International Workshop on Robust Statistics and R

Claudio Agostinelli (Università Ca'Foscari di Venezia),
Peter Filzmoser (Vienna University of Technology),
Matias Salibian-Barrera (University of British Columbia)

Oct 28 – Nov 2, 2007

1 Overview of the Field

Robust Statistics deals with a pressing problem in statistical applications: many classical statistical methods work well only for high-quality data that can be modelled adequately. In many practical applications, however, this is the exception than the rule. When the sample size is moderately small, the sampling variability (i.e. the variation induced by the random nature of the sample obtained from the underlying population) may dominate the error produced from possible model misspecifications. However, the sampling variance decreases when the sample size increases, and for large data sets this variance can be very small, and the overall error may be mainly due to the systematic bias of model misspecification, which is not reduced by larger sample sizes.

In recent years, we have seen an enormous increase in the amount of data being modelled and analyzed. For example automated electronic data collection, and complex data sets, produce data sets for which both the number of cases and variables much exceed the orders of magnitude that were routine only a decade ago. Two problems may arise when analysing these large and complex data sets with classical statistical methodologies: (a) it may not be easy to fit simple and parsimonious models that reflect equally well all the data; and (b) the sampling variability for such large datasets can be very small, to the extent that, as mentioned above, the possible model misspecification bias (which, unlike the variance of most estimators, does not decrease to zero as the sample size grows) dominates the statistical error, and may put into question the validity of the analysis. Furthermore, in some large-scale applications the interest may even be focused particularly in finding whether there is a subset of the data that do not seem to follow the same model as the majority of them.

Robust statistical techniques are natural candidates to deal with the challenges mentioned above. They are designed to perform well both when the data follow the proposed model, but also when a proportion of them may contain "outliers", that is, observations that do not follow the same model as the other points in the sample. The field of Robust Statistics has been extremely active in the last 40 years, with many robust methods proposed and discussed in the best international statistical journals. Unlike the methodologies derived under the assumption that there exists a relatively simple model (at least in terms of mathematically tractability) that fits all data reasonably well, robust techniques are generally characterized by: (a) high computational complexity; (b) relying on asymptotic properties (because their finite-sample properties are typically unknown); and (c) estimators that do not admit a closed-form formula.

After the seminal work of Peter Huber in the early 60's, the field of Robust Statistics enjoyed several years of active development, and faced increasingly complex problems (both of conceptual and computational nature). As mentioned before, the emergence of large-scale applications (where model misspecification be-

comes either a concern or a feature directly related to the scientific question of interest) has given the field a new period of attention. The rapid evolution of computer technologies has also been instrumental in allowing many robust techniques to be disseminated for wider use in the general scientific community. An example of the renewed research activity in this field includes the International Conference on Robust Statistics (ICORS) that has been held annually since 2001. This conference has consistently attracted the top scientists working in the area and also young researchers, which has resulted in many successful collaborations and developments in the field. Additionally, in 2004 the European Science Foundation established a research network supporting 10 European institutions for the development of robust statistical methods for complex data.

Although recent years have witnessed increased research interest in robust statistics, many of these techniques are still not widely used in practice. This is probably due to several reasons, among which we list their being particularly difficult to compute (even for relatively simple models) and the lack of a generally accepted set of inference principles and methods to follow after point-estimators have been either calculated or approximated. In addition to these considerations, note that there are currently very few easy-to-use, coherently structured and well documented computer programs to calculate or approximate robust estimators, and certainly no freely available ones with these characteristics. Given the intrinsic difficulty of calculating or approximating these estimators it is not completely surprising that the observed properties of some of these methods may depend on the particular algorithm used to find the point estimators. In some cases one may even find that "the algorithm is the estimator" (David Rocke, personal communication). Hence, in this case it is particularly desirable to have a collection of computer programs developed by, or with close collaboration of, the leading researchers in the field.

The computational complexity of these estimators cannot be overestimated. Many of the approximating algorithms are based on non-trivial numerical optimization techniques, and thus a non-specialist may find it very challenging to even approximate these estimators for most real-life applications. A simple example of the type of computational difficulties found in this area is given by the Least Median of Squares (LMS) estimator [30] for linear regression. This estimator is defined as the vector of coefficients that minimizes the median of the squared residuals. To compute the LMS estimator we need to solve an optimization problem in many variables with a non-convex and non-smooth objective function. Although some algorithms exist to find an approximate solution their complexity increases exponentially with the dimension of the problem and become unfeasible for high-dimensional problems (as those commonly found in modern statistical applications). Note that the intrinsic challenge here is not only the lack of smoothness of the LMS objective function being optimized, but also that it is not convex, and thus requires some variation of random search. Furthermore, the objective functions of many robust estimators are non-convex. Intuitively, this follows from the fact that, in order to obtain a robust estimator, the corresponding score equations (which involve the first derivative of the objective function being optimized) need to decrease to zero for large values of their argument. This, together with the requirement that the score equations be similar to the likelihood-based ones near zero, makes the corresponding loss function non-convex. For example, this is the case for smooth high-breakdown regression estimators (as the S-estimators of [36]), and similar problems are found when calculating or approximating robust estimators for other models.

For many years these difficulties were put aside in the hope that improved computer technology would provide faster machines that would allow these approximating algorithms to perform a more exhaustive search of the parameter space. However, as mentioned above, together with faster machines, we have also seen an exponential increase in the complexity and size of the data sets that require analysis. Thus, the need for more efficient algorithms and computer implementations has not diminished, and may have in fact increased. Moreover, the last decades have also seen explosive growth in the routine application of relatively sophisticated statistical analyses by non-statistician in a diverse range of subject-matter disciplines. It is of necessity to provide these practitioners with stable, efficient, scalable and easy to use software. We also believe that this software, being of scientific value, should be provided free of charge, and on an open-source platform to be completely transparent, and contribute to the advancement of science and population welfare.

The existing code for robust estimators is of varied origins, quality, ease of use and accessibility (namely, financial cost, restrictions associated with their licenses, etc.) While many international leaders in robust statistical methodology have written their own computer code to calculate or approximate some estimators, these programs have generally been written individually by different researchers or research teams around the world. In many cases, these programs were originally intended for the private use of their author (as opposed to practitioners in general), and run under one of a variety of different and non-compatible environments,

including R [23], S-PLUS, Matlab and SAS.

2 Objectives of the workshop

One of the main objectives of this workshop was to tackle the problem of the lack of high-quality and easily-available and usable computer implementations for robust statistical methods. An important difficulty in trying to organize and coordinate the atomized efforts of isolated researchers developing code for personal use, is that, until recently, there have been no scientific meetings focused on this problem. Inconsistent (but generally low) scientific credit given to the study of these problems by our peers in academic positions resulted in a lack of international high-calibre meetings devoted to the delicate computational challenges faced by robust statistical methods. Without physically bringing together the people working on these topics, it is extremely difficult to coordinate developments over time.

In the last few years the R-project [23] has established itself as a widely available, powerful and versatile computer program for statistical analysis. In particular, R is distributed under the GNU General Public License [12], it is open-source, and has been already widely accepted and adopted by a broad community of students, practitioners and researchers in many disciplines (not only Statistics). Furthermore, many of the isolated individuals and research teams that have been developing "in-house" computer code, have being using R. Thus, R appeared as a natural candidate for a platform to base the coordinated development of computer code for robust methodologies.

Although efforts towards addressing this issue started to materialize around 2002, they have been sporadic. In 2005 the first International Workshop on Robust Statistics and R (Treviso, Italy) attracted a lot of attention in the Robust Statistics community. In this meeting some initial guidelines were agreed upon, researchers with similar interests discussed their specific ideas, and a few basic communication tools were set up (websites, mailing lists, etc.) Shortly after the Treviso meeting the robustbase package [26] for R was put together building on several existing packages and stand-alone code from a number of different authors.

With this workshop we were intending to facilitate and coordinate further development of the robustbase package for robust statistics tools in R [23] and to promote the interaction and collaboration between researchers interested in the computational aspects of Robust Statistics. Some topics we had identified as being of particular interest included: the desired degree of default accessibility for non-experts, the specifics of hierarchical integration with other packages already existing in R, and a discussion on methods and techniques that may need to be incorporated to the package in the near future. Other topics for discussion were updates to algorithms and estimation and inference methods (recent developments in these areas that might need to be incorporated) and the ability of the current version of the library to manage large scale applications (scalability).

Based on these considerations we decided that the best format for our meeting would be one of a proper workshop, namely: a series of round-table open discussions among relatively small groups of researchers working around a common problem or topic, followed by "plenary" discussions of the main points raised by each of these working groups. In our opinion, we could not have efficiently achieved our goals in a "classical" scientific meeting (i.e. in a conference structured as a long sequence of single-speaker presentations followed by short periods of question-and-answers). We had also anticipated that the development of this package would naturally identify important practical and theoretical challenges and become a driving force for new research and activity in the field.

3 Recent Developments and Open Problems

In 2001 a high-quality library of robust methods called "robust" was released for S-PLUS. This library contained up-to-date methodologies, and had been developed in consultation with many international experts in robustness. In particular, it included code directly written by many of the researchers who had proposed and investigated the different techniques. At around the same time, R was starting to assert its place in the statistical community as a reliable, open-source and freely available platform for a diverse range of statistical computing. Furthermore, for many individual users and academic units, the price and license restrictions attached to S-PLUS made the migration to R seem a natural choice, particularly after R had attained a certain degree of stability and development maturity.

In 2002, the International Workshop on Computational Methods for Robust Statistics was held as a satellite meeting of the 2002 International Conference on Robust Statistics (ICORS2002) at The University of British Columbia. This workshop was sponsored by the National Science Foundation and brought together for the first time researchers interested in the computational challenges behind many robust statistical methods. A short time later, after a few very successful early ICORS meetings, there was renewed interest in developing a coordinated package of robust methods for R. This lead, in 2005 to the first International Workshop on Robust Statistics and R (Treviso, Italy) which attracted a lot of attention in the Robust Statistics community. Top researchers in the area participated and the critical importance of this type of meetings was stressed. Groups responsible for different parts of the software development process were identified and guidelines to coordinate future work were agreed upon.

One important principle that was adopted in Treviso was that there should exist a "basic" robustness package implementing some "bread-and-butter" functionality and methods, which in turn could be used as a building block for more specific packages. Shortly after this workshop Martin Maechler took the initiative in putting together the robustbase package [26] that initially merged several existing packages and stand-alone code from a number of different authors. Also in this meeting it was agreed to use the book by Maronna, Martin and Yohai [19] as a guideline to select which methods should be included in this "basic" package, keeping in mind that this package was supposed to contain only building blocks, upon which other R packages can be developed. In particular, it was decided that most multivariate methods should continue to be developed in a separate package, currently rrcov [29], and similarly for robust methods for time series (now in the robust-ts package [27]).

To keep track of the developments and progress on the project a website was open at the R Project main site [37], and during the useR!2006 conference [45] a Focus Session on Robust Statistics was organized, mainly devoted to the presentation of the new available software to the R user community.

3.1 rrcov package

The rrcov package implements robust estimators for multivariate location-scatter models. Additionally, it is particularly focused on exploiting the functionality of the S4 object-oriented capability of R. The object oriented programming paradigm has revolutionized the style of software system design and development. A further step in software reuse is the object oriented framework (see Gamma et al. [9]) which provides technology for reusing both the architecture and the functionality of software components. The rrcov package provides an object oriented framework for robust multivariate analysis. The goals of this framework include: (i) to provide the end-user with a flexible and easy access to newly developed robust methods for multivariate data analysis; (ii) to allow the programming statisticians an extension by developing, implementing and testing new methods with minimum effort, and (iii) to guarantee the original developers and maintainer of the packages a high level of maintainability.

The objected-oriented nature of this package is better illustrated with a simple example. Starting with the generic object model of a multivariate method with all the necessary interfaces and functionalities we build a class hierarchy to represent it, along with the robust methodologies associated with it. The basic idea is to define an abstract S4 class which has as slots the common data elements of the classical method and its robust counterparts (e.g. principal components analysis, PCA). For this abstract class we can implement standard generic R functions like print(), summary(), plot() and maybe also predict(). Then we can derive and implement a concrete class which will represent the classic method, say PCAClassic. Then we derive another abstract class which represents the associated robust method, e.g. PCARobust. This class is abstract because we want to have a placeholder for the robust methods we are going to develop next. The generic functions that we implemented for the class PCA are still valid for PCARobust but whenever necessary we can override them with new functionality. Thus a platform for building new robust methods for PCA is created.

The framework includes an almost complete set of algorithms for computing robust multivariate location and scatter, which are the cornerstones of most multivariate methods, such as Minimum Covariance Determinant (MCD), different S estimators (SURREAL, FAST-S, Bisquare, Rocke-type), and the orthogonalized GnanadesikanKettenring (OGK) estimator of Maronna and Zamar [21]. The next large group of classes are the methods for robust Principal Component Analysis (PCA) including ROBPCA of Hubert, Rousseeuw and Branden [13], Spherical Principal Components (SPC) of Locantore et al. [14], and the projection pursuit algorithms of Croux and Ruiz-Gazen [7] and Croux, Filzmoser and Oliveira [6]. Further applications imple-

mented in the framework are linear and quadratic discriminant analysis (see Todorov and Pires [44], for a review), multivariate tests (Willems, Pison, Rousseeuw and van Aelst [47]; [43]) and outlier detection tools. As a reference for the multivariate methods implemented in this framework, Chapter 6 of Maronna, Martin and Yohai [19] can be used. The package also includes several examples that illustrate the usage of the framework for data analysis by the end user as well as for development of new methods.

3.2 robustbase package

The robustbase package was initially developed after the first workshop held in Treviso, Italy, in 2005. Martin Maechler was particularly active in this project merging several packages already existing and other stand-alone code from a number of different authors. The choice of robust methods included in this package is based on those covered in the recent book by Maronna, Martin and Yohai [19]. The guiding principle is that this package should contain "basic" building blocks, upon which other R packages can be developed.

We can identify three main models for which robust estimation methods are implemented in robustbase: (a) linear models; (b) generalized linear models; and (c) multivariate location-scatter models. In addition, there is also a fit for non-linear regression models, and robust estimators for simple location-scale models. Robust methods currently implemented in robustbase include: robust linear regression: MM-estimators [48] lmrob and LTS estimators [31] ltsReg , robust generalized linear models glmrob (only for binomial and poisson responses, using the robust quasi-likelihood approach of Cantoni and Ronchetti [3]), robust non-linear regression nlrob, location M-estimators with fixed scale huberM, and multivariate location-scatter: the Minimum Covariance Determinant [33], and the Orthogonalized Gnanadesikan–Kettenring [11] and [21].

What follows is a brief description of the functionality implemented in the robustbase package prior to the workshop for each of the main models mentioned above.

- Linear regression models: the main function is lmrob. It currently implements MM-estimators [48] starting from an S-estimator [36]. These estimators are highly resistant to outliers, and also efficient when the errors are normally distributed. By default this function computes a regression estimator with 50% breakdown point and 95% efficiency for normal errors. The S-estimator is computed using the fast-S algorithm proposed in [39], and the local M-solution is found using re-weighted least squares iterations. The standard errors of the estimators reported by summary are based on [5] and valid for the case of both symmetric and asymmetric error distributions (including contaminations). The plot method produces similar plots to those available for the "classical" regression estimators via the plot method for the lm function. Tuning parameters (both for the estimators (controlling their breakdown point and efficiency) and the algorithm (number of random sub-samples for the fast-S algorithm, etc.) are passed using the "control" function lmrob.control.

  The least trimmed squares estimator [31] is implemented in the function ltsReg. This function approximates the estimator using the fast-LTS algorithm proposed in [34] and [35]. Further consistency and finite-sample corrections are applied, see [22].

- Generalized linear models: the main function is glmrob. It implements robust estimators for these models based on the robust quasi-likelihood approach of Cantoni and Ronchetti [3] with monotone score functions. The effect of potential high leverage points can be controlled by downweighting observations according to their Mahalanobis distances using different robust multivariate location-scatter estimators (see [8]). Currently only log-linear and logistic regression models are implemented. As is the case for lmrob, tuning parameters are managed with a control function.

- Multivariate location-scatter estimators are implemented through the functions covMcd and covOGK. The implementation of covMcd uses the Fast MCD algorithm of [32]. The estimator is further corrected applying rescaling factors as in [22]. The Orthogonalized Gnanadesikan-Kettenring is implemented in the function covOGK.

4 Scientific Progress Made

During this workshop we continued the work started in the first International Workshop on Robust Statistics and R (Treviso, Italy, 2005). In particular, the recent package robustbase was formally introduced, and the

discussion focused on what needs to be included in this package in the immediate future. Part of these discussions focused on technical issues (e.g. the use of S4 classes, unified criteria for passing arguments related to both the algorithm and the estimator, guidelines to unify the different implementations, classes, common methods, common graphical displays, etc.)

We took fully advantage of the workshop facilities at BIRS, and decided to divide the participants in focus groups that would meet during the day to have round-table discussions and presentations. Each working group then reported back to the "plenary" session, where we held an open discussion among all the workshop participants. We found the possibility of structuring the work in this way much preferable to the more common style of a series of individual presentations. In our opinion this is the key advantage of BIRS compared with other conference facilities: having several days to interact and work jointly with colleagues. Some short presentations were necessary and very useful, e.g. the working group reports.

There were 31 participants from 11 countries. Following the structure used in the 2005 Workshop in Treviso, several following working groups were identified around pivotal themes for future development. The workshop participants were allowed to join one of these groups to work throughout the week. The membership of the working groups was

1. Linear models / econometrics: Claudio Agostinelli, Kris Boudt, Christophe Croux, Roger Koenker, Kjell Konis, Guixian Lin, Martin Maechler, Alfio Marazzi, Ivan Mizera, Andreas Ruckstuhl, Matias Salibian-Barrera, and Stefan Van Aelst.

2. Descriptive / Exploratory data: Rudolf Dutter, Chris Field, Roy Welsh.

3. Time Series: Roland Fried, Ursula Gather, Peter Ruckdeschel, Bernhard Spangl.

4. Multivariate methods: Peter Filzmoser, Heinrich Fritz, Luis Angel Garcia-Escudero, Marc Genton, Justin Harrington, Christian Henning, Ricardo Maronna, Matthias Templ, Valentin Todorov, David Tyler, Gert Willems.

Note that some participants were interested in more than one topic, and thus may have participated in the deliberations of more than one working group along the length of the workshop.

The main recommendations of each working group were:

1. Linear models / econometrics:

   - the function lmrob currently performs re-descending iterations starting from an S-estimator of regression. It was discussed that the choice of initial estimator can be offered as an argument at the user level, so that other options are easily implemented.
   - the current sub-sampling strategy to calculate the S-estimator breaks if the design matrix is particularly sparse. This happens naturally when there are several categorical variables in the regression model. The MS- algorithm of [20] that is already implemented in the robust library for S-PLUS should be included in robustbase to deal with these cases.
   - Although some robust GLM methods are available, it was agreed that it would be important to implement the proposal of Bianco and Yohai [1].
   - Some model selection functionality is desirable. In particular functions like add1 and drop1. The ensuing discussion centered on which criteria could be used (some options include: robust C_p or AIC [28], a robust R^2 [4], or the robust future prediction error RFPE [19].
   - The addition of an anova function to test for nested models was considered. Here one can use Wald-type tests, or R^2 type test statistics ([17] and [18]). Furthermore, valid approximations to the p-values can be computed with the robust bootstrap [40] and [38]. These approximations hold under more general conditions than the usual asymptotic results.
   - Adding distance-distance plots to plot.lmrob was discussed and agreed that it would be useful to have included.
   - The accuracy of the estimated standard errors currently reported by summary.lmrob was also discussed.

2. Descriptive / Exploratory data:

   This includes mainly methods that are data based and not model based. Accordingly, methods and plots should focus directly on the description and presentation of the data at hand, but not on the model that may be applied to the data. There already exist useful plots, like boxplot presentations, and plots for outlier detection. However a Data Diagnostics and Visualization (DDV) package for the visualization of outliers in multivariate dataset is still missing.

3. Time Series:

   - Existing packages are on robust Kalman filtering and robust signal extraction. Planned packages are robust counterparts to those included in the stats package, i.e. time series functions and FinMetrics routines.
   - The development and maintenance of packages will be done in R-forge. No mailing list should be used, but R-forge will serve for communication.
   - The contents of the planned ts package includes robust counterparts to acf/pacf (Ma and Genton [16]): quadrant, M-estimator; AIC/BIC (Ronchetti and Staudte [28]); ar/arima: GM, tau, diagnostics; arch/garch (Boudt and Croux [2]); filter; Holt-winters (Croux, Gelper and Fried [10]); spec/spectrum (Spangl and Dutter [41]); methods for plot, print, and summary need to be adapted.
   - The input structure of the functions should consider
     - possible input data are (raw) vectors and irregularly spaced time stamps
     - as far as possible the same arguments as for classical routines
     - for the method argument a (vector of) character(s) or an object of S4 class
     - the control argument is a function generating certain control/tuning objects
   - The output structure of the functions should consider
     - time-stamped elements
     - generally at least S3 classes
     - class should "inherit" from corresponding classical return class
     - in case of several methods computed in parallel: a list of corresponding objects

4. Multivariate methods:

   - General structure: For certain functions and estimators we should make use of overall functions, "superfunctions" (wrapper) handling several single functions for different robust estimation methods. The goal is to design the "superfunction" in a way that it has not to be changed if a new method is available. The design of such a function could be made for each class of estimators by taking a "method" parameter and calling the corresponding estimator function. The method parameter can either be a character string, like "mcd", "mve", "M", "ogk", "auto", etc. or a control object. If the method parameter is left empty or is set to "auto", the overall function tries to call the appropriate estimator.
   - Naming conventions: In rrcov the different robust estimators and methods are implemented as functions returning S4 objects. The names of the function and the corresponding S4 class are identical (i.e. the function CovMcd() returns an object of S4 class CovMcd). The names start with a capital letter, like CovMcd, CovMest, CovOgk, CovMve for location/scatter estimators or PcaHubert, PcaGrid, Pcaproj, PcaMaronna, PcaSpherical for PCA, etc. If a function returns an S3 class, its name starts with a lowercase letter.
     In robustbase, as far as the multivariate methods are concerned, same conventions as in rrcov are used. Currently there are the functions covMcd and covOGK, both returning S3 objects. Later all CovXxx and PcaXxx functions will be moved there. The case of the lmXxx family is more complicated since there are already established conventions for the classical methods, but we hope that the Regression working group will come up with a solution. It is important to agree on one convention and to be consistent.
     In the robust library the names start with lower case letters for the estimator (method) and add upper case "Rob", like lmRob. The problem here is that it inherits from S-PLUS and must maintain comparability with S-PLUS.

- General goals: The idea behind the *robust-library* is to include user-friendly functions with "sensible" defaults and no possibility of fine-tuning for experts. While the user-friendliness is desired any other package, these must grant full control to the experienced user.
- Further "superfunctions": Similar to the design of a general function for robust covariance estimation, such a general "superfunction" could exist for:
  - *PCA:* Here one has to distinguish
    * covariance based PCA
    * projection pursuit based PCA

    For covariance based methods an object from robust covariance estimation can be plugged-in. For projection pursuit based PCA some approaches are already available in R (e.g. package "pcaPP").
  - *LDA:* Here it is important to include
    * different methods for covariance estimation (as plug-in)
    * different ways for computing the common covariance matrix
    * and probably additional features for variable selection
  - *QDA:* This is even simpler than LDA since no pooling of the covariances is needed. Thus we need
    * different methods for covariance estimation
    * functions for plotting the results (cf. Uwe Ligges?)
  - *Factor Analysis:* Already the standard function provides the possibility of using a robust covariance matrix.
- Cluster analysis: This topic is much more complicated. Due to the variety of different approaches, no "superfunction" is possible, only for certain classes of methods this is feasible.
- Further methods:
  - *Logistic Regression:* It is desirable to include the Bianco-Yohai estimator [1]. This should be coordinated with the Working Group on Regression.
  - *Discriminant Coordinates:* for cluster validation Christian Hennig has done work
  - *ICA:* already available in R (package ICS). This should probably not be included in robustbase.
  - *Outlier Detection:* could be implemented using a "superfunction". Some methods already exist in R (e.g. package mvoutlier)
  - *Multivariate Tests:* Hotellings T^2 and Wilks Lambda are already implemented by Valentin Todorov.
- Advances topics:
  - *Robust bootstrap:* This can be used as additional functionality within certain multivariate methods (e.g. selection of the optimal number of PCs for PCA). However, this can be done later.
  - *Missings:* There could be an optional parameter for multivariate methods how to treat data with missing values (e.g. Cov-estimation: impute very large values, apply MCD). Again, this can be done later.
- Where to include the functions: The robust estimators of a given class (e.g. location/scatter, PCA, LDA, QDA, Cluster) are implemented either in package robustbase or in a 'higher level' package which depends on robustbase. The overall function must be in the 'highest-level' package in order to avoid dependence of robustbase on other packages. Later, if considered appropriate, the functions can be moved to robustbase and this will remain transparent for the user.

Kjell Konis confirmed that Insightful Co. was releasing their robust library for S-PLUS under an open-source license, and this library was being converted into an R package by himself. This announcement was well received, and opened the door for re-use of some of these functions in the robustbase package, thus

avoiding extensive re-coding. In general, there was general consensus regarding the migration of the package robust originally developed for S-PLUS into R. There was a informative discussion regarding which license applies to this release, and the desired roles of the packages robust and robustbase. It was agreed that, although more modern and efficient algorithms for basic estimators are already included in robustbase, the package robust contains valuable implementations. For example, the fit.model paradigm that allows for side-by-side comparison of different fits (typically robust and non-robust), both graphically and summaries. The concensus was that while robustbase will focus on implementing workhorse functions and allow for direct access to the many tuning parameters and technical options behind each estimator and algorithm, the robust package will provide a layer of user-friendly functionality.

5 Outcome of the Meeting

One important goal of the workshop was to discuss and agree on specific guidelines for future development of the robustbase package. Based on the working-group structure of our workshop, we had each working group identify and report for plenary discussion what was perceived as the desired next steps. Much discussion centered on which robust methods were sufficiently "mature" to be included in a commercial-quality release of the package. Other issues considered were: scalability of the current algorithms and their ability to deal with different types of data that occur frequently in practice.

After the launch of the R-Forge platform [24] (which is the new platform for the R Community) in late 2007, several projects were open by participants of the workshop: (i) RobKalman [25], which implements several robustifications of the classical Kalman filter. A common filtering iterface for all robustifications is provided as well as S4-classes for state space models and filtering results; (ii) robust-ts [27] which is a collaborative project to provide robustifications to the basic time series procedures from package stats. A target will be chapter 8 in [19]; (iii) libRa [15] which implements robust statistics algorithms from the research group Robust Statistics at the Katholieke Universiteit Leuven and the Universiteit van Antwerpen.

A Task View on Robust Statistics was made available after the workshop at [42] which contains a brief description of all available packages in the field. In order to disseminate to the R users community the new advances in both theory and available software attained after this workshop, we plan to organize a tutorial session during the useR!2008 conference [46].

Both organizers and participants felt that the workshop was a great success. We achieved our goals in a very congenial atmosphere. We have also had very lively and productive discussions. The meeting also gave new incentives for future work and collaborations to many researchers. Although nowadays many possibilities for communication exist, face-to-face meetings allow for a much better and more efficient coordination and motivation.

The relaxed atmosphere contributed a great deal to have long exchanges with several researchers at a time, that would not have been possible in the context of a regular scientific conference / meeting. The workshop-type structure allowed by BIRS during these 5 days greatly contributed to foster interactions, and communication between researchers from different institutions. This seemingly unstructured schedule was identified by many participants as a key factor contributing to the advances made at BIRS. In our opinion, the flexibility offered by BIRS is very valuable and should be preserved, since it was in large part because of it that our workshop exceeded all expectations.

References

[1] Bianco, A. and Yohai, V.J. (1996). Robust estimation in the logistic regression model. In: H. Rieder, Editor. *Robust Statistics, Data Analysis, and computer intensive methods. Lecture Notes in Statistics.*, **109**, 17-34. Springer, New York.

[2] Boudt, K. and Croux, C. (2008). Robust M-estimation of multivariate GARCH models. *Revision invited by Econometrics Journal*.

[3] Cantoni, E. and Ronchetti, E. (2001). Robust Inference for Generalized Linear Models. *Journal of the American Statistical Association*, **96** (455), 1022-1030.

[4] Croux, C., and Dehon, C. (2003). Estimators of the Multiple Correlation Coefficient: local robustness and confidence intervals, *Statistical Papers*, **44**, 315-334.

[5] Croux, C., Dhaene, G. and Hoorelbeke, D. (2003) Robust standard errors for robust estimators, *Discussion Papers Series 03.16*, K.U. Leuven, Belgium.

[6] Croux, C., Filzmoser, P. and Oliveira, R. (2007). Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, **87**(218), 225.

[7] Croux, C. and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: The projection-pursuit approach revisited. *Journal of Multivariate Analysis*, **95**, 206-226.

[8] Donoho, D.L. and Huber, P.J. (1983) The notion of breakdown-point. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L Hodges, Jr., eds.) 157-184. Wadsworth, Belmont, California.

[9] Gamma, E., Helm, R., Johnson, R. and Vlissides, J. (1995). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, Reading.

[10] Croux, C., Gelper, S. and Fried, R. (2008). Computational Aspects of Robust Holt-Winters Smoothing based on M-estimation, *Applications of Mathematics*, **53**, 163-176.

[11] Gnanadesikan, R. and John R. Kettenring (1972) Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, **28**, 81-124.

[12] http://www.gnu.org/copyleft/gpl.html

[13] Hubert, M., Rousseeuw, P.J., and Branden, V. (2005). Robpca: a new approach to robust principal component analysis. *Technometrics*, **47**, 64-79.

[14] Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J. and Cohen, K. (1999). Robust principal components for functional data. *Test*, **8**, 1-28.

[15] libra library: http://r-forge.r-project.org/projects/libra/

[16] Ma, Y., and Genton, M. G. (2001), Highly robust estimation of dispersion matrices *Journal of Multivariate Analysis*, **78**, 11-36.

[17] Markatou, M., Hettmansperger, T.P. (1990). Robust bounded-influence tests in linear models. *J. Amer. Statist. Assoc.*, **85**, 187-190.

[18] Markatou, M., Stahel, W., Ronchetti, E. (1991). Robust M -type testing procedures for linear models. In: *Directions in Robust Statistics and Diagnostics, Part I*, IMA Vol. Mathematical Applications **33**. Springer, New York, pp. 201-220.

[19] Martin, D., Maronna, R. and Yohai, V. (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.

[20] Maronna, R. and Yohai, V. (2000). Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference*, **89**, 197-214.

[21] Maronna, R.A. and Zamar, R.H. (2002) Robust estimates of location and dispersion of high-dimensional datasets; *Technometrics*, **44(4)**, 307-317.

[22] Pison, G., Van Aelst, S., and Willems, G. (2002) Small Sample Corrections for LTS and MCD. *Metrika*, **55**, 111-123.

[23] R: A language and environment for statistical computing, 2008. R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org.

[24] R-forge: http://r-forge.r-project.org

[25] robkalman library: http://r-forge.r-project.org/projects/robkalman/

[26] robustbase library: Basic Robust Statistics. 2007. Original code by many authors, notably Peter Rousseeuw, Christophe Croux, Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, and Martin Maechler. R package version 0.2-8. http://r-forge.r-project.org/projects/robustbase/

[27] robust-ts library: http://r-forge.r-project.org/projects/robust-ts/

[28] Ronchetti, E. and Staudte, R.G. (1994). A robust version of Mallow's Cp. *Journal of the American Statistical Association*, **89**, 550-559.

[29] rrcov library: http://cran.r-project.org/web/packages/rrcov/index.html

[30] Rousseeuw, P.J. (1984) Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871-880)

[31] Rousseeuw, P.J. and Leroy, A.M. (1987) *Robust Regression and Outlier Detection*, Wiley.

[32] Rousseeuw, P.J. and van Driessen, K. (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212-223.

[33] Rousseeuw, P.J., van Driessen, K. (1999). A Fast algorithm for the Minimum Covariance Determinant Estimator, *Technometrics*, **41**, 212–223.

[34] Rousseeuw, P.J., van Driessen, K. (2002) Computing LTS regression for large data sets. *Estadistica*, **54**, 163190.

[35] Rousseeuw, P.J., van Driessen, K. (2006) Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, **12**, 2945.

[36] Rousseeuw, P.J. and Yohai, V.J. (1984) Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series*. (J. Franke, W. Hardle and D. Martin, eds.). *Lecture Notes in Statist.*, **26** 256-272. Berlin: Springer-Verlag.

[37] http://www.r-project.org/Robust

[38] Salibian-Barrera, M. (2005). Estimating the p-values of robust tests for the linear model. *Journal of Statistical Planning and Inference*, **128**, 241-257.

[39] Salibian-Barrera, M. and Yohai, V.J. (2006). A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics*, **15**, 414-427.

[40] Salibian-Barrera, M. and Zamar, R.H. (2002). Bootstrapping robust estimates of regression. *The Annals of Statistics*, **30**, 556-582.

[41] Spangl, B. and Dutter, R. (2007). Estimating Spectral Density Functions Robustly *REVSTAT - Statistical Journal*, **5**(1), 41-61.

[42] Robust Task View: http://cran.r-project.org/web/views/Robust.html

[43] Todorov, V. and Filzmoser, P. (2008). Robust statistic for the one-way MANOVA. submitted for publication.

[44] Todorov, V. and Pires, A.M. (2007). Comparative performance of several robust linear discriminant analysis methods. *REVSTAT Statistical Journal*, **5**, 63-83.

[45] useR!2006 conference: http://www.r-project.org/useR-2006/

[46] useR!2008 tutorials: http://www.statistik.uni-dortmund.de/useR-2008/tutorials/maechler.html

[47] G. Willems., G. Pison, P.J. Rousseeuw and S. van Aelst (2002). A robust Hotelling test. *Metrika*, **55**, 125-138.

[48] V.J. Yohai (1987). High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics* **15**, 642-656.