point. In real life applications, abrupt changes in the hazard function are observed due to overhauls, major operations or specific maintenance activities. The proposed estimation procedure is based on certain structural properties and on least squares ideas. A simulation study is carried out to compare the performance of the proposed estimator with two estimators available in the literature: an estimator based on a functional of the Nelson-Aalen estimator and a maximum likelihood estimator. This work is related to research on change-point detection in **WP1**.

1.4.2 Frailty modelling in survival analysis

In 2003 the work with the "frailty working group" focused on the following issues.

We studied applications of frailty models in treatment outcome studies (in collaboration with the European Organisation for Research and Treatment of Cancer (EORTC), Brussels) and in animal breeding experiments (in collaboration with the International Livestock Research Institute, Nairobi, Kenya and with the Institut National de la Recherche Agronomique (INRA), Jouy-en-Josas, France). Also the advantages of the joint modelling of survival times and covariate processes have been studied for animal breeding data. See Nguti, Burzykowski, Rowlands, Renard and Janssen (2003). Also the usefulness of frailty models to model the recurrent asthma event rate over time has been shown in Duchateau, Janssen, Kezic and Fortpeid (2003).

We further showed that the use of splines is very useful in obtaining more flexible frailty models. Examples from veterinary sciences have been used to demonstrate this idea in Duchateau and Janssen (2003). Part of this work is in collaboration with Luc Duchateau (UGent).

Further activities of the frailty working group focused on the study of resampling plans for frailty models needed to estimate the standard error of the estimated heterogeneity parameter (see Massonet, Burzykowski and Janssen (2003)) and on a further study of likelihood ratio and score tests to test for the presence of heterogeneity in the data (to test for the presence of a cluster effect). See Nguti, Claeskens and Janssen (2003). This

work is in collaboration with Gerda Claeskens (UCL).

During 2003 we started investigating the following topics.

We started the study of the asymptotic behaviour of likelihood ratio tests for testing heterogeneity in non-proportional hazard models with random effect.

We show that to get a good feel of the size of the heterogeneity parameter it is useful to study how the heterogeneity present in the data influences important quantities of interest such as the prevalence of a disease or the median survival time (which can be seen as transformations of the random effect).

We started studying frailty models with more than one random effect for use in treatment outcome trials and for prognostic index analysis. To fit models with more than one random effect the existing survkit software (written by Ducrcocq et al.) will be extended to these more complicated models. Following Ducrocq and Casella (1996) the approach followed is a Bayesian one. This work is in collaboration with INRA and EORTC.

1.5 Work package 4: Mixed Models

The research topics treated in this Work Package are subdivided into four related themes showing also a high relationship with other Work Packages.

1.5.1 The implementation of multivariate random effects

The topics below were treated.

Flexible distributions for the random effects part of a linear mixed model.
Members of the KUL-2 team (Ghidey and Lesaffre) have developed, in collaboration with Paul Eilers (University of Leiden) a mixed model with a smooth random effects distribution. The model assumes a flexible random effects distribution that can be well approximated by a smooth function of B-splines or of Gaussian densities. Penalized likelihood maximization delivers the estimated fixed effects and the smoothing coefficients of the random effects distribution. See Ghidey, Lesaffre and Eilers (2004). In the next step the normality assumption of the error distribution is replaced by a smooth distribution of Gaussian distributions in conjunction with a smooth random effects distribution. This work is in development. Up to now, the results show that a mixture of Gaussian distributions on the error distributions combines nicely with a Gaussian random effects distribution, but that, due to a large number of (simple) integrals to be evaluated a different numerical technique is needed.

Joint modelling of multivariate longitudinal profiles.
Random-effects models as a joint modelling approach have been critically investigated using a dataset on hearing thresholds measurements. See Fieuws and Verbeke (2004). Ongoing research focuses on the modelling of multivariate longitudinal profiles in situations where the high number of outcomes makes the random-effects approach computationally extremely demanding or impossible. Borrowing ideas from the multiple imputation framework, the proposed alternative strategy combines results of all possible pairwise (pair of outcomes) joint random-effects models.

1.5.2 The investigation of mixture models as an alternative for approaching the random effects distribution

The following topics were treated.

Allowing for examiner's bias and variability in a logistic random effects model. Mwalili, Lesaffre and Declerck (2003a) developed a method to correct for the examiner's bias and variability when a gold standard and calibration data are available for an ordinal logistic regression model. The method was applied to the geographical distribution of the dmft-score (caries) of seven-year old children in Flanders. Lesaffre, Mwalili and Declerck (2003) have submitted a paper to the Journal of Dental Research explaining the methodology in the previous paper to a dental audience. Further, the same methodology has been implemented by Mwalili and Lesaffre to a Zero-inflated Poisson distribution for the dmft-score. A manuscript is in preparation and will be submitted in the near future. Furthermore, in collaboration with Helmut Kuechenhoff (University of Munich), Mwalili and Lesaffre have developed the SIMEX (Simulation-Extrapolation) method for misclassification.

Reformulation of item-response models as generalized linear mixed models and nonlinear mixed models.
The work has been concentrated on four issues.

Statistical framework for item response models. In comparison with the report of last year, we can mention that manuscript of the planned volume has been submitted and will be out this summer in the Springer series on Statistics for Social Science and Public Policy, entitled "Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach". It is an edited book (De Boeck and Wilson (2004)), but written as a monograph, in a collaboration between researchers involved in **WP4** (Geert Verbeke, Steffen Fieuws), **WP5** (Paul De Boeck, and many others), and **WP6** (Geert Molenberghs) on the one hand, and the educational measurement research group from the UC Berkeley (Mark Wilson). An article on the same topic is published last year. See Rijmen, Tuerlinckx, De Boeck and Kuppens (2003).

Residual dependence. An important issue in the mixed models we use is residual dependence. This is dependence beyond the random effects that are included in the model. We have successfully used conditional modelling to deal with this problem in Smits, De Boeck and Hoskens (2003), and we have compared this conditional approach with a partly marginal approach. See Ip, Wang, De Boeck and Meulders (2004).

MIRID. We have further investigated models for covariate effects that are a function of other covariate effects (so-called MIRIDs) in Smits and De Boeck (2003). Software has been developed for a conditional likelihood estimation of these models, and this estimation method is compared with the more common estimation based on a normal distribution for the random effects. See Smits, De Boeck and Verhelst (2003).

Counting process. Tuerlinckx (2004) has proposed a multivariate counting process with positive dependencies for reaction times that can be approached as a random-effects model

with independent non-homogeneous Poisson processes (conditional on the random effect).

Crossed-random effect models for educational measurement.
Van den Noortgate, De Boeck, Janssen and Meulders (2002) have extended an item-response model for simultaneous random effects of persons and items. This topic is also included in the bilateral project with Geert Molenberghs from the LUC (**WP6**) and the Pontifical University of Chili (PUC, Santiago). The project is a bilateral (international) scientific and technological cooperation between Paul De Boeck (**WP5**), Geert Molenberghs (**WP6**) and the Departement of Statistics of the PUC (Pilar Iglesias, Guido Del Pino, Ernesto San Martin). Part of the collaboration was an international workshop held in Leuven in July 2003.

San Martin and del Pino (PUC, Santiago) and Mwalili and Lesaffre have started to examine the IRT model with random guessing parameters for $L > 1$. Simulations have been performed, but the estimates of the random effects parameters are biasedly estimated. Further exploration is needed to find the cause for the biased estimation, this investigation is currently done.

Relation between diffusion models and latent trait models/random-effect models.
Tuerlinckx and De Boeck (2002) have shown that a well-known latent trait model (the 2PLM) can be derived as the marginal choice probability model from a diffusion model (a Wiener process with constant drift and variance and two absorbing boundaries). On the other hand, we have formulated a random-effect (or latent trait) version of the bivariate diffusion model, implemented it in a computer program and estimated the model for data on personality self-ratings. The paper on these results is in preparation, but partial results can be found in Ratcliff and Tuerlinckx (2002). The work on diffusion models is in collaboration with Roger Ratcliff from Northwestern University.

1.5.3 Extensions to interval-censored data

The topics that were treated here are much related to those of **WP3**, we distinguish the following research topics.

Modelling the emergence times using random effects models for interval, left and right censored data.
The Signal Tandmobiel Study is a longitudinal dental study on about 4500 children. The KUL-2 team (Komárek and Lesaffre) fitted in collaboration with Tommi Härkänen (Helsinki) several survival models with two random effects parameters (frailty parameter and birth of dentition parameter) for emergence the caries experience of the first permanent molars. See Komárek, Lesaffre, Härkänen, Declerck and Virtanen (2003). The above analyses suggested examining a different approach, i.e. AFT-models (accelerated failure time models) with a complex error structure not necessarily assuming classical assumptions like normality. The KUL-2 team, in collaboration with Hilton (UCSF), has developed an AFT model with a smooth error distribution being the mixture of Gaussian distributions. See Komárek, Lesaffre and Hilton (2003). Further, Komárek and Lesaffre have developed a Bayesian model for AFT-models with a flexible error distribution and

with a parametric random effects component; a manuscript is in preparation. In parallel, the similar AFT-model is planned to be developed in a frequentist context in a collaboration with Lambert (**WP3**).

Modelling jointly repeated measurements and survival times.
In the context of a clinical trial or an epidemiological study, there are often repeated measures on a risk factor available which have an impact on the survival response (e.g. serum cholesterol on survival). It is advantageous for the prediction of survival to model the repeated measurements model and the survival model jointly. This is often done by assuming conditional independence of the repeated measurement and the survival outcome conditional on some well-chosen random effects. The IAP doctoral student Dora Kocmanova explored different Bayesian techniques and a manuscript, comparing the different approaches, is in preparation. Dora Kocmanova has been awarded a Marie-Curie Training Site Scheme Grant in the University of Lancaster (UK). Under the supervision of Peter Diggle and Rob Henderson she will explore the repeated measurements-survival techniques from Jan 1, 2004 until June 30, 2004. Her thesis topic is the development of models for jointly modelling an interval censored response with repeated measures taken on continuous/discrete covariates. Further, the existence of latent classes in the data will also be examined, the latter is important for the modelling of HIV data.

Modelling multivariate interval-censored emergence times.
A GEE-method for modelling multivariate interval-censored data has been developed, and has been applied to the emergence times of the permanent teeth as recorded in the Signal Tandmobiel Study. See Bogaerts, Leroy, Lesaffre and Declerck (2002). Furthermore an improvement to the current methodology to calculate the bivariate nonparametric estimate of a survival function for interval-censored data has been proposed by Bogaerts and Lesaffre (2003). Bogaerts and Lesaffre (2003) developed also a smooth bivariate estimate of the survival function, using a penalized likelihood approach, (Bogaerts and Lesaffre, 2003). This methodology can be employed to have an improved estimate of association measures in bivariate survival models. In this respect, Lesaffre and Bogaerts developed, in collaboration with Gijbels (**WP3**) an estimator for Kendall's tau using the bivariate smooth estimate of the survival distribution. The performance of this estimator in small sample sizes has been explored and good results were obtained, the asymptotic behaviour of the estimator is currently under investigation.

Silvia Cecere started the development of a Bayesian model to estimate the correlation in a multivariate survival model when the survival times are interval-censored. In the first step only normally distributed interval-censored data are explored. The method has been applied to the emergence times of some teeth from the children sampled in the Signal Tandmobiel Study. The model will be extended with covariates and this will allow us to see patterns in the emergence times (and associations between them) possibly depending on gender, dietary behaviour and the geographical location in Flandres.

1.5.4 Modelling compliance data

Using data from a lipid-lowering clinical trial Lesaffre and Kocmanova (see Lesaffre et al. (2003)) developed survival models to examine the impact of noncompliance to the administered drug on the time to experiencing a cardiac event.

Rizopoulos, Tsonaka and Lesaffre (KUL-2) developed frequentist and Bayesian methods to analyze U- and J-shaped cross-sectional and repeated measurements data. This kind of data occur frequently in compliance studies as the percentage of days that a patient takes his/her drug correctly. The idea behind the models is to replace the observed proportions by a latent score which has on a transformed scale (logit-scale) approximately a normal distibution. This allows to employ classical (random effects) models (in a Bayesian context) to analyze these data. Two manuscripts are being prepared.

1.6 Work package 5: Classification and mixture models

The progress on the various primary objectives as described in the research proposal is briefly described below.

1.6.1 Studying specific types of mixture models

The specific types of mixture models we have concentrated on are of four types.

PMD models.
Probability matrix decomposition models are models for three-mode binary data, with components for one or two of the modes. The manuscript on how to decide for which modes mixture components are needed is published now. See Meulders, De Boeck and Van Mechelen (2003). Another achievement is the hierarchical extension of the model, with hyperparameters referring to the distribution of the component probabilities as can be found in Meulders, De Boeck, Van Mechelen and Gelman (2003).

Item response models (IRT).
IRT models are models for repeated binary and ordered-category data with a logit or probit link. An important topic of research is the modelling of cognitive data with mixture components as representations of cognitive states. We have concentrated on deductive reasoning (Rijmen and De Boeck (2003)), and on models for transitions between cognitive states (Rijmen, De Boeck and van der Maas (2003)). Finally, we have extended mixture models of the IRT type to a Bayesian framework in De Knop and Van Mechelen (2003), based on a fully Bayesian estimation of basic IRT models (Rasch model and linear logistic test model, both with logit and probit link) (De Knop, Rijmen and Van Mechelen (2003)).

Models with an errant process.
It often occurs in psychological studies that not all data are generated through the assumed process, but instead through a minority process that is considered as an errant process. Guessing is an example. This process can be captured in a mixture component that is differentiated from the dominant normal process. We have extended an extant diffusion process model with such an errant component (Tuerlinckx and Ratcliff (2003)), and we have expanded a Rasch model with random effects for the component probabilities of the errant process component (San Martin, Del Pino and De Boeck (2003)). For

a totally different kind of data (large claims data), Beirlant, Joossens, and Segers (2003) have proposed an extension of the generalized Pareto distribution to model excesses of claim amounts.

Diffusion processes and related stochastic processes have been studied also by the RWTH-Aachen team, although not within a mixture context. They studied the usage of copula for modelling multivariate distributions and for characterizing time-dependence in stochastic processes (Markov processes, Brownian motion, diffusion processes) with a view to applications in finance mathematics. This work has been finished in March 2003. See Schmitz (2003).

On the other hand, also smoothing approaches are developed and investigated for their use in the context of mixture modelling. Ghidey, Lesaffre and Eilers (2004) have developed, in collaboration with Paul Eilers (University of Leiden) a mixed model with a smooth random effects distribution. For instance, the model allows a long-tailed distribution for the random effects, as well as a mixture of normal distributions. The smoothing approach will reveal the underlying mixture structure, though without explicitly detecting the components of the mixture. Penalized likelihood maximization delivers the estimated fixed effects and the smoothing coefficients of the random effects distribution. A similar approach was developed for survival models with left-, right and interval censoring: Komarek, Lesaffre and Hilton (2003) developed an AFT model with a smooth error distribution. Results show that this model nicely reveals the mixture structure of the error distribution, if present. This approach is now being extended to the random effects AFT model (frailty AFT model). Bogaerts and Lesaffre (2003) have developed the approach for fitting a bivariate survival model.

1.6.2 Investigating methods to decide on the number and type of components

We continued our work on Bayesian methods for model selection and model checking, in particular with regard to mixture models (Berkhof, Van Mechelen and Gelman (2003a)) and models that imply latent data (such as mixture component memberships) more in general (Gelman, Van Mechelen, Verbeke, Heitjan, Meulders and Price (2003)). One of the research topics was the enhancement of the power of posterior predictive checks (Berkhof, Van Mechelen and Gelman (2003b)).

1.6.3 Classification techniques other than mixtures

We have further developed a family of three-way hierarchical classes models (with simultaneous three-mode classifications). See Ceulemans and Van Mechelen (2003a, 2004); Ceulemans, Van Mechelen and Leenen (2003), and Ceulemans, Van Mechelen and Kuppens (2004). This has led to a taxonomy of models and an associated model selection strategy, which has been evaluated in a simulation study (Ceulemans and Van Mechelen (2003c)).

For the category of two-way hierarchical classes models, apart from previous developments (Ceulemans and Van Mechelen (2003b); Leenen and Van Mechelen (2004); Lombardi, Ceulemans and Van Mechelen (2003a); Van Mechelen, Lombardi and Ceulemans (2003), a novel constrained model implying a K-class partition has been developed (Lom-

bardi, Ceulemans, and Van Mechelen (2003b)).

The design and mathematical investigation of a 'convexity-based' clustering criterion for analyzing heterogeneity in data sets has been a topic of research of the RWTH-Aachen team. Various forms and specifications of this criterion are possible. A theoretical investigation of special cases and the implementation of a corresponding computer program is under way (diploma thesis by Reinhard Voss).

The previously mentioned type of clustering criterion is directly suited to the simultaneous clustering of the rows and the columns of a contingency table such that the resulting row and column partitions are 'maximally related' to each other. This aspect and corresponding clustering methods have been worked out in detail.

As a major activity and by-product, RWTH-Aachen has elaborated, together with KUL-1 (Van Mechelen, De Boeck), a review article on the large range of 'two-way' or 'two-mode' clustering methods which were scattered around in the literature. Thereby, a classification of classification methods has been tried in order to find a suitable structure for presentation. See Van Mechelen, Bock and De Boeck (2004). See also Bock (2003b, 2004).

Moreover, RWTH-Aachen has investigated and proposed several methods for the classification of 'symbolic data' where the classical data points are replaced by data rectangles in Rp. In particular, they have developed and implemented three methods for constructing Kohonen maps for such data. See Bock (2003a).

1.6.4 Specific cross-links with other work packages

An important part of the work done by the KUL-1 group concerns generalized linear and nonlinear mixed models with a logit link, a topic that is directly related to **WP4** (mixed models) and is also of relevance to **WP6** (latent variables).

It is planned to extend and adapt the developed 'two-way' clustering methods to the case of two-way clustering of micro-array data where genes and samples are clustered simultaneously. This is linked to research work in **WP1**.

1.6.5 Methodological problems

The problem of heterogeneity is an important one for a large variety of models. We have investigated six methods to detect and locate heterogeneity in logistic regression models (individual differences in the intercept and slopes). It was found that nonparametric methods that are commonly used in psychometrics do not work very well, whereas a GEE2 approach turned to be very successful (Balazs, Hidegkuti and De Boeck (2003)).

1.7 Work package 6: Incompleteness and latent variables

The work on missing data and sensitivity analysis can be subdivided into three main categories. First, Kenward, Molenberghs, and Thijs (2003) provided a way to properly classify pattern-mixture models in terms of the time-dependence they induce on the missingness mechanism. Second, with the ever increasing complexity of missing data models, comes the concern of a proper use of such methodology in the context of regulated clinical trials.

Not only do the models have to be correct and optimal in a statistical and scientific sense, they have to allow for their insertion in study protocols and statistical analysis plans in an unambiguous sense. Mallinckrodt et al. (2003a, 2003b) contribute to this debate. While these papers report to the applied audiences, ongoing work is directed towards the applied statistical community. Third, work on sensitivity analysis for incomplete data has been reported in Jansen et al (2003), for categorical data, and in, Molenberghs et al (2003a) and Curran et al (2004), who report on continuous data. Work is ongoing on the proper use of pseudo-likelihood methods when data are incomplete.

Interaction with **WP4**: Verbeke and Molenberghs (2003) have contributed to the proper use of score tests when testing for variance components, a problem common in mixed models.

While the validation of surrogate markers and surrogate endpoints in randomized clinical trials was traditionally done using a single trial, now a meta-analytic approach is becoming ever more popular. To this end, hierarchical and often mixed models have to be used. Contributions to these developments have been made in Tibaldi et al. (2003) who studied the computational implications of this paradigm shift; Molenberghs et al. (2003b) who studied its consequences for practice; Alonso et al. (2003) and Renard et al. (2003b) who studied the case of repeated measures for both the true and the surrogate endpoint. Case studies were considered by Buyse et al. (2003) and Brouwer et al. (2003).

Renard et al. (2003a) formulated a mixed model with serial correlation for repeated binary data, a much needed extension since this flexibility has been reserved, to a large extent, for the case of continuous outcomes.

Repeated measures and longitudinal data models for the context of post-harvesting technology have been reported in Lammertyn et al. (2003) and De Ketelaere et al. (2003).

Wouters et al. (2003) presented a complex multivariate and hierarchical model to explore gene expression data.

Ordinal data are frequently modelled as observable discretizations of latent continuous variables. The so-called polychoric correlations are based on the implicit assumption that the joint distribution of the latent variables is multivariate normal and measures the association between the ordinal variables through the simple correlations among the associated latent variables. Almeida and Mouchart (2003a,b) re-examine the implicit assumption of normality. In Almeida and Mouchart (2003a), the authors notice that the marginal distributions of the latent variables are not identified by the discretizations and that an approach through the copulas appear accordingly to be natural; they re-interpret the normality assumption from the point of view of copulas. Almeida and Mouchart (2003b) consider the general identification problem of discretization models and, making use of the previous paper, they re-interpret the normality assumption through a clear distinction between the identifying aspect and the actually restrictive aspect of the normality assumption.

Many statistical models in social sciences are oriented toward the analysis of individual data and differences among individual are often modelled through individual latent variables, in the spirit of random effect models. Mouchart and San Martin (2003) consider a class of models with several levels of latent variables in an approach not far from multilevel analysis. The authors propose a strategy of progressive specification for this

class of models and pay a particular attention to various identification issues.

Incomplete observability of latent variables typically leads to identification problem. Oulhaj and Mouchart (2003) investigates a duality between the sufficiency for a subparameter and the identification under partial observability.

Vandenhende, Lambert and Ramadan (2003) study statistical models for analyzing longitudinal series of ordinal data. Specific efficacy criteria were defined by the International Headache Society for controlled clinical trails on acute migraine. They are derived from the pain profile and the timing of rescue medication intake. A joint model is proposed enabling to derive success rates in any criteria of interest. Cumulative regression models (based on latent variables) are proposed for each response and combined in a single joint model using a multivariate normal copula. The method is well suited to make predictions based on dose-response trials. More generally, it is a very flexible method to analyse longitudinal series of ordinal data.

2 Network activities

2.1 Web Site

All activities of the IAP-statistics network can be followed very closely from our web site. The address of the website is **http://www.stat.ucl.ac.be/IAP**
The following items can be found on the web pages:

-Our logo;
-Description of the project;
-Call for Applications;
-Research activities (including Meetings, Seminars, Specific Research Projects, etc);
-Technical Report Series and Reprint Series;
-Training and mobility (including short courses, visitors of the network, etc);
-Follow-up Committee;
-Members of the Network;
-Reports of the Scientific Meetings organized by the network;
-Contact details.

2.2 Technical Reports and Reprints Series

Two publication series, available via the website, report on scientific results obtained within the IAP-statistics network: the Technical Report Series and the Reprint Series. The IAP-statistics Technical Report Series groups all papers written under the IAP-statistics network. Each paper in this series has been submitted for publication in an international journal. Once a paper has been accepted for publication in an international journal and has been printed, we will list it into the IAP-statistics Reprint Series.

For the IAP-statistics Technical Report Series we list (title and authors) all papers on our website and for each paper we post a document (ps file or pdf file of the paper) that

can be downloaded from the site. For the IAP-statistics Reprint Series we provide on the website a list containing titles, authors and abstracts of published papers.

In 2003 we had a total of 48 papers written by members of the IAP-statistics network that were put into the IAP-statistics Technical Report Series. Furthermore many papers have been published in international journals. See the lists in Section 3.

2.3 Scientific Meetings

2.3.1 International Workshop

An International workshop on the theme "Statistical Modelling for Complex Data" took place at the Limburgs Universitair Centrum, Diepenbeek, from March 31 till April 2, 2003, and was organized by the LUC, in collaboration with colleagues from the KUL and the UCL. There were tutorial lectures, invited presentations and contributed papers (oral/poster) on functional estimation, mixed models, classification and mixture models, incompleteness and latent variables. The tutorial lectures were meant to give a good introduction and review to a topic, whereas the invited lectures where focused on more specific research topics. The Workshop was attended by many members of the network, among others. For the detailed programme and more information, see the additional information (report) on this meeting on our website.

2.3.2 Special activities by Young Researchers

A group of PhD students of the Institute of Statistics, UCL, organized a "First Young Researchers Day" on May 19, 2003, at the Institute of Statistics, UCL. Among the main objectives of this day was to exchange information around common research interests of PhD students in statistics in Belgium. The one-day meeting was organized around two topics: "Nonparametric Approaches in Survival Analysis" and "Synchronisation and Shape Analysis in Biostatistics". Apart from two invited talks by well-established researchers, almost all talks were presented by doctoral students from Belgian universities, or foreign PhD students visiting Belgium universities. The one-day meeting was also supported by the Graduate School of the Institute of Statistics (UCL), as well as by the National Science Foundation (FNRS), and the SAS Institute.
Such a meeting is a real encouragement for PhD students, and enforces communication between them. For the programme of this meeting, and other details, see the website http: //www.stat.ucl.ac.be/YRD/YRD1/mainpage.html

The second Young Researchers Day will be organized on April 30, 2004.

2.4 Organization of the network: Administrative meeting

The annual administrative meeting with a representative from the federal Office OSTC-Brussels (Ms Lejour), a member of the Follow-Up-Committee (Prof. T. Snijders), promotors of the network, the coordinators L. Simar and I. Gijbels and Ms D. André (head of the administration Institute of Statistics, UCL) took place on October 15, 2003.

2.5 Collaborations, Working groups and Seminars

2.5.1 Collaborations

A lot of collaborations are going on in the network. We only mention here a few examples of collaborations between members of different teams of the network.

The use of splines P-splines in nonparametric regression.
There are a few collaborations going on on this subject between members from the UJF (Anestis Antoniadis), the KUL-2 (Emmanuel Lesaffre, Arnost Komarek, Kris Bogaert, ...) and the UCL (I. Gijbels and P. Lambert).

Frailty Models and Inference.
A very intensive collaboration is continuing between several members of the network on modelling heterogeneity via frailty. The coordinating team is here the LUC-team (headed by P. Janssen) and participating teams are KUL-2 (E. Lesaffre, ...) and UCL (P. Lambert).

Simultaneous two-mode clustering methods.
The KUL-1 team (I. Van Mechelen) and the RWTH-team (H.-H. Bock) worked on a review on simultaneous two-mode clustering methods, for example during a meeting on October 15, 2003. This resulted into a joint publication.

Inverse problems.
A collaboration between members of the UJF-partner (Anestis Antoniadis) and members of the UCL (J. Bigot and R. von Sachs) has been set up, in order to investigate wavelet-based approaches to tackle ill-posed inverse problems involving Poisson data with an application to PET tomography.

2.5.2 Working groups

Frailty Working Group:
The "Frailty Working Group" continued their research work (mainly in Work Package 3). In 2003 the group met 8 times: on February 19, March 21, May 15, June 16, September 8, October 10, November 24 and December 4, 2003. Participants to these meetings on the "Frailty Working Group" were: Tomasz Burzykowski (LUC), Jose Cortinas (LUC), Luc Duchateau (UGent), Paul Janssen (LUC), Catherine Legrand (EORTC), Goele Massonnet (LUC), Rosemary Nguti (LUC-UON), Richard Sylvester (EORTC), and (from time to time) Vincent Ducrocq (INRA).

2.5.3 Seminars

Each of the participating partners organizes on a regular basis statistics seminars at their universities. Announcements of these seminars are sent out to most of the Belgian statisticians, including these participating in the network.

Apart from the regular statistics seminars at the universities involved, several other seminars have been organized under the IAP-statistics network:

- A statistics seminar on May 7, 2003, at the LUC: Arnošt Komárek, (Katholieke Universiteit Leuven), "Accelerated failure times model for arbitrarily censored data with smoothed error distribution".

- An afternoon on "Modelling and Inference for frailties", on Friday May 9, 2003, at the Institute of Statistics, UCL. Speakers were Prof Paul Janssen (LUC Diepenbeek), "The shared frailty model"; and Prof Yi Li (Harvard University, USA), "Inference on clustered survival data using imputed frailties";

- A statistics seminar on May 14, 2003, at the LUC: Geert Molenberghs (LUC Diepenbeek), on "The use of score tests for inference on variance components";

- June 18, 2003, at the LUC: Kris Bogaerts (Katholieke Universiteit Leuven), " A smooth estimate of the bivariate survival density in the presence of left, right and interval censored data".

2.6 Short Courses and Graduate Schools

Several short (intensive) courses have been organized within the framework of the IAP-statistics network. These courses were intended for all members of the network, and in particular (but not exclusively) for the PhD-students. The announcements were each time sent out to all members and posted on the website. No (or reduced) registration fees were required for IAP-members.

A list of the short courses organized during the working year 2003 is given below.

- Professor Alois Kneip, University of Mainz, Germany, gave a short course on "Functional data analysis with applications in biometrics and econometrics"; course of 7.5 hours, in March-April 2003, at the Institute of Statistics, UCL;

- Professor Joel Horowitz, Northwestern University, Illinois, USA, presented a short course on "Bootstrap methods for cross-sectional and time series data"; course of 7.5 hours in May 2003, at the Institute of Statistics, UCL;

- Professor Petros Dellaportas, Athens University of Economics and Business, Greece and Imperial College, London, UK, gave an intensive course of two days on "Advanced Use of MCMC methods", on September 25 and 26, 2003, at the Katholieke Universiteit Leuven, Centrum voor Biostatistiek (organized by Emmanuel Lesaffre).

- Two lectures by Professor Ray Carroll, Deparment of Statistics, Texas A&M University, USA. A first lecture on "Nonparametric and semiparametric regression for longitudinal and clustered data", on Wednesday September 24, 2003, at the Institute of Statistics, UCL, Louvain-la-Neuve.
  A second lecture on "Functional data analysis for colon carcinogenesis experiments", on October 1, 2003, at the Institute of Statistics, UCL, Louvain-la-Neuve.

- Three lectures on time series topics by Professor Jiti Gao, University of Western Australia, Perth, Australia (visiting UCL). All lectures took place at the Institute of Statistics, UCL, Louvain-la-Neuve.
  A first lecture "Recent Developments in Semiparametric Time Series Regression: A

Personal Overview ", on October 2, 2003.
A second lecture on "Simultaneous Model Specification Testing in Nonparametric and Semiparametric Time Series Econometrics", on October 8, 2003.
A third lecture on "Nonparametric Estimation and Comparisons in Stochastic Short-Term Interest Rate Models", on October 9, 2003.

- A short course by Professors Molenberghs, G. (LUC) , Tuerlinckx, F. (KUL) and Verbeke, G. (KUL) (2003) on "Numerical Techniques for Statisticians". This intensive course was organized jointly by the "Interuniversity Graduate School of Psychometrics and Sociometrics" and the IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data" (organizers: the Katholieke Universiteit Leuven and the Limburgs Universitair Centrum). The course took place on November 13 and 14, 2003, at the Katholieke Universiteit Leuven.

- A short course by Professor Helmut Kuechenhoff, University of Munich, Germany on "Measurement error in epidemiologic studies", on November 17 and 18, 2003, at the Katholieke Universiteit Leuven, Centrum voor Biostatistiek (organized by Emmanuel Lesaffre).

- A short course by Professor Paul Eilers, University of Leiden, The Netherlands, on "The Power of Penalties", on December 10 and 11, 2003, at the Institute of Statistics, UCL, Louvain-la-Neuve.

The short courses organized by the UCL were also part of the doctoral programme of the Graduate School in Statistics of the UCL.

2.7 Postdoctoral Researchers and Return Grants

Two post-doctoral researchers have been appointed within the network. Dr Jeremie Bigot, a former PhD-student at UJF, Grenoble, is now a post-doctoral researcher at the UCL. He is working mainly on topics in functional estimation, and more particularly on landmark detection, functional estimation based on samples of curves, and estimation under constraints. Dr Seog-Oh Jeong, from Korea, is collaborating with other members of the network on problems in frontier estimation. Both post-doctoral researchers are affiliated to the Institute of Statistics at the UCL.

Another new participating member of the network is Dr Gerda Claeskens. She obtained a "Return Grant" from the OSTC-Brussels and will spend two years at the UCL (starting August 2003). She is mainly working on function estimation, with special interest in model selection issues.

3 Technical Reports and Publications

Below we provide in each of the subsections two lists of scientific works related to the IAP-statistics network:

A. List of Technical Reports:
This list contains all Technical Reports that have been written in 2003, and **have been submitted for publication to an international journal**. These reports are also available on our web site and the number listed refers to this electronic IAP-Statistics Technical Report Series.

B. List of Publications:
This list contains all publications in international journals (with refereeing system), including also papers that are accepted for publication and are 'in press'. This list also includes papers that have been published in Proceedings and have undergone a peer review (i.e. full length papers). See also the IAP-statistics Reprint Series on our web site.

The list of Technical Reports is included since it allows us to provide a more complete overview of the achieved research results.

3.1 List of publications per research unit/partner

3.1.1 Université catholique de Louvain, UCL partner

A.List of Technical Reports

Almeida, C. and Mouchart, M. (2003a). A Note on a copula approach to polychoric correlations. IAP-statistics Technical Report Series TR # 0327.

Almeida, C. and Mouchart, M. (2003b). Identification of polychoric correlations: a copula approach. IAP-statistics Technical Report Series TR # 0326.

Badin, L. and Simar, L. (2003). Confidence intervals for DEA-type efficiency scores: How to avoid the computational burden of the bootstrap? IAP-statistics Technical Report Series TR # 0322.

Bouezmarni, T., Mesfioui, M. and Rolin, J.-M. (2003). L_1 rate of convergence of Asymmetric Kernel Density Estimators and Smoothed Histograms. IAP-statistics Technical Report Series TR # 0339.

Bouezmarni, T. and Rolin, J.-M. (2003b). Bernstein Estimator For Unbounded Density Function. IAP-statistics Technical Report Series TR # 0311.

Bouezmarni, T. and Scaillet, O. (2003). Consistency of Asymmetric Kernel Density Estimators and Smoothed Histograms with Application to Income Data. IAP-statistics Technical Report Series TR # 0306.

Daouia, A. and Simar, L. (2003). Robust Nonparametric Estimators of Monotone Boundaries. IAP-statistics Technical Report Series TR # 0335.

Delaigle, A. and Gijbels, I. (2003). Boundary estimation and estimation of discontinuity points in deconvolution problems. IAP-statistics Technical Report Series TR # 0325.

Delouille, V., Jansen, M. and von Sachs, R. (2003). Second generation wavelet methods for denoising of irregularly spaced data in two dimensions. IAP-statistics Technical Report Series TR # 0303.

Delouille, V. and von Sachs, R. (2003b). Properties of design-adapted wavelet transforms of nonlinear autoregression models. *Revised version of* IAP-statistics Technical Report Series TR # 0217.

Einmahl, J. and Van Keilegom, I. (2003). Goodness-of-fit tests in nonparametric regression. IAP-statistics Technical Report Series TR # 0341.

Heuchenne, C. and Van Keilegom, I. (2003). Polynomial regression with censored data based on preliminary nonparametric estimation. IAP-statistics Technical Report Series TR # 0340.

Jeong, S.O. and Park, B.U. (2003). Limit distribution of convex hull estimators for boundaries.

Kneip, A, Simar, L. and Wilson, P.W. (2003). Asymptotics for DEA Estimators in Nonparametric Frontier Models. IAP-statistics Technical Report Series TR # 0323.

Mouchart, M. and Rombouts, J. (2003). Clustered panel data models: an efficient approach for nowcasting from poor data. Discussion Paper # 0330, Institut de statistique, UCL.

Mouchart, M. and Vandresse, M. (2003). A measure of market imperfection by frontier analysis IAP-statistics Technical Report Series TR # 0328.

Ombao, H., von Sachs, R. and Guo, W. (2003). SLEX Analysis of multivariate nonstationary time series. IAP-statistics Technical Report Series TR # 0337.

Park, B., Sickles, R. and Simar, L. (2003b). Semiparametric Efficient Estimation in Dynamic Panel Data Models. IAP-statistics Technical Report Series TR # 0321.

Simar, L. (2003b). How to Improve the Performances of DEA/FDH Estimators in the Presence of Noise. IAP-statistics Technical Report Series TR # 0328.

Simar, L and Wilson, P. (2003b). Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes. IAP-statistics Technical Report Series TR # 0310.

Simar, L. and Zelenyuk, V. (2003). Statistical Inference for Aggregates of Farrell-type Efficiencies. IAP-statistics Technical Report Series TR # 0332.

Steinmann, L. and Simar, L. (2003). On the comparability of efficiency scores in nonparametric frontier models. IAP-statistics Technical Report Series TR # 0324.

Van Bellegem, S. and von Sachs, R. (2003a). Locally adaptive estimation of sparse evolutionary wavelet spectra. IAP-statistics Technical Report Series TR # 0316.

Van Bellegem, S. and von Sachs, R. (2003b). On adaptive estimation for locally stationary wavelet processes and its application. IAP-statistics Technical Report Series TR # 0336.

B. List of Publications

Akritas, M. G. and Van Keilegom, I. (2003). Estimation of the bivariate and marginal distributions with censored data. *Journal of the Royal Statistical Society - Series B*, **65**, 457–471.

Beguin, Cl. and Simar, L. (2004). Analysis of the Expenses Linked to Hospital Stays: How to Detect Outliers. *Health Care Management Science*, to appear.

Bouezmarni, T. and Rolin, J.-M. (2003a). Consistency of Beta Kernel Density Function Estimator. *Canadian Journal of Statistics*, **31**, No 1, 89–98.

Chen, X., Linton, O. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, **71**, 1591–1608.

Claeskens, G. and Hjort, N.L. (2003). The Focussed Information Criterion. *Journal of the American Statistical Association*. 2003, **98**, 900–916. With Discussion.

Claeskens, G. and Hjort, N.L. (2004). Goodness of fit via nonparametric likelihood ratios. *Scandinavian Journal of Statistics*, to appear.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics*, **31**, 1852–1884.

Daraio, C. and Simar, L. (2004). Introducing environmental variables in nonparametric frontier models: a probabilistic approach. *Journal of Productivity Analysis*, to appear.

Delaigle, A. and Gijbels, I. (2004a). Practical bandwidth selection in deconvolution kernel density estimation. *Computational Statistics and Data Analysis*, **45**, Number 2, 249–267.

Delaigle, A. and Gijbels, I. (2004b). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *The Annals of the Institute of Statistical Mathematics*, **56**, to appear.

Delouille, V., Simoens, J. and von Sachs, R. (2004). Smooth design-adapted wavelets for nonparametric stochastic regression. *Journal of the American Statistical Association*, to appear.

Delouille, V. and von Sachs, R. (2003a). Smooth design-adapted wavelets for half-regular designs in two dimensions. In Proceedings of the SPIE 2003, San Diego.

De Macq, I. and Simar, L. (2004). Hyperrectangular Space Partitioning Trees, a practical approach. *Computational Statistics*, to appear.

Donoho, D., Mallat, S., von Sachs, R. and Samuelides, Y. (2003). Signal and Covariance Estimation with Macrotiles. *IEEE Transactions on Signal Processing*, **51**, 3, 614–627.

Du, Y., Akritas, M. G. and Van Keilegom, I. (2003). Nonparametric methods for analysis of covariance with censored data. *Biometrika*, **90**, 269–287.

Flahaut, B., Mouchart, M., San Martin, E. and Thomas, I. (2003). The local spatial autocorrelation and the kernel method for identifying black zones: a comparative approach. *Accident Analysis and Prevention*, **35**, 991–1004.

Florens, J.P. and Simar, L. (2004). Parametric Approximations of Nonparametric Frontier. *Journal of Econometrics*, to appear.

Fryźlewicz, P., Van Bellegem, S. and von Sachs, R. (2003). Forecasting non-stationary time series by wavelet process modelling. *Annals of the Institute of Statistical Mathematics* (2003), **55**, 737–764.

Gijbels, I. (2003). Inference for nonsmooth regression curves and surfaces using kernel-based methods. In *Recent Advances and Trends in Nonparametric Statistics*, Eds. Michael G. Akritas and Dimitris N. Politis. Elsevier Science (North Holland), The Netherlands, pp 183–201.

Gijbels, I. (2004). Monotone regression. Institut de Statistique, Université catholique de Louvain, *Discussion Paper*. To appear in *Encyclopedia of Statistical Sciences*. Wiley, New York.

Gijbels, I. and Goderniaux, A.-C. (2004a). Bandwidth selection for change point estimation in nonparametric regression. *Technometrics*, **46**, Number 1, 76–86.

Gijbels, I. and Goderniaux, A.-C. (2004b). Data-driven discontinuity detection in derivatives of a regression function. *Communications in Statistics–Theory and Methods*, **33**, Number 4, 851–871.

Gijbels, I. and Goderniaux, A.-C. (2004c). Bootstrap test for change points in nonparametric regression. *Journal of Nonparametric Statistics*, to appear.

Gijbels, I. and Gürler, Ü. (2003). Estimation in change point models for hazard function with censored data. *Lifetime Data Analysis*, **9**, Number 4, 395–411.

Gijbels, I., Hall, P. and Kneip, A. (2004). Interval and band estimation for curves with jumps. *Journal of Applied Probability*, special issue "Stocastic Methods and Their Applications", in honour of Chris Heyde, Volume 41A, to appear.

Gijbels, I. and Heckman, N. (2004). Nonparametric testing for a monotone hazard function via normalized spacings. *Journal of Nonparametric Statistics*, to appear.

Guo, W., Dai, M., Ombao, H. and von Sachs. R. (2003). Smoothing Spline ANOVA for Time-Dependent Spectral Analysis. *Journal of the American Statistical Association* (2003), **98**, 463, 643–652.

Hall, P. and Van Keilegom, I. (2003). Using difference-based methods for inference in nonparametric regression with time-series errors. *Journal of the Royal Statistical Society - Series B*, **65**, 443–456.

Hjort, N.L. and Claeskens, G. (2003a). Frequentist model average estimators. *Journal of the American Statistical Association*, **98**, 879–899. With Discussion.

Hjort, N.L. and Claeskens, G. (2003b). Rejoinder to "The focussed information criterion" and "Frequentist model average estimators". *Journal of the American Statistical Association*, **98**, 938–945. (Discussion: pp. 917–938).

Mouchart, M. and San Martin, E. (2003). Specification and Identification Issues in Models involving a latent Hierarchical structure. *Journal of Statistical Planning and Inference,* **111**, 143–163.

Oulhaj, A. and Mouchart, M. (2003). Partial Sufficiency with Connection to the Identification Problem. *Metron*, **LXI**, 2, 267–283.

Park, B., Sickles, R. and Simar, L. (2003a). Semiparametric Efficient Estimation of AR(1) Panel Data Models. *Journal of Econometrics*, **117**, 2, 279–309. Corrigendum to "Semiparametric-efficient estimation of AR(1) panel data models".*Journal of Econometrics*, **117**, 2, 311.

Simar, L. (2003a). Detecting Outliers in Frontiers Models: a Simple Approach. *Journal of Productivity Analysis*, **20**, 391–424.

Simar, L., and Wilson, P.W. (2003a). Performance of the bootstrap for DEA estimators and iterating the principle, In *Handbook on Data Envelopment Analysis*, W.W. Cooper and J. Zhu, editors, Kluwer, to appear.

Tilquin, P., Van Keilegom, I., Coppieters, W., Le Boulengé, E. and Baret, P.V. (2003). Non-parametric interval mapping in half-sib designs: use of midranks to account for ties. *Genetical Research*, **81**, 221–228.

Van Bellegem, S. and von Sachs, R. (2004). Forecasting economic time series using models of nonstationarity.*International Journal of Forecasting*, to appear.

Vandenhende, F., Lambert, P. and Ramadan, N. (2003). Statistical models for the analysis of controlled trials on acute migraine. *Pharmaceutical Statistics*, **2**, 199–210.

Van Keilegom, I. (2004). A note on the nonparametric estimation of the bivariate distribution under dependent censoring. *Journal of Nonparametric Statistics*, to appear.

Zhang, J. and Gijbels, I. (2003). Sieve Empirical Likelihood and Extensions of the Generalized Least Squares. *Scandinavian Journal of Statistics*, **30**, 1–24.

3.1.2 Katholieke Universiteit Leuven, KUL-1 partner

A.List of Technical Reports

Balazs, K., Hidegkuti, I. and De Boeck, P. (2003). A comparison of methods to detect heterogeneity in logistic regression models.

Berkhof, J., Van Mechelen, I. and Gelman, A. (2003b). Enhancing the power of a posterior predictive check.

Ceulemans, E., and Van Mechelen, I. (2003c). Hierarchical classes models for three-way three-mode binary data: Interrelations and model selection.

De Knop, S., Rijmen, F. and Van Mechelen, I. (2003). Fully Bayesian estimation of basic IRT models.

Lombardi, L., Ceulemans, E. and Van Mechelen, I. (2003b). K-centroids hierarchical classes analysis.

Meulders, M., De Boeck, P., Van Mechelen, I. and Gelman, A. (2003). Probabilistic feature analysis of facial perception of emotions.

Rijmen, F., De Boeck, P. and Van der Maas, H.L.J. (2003). An IRT model with a parameter-driven process for change.

San Martin, Del Pino, G. and De Boeck, P. (2003). A random-effects model for individual differences in guessing.

Van Mechelen, I., Lombardi, L. and Ceulemans, E. (2003). Hierarchical classes modeling of rating data.

B. List of Publications

Beirlant, J., Joossens, E. and Segers, J. (2003). A new model for large claims. *North American Actuarial Journal*, to appear.

Berkhof, J., Van Mechelen, I. and Gelman, A. (2003a). A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica*, **13**, 423–442.

Ceulemans, E. and Van Mechelen, I. (2003a). Uniqueness of the N-way N-mode hierarchical classes models. *Journal of Mathematical Psychology*, **47**, 259–264.

Ceulemans, E. and Van Mechelen, I. (2003b). An algorithm for the HICLAS-R model. In M. Schader, W. Gaul and M. Vichi (Eds.), *Between data science and applied data analysis: Studies in classification, data analysis, and knowledge organization* (pp. 173–181). Heidelberg: Springer.

Ceulemans, E. and Van Mechelen, I. (2004). Tucker2 hierarchical classes analysis. *Psychometrika*, to appear.

Ceulemans, E., Van Mechelen, I. and Kuppens, P. (2004). Adapting the formal to the substantive: Constrained Tucker3-HICLAS. *Journal of Classification*, to appear.

Ceulemans, E., Van Mechelen, I. and Leenen, I. (2003). Tucker3 hierarchical classes analysis. *Psychometrika*, **68**, 413–433.

De Boeck, P. and Wilson, M. (Eds.) (2004). *Explanatory item response model: A generalized linear and nonlinear approach*. New York: Springer, to appear.

De Knop, S. and Van Mechelen, I. (2003). Estimation of mixture models: A Bayesian approach. Paper presented at the 13th International Meeting of the Psychometric Society, Cagliari, Italy.

Ip, E.H., Wang, Y.J., De Boeck, P. and Meulders, M. (2004). Locally dependent latent trait models for polytomous responses. *Psychometrika*, to appear.

Leenen, I. and Van Mechelen, I. (2004). A conjunctive parallelogram model for pick any/n data. *Psychometrika*, to appear.

Lombardi, L., Ceulemans, E. and Van Mechelen, I. (2003a). A hierarchical classes approach to discriminant analysis. In M. Schader, W. Gaul and M. Vichi (Eds.), *Between data science and applied data analysis: Studies in classification, data analysis, and knowledge organization* (pp. 296–304). Heidelberg: Springer.

Meulders, M., De Boeck, P. and Van Mechelen, I. (2003). A taxonomy of latent structure assumptions for probability matrix decompositions. *Psychometrika*, **68**, 61–77.

Rijmen, F. and De Boeck, P. (2003). Reasoning correlates of individual differences in the interpretation of conditionals. *Psychological Research*, **67**, 219–231.

Rijmen, F., Tuerlinckx, F., De Boeck, P. and Kuppens, P. (2003). A nonlinear mixed model framework for IRT models. *Psychological Methods*, **8**, 18–205.

Smits, D. and De Boeck, P. (2003). A componential IRT model for guilt. *Multivariate Behavioral Research*, **38**, 161–188.

Smits, D., De Boeck, P. and Hoskens, M. (2003). Examining the structure of concepts: Using interactions between items. *Applied Psychological Measurement*, **27**, 41–439.

Smits, P., De Boeck, P. and Verhelst, N. (2003). Estimation of the MIRID: A program and a SAS based approach. *Behavior Research Methods, Instruments, & Computers*, **35**, 537-549.

Tuerlinckx, F. (2004). A multivariate counting process with Weibull distributed first-arrival times. *Journal of Mathematical Psychology*, **4**, 65–79.

Tuerlinckx, F. and Ratcliff, R. (2003). Handling outliers in estimating the diffusion model for reaction time data. Paper presented at the 13th International Meeting of the Psychometric Society, Sardinia, Italy.

3.1.3 Katholieke Universiteit Leuven, KUL-2 partner

A.List of Technical Reports

Andries E., Croes K., Verbeke G., De Schepper L. and Molenberghs G. (2003). Likelihood estimation of finite mixtures.

Bogaerts, K. and Lesaffre, E. (2003). A new fast algorithm to find the regions of possible support for bivariate interval censored data. IAP - statistics Technical Report Series TR #0312.

Fieuws, S. and Verbeke, G. (2004). Modelling multivariate longitudinal profiles: Pitfalls of the random-effects approach. Conditionally accepted by *Statistics in Medicine*.

Fieuws S., Verbeke G., and Brant L.J. (2003). Classification of longitudinal profiles using nonlinear mixed-effects models.

Ghidey, W., Lesaffre, E. and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. (*revised version*).

Komárek, A., Lesaffre, E., Hilton, J. F. (2003). Accelerated failure times model for arbitrarily censored data with smoothed error distribution.

Lesaffre, E., Mwalili S. M., and Declerck, D. (2003). Analysis of caries experience taking inter-observer bias and variability into account.

B. List of Publications

Brant L.J., Sheng S.L., Morrell C.H., Verbeke G., Lesaffre E., and Carter H.B. (2003). Heterogeneous random-effects models for screening of prostate cancer. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, **166**, 51–62.

Bogaerts, K., Leroy, R., Lesaffre, E., and Declerck, D. (2002). Modelling tooth emergence data based on multivariate interval-censored data. *Statistics in Medicine*, **21**, 3775–3787.

Bogaerts, K. and Lesaffre, E. (2003). A new fast algorithm to find the regions of possible support for bivariate interval censored data. *Journal of Computational and Graphical Statistics*, to appear.

Bogaerts, K and Lesaffre, E. (2003). A smooth estimate of the bivariate survival density in the presence of left, right and interval censored data. In Alexandria, VA (Ed): Proceedings of the American Statistical Association, Biometrics Section. American Statistical Association. 2003.

Feys H., De Weerdt W., Peeraer L., Verbeke G., Verlinden B. and Nieuwboer A. (2003). Foot loading and roll-off pattern during walking: A comparison between stroke patients and age- and sex-matched control subjects. *Stroke*, to appear.

Fieuws, S. , Spiessens, B. and Draney, K. (2004). Mixture Models. In: Explanatory item response models. A generalized linear and nonlinear approach. De Boeck, P. and Wilson, M. (eds). New York: Springer, to appear.

Komárek, A., Lesaffre, E., Härkänen, T., Declerck, D., and Virtanen, J. I. (2003). A Bayesian analysis of multivariate doubly interval censored dental data. *Biostatistics*, to appear.

Komárek, A., Lesaffre, E., Hilton, J. F. (2003). Accelerated failure times model for arbitrarily censored data with smoothed error distribution, Proceedings of the 18th International Workshop on Statistical Modelling, Leuven, Belgium, Verbeke, Molenberghs, Aerts, Fieuws (eds.), 233–238.

Lesaffre, E., Kocmanová, D., .Lemos, P.A., Disco, C.M.C. and Serruys, P.W. (2003). A Retrospective Analysis of the Effect of Noncompliance on Time to First Major Adverse Cardiac Event in LIPS. *Clinical Therapeutics*, **25**, 2431–2447.

Morrell C.H., Brant L.J., Pearson J.D., Verbeke G., and Fleg J.L. (2003). Applying linear mixed-effects models to the problem of measurement error in epidemiologic studies. *Communications in Statistics: Simulation and Computation*, **32**, 437–459.

Mwalili S. M., Lesaffre, E. and Declerck, D. (2003a). Correcting for Inter-observer Variability in a Geographical Oral Health Study. Proceedings of the 18th International Workshop on Statistical Modelling, Leuven, Belgium, Verbeke, Molenberghs, Aerts, Fieuws (eds.), 329–334.

Mwalili S. M., Lesaffre, E. and Declerck, D. (2003b). A Bayesian ordinal logistic regression model to correct for inter-observer measurement error in a geographical oral health study. *Journal of the Royal Statistical Society, Series C*, to appear.

Spiessens B., Lesaffre E. and Verbeke G. (2003). A comparison of group sequential methods for binary longitudinal data. *Statistics in Medicine*, **22**, 501–515.

3.1.4 Limburgs Universitair Centrum, LUC partner

A.List of Technical Reports

Braekers, R. and Veraverbeke, N. (2003). A copula-graphic estimator for the conditional survival function under dependent censoring. IAP-statistics Technical Report Series TR # 0315.

Cao, R., Janssen, P. and Veraverbeke, N. (2003). Relative hazard rate estimation for censored and left truncated data. IAP-statistics Technical Report Series TR # 0308.

Cao, R., Lopez-de-Ullibarri, I., Janssen, P. and Veraverbeke, N. (2003). Presmoothed Kaplan-Meier and Nelson-Aalen estimators. IAP-statistics Technical Report Series TR # 0345.

Duchateau, L. and Janssen, P. (2003). Penalized partial likelihood for frailties and smoothing splines in time to first isemination models for dairy cows. IAP-statistics Technical Report Series TR # 0344.

Duchateau, L., Janssen, P., Kezic, I. and Fortpied, C. (2003). Evolution of recurrent asthma event rate over time in frailty models. IAP-statistics Technical Report Series TR # 0307.

Janssen, P., Swanepoel, J. and Veraverbeke, N. (2003). Bootstrapping modified goodness-of-fit statistics with estimated parameters. IAP-statistics Technical Report Series TR # 0347.

Massonet, G., Burzykowski, T. and Janssen, P. (2003). Resampling plans for frailty models. IAP-statistics Technical Report Series TR # 0348.

Nguti, R., Burzykowski, T., Rowlands, J., Renard, D. and Janssen, P. (2003). Joint modelling of repeated traits of lambs bred in sub-humid tropics. IAP-statistics Technical Report Series TR # 0346.

Alonso, A., Geys, H., Kenward, M., Molenberghs, G. and Vangeneugden, T. (2003). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements. *Biometrical Journal*, **45**, 931–945.

Beutels, P., Shkedy, Z., Mukomolov, S., Aerts, M., Shargorodskaya, E., Plotnikova, V., Molenberghs, G., and Van Damme, P. (2003). Hepatitis B in Saint Petersburg, Russia (1994-1999): a descriptive epidemiological analysis. *Journal of Viral Hepatitis*, **10**, 141–149.

Braekers, R., Veraverbeke, N. (2003). Testing for the partial Koziol-Green model with covariates. *Journal of Statistical Planning and Inference*, **115**, 181–192.

Buntinx, F., Geys, H., Lousbergh, D., Broeders, G., Cloes, E., Dhollander, D., Op De Beeck, L., Vandenbrande, J., Van Waes, A., Molenberghs, G. (2003). Geographical differences in cancer incidence in the Belgian province of Limburg. *European Journal of Cancer*, **39**, 2058–2072.

Burzykowski, T., Molenberghs, G., Abeck, D., Haneke, E., Hay, R., Katsambas, A., Roseeuw, D., van der Kerkhof, P., and Marynissen, G. (2003). High prevalence of foot diseases in Europe: results of the Achilles project. *Mycoses*, **46**, 495–505.

Burzykowski, T., Szubiakowski, J., Ryden, T. (2003). Analysis of photon count data from single-molecule fluorescence experiments. *Chemical Physics*, **288**, 291–307.

Buyse, M., Burzykowski, T., Parmar, M., Torri, V., Omura, G., Colombo, N., Williams, C., Conte, P., Vermorken, J. (2003). Using the "expected" survival to explain differences between the results of randomized trials: a case in advanced ovarian cancer. *Journal of Clinical Oncology*, **21** (9), 1682–1687.

De Ketelaere, B., Lammertyn, J., Molenberghs, G., Nicolaï, B., De Baerdemaeker, J. (2003). Statistical models for analyzing repeated quality measurements of horticultural products. Model evaluation and practical example. *Mathematical Biosciences*, **185**, 169–189.

Duchateau, L., Janssen, P., Kezic, I. and Fortpied, C. (2003). Evolution of recurrent asthma event rate over time in frailty models. *Applied Statist., Journal of the Royal Statistical Society C*, **52**, 355–363.

Faes, C., Geys, H., Aerts, M., Molenberghs, G. (2003). Use of fractional polynomials for dose response modelling and quantitative risk assessment in developmental toxicity studies. *Statistical Modelling*, **3**, 109–125.

Fischler, B., Tack, J., De Gucht, V., Shkedy, Z., Persoons, P., Broekaert, D., Molenberghs, G. and Janssens, J. (2003). Heterogeneity of symptom pattern, psychosocial factors, and pathophysiological mechanisms in severe functional dyspepsia. *Gastroenterology*, **124**, 903–910.

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.

Kenward, M.G., Molenberghs, G., and Thijs, H. (2003). Pattern-mixture models with proper time dependence. *Biometrika*, **90**, 53–71.

Lammertyn, J., De Ketelaere, B., Marquenie, D., Molenberghs, G., Nicolaï, B.M. (2003). Mixed models for repeated multicategorical response: modelling the time effect of physical treatments on strawberry sepal quality. *Postharvest Biology and Technology*, **30**, 195-207.

Mallinckrodt, C.H., Carroll, R.J., Debrota, D.J., Dube, S., Molenberghs, G., Potter, W.Z., Sanger, T.D., and Tollefson, G.D. (2003a). Assessing and interpreting treatment effects in longitudinal clinical trials with subject dropout. *Biological Psychiatry*, **53**, 754–760.

Mallinckrodt, C.H., Scott Clark, W., Carroll, R.J. and Molenberghs, G. (2003b). Assessing response profiles from incomplete longitudinal clinical trial data with subject dropout under regulatory conditions. *Journal of Biopharmaceutical Statistics*, **13**, 179–190.

Molenberghs, G., Burzykowski, T., Alonso, A., and Buyse, M. (2003). A perspective on surrogate endpoints in controlled clinical trials. *Statistical Methods in Medical Research*, to appear.

Molenberghs G., Cuijpers, C., Goetghebeur, E.J.T., Passchier, W.F., and Pieters, J. (2003). Comment on Den Hond, E., Roels, H.A., Hoppenbrouwers, K., Nawrot, T., Thijs, L., and Vandermeulen, C. et al. Sexual maturation in relation to polychlorinated aromatic hydocarbons: sharpe and Skakkebaek's hypothesis revisited. *Environmental Health Perspectives* (2002); 110, 771-776; 111, A12.

Nguti, R., Janssen, P., Rowlands, G.J., Audho, J.O. and Baker, R.L. (2003). Survival of Red Masaai, Dorper and crossbred lambs in the sub-humid tropics. *Animal Science*, **76**, 3–17.

Pierik, M., Vermeire, S., Van Steen, K., El Housni, H., Devière, J., Rutgeerts, P., Franchimont, D. (2003). Deficient host - bacteria interaction in inflammatory bowel disease (IBD): The toll-like receptor (TLR)-4 Asp299Gly polymorphism is associated with Crohn's disease (CD) and ulcerative colitis (UC). *Gut*, **52** (suppl IV): A36.

Pierik, M., Vermeire, S., Van Steen K., Joossens, S., Claesses, G., Vlietinck, R., Rutgeerts, P. (2003). TNF-alpha receptor 1 and 2 (TNFR1 and TNFR2) polymorphisms in UBD and their association with response to infliximab. *Gut*, **52** (suppl IV): A46.

Regula, J., Hennig, E., Burzykowski, T., Orlowska, J., Prytulski, K., Pollowski, M., Dziurkowska-Marek, A., Marek, T., Nowak, A., Butruk, E., Ostrowski, J. (2003). Multivariate analysis of risk factors for development of duodenal ulcer in Helicobacter pylori-infected patients. *Digestion*, **67**, 25–31.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., Buyse, M., Vangeneugden, T. and Bijnens, L. (2003b). Validation of longitudinally measured surrogate marker for a time-to-event endpoint. *Journal of Applied Statistics*, **30**, 235–247.

Renard, D., Molenberghs, G., and Geys, H. (2003a). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis*, **44**, 649–667.

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

Shkedy, Z., Aerts, M., Molenberghs, G., and Beutels, P., and Van Damme, P. (2003) Modeling forces of infection using monotone local polynomials. *Applied Statistics*, **52**, 469–485.

Speybroeck, N., Boelaert, F., Renard, D., Burzykowski, T., Mintiens, K., Molenberghs, G., Berkvens, D.L. (2003). Design-based analysis of surveys: a bovine herpesvirus 1 case study. *Epidemiology and Infection*, **131**, 991–1002.

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. *International Journal of Public Health, Hints and Kinks*, **48**(4).

Tibaldi, F.S, Cortiñas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R. (2003). Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation*, **73**, 643–658.

Van Vlierberghe, H., Leroux-Roels, G., Adler, M., Bourgeois, N., Nevens, F., Horsmans, Y., Brouwer, J., Colle, I., Delwaide, J., Bastens, B., Henrion, J., Vries, R.A., Galocsy, C., Michielsen, P., Robaeys, G., 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.

Vinh-Hung, V., Burzykowski, T., Cserni, G., Voordeckers, M., Van De Steene, J., Storme, G. (2003). Functional form of the effect of the numbers of axillary nodes on survival in early breast cancer. *International Journal of Oncology*, **22**, 697–704.

Vinh-Hung, V., Cserni, G., Burzykowski, T., Van De Steene, J., Voordeckers, M., Storme, G. (2003). Effect of the number of uninvolved nodes on survival in early breast cancer. *Oncology Reports*, **10**, 363–368.

Wouters, L., Göhlmann, H.W., Bijnens, L., Kass, S.U., Molenberghs, G., and Lewi, P.J. (2003). Graphical exploration of gene expression data: a comparative study of three multivariate methods. *Biometrics*, **59**, 1133–1141.

3.1.5 Université Libre de Bruxelles, ULB partner

A.List of Technical Reports

Azrak, R. and Mélard, G. (2003). Asymptotic Properties of Quasi-Maximum Likelihood Estimators for ARMA Models with Time-Dependent Coefficients.

Defrise, M. and De Mol, C. (2003). Inverse Imaging with Mixed Smoothness and Sparsity Constraints.

Doz, C., Giannone, D. and Reichlin, L. (2003). A Maximum Likelihood Approach to Dynamic Factor Analysis in Large Panels. *ECARES mimeo, September 2003.*

Dufour, J.-M., Farhat, A. and Hallin, M. (2003). Distribution-Free Bounds for Serial Correlation Coefficients in Heteroskedastic Symmetric Time Series. IAP-statistics Technical Report Series TR # 0343.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2003b). The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting. *Revised version of* IAP-statistics Technical Report Series TR # 0205.

Forni, M., Lippi, M. and Reichlin, L. (2003). Opening the Black Box: Structural Factor Models versus Structural VARs. *CEPR working paper, 2003.*

Giannone, D., Reichlin, L. and Sala, L. (2003a). Tracking Greenspan: systematic and unsystematic monetary policy revisited. *CEPR Working Paper 3550.*

Giannone, D., Reichlin, L. and Sala, L. (2003b). VARs, common factors and the empirical validation of equilibrium business cycle models. *CEPR working paper 3701.*

Hallin, M. and Paindaveine, D. (2003a). Asymptotic Linearity of Serial and Nonserial Multivariate Signed Rank Statistics. IAP-statistics Technical Report Series TR # 0243.

Hallin, M. and Paindaveine, D. (2003b). Optimal Rank-Based Tests for Sphericity. IAP-statistics Technical Report Series TR # 0404.

Hallin, M., Vermandele, C. and Werker, B. (2003a). Linear Serial and Nonserial Sign-and-Rank Statistics: Asymptotic Representation and Asymptotic Normality. IAP-statistics Technical Report Series TR # 0305.

Hallin, M., Vermandele, C. and Werker, B. (2003b). Semiparametrically Efficient Sign-and-Rank Statistics for Median Restricted Models. IAP-statistics Technical Report Series TR # 0403.

Paindaveine, D. (2003a). Should We Throw Away Student and Fisher? An Elementary Proof of Multivariate Chernoff-Savage Results.

Paindaveine, D. (2003b). Chernoff-Savage and Hodges-Lehmann Results for Wilks' Test of Multivariate Independence.

B. List of Publications

Akharif, A. and Hallin, M. (2003). Efficient Detection of Random Coefficients in AR(p) Models. *Annals of Statistics*, **31**, 675-704.

Azrak, R., Mélard, G. and Njimi, H. (2003). Forecasting in the analysis of mobile telecommunication data: corrections for outliers and replacement of missing observations. *CoPSTIC'03, Actes de la première Conférence en Sciences et Techniques de l'Information et de la Communication*, Université Mohamed V, Rabat, 69–72.

Bertero, M., Boccacci, P., Custo, A., Demol, C. and Robberto, M. (2003). A Fourier-Based Method for the Restoration of Chopped and Nodded Images. *Astronomy and Astrophysics*, **406**, 765-772.

Cohen, A., Mélard, G. and Ouakasse, A. (2003a). Une expérience de télé-enseignement en statistique pour une banque centrale : aspects techno-logiques. *CoPSTIC'03, Actes de la première Conférence en Sciences et Techniques de l'Information et de la Communication*, Université Mohamed V, Rabat, 19–22.

Cohen, A., Mélard, G. and Ouakasse, A. (2003b). Emploi d'un tableur dans un cours d'analyse de séries temporelles. *Actes des XXXVèmes Journées de Statistique*, Lyon, France, May 13-17, Société Française de Statistique, Vol. **1**, 341-344.

Cristadoro, R., Forni, M., Reichlin, L. and Veronese, G. (2004). A Measure of Core Inflation for the EURO Area. *Journal of Money, Credit and Banking*, to appear.

Daubechies, I., Defrise, M. and De Mol, C. (2003). An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint. *Communications on Pure and Applied Mathematics*, to appear.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2003a). Do Financial Variables Help Forecasting Inflation and Real Activity in the Euro Area. *Journal of Monetary Economics*, **50**, 1243-1255.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2004). The Generalized Dynamic Factor Model: Consistency and Rates. *Journal of Econometrics*, **119**, 231–255.

Hall, P.G., Hallin, M. and Roussas, G. G. (2003). Madan Lal Puri, Selected Collected Work. V.S.P., Utrecht and Boston, 3 volumes, xvi + 787, xi + 743, and xvi + 773 pages.

Hallin, M. and Lotfi, S. (2004). Optimal detection of periodicities in vector autoregressive models. In *Statistical Modeling and Analysis for Complex Data Problems*, Duchesne, P. and Rémillard, B. (Eds), Kluwer, to appear.

Hallin, M., Lu, Z. and Tran, L.T. (2003). Kernel Density Estimation for Spatial Processes: the L_1 Theory. *Journal of Multivariate Analysis*, **88**, 61-75.

Hallin, M., Lu, Z. and Tran, L. T. (2004). Local Linear Spatial Regression. *Annals of Statistics*, to appear.

Hallin, M and Paindaveine, D. (2003). Affine-Invariant Linear Hypotheses for the Multivariate General Linear Model with ARMA Error Terms. In *Mathematical Statistics and Applications*, Froda, S., Léger, Chr. and Moore, M. (eds), Festschrift for Constance van Eeden, I.M.S. Lecture Notes-Monograph Series, 417-434.

Hallin, M. and Paindaveine, D. (2004a). Multivariate Signed Rank Tests in Vector Autoregressive Order Identification. *Statistical Science*, to appear.

Hallin, M. and Paindaveine, D. (2004b). Affine-Invariant Aligned Rank Tests for the Multivariate General Linear Model with ARMA Errors. *Journal of Multivariate Analysis*, to appear.

Hallin, M. and Paindaveine, D. (2004c). Rank-Based Optimal Tests of the Adequacy of an Elliptic VARMA Model. *Annals of Statistics*, to appear.

Hallin, M. and Saidi, A. (2004). Testing Non-Correlation Between Two Multivariate ARMA Time Series. *Journal of Time Series Analysis*, to appear.

Hallin, M. and Werker, B. (2003). Semiparametric Efficiency, Distribution-Freeness, and Invariance. *Bernoulli*, **9**, 137-165.

Klein, A. and Mélard, G. (2004). An algorithm for computing the asymptotic Fisher information matrix for seasonal SISO models. *Journal of Time Series Analysis*, to appear.

Mélard, G., Njimi, H. and Pasteels, J.-M. (2003). Modélisation SARIMA assistée. *Actes des XXXVèmes Journées de Statistique*, Lyon, France, May 13-17, Société Française de Statistique, Vol. **2**, 731-734.

Mélard, G. and Ouakasse, A. (2003). Estimation en ligne pour des modèles dynamiques. *Actes des XXXVèmes Journées de Statistique*, Lyon, France, May 13-17, Société Française de Statistique, Vol. **2**, 965-968.

Oja, H. and Paindaveine, D. (2004). Optimal Signed-Rank Tests Based on Hyperplanes. *Journal of Statistical Planning and Inference*, to appear.

Paindaveine, D. (2004). Procédures optimales fondées sur les rangs multivariés. *Journal de la Société Française de Statistique*, to appear.

Reichlin, L. (2003). Factor Models in Large Cross Sections of Time Series, in *Advances in Economics and Econometrics: Theory and Applications*, Vol **III**, Dewatripont, M., Hansen, P.L. and Turnowsky, S. (eds), 8th World Congress of the Econometric Society, Cambridge University Press, 47–86.

3.1.6 Aachen Technical University, RWTH partner

A.List of Technical Reports

Schmitz, V. (2003). Copulas and stochastic processes. *Doctoral Dissertation*, RWTH Aachen, 108 pp.

Nordhoff, O. (2003). Erwartungswerte zufälliger Quader (Expectation of random rectangles). *Diploma thesis*, RWTH Aachen, 84 pp.

B.List of Publications

Bock, H.-H. (2003a). Clustering algorithms and Kohonen maps for symbolic data. *Journal of the Japanese Society of Computational Statistics*, **15**, 217–229.

Bock, H.-H. (2003b). Two-way clustering for contingency tables: Maximizing a dependence measure. In: M. Schader, W. Gaul, and M. Vichi (eds.): *Between data science and applied data analysis*. Springer Verlag, Heidelberg, 143–154.

Bock, H.-H. (2004). Convexity-based clustering criteria: theory, algorithms, and applications in statistics. *Statistical Methods and Applications*, **12**, to appear.

3.1.7 Université Joseph Fourier, UJF-LMC-IMAG partner

A.List of Technical Reports

Antoniadis, A., Amato, U. and De Feiss, I. (2003). Dimension Reduction in Functional Regression with Applications.

Bigot, J. (2002). A scale-space approach with wavelets to landmark detection. IAP-statistics Technical Report Series TR # 0246. (conditionally accepted in ESAIM: P & S).

Bigot, J. (2003b). Landmark-based registration of 1D curves and functional analysis of variance with wavelets. IAP-statistics Technical Report Series TR # 0333.

Fort, G. and Lambert-Lacroix, S. (2003). Classification using Partial Least Squares with Penalized Logistic Regression. IAP-statistics Technical Report Series TR # 0331.

B.List of Publications

Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003). Effective Dimension Reduction Methods for Tumor Classification using Gene Expression Data. *Bioinformatics*, **19**, No 5, 563–570.

Antoniadis, A., Peyre, J. et al. (2003). Biological detection of low radiation doses by combining results of two microarray analysis methods. *Nucleid Acid Research*, to appear.

Bigot, J. (2003a). Automatic landmark registration of 1D curves. In M. Akritas and D.N. Politis (eds.), *Recent advances and trends in nonparametric statistics*, Elsevier.

3.2 List of joint publications

A.List of joint Technical Reports

Nguti, R., Claeskens, G. and Janssen, P. (2003). Likelihood ratio and score tests for a shared frailty model: a non-standard problem. IAP-statistics Technical Report Series TR # 0309. (LUC and UCL).

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D.F. and Meulders, M. (2003). Completed-data plots for model diagnostics with missing and latent data. (KUL-1 and KUL-2).

Gelman A., Van Mechelen I., Verbeke G., Heitjan D.F., Meulders M. and Price P.N. (2003). Bayesian model checking for missing and latent data problems using posterior predictive checks. (KUL-1 and KUL-2).

B. List of joint Publications

Claeskens, G., Aerts, M., and Molenberghs, G. (2003). A quadratic bootstrap method and improved estimation in logistic regression. *Statistics and Probability Letters*, **61**, 383–394. (LUC and UCL).

Curran, D., Molenberghs, G., Thijs, H. and Verbeke, G. (2004). Sensitivity analysis for pattern mixture models. *Journal of Biopharmaceutical Statistics*, **14**, 125–143. (LUC and KUL-2).

Hens N., Aerts M., Molenberghs G., Thijs H., and Verbeke G. (2004). Kernel weighted influence measures. *Computational Statistics and Data Analysis*, to appear. (LUC and KUL).

Molenberghs, G., Thijs, H., Kenward, M.G., and Verbeke, G. (2003). Sensitivity analysis for continuous incomplete longitudinal outcomes. *Statistica Neerlandica*, **57**, 112–135. (LUC and KUL-2).

Molenberghs G. and Verbeke G. (2003b). Meaningful statistical model formulations for repeated measures. *Statistica Sinica*, to appear. (LUC and KUL-2).

Molenberghs G. and Verbeke G. (2004). An Introduction to (Generalized) (Non-)Linear Mixed Models,' In: Explanatory item response models: A generalized linear and nonlinear approach, P. De Boeck and M. Wilson (Eds.), Springer-Verlag, New-York, in press. (LUC and KUL-2).

Van Mechelen, I., Bock, H.-H. and De Boeck, P. (2004). Two-mode clustering methods: A structured overview. *Statistical Methods in Medical Research*, to appear. (KUL-1 and RWTH-Aachen).

Verbeke, G., and Molenberghs, G. (2003a). The use of score tests for inference on variance components. *Biometrics*, **59**, 254–262. (LUC and KUL-2).

Verbeke G., and Molenberghs G., (2004). Modelling Gaussian and non-Gaussian longitudinal data. *Yearbook of the Finnish Statistical Society*, in press. (LUC and KUL-2).