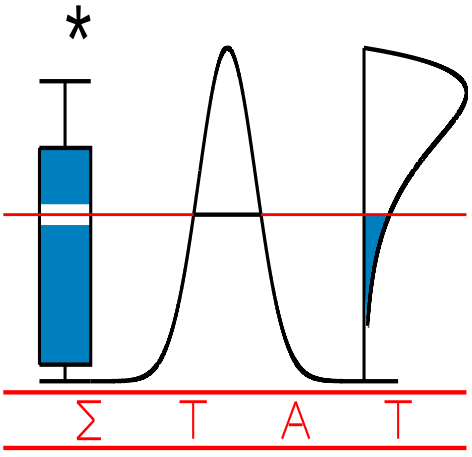
Progress Report 2002

IAP-network in Statistics

Contract P5/24

March 27, 2003



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:

* Bootstrap procedures and robust estimation procedures in frontier estimation have been developed. Nonparametric procedures for estimating the boundary of a density in the deconvolution context have been proposed and the issue of practical bandwidth selection in the deconvolution context has been studied in detail.
  For detecting abrupt changes in regression curves and surfaces the research was focused on the choice of error criteria and diagnostic functions, and practical choices of smoothing parameters. Furthermore, wavelet-based methods for landmark detection have been developed.

* New results have been obtained for the inference on regression models with survival data. These results include data that are subject to informative censoring or interval censoring. Moreover they also allow for complex designs by considering frailty models and accelerated failure time models. The study of likelihood ratio hetereogeneity tests and score tests in shared frailty models has been initiated.

* Local modelling of non-stationarity based on wavelet methods has been further developed. A dynamic factor methodology has been worked out and has been applied very succesfully in practice. Furthermore rank-based procedures in a general class of time series models have shown to allow for semiparametric efficiency while keeping distribution-freeness. Also an important step in the direction of optimal rank-based multivariate method in time series problems has been accomplished.

* A reformulation of a broad range of psychometric models of the item-response model type into the framework of generalized linear mixed models and nonlinear mixed model has been worked out. A further development of binary decomposition models for classification (hierachical classes models), parallel to the extant three-way PCA models has been carried out. The development of mixture models with two levels of latent binary variables, based on a stochastic version of the hierachical classes models has been taken care of.

* Methods for sensitivity analysis in incomplete longitudinal data have been proposed, as well as methods for analyzing incomplete longitudinal and survey data and methods for the validation of surrogate markers in clinical trials.
  The linear mixed model has been extended with a flexible random effects model and the accelerated failure time model has been studied with a smooth error distribution. A smooth bivariate survival model for interval censored data has been proposed and joint modelling of multivariate longitudinal profiles has been focused on.

1.2 Work package 1: Functional estimation

1.2.1 Nonparametric estimation of a frontier function and deconvolution problems

In deterministic frontier models, the most popular nonparametric estimators are based on envelopment estimators. Simar and Wilson (2002) propose a methodology for testing returns to scale in the production process. Cazals, Florens and Simar (2002) suggest a new concept of frontier (order-m frontier) with a nonparametric estimator, more robust than the traditional ones to outliers and/or to extreme values. Florens and Simar (2002) propose a new method for estimating parametric approximations of nonparametric frontiers. Beguin and Simar (2002) use the order-m frontier concept to detect outliers in the field of Hospital's expenses. C. Beguin defended her PhD thesis on this topic in 2002. The research on nonparametric frontier estimation will be pursued in several directions: how to explain the inefficiencies of firms by environmental factors; how to avoid bootstrap for doing inference in these envelopment models; is the bootstrap consistent in this framework?

Members from the UJF-team (Bouchard, Girard, Iouditski and Nazin) have developed a new method for estimating a frontier function. It is a linear combination of kernel functions enveloping all the observed points. The asymptotic behavior for the L_1 error is provided.

Hall and Simar (2002) have proposed a new way of handling situations where some noise can pertubate the data in a boundary problem: this can help to analyze stochastic nonparametric models. Future work will analyze how this basic procedure can be adapted

to multidimensional problems. This situation of noise pertubating the data, can be viewed as a deconvolution problem. Delaigle and Gijbels (2001, 2002a,b) studied in detail the bandwidth selection issue in kernel type estimation of a density in deconvolution problems. This work was part of the PhD thesis of Aurore Delaigle (see Delaigle (2003)).

When a panel of firms is available, stochastic semiparametric modelling is possible: Park, Sickles and Simar (2003) continue to generalize previous work to more complex data generating processes. Future work will involve dynamic models.

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

Fully data-driven procedures for detecting changes in a regression function or its derivatives were developed by Gijbels and Goderniaux (2000, 2001a). Their work discusses two crucial issues: how to choose the parameters in the procedure proposed by Gijbels, Hall and Kneip (1996) and related procedures, and how to determine, at the same time, the number of change-points in a regression function. Gijbels and Goderniaux (2001b) also developed testing procedures for testing continuous versus discontinuous regression functions, as well as similar testing procedures concerning derivatives of regression functions. Parts of this work have been revised.
The techniques developed for the univariate case have also been generalized to detection of changes in regression surfaces (the bivariate case). Here additional interesting questions arise: which error criterion to use, how to deal with selection of smoothing parameters, what are appropriate diagnostic functions, etc. Some partial answers were obtained by members of the UCL team (I. Gijbels, and A. Lambert, a doctoral student of the IAP-statistics network). Parts of the results can be found in the DEA-thesis of A. Lambert. See Lambert (2002). A manuscript is under preparation.

Antoniadis and Gijbels (2002) considered wavelet-based methods for detecting irregularities in a regression function. This procedure also allows to estimate the number of discontinuities. Taken into account the estimated change-points it was shown that the resulting non-smooth regression estimate achieves the optimal rate of convergence.

The team of UJF (J. Bigot and A. Antoniadis) developed a scale-space approach with wavelets for landmark detection. Bigot (2002) is concerned with the problem of determining the typical features of a curve when it is observed with noise for the purpose of landmark-based matching. It has been shown that one can characterize the local structure of a signal by following the propagation across scales of the zero-crossings and the modulus maxima of its continuous wavelet transform. In his work a nonparametric approach is proposed to estimate the zero-crossings and the wavelet maxima of a signal observed with noise at various scales. In order to identify the landmarks of the unknown signal, he introduces a new tool that computes the "density" of the location of the zeros and the modulus maxima of a wavelet representation along various scales. Combined with bagging this approach is shown to be an effective technique for detecting significant features of a signal corrupted by noise and for removing spurious estimates. The asymptotic properties of the resulting estimators are studied and illustrated by simulations. An application to some real data sets is also proposed.

1.2.3 Modelling of heterogeneous regularities

A first item of research has been the development of design-adapted wavelets for stochastic regression and autoregression. Delouille and von Sachs (2002a,b) and Delouille, Jansen and von Sachs (2003) developed and studied a new algorithm treating non-parametric regression and non-linear autoregression by wavelet thresholding estimators. This new methodology is based on the use of "Lifting Schemes" to be able to generalize existing wavelet methods to non-regular, stochastic designs of arbitrary (i.e. non-dyadic) sample size, and being automatically adapted to the interval of observation. Starting from a non-balanced orthogonal Haar basis, more regular biorthogonal bases are constructed and successfully applied to both stochastic regression and non-linear autoregression models. See Delouille and von Sachs (2002a). This allows interesting applications in financial time series analysis, including non-parametric ARCH type models. In a second phase, two-dimensional generalisations of the previously univariate approach have been achieved. Delouille and von Sachs (2002b) covers the case of a random design on a cartesian product, whereas Delouille, Jansen and von Sachs (2003) treats the more general case of an arbitrary two-dimensional design. For the latter one, a completely different construction of lifting has to be used, which is based on interpolating Lagrange polynomials. A Bayesian thresholding scheme allows to achieve the same good performance as in the classical case of regular, i.e. equidistant and non-random, design. The proposed algorithms parallel the existing ones for the classical case as for their computing complexity.

1.2.4 Functional estimation for microarray data

Here the focus is on microarray data, and one of the main issues is to classify such complex data. There are cross links with **WP5** where one also deals with classification techniques for other complex data structures.

With respect to developing nonparametric methods for analyzing and classifying microarray data, a technical report by Peyre (2001) reviews all available methods for normalizing such data. The conclusion is to use a new normalization method based on ranks which is the method adopted by the UJF group for all subsequent analyses related to microarray data.

Another particular application of microarray data, is to uncover the molecular variation among cancers. One feature of microarray studies is the fact that the number n of samples collected is relatively small compared to the number p of genes per sample which are usually in the thousands. In statistical terms this very large number of predictors compared to a small number of samples or observations makes the classification problem difficult. An efficient way to solve this problem is by using dimension reduction statistical techniques in conjunction with nonparametric discriminant procedures. Antoniadis, Lambert-Lacroix and Leblanc (2003) view the classification problem as a regression problem with few observations and many predictor variables. They use an adaptive dimension reduction method for generalized semi-parametric regression models that allows them to solve the 'curse of dimensionality problem' arising in the context of expression data.

1.2.5 General methodological issues

Several methodological issues have been studied under this Work Package, among which issues such as bandwidth selection, construction of testing procedures and confidence bands. Furthermore, research has been devoted to develop a general methodology for nonparametric regression via regularization, as well as to develop classification trees based on general splitting rules (as opposed to the classical binary splitting rules).

D. Climov defended her PhD thesis in 2002 on the topic of semiparametric estimation of Poisson regressions. Climov, Hart and Simar (2002) and Climov, Delecroix and Simar (2002) analyze the properties of the estimators and discuss the bandwidth selection issue. Testing issues will be analyzed in a near future.

M-estimators in non-standard contexts has been considered by Chen, Linton, and Van Keilegom (2002). 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.

Hall and Van Keilegom (2002) consider 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 provided.

Claeskens and Van Keilegom (2002) have been studying the construction of confidence bands for a regression function and its derivatives up to order p, using local polynomial estimation of order p. Two types of confidence bands are obtained: those based on asymptotic normality and those based on a smoothed bootstrap approach.

The use of a nonparametric approach based on a generalization of the Wilcoxon test in the context of quantitatve trait loci methods has been studied by Tilquin, Van Keilegom, Coppieters, Le Boulengé and Baret (2002).

Gijbels and Heckman (2000) investigate the problem of testing for a monotone increasing or decreasing hazard function. Their technique is based on local versions of an existing global test using normalized spacings. The advantage is that no continuous smoothing parameter is involved in these tests. Hall and Van Keilegom (2002) propose a new test for the hypothesis that the hazard rate of a certain population is an increasing function. The test statistic and the way to calibrate it do not suppose that the distribution under the null hypothesis is exponential.

Wang, Akritas and Van Keilegom (2002) consider a heteroscedastic regression model and propose a new test statistic for the hypothesis of a constant regression function. The test statistic is inspired by the classical test for this hypothesis in the context of repeated measurements.

With respect to developing a general methodology for nonparametric regression via regularization, ongoing work between the UCL (I. Gijbels) and the UJF partner (A. Antoniadis) and collaborators is devoted to penalized likelihood regression for generalized linear models with nonquadratic penalties. Indeed, one popular method for fitting a regression function from data measurements is regularization: minimize an objective function which enforces a roughness penalty in addition to coherence with the data. This is the case when formulating penalized likelihood regression for exponential families models. Most existing

smoothing methods employ quadratic penalties that lead to linear estimates which are easy to compute and enjoy good asymptotic properties. However, such smoothing methods are in general incapable of recovering discontinuities or other important attributes in the regression function. In contrast, nonlinear estimates are generally more accurate. In their work they focus on nonparametric penalized likelihood regression methods involving the use of spline spaces and a variety of nonquadratic penalties, pointing out some basic principles they have in common. They also present an asymptotic analysis of convergence rates that justify the approach. This project concernes the use of P-splines. Similar techniques are used in **WP 4** and **WP 5** in the context of mixture models and modelling heterogeneity. The study here is more theoretical of nature and hopes to provide the users with a clear recommendation regarding, for example, the choice and impact of the penalty function.

Members of the UCL-team (I. De Macq and L. Simar) are working on classification trees. An exact algorithm based on hyper-rectangular partitioning trees has been implemented: the computing complexity limits its applicability. A new algorithm, based on the quantiles for building the splitting rules is now implemented: its performances will be compared with more classical classification algorithms. This fairly new technique of classification trees is a generalization of the classical classification tree and this item has links with **WP5**.

1.3 Work package 2: Time series

1.3.1 Approaches to nonstationarity in combination with dimension reduction

Although sophisticated, the theory of multivariate time series still is often defeated in the presence of two major complexities: nonstationarity and high-dimensional data (as opposed to long series lengths). Such complexities unfortunately are the rule rather than the exception in a number of applications, including stock exchange data, EEG-curves, macroeconomic data, environmental data or data from clinical monitoring.

With this project we address these two important problems by means of (i) the development of applicable approaches to non-stationarity, and (ii) adequate methods of dimension reduction. The common thread connecting these two aspects is a search for an appropriate decomposition of the data, based on techniques such as multivariate (wavelet, localized Fourier) basis representations for stochastic processes, spectral Principal Component Analysis techniques, and alike. In particular we further develop and compare the methodology proposed by partners from the UCL team, using a specifically localized Fourier basis (SLEX: Smoothed Localized complex EXponentials), see Ombao et al. (2002). This approach bas recently been rendered truly multivariate by Ombao et al. (2003) in order to be compared in the next stage of the project with the approach of the ULB team, which uses the dynamic factor modelling approach.

Methods of dimension reduction are appropriate when the number of channels is high (as high, for instance, as the number of observations per channel). Under this, still stationary, approach, the observations are decomposed into two mutually orthogonal components, a "common" and an "idiosyncratic" one, where the common component results

of a "low-dimensional" unobserved series of common shocks—the "factors". A consistent method of estimation of these common components, based on a dynamic factor-analytic technique, has been proposed by partners from the ULB team. This method has been successfully applied to large cross-sections of time series data, with dimension as high as 1,000. In Forni, Hallin, Lippi and Reichlin (2002, 2002b), this approach has been successfully used for prediction and explaining of financial data such as inflation and activity in the Euro zone. It is hence an interesting though still stationary alternative to Van Bellegem and von Sachs (2002), who use results from Fryzlewicz, Van Bellegem and von Sachs (2002), on forecasting economic time series using the aforementioned (wavelet based) models of nonstationarity.

1.3.2 Dynamic factor methodology and the analysis of large panels of time series data

The study of the generalized dynamic factor methodology proposed by Forni, Hallin, Lippi, and Reichlin (2000) has been pursued further.

In Forni, Hallin, Lippi, and Reichlin (2002a), the rates of consistency for the method are investigated as both the cross-sectional dimension n and the series lengths T are tending to infinity along some path $(n, T(n))$. The results show that, under suitable assumptions, consistency requires $T(n)$ to be at least of the same order as n, whereas an optimal rate of $\sqrt{n}$ is reached for $T(n)$ of the order of n^2. If convergence to the space of common components is considered, consistency holds irrespective of the path ($T(n)$ thus can be arbitrarily slow); the optimal rate is still $\sqrt{n}$, but only requires $T(n)$ to be of the order of n.

Forni, Hallin, Lippi, and Reichlin (2002b) illustrates the applicability of the method and its efficiency on real-world data. The paper uses a large data set, consisting of 447 monthly macroeconomic time series concerning the main countries of the Euro area to simulate out-of-sample predictions of the Euro area industrial production and the harmonized inflation index and to evaluate the role of financial variables in forecasting. Two competing models are considered : Forni, Hallin, Lippi, and Reichlin (2000, 2001b) and Stock and Watson (1999). The performances of both models are compared to those of a simple univariate AR model. Results show that multivariate methods outperform univariate ones in the forecast of inflation at one, three, six, and twelve months, and in the forecast of industrial production at one and three months. It is found that financial variables do help forecasting inflation, but do not help forecasting industrial production.

The method described in Forni, Hallin, Lippi, and Reichlin (2000) separates the common shock and idiosyncratic spaces via two-sided infinite-order estimated filters acting on the observations. Therefore, its performance at the end of the sample, hence in forecasting problems, is poor. A forecasting method, which uses the same methodology only as a first step, is developed in Forni, Hallin, Lippi, and Reichlin (2000); two other related technical papers have been prepared by Giannone, Reichlin, and Salaa, and appeared as CEPR Working Papers.

1.3.3 Nonparametric and semiparametric approaches to time series problems

This theme is organized along three main axes: (a) (auto)regression quantile and rank score methods for time series problems, (b) multivariate extensions of ranks and signs,

and (c) kernel-based methods for random fields (which constitute a nontrivial multi-index generalization of time series).

(a) Classical rank-based estimators (R-estimators) have been investigated, mainly, in the context of linear models with independent observations. This method of R-estimation has been extended to the context of autoregressive time series models by Koul and Sen (1991) and Koul and Saleh (1993). El Bantli and Hallin (2002a) propose a new class of estimates of the autoregression parameter in AR(p) models, based on *autoregression rank scores*. These estimators are based on linear programming algorithms, combined with a discrete numerical optimisation step. They are shown to be asymptotically equivalent to the R-estimators of autoregressive parameters proposed by Koul and Saleh (1993). In contrast with the latter, however, autoregression rank score estimators are *autoregression invariant*, so that each component of the parameter can be estimated separately. This property allows for substituting p one-dimensional discrete optimisation steps for a unique p-dimensional one, which is computationally simpler.

The dual linear programs yield the so-called autoregression quantiles. In El Bantli and Hallin (2002b), these autoregression quantiles are used in order to compute an estimation of the quantile function or *sparsity function* $\alpha \mapsto f(F^{-1}(\alpha))$ associated with the innovation density f of an autoregressive model of order p. Being based on autoregression quantiles, these estimates do not require estimating the parameters of the model. Contrary to more classical estimates based on estimated residuals, they are *autoregression-invariant* and scale equivariant. Their asymptotic behavior is derived from a uniform Bahadur representation for autoregression quantiles.

This investigation of inference methods for time series based on autoregression quantiles and rank scores is part of the doctoral dissertation of Faouzi El Bantli, under the supervision of Marc Hallin.

(b) This second axis of research was the subject of D. Paindaveine's doctoral dissertation (defended in September of 2002; promotor M. Hallin). Four publications (Hallin and Paindaveine 2002a, b, and c) already appeared or were accepted from that thesis, which since then (January 2003) has been awarded (for the three-year period 2000-2002) the Prix de la meilleure thèse francophone of the French statistical Society (Société Française de Statistique).

A class of multivariate signed rank tests is developed for the general linear model with VARMA error terms. This model includes, as particular cases, one- and m-sample location, multiresponse ANOVA and regression, and VARMA models with a linear trend. Two types of multivariate signs and ranks are considered:

(i) hyperplane-based signs (namely, Randles (1989)'s concept of interdirections), and ranks of *lift-interdirections* (a related concept of distances between pairs of points in $\mathbb{R}^k$), or

(ii) spatial signs of sphericized residuals, and the ranks of pseudo-Mahalanobis distances (distances between the sphericized residuals and the origin in $\mathbb{R}^k$), where the sphericization is performed via a square-root of the multivariate M-estimator of scatter due to Tyler (1987).

The resulting tests, which generalize univariate serial and nonserial signed-rank tests, are strictly affine-invariant, and asymptotically invariant under a group of monotone radial transformations acting on the residuals, hence asymptotically distribution-free. They are valid under the class of elliptically symmetric densities, without any moment assumption. Depending on the score function considered (van der Waerden, Laplace, ...), they allow for locally asymptotically optimal (à la Le Cam-Hájek) tests at selected densities (multivariate normal, multivariate double-exponential, ...). Local powers and asymptotic relative efficiencies are derived—with respect to the corresponding optimal parametric Gaussian tests, and with respect to some well-known competitors (Randles (1989)'s sign test and Peters and Randles (1990)'s Wilcoxon-type signed rank tests, Oja median tests, etc). Two famous (location) univariate results are extended to the multivariate case (both in the location and serial cases): (i) we establish a multivariate version of the traditional Chernoff-Savage (1958) property, showing that the traditional Gaussian procedures (used in daily practice) are uniformly dominated, in the Pitman sense, by the van der Waerden version of our tests, and (ii) we generalize the celebrated Hodges-Lehmann (1956) ".864 result", providing, for any fixed space dimension k, the lower bound for the asymptotic relative efficiency of Wilcoxon-type (or Spearman-type, in the serial case) tests with respect to the Gaussian tests.

(c) Research on this topic was carried out by M. Hallin, in collaboration with Zudi Lu (Chinese Academy of Sciences, Beijing) and Lanh T. Tran (Indiana University, Bloomington). The paper Hallin, Lu, and Tran (2002) deals with the L_1 theory for kernel density estimation in random fields. The purpose of this paper is to investigate kernel density estimators for spatial processes with linear or nonlinear structures. Sufficient conditions for L_1 consistency are obtained under extremely general, verifiable conditions. The results hold for mixing as well as for non-mixing processes. Potential applications include testing for spatial interaction, the spatial analysis of causality structures, the definition of leading/lagging sites, the construction of clusters of comoving sites, etc. This line of research is still active, and the same authors are presently working on local linear fitting for random fields.

1.3.4 Problems in multivariate time series

Testing independence or noncorrelation (depending on the asumptions made) between two multivatiate time series is a problem of fundamental practical importance. Haugh (1976) developed an approach to the problem of testing non-correlation (at all leads and lags) between two univariate time series. His tests however have low power for two series which are related over a long distributed lag when individual lag coefficients are relatively small. As a remedy, Koch and Yang (1986) proposed an alternative method that performs better than Haugh's under such dependencies. A multivariate extension of Haugh's procedure was proposed by El Himdi and Roy (1997), but suffers the same weaknesses as the original univariate method. In Hallin and Saidi (2002), an asymptotic test generalizing Koch and Yang's to the multivariate case is developed. The method includes El Himdi and Roy's as a special case. Based on the same idea, a generalization of the El Himdi and Roy procedure for testing causality in the sense of Granger (1969) between two multivariate series is also proposed. This line of research is pursued in Abdessamad Saidi's

doctoral dissertation under the supervision of Marc Hallin; a paper on optimal procedures in the Le Cam sense is under progress.

1.4 Work package 3: Survival Analysis

In this work package we focus on nonparametric estimation when dealing with complex censoring mechanisms, discontinuities and heterogeneity. There are many cross links with **WP1** where we studied more general methodological aspects of nonparametric regression, density and hazard estimation. Here the focus is more on a particular class of complex data structures, namely these encountered in survival analysis.

1.4.1 Nonparametric estimation with interval censored data

Regression models in which the response variable is subject to censoring by two types of censoring variables: an informative and a non-informative one, has been studied by Braekers and Veraverbeke (2002a). For this model with partially informative censoring, they established the validity of a bootstrap approximation. They also used profile likelihood methods to study Cox's regression model with data subject to partially informative censoring. See Braekers and Veraverbeke (2002b).

Li and Van Keilegom (2002) and Van Keilegom and Veraverbeke (2002) consider a nonparametric regression model, where the response Y is subject to right censoring and the covariate X is always observed. In this context Van Keilegom and Veraverbeke (2002) study the estimation of the hazard function of Y conditionally on X and they suppose that the vector (X, Y) follows a heteroscedastic model. In Li and Van Keilegom (2002) the authors construct confidence intervals and bands for the conditional distribution function of Y given X. They use an empirical likelihood approach, which has several important advantages: it always produces intervals in $[0, 1]$, it does not require the estimation of the variance and it might produce asymmetric intervals.

Members of the LUC team (P. Janssen and N. Veraverbeke), in collaboration with R. Cao (La Coruna, Spain) studied the efficiency of a presmoothed Kaplan-Meier and Nelson-Aalen estimator and the estimation of the relative hazard function with left truncated and right censored data.

Van Keilegom and Hettmansperger (2002) consider two random variables which are both subject to random right censoring, and they construct M-estimators for these two variables. A special case, which is studied in more detail, is the bivariate L_1 median.

As towards the problem of detecting abrupt changes in hazard functions when data are subject to censoring, Gijbels and Gürler (2001) considered a simple piecewise constant hazard model. They studied methods for estimating the location of the discontinuity, as well as the size. These methods have ingredients in common with the results discussed in **WP1** on detecting discontinuities in a regression function for non-censored data. This work has been revised recently.

1.4.2 Extending and improving inference for frailty models

An other field of activity is that of frailty modelling in survival analysis. Here a collaboration has been setup between the LUC-team, the KUL-2 team, and the UCL-team.

Frailty models provide a powerful tool to understand time-to-event data in different fields of applications. The applied domains covered are animal breeding studies (the survival of lambs in the sub-humid tropics) and treatment of outcome studies (understanding the heterogeneity in survival data from patients that receive the same treatment). The study of the use of frailty models in animal breeding is collaborative research with the International Livestock Research Institute (ILRI, Nairobi, Kenya); the treatment outcome research is in collaboration with the European Organisation for Research and Treatment of Cancer (EORTC) in Brussels.
At the moment there is collaboration on joint modelling of longitudinal data and time-to-event data (for animal breeding data, ILRI) and on the use of frailty models in prognostic index analysis (for clinical studies, EORTC). Also a number of challenging methodological topics need to be studied for frailty models. The current research topics we are dealing with are: the study of the asymptotic distributional behaviour of likelihood ratio tests (LRT) and score tests for the heterogeneity parameter in a shared frailty model and modelling recurrent event data.
It is interesting to note that the behaviour of the LRT for variance components is well studied within mixed models (the theme of **WP4**), but that the parallel theory for frailty models is not yet well understood. Recenlty a study started on the use of resampling techniques to estimate the standard error of the estimate of the heterogeneity parameter and a study on the use of splines to model the baseline hazard and/or the linear component of the loghazard in a more flexible way.

In a study regarding evaluation of kidney graft survival, Lambert, Collett, Kimber and Johnson (2002) model the random effects as a frailty component in an accelerated failure time model. They also briefly discuss the advantage of working under the framework of an accelerated failure time model instead of a proportional hazards model in this context.

1.5 Work package 4: Mixed Models

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

1.5.1 The implementation of multivariate random effects

The following topics were treated:

1. *Flexible distributions for the random effects part of a linear mixed model.*
   Members of the KUL-2 team (Ghidey and Lesaffre) have developed a mixed model with a smooth random effects distribution. The model is based on the assumption that the random effects distribution can be well approximated by a smooth function of B-splines. The estimation is done using the P-splines approach of Eilers and Marx (1996). In a second step the B-spline is replaced by an approximating normal density. The two approaches were presented at international conferences,

and the results will be reported on in a forthcoming manuscript. In the next step the normality assumption of the error distribution is replaced by a smooth distribution in conjunction with the smooth random effects distribution. This work is in preparation.

2. *Joint Modelling of Multivariate Longitudinal Profiles.*
   Random-effects models as a joint modelling approach have been critically investigated using a dataset on hearing tresholds measurements. A paper on these results is in preparation.

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

The following topics were treated:

1. *Allowing for examiner's bias and variability in a logistic random effects model.*
   Mwalili, Lesaffre and Declerck (2002) proposed a method to correct for the examiner's bias and variability when a gold standard and calibration data are available. Further, we currently investigate the necessary size of the calibration data set.

2. *Reformulation of item-response models as generalized linear mixed models and nonlinear mixed models.*
   A framework for the reformulation of item-response models as generalized linear mixed models and nonlinear mixed models is described by Rijmen, Tuerlinckx and De Boeck (2003), and an example is studied more in depth by Rijmen and De Boeck (2002). On the same topic a book to be published by Springer is planned, in collaboration with Mark Wilson (UC, Berkeley), and with Geert Molenberghs (**WP6**) and Geert Verbeke (WP4). Other members of WP4 (Bart Spiessens and Steffen Fieuws) are contributing a chapter devoted to mixture modelling. A SAS macro developed to allow for finite mixtures of random-effects models has been applied in the context of item-response theory. See Spiessens, Verbeke, Fieuws and Rijmen (2002). Using the macro on a dataset on complex reasoning problems, the hypothesis that there are three latent classes of reasoners has been tested.

3. *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 is an international workshop to be held in Leuven in July 2003.

4. *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:

1. *Modeling the emergence times using random effects models for interval censored data.*
   The Signal Tandmobiel Study is a longitudinal dental study on about 4500 children. The KUL-2 team (Komarek 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 the emergence of the first permanent molars. See Komarek, Lesaffre, Härkänen, Declerck and Virtanen (2002). The above analyses suggested examining a different approach, i.e. AFT-models (accelerated failure time models) with a complex random effects structure not necessarily assuming classical assumptions like normality. The KUL-2 team started in this respect, in collaboration with Hilton (UCSF), the development of an AFT model with a smooth random effects distribution. A manuscript is in preparation.

2. *Modeling 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 started recently the literature on this topic. The purpose is to develop a method for jointly modeling an interval censored response and continuous/discrete covariates. Further, the existence of latent classes in the data will also be examined. The latter is important for the modeling of HIV data. This research is also done in collaboration with Hilton (UCSF).

3. *Modeling multivariate interval-censored emergence times.*
   A GEE-method for modeling 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 non-parametric estimate of a survival function for interval-censored data has been proposed by Bogaerts and Lesaffre (2003). Under current investigation is a smooth bivariate estimate of the survival function. The approach of Eilers and Marx (1996) is employed for this.

1.6 Work package 5: Classification and mixture models

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

1.6.1 Studying specific types of mixture models

Extension of PMD models (mixture models with multiple binary latent variables and restrictions on the conditional and marginal probabilities).
The work was concentrated on two topics. Firstly we studied three-way models, with two levels, one for the observed variables and one for the latent binary variables. The paper by Meulders, De Boeck, Kuppens and Van Mechelen (2002) reporting on this was described in the editorial of the publishing journal as ".. builds a nice bridge between the world of latent class analysis and the deterministic worlds of overlapping clusters and three-way methods". Preceeding results can be found in Meulders, De Boeck, and Van Mechelen (2002). Secondly, we investigated assumptions regarding the sampling of the latent binary variables. See Meulders, De Boeck and Van Mechelen (2003).

Mixture Rasch models.
First, we started extending these models with a component of change, so that individuals can move from one component to the other. See Rijmen, De Boeck, and van der Maas (2002), which was part of the doctoral dissertation of Frank Rijmen (defended in December 2002). Second, we are studying the relation between mixture Rasch models and multidimensional two-parameter logistic models. See Rijmen and De Boeck (2002a). Third, we apply these models in our own research on individual differences in the domains of cognition and emotion in Rijmen and De Boeck (2002b) and in Tuerlinckx, De Boeck and Lens (2002). Related to applications of mixture models (but also mixed-effect models) to the domain of emotion, a small-scale international meeting (Leuven, November 1 to 2, 2002) has been organized with invited speakers from the US and Germany presenting models for the study of emotion data.

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

The work on this topic is to investigate Bayesian methods for model estimation and model selection. See Berkhof, Van Mechelen and Gelman (2002) and De Knop and Van Mechelen (2002). We have developed a collaboration with Andrew Gelman (Columbia University, New York) on Bayesian approaches in general

1.6.3 Classification techniques other than mixtures

First, an important line of work was the development of a family of three-way Boolean decomposition (hierarchical classes) models that parallels the family of three-way PCA models (Ceulemans and Van Mechelen (2002) and Ceulemans, Van Mechelen and Leenen (2003)). Furthermore, various constrained versions of these models have been developed, the constraints reflecting various kinds of substantive considerations. See Ceulemans and Van Mechelen (2002). Also uniqueness properties of the models (as well as of a generalization to N-way N-mode models) have been investigated by Ceulemans and Van Mechelen (2003b).

Second, the two-way hierarchical classes model for binary data (as well as the associated algorithm) have been extended to rating data by Ceulemans and Van Mechelen (2003a) and Van Mechelen, Lombardi and Ceulemans (2002).

Third, hierarchical classes analogues of other models than PCA have been developed, for example, of the discriminant analysis model in Lombardi, Ceulemans and Van Mechelen (2003) and Lombardi and Van Mechelen (2002). This topic is also part of the doctoral dissertation of Luigi Lombardi, who has prepared his dissertation (defended in January 2003 at the University of Padova) in the KUL-1 group.

Fourth, the RWTH-partner has investigated the analysis of a general type of clustering criteria which involves a convex function of the class means. Clustering algorithms were derived which use the concept of "maximum-support-plane partitions" as analogues to "minimum-distance partitions" in the classical theory. It appears that the results can be applied, in particular, for two-way clustering of a (large) contingency table such that the dependence between classes of rows and classes of columns is optimally reproduced (Bock, (2003)). A possible collaboration on this topic between the RWTH-partner, the KUL-1 team and Jean-Paul Rasson from the FUNDP, was discussed in a meeting in Aachen on May 24 of 2002, where also topics for collaboration in general have been discussed.

Fifth, we started the preparation of an invited review article on simultaneous two-mode clustering methods. This involves the network members Van Mechelen and De Boeck from the KUL-1 team and Bock from the RWTH-partner.

1.6.4 Specific cross-links with other work packages

Specific cross-links with **WP1** and **WP4** have been realized. We developed a penalized spline approach for the visualization of interactions (work by Bollaerts). Also methods for non-parametric density estimation are investigated by Beirlant, Berlinet, Biau and Vajda (2002) and Beirlant, Berlinet and Biau (2002).

A cross-link with **WP1** and **WP3** is the work by the RWTH-partner on copulas. The usage of copulas for multivariate has been investigated for characterizing time-dependence in stochastic processes (Markov processes, Brownian motions, diffusion processes) with a view to applications in finance mathematics. Volker Schmitz defended his PhD thesis on "Copulas and stochastic processes" at Aachen University. See also Schmitz (2003).

A further cross-link with **WP4** is related to "Mixed models for psychometrics" which is one of our main topics. This work is described in the report of **WP4** and **WP6**. The major achievement in this domain is the reformulation of item-response models as generalized linear mixed models and nonlinear mixed models (work by De Boeck and Wilson and Rijmen, Tuerlinckx and De Boeck (2003)). In collaboration between **WP4** and **WP5**, the SAS macro for NLMIXED developed within **WP4**, for estimating mixtures of random effects, has been investigated for its use in item-response modeling. See Spiessens, Fieuws and Rijmen (2002).

A strong cross-link is active with **WP6** (see also above) since we consider the random person effects as latent variables. Further, our work on PMD models is directly relevant to **WP6**, because of the latent binary variables that are included and that are treated as missing observations to estimate the models.

1.6.5 Methodological problems

Two methodological problems have been studied. First, we have made progress in the difficult issue of identification of the latent class model, using a Bayesian approach. We report on this in San Martin and De Boeck (2002). Second, we are investigating different methods to explore the heterogeneity of parameters in the logistic regression model. This is work in progress by Katalin Balazs and Istvan Hidegkuti. We started with the logistic regression model and hope to extend the approach to other models, for example also time-series models (cross-link with **WP2**).

1.7 Work package 6: Incompleteness and latent variables

1.7.1 Sensitivity analysis and missingness mechanism

To a large extent we have focused on fundamental methods for incomplete data, modelling incomplete data in practice, and incomplete data in clinical trials. There has been a large activity in the area of sensitivity analysis for incomplete data. On these topics, a number of papers have appeared (1) from the LUC research group, (2) in collaboration with partners from KUL-2 (G. Verbeke), and (3) in collaboration with international partners (including Harvard School of Public Health, London School of Hygiene and Epidemiology, Texas A&M University, Janssen Pharmaceutical Research and Development, Eli Lilly & Company).

Geert Molenberghs and Mike Kenward (London School of Hygiene and Tropical Medicine) are preparing a monograph on handling incomplete data in clinical studies. The book is expected to appear in 2004.

1.7.2 Evaluation of latent variables and mixed-effect models

The network members Paul De Boeck (**WP5**), Geert Verbeke (**WP4**), Geert Molenberghs (WP6), and colleagues within their teams, are working on an edited Springer-Verlag volume on the use of generalized linear mixed models in psychometric applications and testing theory. Editors are Paul De Boeck (KULeuven) and Mark Wilson (University of California at Berkeley). A planning meeting took place in September 2002, in Berkeley, California. The book is in the second draft stage and is expected to be finished in 2003.

Geert Verbeke (**WP4**) and Geert Molenberghs (WP6) are preparing a monograph on Discrete Repeated Measures, with Springer-Verlag. The book is due 2004 and follows on their 2000 Springer book on Linear Mixed Models for Longitudinal Data.

A lot of work has been done in the area of surrogate marker validation in clinical trials by the LUC partner. The methodology combines elements from **WP4** (mixed models) with WP6 (latent variables). The co-authors of the LUC team include national and international partners. A book project with Springer-Verlag, edited by Geert Molenberghs, Tomasz Burzykowski, and Marc Buyse, is under way.

Geert Verbeke (**WP4**) and Geert Molenberghs (WP6) collaborate on the PhD project of Ellen Andries (LUC, Institute for Material Physics), grantee of the IWT, on the use of mixed and mixture models in the context of material-physical reliability studies.

Geert Verbeke (**WP4**) and Geert Molenberghs (WP6) collaborate with Arnost Komarek (KULeuven, Biostatistical Centre) on implementation of mixed and mixture models in standard statistical software, such as SAS.

2 Network activities

2.1 Web Site

All activities of the IAP-statistics network can be followed very closely from our web site which was created in January 2002.
The address of the website is **http://www.stat.ucl.ac.be/IAP**
The following items can be found on our 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

A Technical Report Series and a Reprint Series have been created and are available via the website. 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 2002 we had a total of 47 papers written by members of the IAP-statistics network that were put into the IAP-statistics Technical Report Series. For 2002 there are 27 papers listed in the IAP-statistics Reprint Series.

2.3 Scientific Meetings

In 2002 the IAP-statistics network organized three meetings, listed below. Detailed reports on the objectives of these meetings, the programmes and the final event can be found on our web site (see item Reports). The organized meetings were:

- 22 February, 2002, UCL: one day meeting for IAP-statistics members;
- 21-23 May, 2002, UCL: International Workshop on "Statistical Modelling and Inference for Complex Data Structures", organized by UCL in collaboration with LUC (Noel Veraverbeke) and UJF (A. Antoniadis);
- 24 May, 2002, UCL: one day meeting for IAP-statistics members, open to participants of the preceeding International Workshop.

2.4 Organization of the network: Administrative meetings

A very first administrative meeting took place on April 18, 2001. The aim of this meeting was to prepare the research proposal for the network and to work out the organization of the network-to-be. For running the network the team of promotors, including the coordinator L. Simar and the scientific coordinator I. Gijbels met several times. At each of the scientific meetings above we had attached to it a brief administrative meeting. In addition to that we organized, in 2002, two purely administrative meetings at the UCL:

- 5 September, 2002: A meeting with all promotors and the administrative and scientific coordinators for discussing the candidates that applied for the research and post-doctoral positions that were announced (call for applications) early June 2002.

- 16 October, 2002: The annual administrative meeting with a representative from the federal office OSTC-Brussels (Ms Henry), members of the Follow-Up-Committee (Prof. A. Albert and Prof. T. Snijders), promotors of the network, the coordinators L. Simar and I. Gijbels and Ms D. Andre (head of the administration Institute of Statistics, UCL).

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 the collaborations between members of different teams of the network.

General methodology for nonparametric regression via regularization.
A collaboration has been setup between I. Gijbels (UCL) and A. Antoniadis (UJF) on one side and an external collaborator M. Nikolova (ENST Paris) to contribute into a development of a general methodology for nonparametric regression via regularization.

Frailty Models and Inference.
A very intensive collaboration has been started up 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) are preparing an invited review article on simultaneous two-mode clustering methods.

Estimating mixtures of random effects, item-response modelling and SAS-macro.
The teams of KUL-1 and KUL-2 have been collaborating on estimating mixtures of random effects. The KUL-2 team developed a SAS-macro which has been applied to item-response models encountered mainly by the KUL-1 team.

Generalized linear mixed models in psychometric applications and testing theory.
An intensive collaboration between the LUC-team and the teams from the KUL-1

and KUL-2 on the use of generalized linear mixed models in psychometric applications and testing theory is going on. An edited book with authors/editors from the various teams is in progress.

Discrete Repeated Measures.
Members from the LUC-team (Geert Molenberghs) and the KUL-2 team (Geert Verbeke) are continuing their collaboration on linear mixed models for longitudional data. A book on Discrete Repeated Measures is in progress.

Use of mixed and mixture models in reliability studies.
Geert Verbeke from KUL-2 and Geert Molenberghs from LUC collaborate on a doctoral research project of Ellen Andries (LUC, Institute for Material Physics) on the use of mixed and mixture models in the context of material-physical reliability studies.

Implementation of mixed and mixture models in standard statistical software, such as SAS.
Members from the KUL-2 team (Geert Verbeke and Armost Komarek) and from the LUC-team (Geert Molenberghs) are jointly working on implementation of mixed and mixture models in standard statistical software, such as SAS.

2.5.2 Working groups

Frailty Working Group:
A "Frailty Working Group" has been created under the initiative of the LUC and the KUL (group of Emmanuel Lesaffre) and with collaboration of the UCL. This working group met 8 times in 2002 at the LUC. The working group concentrates on research issues under Work Package 3.

2.5.3 Seminars

The "Frailty Working Group" has organized 8 seminars on the topic of "Frailty models and inference". All seminars took place at the LUC. Speakers and titles of the talks are provided below:

- 15 March 2002: Luc Duchateau (Univ. Gent), "Time evolution of recurrent event rate in frailty models".

- 27 March 2002: Paul Janssen (LUC), "Likelihood inference for multiplicative frailty models".

- 27 March 2002: José Cortiñas Abrahantes (LUC), "A version of EM algorithm for multivariate frailty models".

- 10 April 2002: Catherine Legrand (EORTC, Brussels), "Frailty models: confidence intervals for the heterogeneity parameter".

- 24 April 2002: Jim Lindsey (Univ. Liège and LUC). "Compartment models for pharmacokinetics and event histories".

- 21 October 2002: Vincent Ducrocq (INRA, Jouy-en-Josas, France), "Frailty models: Bayesian estimation and application to animal breeding problems".

- 4 November 2002: Florin Vaida (Harvard School of Public Health, USA), "Random effects models and the accelerated failure time paradigms".

Each of the participating partners organizes on a regular basis statistics seminars at their universities. Annoucements 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, two other seminars have been organized under the IAP-statistics network:

- 23 October 2002, LUC: Denis Belomestny (University of Bonn, Germany), "Statistical inferences based on transformed data: identifiability aspects";

- 27 November 2002, UCL: Farida Enikeeva (Moscow State University), "Asymptotically Minimax Estimation in the Wicksell Problem".

2.6 Short Courses and Graduate Schools

A short course on "Goodness-of-fit tests and survival analysis", given by Professor Ricardo-Cao (La Coruna, Spain) has been offered to IAP-members. This course took place in October 2002, and was could also be taken as a part of the doctoral program of the Graduate School in Statistics of the UCL.

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 2002, 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

Beguin, Cl. and L. Simar (2002). Analysis of the Expenses Linked to Hospital Stays: How to Detect Outliers? IAP-statistics Technical Report Series TR # 0213.

Chen, X., Linton, O. and Van Keilegom, I. (2002). Estimation of semiparametric models when the criterion function is not smooth. IAP-statistics Technical Report Series TR # 0201. *Under revision.*

Claeskens, G. and Van Keilegom, I. (2002). Bootstrap confidence bands for regression curves and their derivatives. IAP-statistics Technical Report Series TR # 0210. *Under revision.*

Delaigle, A. and Gijbels, I. (2001). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. Institut de Statistique, Université catholique de Louvain, *Discussion Paper* # 0116. *Under revision.*

Delouille, V. and von Sachs, R. (2002a). Properties of design-adapted wavelet transforms of nonlinear autoregression models. IAP-statistics Technical Report Series TR # 0217.

Delouille, V. and von Sachs, R. (2002b). Smooth design-adapted wavelets for half-regular designs in two dimensions. IAP-statistics Technical Report Series TR # 0218.

Denuit, M. and Lambert, P. (2002). Constraints on concordance measures in bivariate discrete data. IAP-statistics Technical Report Series TR # 0211.

Florens, J.P. and Simar, L. (2002). Parametric Approximations of Nonparametric Frontier. IAP-statistics Technical Report Series TR # 0214.

Fryzlewicz, P., Van Bellegem, S. and von Sachs, R. (2002). Forecasting non-stationary time series by wavelet process modelling. IAP-statistics Technical Report Series TR # 0204.

Gijbels, I. and Goderniaux, A.-C. (2000). Bandwidth selection for change point estimation in nonparametric regression. Institut de Statistique, Université catholique de Louvain, *Discussion Paper* # 0024. *Under revision.*

Gijbels, I. and Goderniaux, A.-C. (2001a). Data-driven discontinuity detection in derivatives of a regression function. Institut de Statistique, Université catholique de Louvain, *Discussion Paper* # 0130.

Gijbels, I. and Goderniaux, A.-C. (2001b). Bootstrap test for change points in nonparametric regression. Institut de Statistique, Université catholique de Louvain, *Discussion Paper* # 0144. *Under revision.*

Gijbels, I. and Gürler, Ü. (2001). Estimation in change point models for hazard function with censored data. Institut de Statistique, Université catholique de Louvain, *Discussion Paper* # 0114. *Under revision.*

Gijbels, I. and Heckman, N. (2000). Nonparametric testing for a monotone hazard function via normalized spacings. Institut de Statistique, Université catholique de Louvain, *Discussion Paper* # 0028. *Under revision.*

Hall, P. and Van Keilegom, I. (2002). Testing for monotone increasing hazard rate. IAP-statistics Technical Report Series TR # 0240.

Lambert, A. (2002). Automatic jump detection in regression surface. *Mémoire de DEA (Master Thesis)*, Institut de Statistique, Université catholique de Louvain, Louvain-la-Neuve, Belgium.

Lambert, P., Collett, D., Kimber, A. and Johnson, R. (2002). Evaluation of kidney graft survival using accelerated failure time models with random effects. IAP-statistics Technical Report Series TR # 0215.

Oulhaj, A. and Mouchart, M. (2002). The role of the exogenous randomness in the identification of conditional models. IAP-statistics Technical Report Series TR # 0202.

Van Bellegem, S. and von Sachs, R. (2002). Forecasting economic time series using models of nonstationarity. IAP-statistics Technical Report Series TR # 0219.

Vandenhende, F. Lambert, P. and Ramadan, N. (2002). Statistical models for the analysis of controlled trials on acute migraine. IAP-statistics Technical Report Series TR # 0241.

Wang, L., Akritas, M.G. and Van Keilegom, I. (2002). Nonparametric goodness-of-fit test for heteroscedastic regression models. IAP-statistics Technical Report Series TR # 0244.

B. List of Publications

Cazals, C., Florens, J.-P. and Simar,L. (2002). Nonparametric Frontier Estimation: a Robust Approach. *Journal of Econometrics*, **106**, 1–25.

Climov, D., Delecroix, M. and Simar, L. (2002). Semiparametric Estimation in Single index Poisson Regression: a practical approach. *Journal of Applied Statistics*, **29**, 1047–1070.

Climov, D., Hart, J. and Simar, L. (2002). Automatic Smoothing and Estimation in Single Index Poisson Regression. *Journal of Nonparametric Statistics*, **14**, 307–323.

Delaigle, A. (2003). *Kernel Estimation in Deconvolution problems.* PhD Dissertation. Université catholique de Louvain.

Delaigle, A. and Gijbels, I. (2002a). Estimation of integrated squared density derivatives from a contaminated sample. *Journal of the Royal Statistical Society*, Series B, **64**, 869–886.

Delaigle, A. and Gijbels, I. (2002b). Practical bandwidth selection in deconvolution kernel density estimation. IAP-statistics Technical Report Series TR # 0212 (under title "Comparison of data-driven bandwidth selection procedures in deconvolution kernel density estimation"). *Computational Statistics and Data Analysis*, to appear.

Hall, P. and Simar, L. (2002). Estimating a Changepoint, Boundary or Frontier in the Presence of Observation Error, *Journal of the American Statistical Association*, **97**, 523–534.

Hall, P. and Van Keilegom, I. (2003). Using difference-based methods for inference in nonparametric regression with time-series errors. IAP-statistics Technical Report Series TR # 0203. *Journal of the Royal Statistical Society - Series B*, to appear.

Lambert, P. (2002). A mixture model to assess the effect of hormonal stimulation on the development of follicles in prepubertal heifers. *Journal of the Royal Statistical Society, Series C*, **51**, 405–420.

Lambert, P. and Vandenhende, F. (2002). A copula-based model for multivariate non-normal longitudinal data : analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine*, **21**, 3197–3217.

Li, G. and Van Keilegom, I. (2002). Likelihood ratio confidence bands in non-parametric regression with censored data. *Scandinavian Journal of Statistics*, **29**, 547–562.

Mouchart, M. and Rolin, J.-M. (2002). Competing risks models: problems of modelling and of identification. *Life Tables, Modelling Survival and Death*, edited by G. Wunsch, M. Mouchart and J. Duchêne, Dordrecht: Kluwer Academic Publishers, 245–267, 2002.

Ombao, H., Raz, J., von Sachs, R. et Guo, W. (2002). The SLEX model of a non-stationary random process. *Annals of the Institute of Statistical Mathematics*, **54**, 171–200.

Park, B., Sickles, R. and Simar, L. (2003). Semiparametric Efficient Estimation of AR(1) Panel Data Models. *Journal of Econometrics*, to appear.

Simar, L. and Wilson, P. (2002). Nonparametric Test of Return to Scale, *European Journal of Operational Research*, **139**, 115–132.

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. IAP-statistics Technical Report Series TR # 0216. *Genetical Research*, to appear.

Vandenhende, F. and Lambert, P. (2002). On the joint analysis of longitudinal responses and early discontinuation in rendomized trials. *Journal of Biopharmaceuticals Statistics*, **12**, 425–440.

Van Keilegom, I., Hettmansperger, T.P. (2002). Inference on multivariate M-estimators based on bivariate censored data. *Journal of the American Statistical Association*, **97**, 328–336.

Zhang, J. and Gijbels, I. (2003). Sieve Empirical Likelihood and Extensions of the Generalized Least Squares. *Scandinavian Journal of Statistics*, to appear.

3.1.2 Katholieke Universiteit Leuven, KUL-1 partner

A.List of Technical Reports

Beirlant, J., Berlinet, A., and Biau, G. (2002). Higher order estimation at Lebesgue points with applications to density estimation. Technical report L.S.T.A., Univ. Paris VI (2002).

Ceulemans, E., and Van Mechelen, I. (2002). Tucker2 hierarchical classes analysis. *Psychometrika, under revision.*

Lombardi L., and Van Mechelen, I. (2002). Conjunctive prediction of an ordinal criterion on the basis of binary predictors. *Discrete Applied Mathematics, under revision.*

Rijmen, F. and De Boeck, P. (2002a). The relation between a between-item multidimensional model and the mixture-Rasch model.

Rijmen, F., De Boeck, P. and van der Maas. (2002). An IRT model with a parameter-driven process for change.

San Martin, E., and De Boeck, P. (2002). On the identifiability and estimability of latent class models: A Bayesian analysis.

Tuerlinckx, F., and De Boeck, P. (2002). A note on the interpretation of the discrimination parameter.

Van Mechelen, I., Lombardi, L., and Ceulemans, E. (2002). HICLAS-R: Hierarchical classes models for rating data.

Van den Noortgate, W., De Boeck, P., Janssen, R. and Meulders, M. (2002). Cross-classification multilevel logistic models in psychometrics.

B.List of Publications

Beirlant, J., Berlinet, A., Biau, G., Vajda, I. (2002). Divergence-type errors of smooth Barron-type density estimators. *Test*, **11**, 191–217.

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

Ceulemans, E., and Van Mechelen, I. (2003a). An algorithm for HICLAS-R models. *Proceedings of the 26th Annual Conference of the Gesellschaft fr Klassifikation*, to appear.

Ceulemans, E., and Van Mechelen, I. (2003b). Uniqueness of N-way N-mode hierarchical classes models. *Journal of Mathematical Psychology*, to appear.

Ceulemans, E., Van Mechelen, I., and Leenen, I. (2003). Tucker3 hierarchical classes analysis. *Psychometrika*, to appear.

Lombardi, L., Ceulemans, E., and Van Mechelen, I. (2003). A hierarchical classes approach to discriminant analysis. *Proceedings of the 26th Annual Conference of the Gesellschaft für Klassifikation*, to appear.

Meulders, M., De Boeck, P., and Van Mechelen, I. (2003). A taxonomy of latent structure assumptions for probability matrix decomposition. *Psychometrika*, to appear.

Meulders, M., De Boeck, P., Kuppens, P., and Van Mechelen, I. (2002). Constrained latent class analysis of three-way three-mode data. *Journal of Classification*, **19**, 277–302.

Meulders, M., De Boeck, P., and Van Mechelen, I. (2002). Rater classification on the basis of latent features in responding to situations. In W. Gaul, & G. Ritter (Eds.), *Classification, Automation, and New Media*. Proceedings of the 24-th Annual Conference of the Gesellschaft für Klassifikation, University of Passau (pp. 453–461). Berlin: Springer-Verlag.

Ratcliff, R., and Tuerlinckx, F. (2002). Estimating the parameters of the diffusion model. *Psychonomic Bulletin & Review*, **9**, 438–481.

Rijmen, F., and De Boeck, P. (2002b). The random weights linear logistic test model. *Applied Psychological Measurement*, **26**, 271–285.

Rijmen, F., and De Boeck, P. (2003). Reasoning correlates of individual differences in the interpretation of conditionals. *Psychological Research*, to appear.

Rijmen, F., Tuerlinckx, F., and De Boeck, P. (2003). A nonlinear mixed model framework for IRT models. *Psychological Methods*, to appear.

Tuerlinckx, F., De Boeck, P, and Lens, W. (2002). Measuring needs with the Thematic Apperception Test: A psychometric study. *Journal of Personality and Social Psychology*, **82**, 448–461.

3.1.3 Katholieke Universiteit Leuven, KUL-2 partner

A.List of Technical Reports

Komarek, A., Lesaffre, E., Harkanen, T., Declerck, D., and Virtanen, J. (2002). A Bayesian analysis of multivariate doubly interval censored dental data.

Mwalili S., Lesaffre, E. and Declerck, D. (2002). Correcting for inter-observer effects in a geographical oral health study.

B.List of Publications

Bogaerts, K., Leroy, R., Declerck, D. and Lesaffre, E. (2002), Modeling 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 for Computational and Graphical Statistics*, to appear.

Ghidey W., Lesaffre E., Eilers P. and Verbeke, G. (2002). P-spline smoothing for random effects distribution estimation. Published at conference proceeding of the 17th International Workshop on Statistical Modelling.

3.1.4 Limburgs Universitair Centrum, LUC partner

A.List of Technical Reports

Braekers, R. and Veraverbeke, N. (2002a). Bootstrapping the conditional survival estimator in the partial Koziol-Green model. IAP-statistics Technical Report Series TR # 0209.

Braekers, R. and Veraverbeke, N. (2002b). Cox's regression model under partially informative censoring. IAP-statistics Technical Report Series TR # 0245.

De Ridder, J., Molenberghs, G. and Aerts, M. (2002). Revisiting the moment method as a mode identification technique. IAP-statistics Technical Report Series TR # 0239.

Faes, C., Geys, H., Aerts, M. and Molenberghs, G. (2002). On the use of fractional polynomial predictors for quantitative risk assessment in developmental toxicity studies. IAP-statistics Technical Report Series TR # 0223.

Faes, C., Geys, H., Aerts, M., Molenberghs, G. and Catalano, P. (2002). Modelling combined continuous and ordinal outcomes in a clustered setting. IAP-statistics Technical Report Series TR # 0232.

Jansen,I., Van Steen, K., Molenberghs, G., De Wit, M. and Peeters, M. (2002). A similarity measure and test between two DNA sequences based on mahalanobis distance between word frequencies. IAP-statistics Technical Report Series TR # 0236.

Laenen, A., Geys, H., Vangeneugden, T. and Molenberghs, G. (2002). Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements. IAP-statistics Technical Report Series TR # 0225.

Lipsitz, S.R., Molenberghs, G., Fitzmaurice, G.M. and Ibrahim, J.G. (2002). A protective estimator for linear regression with nonignorably missing Gaussian outcomes. IAP-statistics Technical Report Series TR # 0230.

Molenberghs, G., Thijs, H., Carroll, R.J. and Kenward, M.G. (2002). Analyzing incomplete longitudinal clinical trial data. IAP-statistics Technical Report Series TR # 0229.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P. and Van Damme, P. (2002). Modelling age dependent force of infection from prevalence data using fractional polynomials. IAP-statistics Technical Report Series TR # 0233.

Tibaldi, F., Molenberghs, G., Burzykowski, T. and Geys, H. (2002). Pseudo-likelihood estimation for a marginal multivariate survival model. IAP-statistics Technical Report Series TR # 0228.

Tibaldi, F., Torres Barbosa, F. and Molenberghs, G. (2002). Modelling associations between time-to-event responses in pilot cancer clinical trials using a Plackett-Dale model. IAP-statistics Technical Report Series TR # 0238.

Torres, F., Tibaldi, F.S., Cortiñas, J., Geys, H., González, G. and Molenberghs, G. (2002). Surrogacy evaluation of immunological parameters in pilot cancer clinical trials. IAP-statistics Technical Report Series TR # 0226.

Van Steen, K., Molenberghs, G., De Wit, M. and Peeters, M. (2002). Comparing DNA sequences using generalized estimating equations and pseudo-likelihood. IAP-statistics Technical Report Series TR # 0224.

Wouters, L., Göhlmann, H.W., Bijnens, L., Kass, S.U., Molenberghs, G. and Lewi, P.J. (2002). Graphical exploration of gene expresion data: a comparative study of three multivariate methods. IAP-statistics Technical Report Series TR # 0237.

B. List of Publications

Aerts, M., Claeskens, G., Hens, N., and Molenberghs, G. (2002). Local multiple imputation. *Biometrika*, **89**, 375–388.

Aerts, M., Claeskens, G. and Wand, M.P. (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference*, **103**, 455–470.

Alonso, A., Geys, H., Kenward, M.G., Molenberghs, G., Vangeneugden, T. (2002). Validation of surrogate markers in multiple randomized clinical trials with repeated measures. IAP-statistics Technical Report Series TR # 0234. *Proceedings of the 17th International Workshop on Statistical Modelling*, Crete, Stasinopoulos, M. (ed), 95–104.

Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T. (2002). Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. IAP-statistics Technical Report Series TR # 0220. *Journal of Biopharmaceutical Statistics*, to appear.

Arbyn, M., Van Oyen, H., Sartor, F., Tibaldi, F., and Molenberghs, G. (2002). Description of the influence of age, period and cohor effects on cervical cancer mortality by loglinear Poission models (Belgium, 1955–1994). *Archives of Public Health*, **60**, 73–100.

Beutels, P., Shkedy, Z., Mukomolov, S., Aerts, M., Shargorodskaya, E., Plotnikova, V., Molenberghs, G., and Van Damme, P. (2002). Hepatitis B in Saint Petersburg, Russia (1994–1999): a descriptive epidemiological analysis. *Journal of Viral Hepatitis*, to appear.

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.

Faes, C., Geys, H., Aerts, M., Catalano, P., and Molenberghs, G. (2002). Modelling combined continuous and ordinal outcomes from developmental toxicity studies. *Proceedings of the 17th International Workshop on Statistical Modelling*, Crete, Stasinopoulos, M. (ed), 247–254.

Geys, H., Molenberghs, G., and Williams, P. (2002). Analysis of clustered binary data with covariates specific to each observation. *Journal of Agricultural, Biological, and Environmental Statistics*, **7**, 1–15.

Hens, N., Bruckers, L., Arbyn, M., Aerts, M., and Molenberghs, G. (2002). Classification trees and its application to cervix cancer screening in the Belgian Health Interview Survey 1997. *Archives of Public Health*, **60**, 275–294.

Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., and Van Steen, K. (2002). A Local Influence Approach applied to Binary Data from a Psychiatric Study. *Biometrics*, to appear.

Janssen, P., Swanepoel, J. and Veraverbeke, N. (2002). The modified bootstrap error process for Kaplan-Meier quantiles. *Statistics and Probability Letters*, **58**, 31–39.

Kenward, M.G., Molenberghs, G., and Thijs, H. (2002). Pattern-mixture models with proper time dependence. *Biometrika*, to appear.

Mallinckrodt, C.H., Carroll, R.J., Debrota, D.J., Dube, S., Molenberghs, G., Potter, W.Z., Sanger, T.D., and Tollefson, G.D. (2002). Assessing and interpreting treatment effects in longitudinal clinical trials with subject dropout. IAP-statistics Technical Report Series TR # 0222. *Biological Psychiatry*, to appear.

Mallinckrodt, C.H., Scott Clark, W., Carroll, R.J., and Molenberghs, G. (2002). Response profiles for longitudinal clinical trial data with subject dropout under regulatory considerations. IAP-statistics Technical Report Series TR # 0221. *Journal of Biopharmaceutical Statistics*, to appear.

Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T., and Thijs, H. (2002). Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout. *Statistics in Medicine*, **21**, 1023–1041.

Molenberghs, G., Buyse, M., Geys, H., Renard, D., and Burzykowski, T. (2002). Challenges in the methodology for the validation of surrogate endpoints in randomized trials. *Proceedings of the 17th International Workshop on Statistical Modelling*, Crete, Stasinopoulos, M. (ed), 475–482.

Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykowski, T., and Alonso, A. (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials*, **23**, 607–625.

Molenberghs, G., Kenward, M.G., and Goetghebeur, E. (2002). "Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case," in: *2002 Proceedings of the Biopharmaceutical Section (American Statistical Association)*, to appear.

Molenberghs, G., Williams, P.L. and Lipsitz, S.R. (2002). Prediction of survival and opportunistic infections in HIV infected patients: a comparison of imputation methods of incomplete CD4 counts. *Statistics in Medicine*, 21, 1387–1408.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., and Buyse, M. (2002). Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical Journal*, **44**, 1–15.

Renard, D., Molenberghs, G., and Geys, H. (2002) A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis*, to appear.

Speybroeck, N., Boelaert F., Renard, D., Burzykowski T., Mintiens, K., Molenberghs, G., and Berkvens D.L. (2002). Important statistical issues in a design-based analysis of surveys: a Bovine Herpesvirus 1 case study. *Epidemiology and Infection*, to appear.

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

Van Steen, K., Curran, D., Kramer, J., Molenberghs, G., Van Vreckem, A., and Sylvester, R. (2002). Multicollinearity in prognostic factor analysis using the EORTC QLQ-C30: identification and impact on model selection. *Statistics in Medicine*, to appear.

3.1.5 Université Libre de Bruxelles, ULB partner

A.List of Technical Reports

Forni, M, Hallin, M., Lippi, M. and Reichlin, L. (2002). The generalized dynamic factor model: one-sided estimation and forecasting. IAP-statistics Technical Report Series TR # 0205.

Hallin, M., Lu, Z., and L.T. Tran. (2002). Kernel density estimation for spatial processes: the L_1 theory. IAP-statistics Technical Report Series TR # 0207.

Hallin, M. and D. Paindaveine. (2002). Rank-based optimal tests of the adequacy of an elliptic VARMA model. IAP-statistics Technical Report Series TR # 0208.

Hallin, M., Lu, Z., and L. T. Tran. (2002). Local linear spatial regression. IAP-statistics Technical Report Series TR # 0242.

Hallin, M. and D. Paindaveine. (2002). Asymptotic linearity of serial and nonserial multivariate signed rank statistics. IAP-statistics Technical Report Series TR # 0243.

Hallin, M. and D. Paindaveine. (2002). Affine-invariant aligned rank tests for the multivariate general linear model with ARMA errors. IAP-statistics Technical Report Series TR # 0247.

Giannone, D., L. Reichlin, and L. Sala. (2002). Tracking Greenspan: systematic and unsystematic monetary policy revisited. CEPR working paper.

Giannone, D., L. Reichlin, and L. Sala. (2002). VARs, common factors and the empirical validation of equilibrium business cycle models. CEPR working paper.

B. List of Publications

Akharif, A. and M. Hallin (2002). Efficient detection of random coefficients in $AR(p)$ models. *Annals of Statistics*, to appear.

Allal, J., A. Kaaouachi, and D. Paindaveine (2002). R-estimation for ARMA models. *Journal of Nonparametric Statistics* **13**, 815–831.

De Mol, C. and M. Defrise (2002). A note on wavelet-based inversion algorithms. In M. Nashed and O. Scherzer, Eds, *Inverse Problems, Image Analysis, and Medical Imaging*. Providence, R.I.: American Mathematical Society, to appear.

El Bantli, F. and M. Hallin (2002a). Estimation in autoregressive models based on autoregression rank scores. *Journal of Nonparametric Statistics*, **13**, 667-697.

El Bantli, F. and M. Hallin (2002b). Estimation of the innovation quantile density function of an $AR(p)$ process, based on autoregression quantiles. *Bernoulli*, **8**, 255–274.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2002a). The generalized dynamic factor model: consistency and rates. *Journal of Econometrics*, to appear.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2002b). Do financial variables help forecasting output and inflation in the EURO area ? IAP-statistics Technical Report Series TR # 0206. *Journal of Monetary Economics*, to appear.

Hallin, M. (2002). Chernoff-Savage Theorems, Contiguity, Differentiability in Quadratic Mean, Hoeffding's U Statistics, Lebesgue Decomposition, Le Cam's First Lemma, Le Cam's third Lemma, Local Asymptotic Mixed Normality, Local Asymptotic Normality, o_P and O_P Notation, Rank Autocorrelation Coefficients, Serial Rank Statistics, U Statistics. In *A Dictionary of Statistical Terms* (sixth edition). Harlow, U.K.: Longman.

Hallin, M., Z. Lu, and L.T. Tran (2002). Kernel density estimation for spatial processes: the L_1 theory. *Journal of Multivariate Analysis*, to appear.

Hallin, M. and D. Paindaveine (2002a). Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *Annals of Statistics*, **30**, 1103–1133.

Hallin, M. and D. Paindaveine (2002b). Multivariate signed ranks: Randles' interdirections or Tyler's angles? In Y. Dodge, Ed., *Statistical Data Analysis Based on the L_1 norm and related methods*, 271–282. Basel: Birkhäuser.

Hallin, M. and D. Paindaveine (2002c). Optimal procedures based on interdirections and pseudo-Mahalanobis ranks for testing multivariate elliptic white noise against ARMA dependence. *Bernoulli*, **8**, 787–815.

Hallin, M. and D. Paindaveine (2002d). Affine invariant linear hypotheses for the multivariate general linear model with ARMA error terms. In *Mathematical Statistics and Applications: Festschrift for Constance van Eeden*, IMS Monograph-Lecture Note Series, to appear.

Hallin, M. and A. Saidi (2002). Testing non-correlation between two multivariate ARMA time series. *Journal of Time Series Analysis*, to appear.

Hallin, M. and B. Werker (2002). Semiparametric efficiency, distribution-freeness, and invariance. *Bernoulli*, to appear.

Reichlin, L. (2002). Extracting business cycle indexes from large data sets: aggregation, estimation, identification. In M. Dewatripont, L. Hansen, and S. Turnovsky, Eds, *Advances in Economics and Econometrics: Theory and Applications*. Cambridge: Cambridge University Press, to appear.

3.1.6 Aachen Technical University, RWTH partner

A.List of Technical Reports

B.List of Publications

Bock, H.-H. (2002). Clustering methods: from classical models to new approaches. *Statistics in Transition*, **5**, 725–758.

Bock, H.-H. (2002). Convexity-based clustering criteria: a new approach. Cracow University of Economics, *Rector's Lectures* **5**, 1–14.

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

Schmitz, V. (2003). Revealing the dependence structure between $X_{(1)}$ and $X_{(n)}$. *Journal of Statistical Planning and Inference*, to appear.

3.1.7 Université Joseph Fourier, UJF-LMC-IMAG partner

A.List of Technical Reports

Peyre, J. (2001). A comparative study of Normalization methods for cDNA microarray data. Mémoire de DEA.

Bigot, J. (2002). A scale-space approach with wavelets to landmark detection. IAP-statistics Technical Report Series TR # 0246.

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**, to appear. (see http: //www-lmc.imag.fr/SMS/software/microarrays/ for freely available software that implements the procedures of this paper).

3.2 List of joint publications

A.List of joint Technical Reports

Komarek, A., Verbeke, G. and Molenberghs, G. (2002). An approximate approach to fit a linear mixed model with a finite normal mixture as random-effects distribution and its SAS implementation. IAP-statistics Technical Report Series TR # 0227. (KUL-2 & LUC)

Spiessens, B., Verbeke, G., Fieuws, S. and Rijmen, F. (2002). Classification of clustered data using a SAS-macro: an application to latent class models. (KUL-1 & KUL-2)

Verbeke, G. and Molenberghs, G. (2002). The use of score tests for inference on variance components. IAP-statistics Technical Report Series TR # 0231. (KUL-2 & LUC)

B.List of joint Publications

Antoniadis, A. and Gijbels, I. (2002). Detecting abrupt changes by wavelet methods. *Journal of Nonparametric Statistics*, **14**, 7–29. (UCL & UJF)

Hens, N., Aerts, M., Molenberghs, G., Thijs, H., and Verbeke, G. (2002). Kernel weighted influence. IAP-statistics Technical Report Series TR # 0235. *Proceedings of the 17th International Workshop on Statistical Modelling*, Crete, Stasinopoulos, M. (ed), 329–333. (KUL-2 & LUC)

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics*, **3**, 245–265. (KUL-2 & LUC)

Molenberghs, G., Renard, D., and Verbeke, G. (2002). A review of generalized linear mixed models. *Journal de la Société Française de Statistique*, **143**, 53–78. (KUL-2 & LUC)

Molenberghs, G., Thijs, H., Kenward, M.G., and Verbeke, G. (2002). Sensitivity analysis for continuous incomplete longitudinal outcomes. *Statistica Neerlandica*, to appear. (KUL-2 & LUC)

Molenberghs, G., Thijs, H., Kenward, M.G., and Verbeke, G. (2002). How meaningful and sensitive are selection models and pattern-mixture models? In: *Proceedings of the XXIth International Biometric Conference*, Invited Papers, pp. 295–315. (KUL-2 & LUC)

Van Keilegom, I. and Veraverbeke, N. (2002). Density and hazard estimation in censored regression models. *Bernoulli*, **8**, 607–625. (LUC & UCL)