Scientific highlights of the network

- *Incomplete data and sensitivity analysis*
  WP3 has contributed to number of high-profile references. The book by [1] promotes the proper use of methodology for incomplete data in the context of clinical studies and, more broadly, in settings with life sciences data. It synthesizes a large amount of work undertaken by the members of WP3, as well as methods from other international schools. It ranges from critiques on simple methods, via the promotion of broadly valid yet easy-to-use state of the art methods, to finally sensitivity analysis tools. In the book by [2], several chapters, written by world renowned experts, are devoted to incomplete data and sensitivity analysis. [3] devote a lot of attention in their book to missing data issues in clinical trials. [4] wrote a monograph sponsored partly by the UK National Health Service Coordinating Committee for Research Methodology. This monograph develops and expounds methods for the analysis of longitudinal clinical trial data with dropout. It covers methods for continuous and discrete data, including sensitivity analysis. [5] is a handbook on non-response in longitudinal studies. In a related fashion, several contributions have been made to encyclopedias and other reference works, such as by [6], who provide a general overview of the missing-data field. [7] contributed a chapter to the Handbook of Epidemiology on incomplete data. Also, discussion contributions have been produced, including by [8], at the occasion of a paper by Peter Diggle (Lancaster University) and colleagues, which discusses existing missing-data methodology and offer further classification schemes. To add to these efforts [9] wrote an editorial to the Journal of the Royal Statistical Society, Series A, dedicated to the state-ofthe-art and future directions of research in incomplete data. [10] wrote an invited review for the Japanese statistical journal. Similarly, [11] and [12] wrote an invited contribution about incomplete data in clinical trials, with discussion. Conversely, [13] wrote a discussion contribution to a work by Lee and Nelder on unobservables, including missing data. [14] prepared an invited review about analysis and sensitivity analysis of incomplete data. [15] discuss the Points to Consider Document on Missing Data, as adopted by the Committee of Health and Medicinal Products (CHMP). In 2007, the CHMP issued a recommendation to review the document, with particular emphasis on summarizing and critically appraising the pattern of drop-outs, explaining the role and limitations of the last observation carried forward method relative to the use of mixed models. The initiative was taken by a working party of the organization, Statisticians in the Pharmaceutical Industry. [16] offer a tutorial on handling incomplete data in the context of survival analysis.

- *Frailty models*
  The research undertaken by the frailty model workgroup, with contributions of most of the partners of this IAP network, has led to the development of new techniques (with papers in, for instance, Biometrics, JABES, JSPI, Am. Statistician, and JRSS C), which makes that the group is considered to be one of the major research groups on frailty models internationally. This is emphasized further by the publication of the standard work on frailty models, entitled ‘The frailty model’, published by Springer Verlag, and authored by L. Duchateau and P. Janssen.

- *Modeling dependencies through copulas*
  Regarding modeling dependencies via copulas there is a large and very important output in the network. Fundamental theoretical contributions were obtained by several members of the network, including members from KUL-1, UHasselt and UCL, and published in top journals. An example is the joint paper by members I. Gijbels (KUL-1), N. Veraverbeke (UH) and M. Omelka (postdoc IAP, KUL-1 and UH) published in The Annals of Statistics. Copula functions are also used in many application areas covered by the network. The research on copulas is on the forefront of research in that field worldwide.

- *Flexible modeling and regularization techniques.*
  When analyzing complex data, flexible modeling is a key issue. An interesting method that allows for such a flexible modeling are P-splines. Several members of the network (e.g. KUL-1, UJF, Utrecht) have made seminal contributions to the research area, including for example data-driven choices of the penalty function, flexible modeling of mean and variance/dispersion functions, and variable selection methods. The methods developed on estimation of mean and dispersion function, in an extended generalized linear models setup are among the first to be able to deal with over dispersed as well as under dispersed data, and even crossovers between the two situations. The research currently carried out on variable selection and additive modeling opens very interesting new research perspectives.

- *Frontier estimation*
  The use of nonparametric frontier models remains a hot research topic in efficiency and productivity analysis because they rely on very few behavioral assumptions. Statistical inference covers now most of the underlying economic models (assumptions on returns to scale, etc.) and the theory behind the bootstrap is now translated in efficient and easy to implement algorithms. Extensions have been proposed which allow to handle noise on the data, to condition on environmental variables and to be robust to extremes and outliers. The links with extreme value theory allowed to extend previous asymptotic results on nonparametric frontier estimators but also provided new estimators of the frontier function having asymptotic normal distributions. The asymptotics for conditional efficiency scores has also been provided, with appropriate bandwidth selection procedures. Many papers were published in the best journals ranging from theoretical statistics, theoretical econometrics but also to more applied papers in the field (Journal of Econometrics (5), Annals of Statistics (1), Bernoulli (1), Journal of Statistical Planning and Inference (1), Econometric Theory (2), Econometric Review (2), Journal of Applied Econometrics (1), Journal of Productivity Analysis (5), Annals of Operations Research (1) and European Journal of Operational Research (1)). Some of these methods were also presented in a book published at Springer, New-York (Daraio and Simar, 2007). L. Simar was also invited, as Quality Manager, to join an European project (EUMIDA) to build micro-data on the activities of European Universities. The idea will be in a second phase to implement these methods to facilitate the relative positioning of the European universities.

- *Non- and semiparametric regression with censored data*
  Lots of research has been done by the UCL and UH partner in the context of semi- and nonparametric regression with censored data. This is considered as an important research area, as in many practical situations the classical Cox and accelerated failure type models are too restrictive. In particular, research has been carried out in the context of generalized conditional linear models with censored data, nonlinear regression with censored data, semiparametric transformation models, nonparametric regression with dependent censored data, goodness-of-fit tests for parametric and semiparametric models in multiresponse regression, single index regression models in the presence of censoring, etc.

- *Inference for semiparametric Z-estimators*
  The UCL partner has made lots of progress on inference and asymptotics for semiparametric estimation problems, in particular for estimators that can be written as approximate Z-estimators. The UCL has also collaborated with the USC and UH partner on this type of problems. The methodology for this type of estimators has been applied to a number of novel research ideas, coming from a broad range of areas in statistics (copula estimation in nonparametric regression, semiparametric regression with missing data, semiparametric transformation models, comparison of semiparametric backfitting and profiling procedures, semiparametric location-scale regression models, ...). The asymptotics for this type of estimators is often highly complex, especially when the criterion function is non smooth in the nonparametric component of the model.

- *Nonparametric location-scale models*
  The study of nonparametric location-scale models remains an important research area in the network, especially for the UCL, USC and UH partners. During Phase VI of the network, a wide range of testing problems has been studied in this model, like (among others) goodnessof-fit tests for the mean and variance function based on the comparison of two estimators of the error distribution (one obtained under the null hypothesis and one under the alternative), tests for the independence between the error and the covariate in this model, change-point tests and the comparison of regression curves. Moreover, the nonparametric location-scale models have also been studied in the presence of censored successive survival times and when the response variable is subject to selection bias. Finally, the estimation of conditional ROC curves has been investigated, copulas have been used to model the (possible) relation between the error and the covariate in the model, and the important extension to multiple regression has been developed. The latter result offers a wide range of possibilities for further research, since so far all results developed in the literature were restricted to one-dimensional covariates, the case of more-dimensional covariates being in fact substantially harder to analyze theoretically.

- *Robust model selection*
  Large datasets are often of unequal quality which leads to outliers and other data anomalies. Building reliable models from such data sets is a challenging task of great importance. The problem of robust model selection appears in all multivariate analysis settings, such as regression, classification and clustering. The selected models have to be stable on the one hand and obtained in a reasonable amount of time on the other hand. The UG team (S. Van Aelst) together with international collaborators have developed and studied several robust, computationally efficient procedures to select stable models in regression and clustering contexts. This research has led to publications in high quality journals such as the Journal of the American Statistical Association, the Journal of the Royal Statistical Society-Series B, and Computational Statistics and Data Analysis.

- *Robustness in high-dimensional data*
  The standard robustness model of regular (clean) and contaminated observations becomes less realistic and useful in high-dimensional data settings. In such settings there are too few observations to discard whole observations when only few of their components are contaminated. On the other hand, one can not assume that there is a majority of completely regular observations without anomaly in any of its components. The UG team (S. Van Aelst) together with international collaborators have proposed new contamination models for highdimensional data and shown the lack of robustness of the current high-breakdown robust methods under this new framework (published in Annals of Statistics). This shows the need to develop new methods that yield robust results when analyzing high-dimensional data.

- *Data fusion*
  As a consequence of our information society, not only more and larger data sets become available, but also data sets that include multiple sorts of information regarding the same system. Such data sets can be denoted by the terms coupled, linked, or multiset data, and the associated data analysis can be denoted by the term data fusion. The KUL-1 team (along with international partners) has contributed significantly to this challenging area through the development of novel models, through the study of model interrelations (which are of utmost relevance in model selection), and through the study of various strategies for optimally integrating information from various linked data blocks (that may differ in size, amount of noise, redundancy, relatedness to other blocks etc.). This work, which resulted in several top journal papers, meets a broad need as nowadays problems of data fusion are ubiquitous in many research disciplines.

References

[1] Molenberghs, G. and M.G. Kenward. Missing Data in Clinical Studies. John Wiley, New York, 2007 (UH, LSHTM). B07004.

[2] Fitzmaurice, G., Davidian, M., Molenberghs, G. and G. Verbeke, Longitudinal Data Analysis, Handbooks of Modern Statistical Methods. New York: Chapman and Hall, 2009. B09002. (KUL-2, UH)

[3] Dmitrienko, A., Molenberghs, G., Christy Chuang-Stein, J.L. and W.W. Offen, Analysis of Clinical Trial Data using SAS: A Practical Guide. (Japanese Translation), Kodansya Scientific and SAS Press, 2009. B09004.

[4] Carpenter, J.R. and M.G. Kenward, Missing data in clinical trials - a practical guide, 2007. TR07082.

[5] Carpenter, J.R. and I. Plewis, Coming to terms with non-response in longitudinal studies. SAGE handbook of Methodological Innovation, M. Williams and P. Vogt (eds.), London, SAGE, 2010. R10006.

[6] Molenberghs, G. and E. Lesaffre, Missing Data. Wiley Encyclopedia of Clinical Trials, 2007 (KUL-2, UH). B07005.

[7] Molenberghs, G., Beunckens, C., Jansen, I., Thijs, H., Verbeke, G. and M.G. Kenward, Missing data. In: Handbook of Epidemiology, I. Pigeot and W. Ahrends (eds.), Heidelberg: Springer, 2009. R09122. (KUL-2, UH, LSHTM)

[8] Molenberghs, G. and G. Verbeke, Discussion of Diggle, P., Farewell, D. and Henderson, R.: "Analysis of longitudinal data with drop-out: objectives, assumptions, and a proposal». Journal of the Royal Statistical Society, Series B, 56, 542, 2007 (KUL-2, UH). R07107.

[9] Molenberghs, G., Editorial: What to do with missing data ? Journal of the Royal Statistical Society, Series A, 170, 861-863, 2007. R07002.

[10] Molenberghs, G. and G. Verbeke, Longitudinal and incomplete clinical studies. Journal of the Japanese Society of Computational Statistics (to appear), 2009. RP09035. (KUL-2, UH)

[11] Molenberghs, G., Incomplete data in clinical studies: Analysis, sensitivity, and sensitivity analysis (with discussion). Drug Information Journal, 43, 409-446, 2009. R09100.

[12] Molenberghs, G., Incomplete data in clinical studies: Analysis, sensitivity, and sensitivity analysis. Rejoinder. Drug Information Journal, 43, 447-448, 2009. R09101.

[13] Molenberghs, G., Kenward, M.G. and G. Verbeke, Discussion of Lee, Y. and Nelder, J.A.: Likelihood inference for models with unobservables: another view. Statistical Science (to appear), 2009. RP09048. (KUL-2, UH, LSHTM)

[14] Ibrahim, J.G. and G. Molenberghs, Missing data methods in longitudinal studies: a review (with discussion and rejoinder). Test, 18, 68-., 2009. R09110.

[15] Burzykowski, T., Carpenter, J. Coens, C., Evans, D., France, L., Kenward, M., Lane, P. , Matcham, J., Morgan, D., Philips, A., Roger, J., Sullivan, B., White, I. and L.M. Yu, of the PSI missing data expert group, Missing data: discussion points from the PSI missing data expert group. Pharmaceutical Statistics (to appear), 2009. RP10061. (UH, LSHTM)

[16] Nur, U., Shack, L.G., Rachet, B., Carpenter, J.R. and M.P. Coleman, Modelling relative survival in the presence of incomplete data: a tutorial. International Journal of Epidemiology, 39, 118-128, 2010. R10007.