Minutes from Robust Statistics and R
Treviso, 2005

Minutes of the Discussion Group "Regression"

Martin gave a brief overview about implementations of **robust methods available on CRAN** (not necessarily in S-Plus!).

He prepared some notes. Discussion of it:

- Matías is implementing Roblm (MM-regression estimators) (cf. his slides)
- What is regression? Ordinary regression, logistic regression, Poisson regression, Gamma regression, Weibull regression, local regression, nonlinear regression, multilevel regression, ... Maybe, we should restrict ourselves to linear regression and maybe to GLMs.
- Alfio Marazzio's ROBETH has implemented quite a lot of these things. On this web site there is an R interface (similar to his S-Plus interface to ROBETH)
- Constrained M-estimators (CM-estimator) have been implemented in MATLAB by Ove Edlund.
- P-estimators have very good asymptotic bias properties, but there isn't any publicly available implementation, and there are no known inferential methods for them. (cf. Victor Yohai).
- Robust model selection: Elvezio has some S-Plus-implementation of a robust Cp.
- Some important methods for linear regressions are implemented in the robust library of S-Plus (i.e. there is a copyright with Insightful). Example: linear regression with a mixture of numerical and factor explanatory variables.
- Matías has code for Markatou's score test. The alternate M/S estimator for a mixture of numerical and factor explanatory variables is described in a Maronna-Yohai paper and could be implemented in R.
- Ricardo has some slow experimental code for robustified quantile regression.
- Random effects? There're some recent proposals in journal papers.
- Victor has MATLAB code to calculate the covariance matrix even if the errors are dependent (estimator is consistent but not efficient) (proposal of Christophe Croux and co-authors). This covariance matrix for the case of non-symmetric errors is implemented in Matías's roblm package.

Input-output of the functions/procedure and general programming "best practices":

- It's important to define a **class** which may inherit from the class of the classical method.
- Make use of methods (sometimes just extractor functions): coef, summary, residuals, fitted, predict, print, print.summary, plot, vcov, ...(anova? see below)
- Is anova() "Analysis of Variance" really the proper name for comparison of robustly fit nested models?
- What type of residuals? (linear Regression) Raw residuals (for normal plot as well).
- We discussed the issue that the robustness weights should be part of the returned object.
- What should be in the return list of the fitting function and what should be (maybe again) calculated in an extractor function like summary or plot?
  Doug Bates (in R package 'Matrix') introduced an idea to add elements to the returned object later, when very expensive calculations had to be done. This could be done when computing robust Mahalanobis distances for a diagnostic plot of robust regression objects.
- Martin: Do not mix things: separate the main function, calculation, printing, summary and plotting in separate functions (the latter two typically using methods).
- Is there a goodness of fit measure in robust fitting (replacing somehow R-squared)? – Just the estimated scale?

- Victor: S-PLUS implemented a non-equivariant S-estimator (Pena & Yohai), where re-sampling candidates are deterministic – it is a fast initial estimator that does not need re-sampling, it may have high breakdown point, but there are no theoretical results that prove it (JASA, 1997?)
- Andreas: we could relatively easily build a complete Linear Model package.
- Matías agrees to get involved in this, but does not want to be "alone". Andreas volunteered to help with testing... – mailing lists?

plans for a package ...

- Owe: maybe we should aim to developing a very good package that gets into the standard distribution of R (a "recommended" package). Andreas raised the question that this will involve deciding which methods are included, and that this could be problematic. Martin thinks that this can still be realistic if we can agree on a subset of methods to be included. Many recommended packages are based on books, and we have a few books out now to pick from. Having the package based on a widely recognized book helps to get it "recommended". The upcoming book by Ricardo, Victor and Doug Martin seems ideal.
- Martin: a package needs a maintainer and it many have several authors
- Andreas: methods for linear models have been around long enough, that it should be easy to agree upon
- Ricardo: what do we want from a good robust package?
- Andreas: all techniques that are included on a linear regression book – check their classical implementations in R, and mimic their user interfaces... then we'll have a package that we can use.
- Question: do we include GLM? Andreas thinks that we should start with linear regression
- D. Martin, R. Maronna, and V. Yohai's book (Robust Statistics: Theory and Methods) covers: GLM – linear regression – time series – multivariate methods.
- Following a book will also help with defining the scope of the package.
- how about we include only methods for which you can do inference?
- How do you call the functions? Using "...rob" may be a good way, but it's not clear. With name spaces, having different functions coexist with the same name should not be a big problem. Martin proposed to talk to Brian Ripley and Bill Venables about moving rlm() out of MASS and into our package...
- Martin: one extreme possibility would be to use a name space and hide everything so that the user is forced to type "<package>::<function>(....)" to use its functions, so that they can be differentiated from the classical ones.
- We need to think about our naming conventions.
- Volunteers for each part of the package: Matías has MM, for model selection, Elvezio and Eva may have code for robust Cp and cross-validation (Claudio has old Fortran code from Blanchard on cross-validation), Victor: can contribute variable selection criteria (backward stepwise procedure), Claudio can contribute code for an extension of Ronchetti and Staudte's Cp, Victor: code for the fast initial estimator for large dimensions (Pena Yohai), Andreas – Eva – Martin can contribute the robust GLM, Mallow's approach should be ready soon, only works for canonical links - Victor can contribute code for binomial data with arbitrary link-function (Croux may have code for this on his web-page), this is the Bianco-Yohai S estimator improved by Croux.
- Andreas (& Martin) has code for robust non-linear regression RWLS with Huber weights on top of nls.

- "forward" package of Riani: it is more about diagnostics and about selecting observations rather than selecting models; seems out of scope for the above package.
- Owe's code for CM: he wants to do it but may not have time to do it... It is pure Matlab code that only uses linear algebra. Victor suggested that the sub-sampling part be done in C, Martin thinks that we can have an "initial.estimator" function that could be re-used by other functions. Many building blocks are already available in R... Owe would help somebody else do it.
- More generally, Owe thinks that we should use Newton-Raphson iterations to find the roots of the $\sum_i \psi((x_i, y_i); \theta)$ equation instead of IRLS. It does require care, but really is faster with quadratic convergence in the end.
- Claudio asks whether we should have a function that returns a list of different local minima. Ricardo thinks that this should only be offered to sophisticated users...
- Ricardo and Victor have noticed that sometimes local minima are better than the global one.
- We'll do lm, glm and multivariate methods from Victor, Ricardo and Doug's book.

05/10/28: Morning

- we are joined by the "Econometric" working group, notably Eva Cantoni & Alfio Marazzi
- Martin gives a *summary of Thursday*'s session.The basic reference for a robust package will be Victor-Ricardo-Doug (MMV) book. The book gives the guide of the spirit of robustness.
- Andreas,Martin and Eva will make available a general function for lm, including GLM, as an example for including other methods. The programs of Matías (e.g., the fast S estimate) will be included into this general "lm" function.
- Contributors have to be able to write help pages. A documentation is available for this purpose.
- It will be useful to add standard data sets to the robust package for the examples. The book also includes data sets. These data sets will be added to the package.
- (The book does not include econometric tools.)
- For the moment, the package will not include time series programs.

Eva gives a *summary* of the work of *the Econometrics group*.

- Robust methods are little used because good recent software is not available. Some Econometricians are starting to use R.
- A list of papers has been produced about robust method in econometrics. These are potential programs for R. Little code (especially R code) is already available. (There exists some Matlab code). None of us is author of these methods, so it is not realistic to expect this code in R.
- It is realistic to expect a package of regression models which are also interesting for econometricians.
- At the moment, we cannot commit other people to write programs about other methods used in econometrics.
- The discussion moves to **function arguments** (called "parameters" by some).
- The control parameters should be included in a large control list.
- The "method" will be a main (separate) parameter.
- Parameter *groups* (cf. Matías Salibian's talk): Methods/Estimator, Algorithm, Inference & Implementation; Main parameter/control parameter. Very few important arguments could be left in the main list.

- The summary should include all information about the method used, and 'print.summary()' should print it (at least partially; careful "output readers" will become aware that there could have been other methods)
- Important basic low level functions (such as fast S, IRLS) should be documented in order to encourage other R function authors to use them as building blocks.
- Eva announces her project to write a book (about robust biostatistics) with Stephan Heritier, Maria-Pia Victoria-Feser etc. The book will include Glm, linear mixed effect model, ... There will be R code. Note: that R code could partly rely on our "robust base" package and otherwise on their own code.
- Victor reminds us: the package should also include some model selection function and some heuristic algorithm for large data sets with large number of variables (essentially alternatives to sub-sampling search for a good initial estimate).

Later notes – by Martin Maechler, only

- The datasets from the MMV are not yet (Nov.2005) ready to be given to me, unfortunately.

  I see that Valentin has a few relevant data sets in the `rrcov` package.

  Since the data sets from the Rousseeuw+Leroy book have been public (on their web site, too) for so long, it should not be a problem to add several (most?) of them to the "basic robust statistics" package.
- Proposed names for the "basic robust statistics" package:

  1. `robstats`
  2. `robbase`
  3. `baseRob`

  I'm choosing the first one for now.
- I've added (Rousseeuw & Croux)'s Q_n and S_n estimators of scale, based on an R package I had started in 2002, which is based on the Fortran code from Antwerpen.
- I plan to add "*psi - Function*" objects, using an S4 class, with instances for Huber, Hampel, Biweight, etc, at least for the M-estimators (of location / regression). Such an object should contain ρ, ψ, $w(x) := \psi(x)/x$, ψ' $(= d/\mathrm{dx}\psi)$ functions, default values for the tuning constants, and functionals such as $E_X[\psi(X)^2]$, $E_X[\psi'(X)]$, for $X \sim N(0,1)$ (these functionals are still *functions* of the tuning parameters).
  Very similarly, I'm interested also in the χ functions used for B-/V- optimal M-estimators of *scale*, both the monotone and the redescending ones.
  In general, I think we should add "**basic M-estimation**" things to the package as well, similar to, but more general than `MASS::huber`; note that I'd like to contribute `sfsmisc::huberM` (my 'improved' `huber`) to the `robstats` as well.

2005 Trevisio Workshop on "Robustness and R" Working Group "Econometrics"

Participants

Matteo Fornasier (Venice), Emilio ?? (Trieste),
Marco Gasparotto (Treviso),
Simone Padovan (Padova), Federico Tedeschi (Padova)
Catherine Dehon (Bruxelle)

Coordinators : Eva Cantoni (Geneva), Alfio Marazzi (Lausanne)

Some papers with software (?) but little in R

ARMA
 Bustos O.H, Yohai V.J, Robust estimates of ARMA models, JASA, 1987.
 Muler N., Yohai V.J. (look in Google)

GARCH
 Muler N., Yohai V., Robust estimation for GARCH models (look in Google)

Extreme Values
 Mia Hubert, A robust estimator of the tail index of Pareto-type distributions (draft)

GLM
 Eva Cantoni & Ronchetti, Binomial and Poisson
 Eva Cantoni, Gamma, Journal Health Economics
 Croux & Heasbroeck, CS&DA (Logistic)
 Agostinelli ?
 Markatou ?
 Victor & Bianco & Boente, Gamma, Logistic
 Mills & Field & Dupuis, logistic GLMM, Biometrics 2002/3

GMM
  Ronchetti & Troiani (2001)

LM with asymmetric responses
 Marazzi A., Yohai V., Adaptively truncated maximum likelihood regression with asymmetric errors,

Robust Box-Cox
 Marazzi A., Yohai V., Robust Box-Cox transformations based on minimum residual autocorrelation,

Cox-Models
 Bednarski, Robust estimation of the Cox regression model, Scan J of Statistics, 1993.

Simultaneous equations
  Ronchetti & Krishnakumar 1997, Matlab
  Maronna & Yohai, JSPI, 1997, Matlab ?
  Krasker & Welsh, Econometrics, 1985

Longitudinal Data
  GEE,Generalized Estimating Equation,
  Eva Cantoni 2004, GEE, Fixed effects, Splus
  Preisser & Qagish, 1999, Econometrics;
  Croux e Bramati, 2005, ?

Mixed linear models
  Copt & Victoria-Feser, 2005, to appear in JASA, R code from the authors.

Tobit Models
  Censored models
  Peracchi, F. 1990, Jour of Econometrics, Bounded Influence Estimator for the Tobit model.

Censored data
  Salibian-Barrera M., Yohai V., High breakdownpoint robust regression with censored data, submi
  Marazzi A., Yohai V., Robust AFT models, under work.
  Bo Honoré, CLAD estimator (Censored Least Absol Dev)
  Powell ?
  Li & Haizheng, Symmetrically censored

Latent variable models
  Moustaki & Victoria Feser (2005) to appear in JASA

Discriminant analysis, Canonocal correlation, Robust R^2
  Dehon (Matlab + Gauss)

PCA
  Croux & Haesbroeck
  Croux, Filzmoser

Error in variables models
  Anne Ruiz (matlab ?)

Robust indirect inference
  Genton & Ronchetti (2003,JASA)

R code

Classical version of sumultaneous equations (system fit)

Agostinelli C., Weighted Likelihood Package for R without covariates
  Robust estimation of normal location and scale univariate and multivariate
                       binomial
                       gamma
                       Poisson
  Linear regression
  One step estimation
  Robust model selection based on Mallows Cp, AIC, Cross-validation, stepwise, t-test, variance-
  Robust estimation of circular data
  Univariate normal mixture models, seasonal autoregressive models and fractional models

Croux C., MM with heteroskedastic errors (Matlab, using fast S)
Croux C., Haesbroeck, 2003, CS&DA, Robust logistic regression using the Bianco and Yohai estimat
Croux C., Filzmoser P. ... ?

Cantoni E., Robust GLM, soon
Cantoni E., Robust GEE, planned

LM with asymmetric responses
 Marazzi A., Yohai V., Adaptively truncated maximum likelihood regression with asymmetric errors,
 Marazzi A., Yohai V., Robust Box-Cox transformations based on minimum residual autocorrelation,

Mixed linear models
 Copt S., Victoria-Feser M.P, 2005, High breakdownto inference in the mixed linear model,
 to appear in JASA, R code from the author

Report of group work at the Workshop on Robust Statistics and R

Multivariate

P. Filzmoser, K. Joossens, H. Oja, S. Pagnotta, M. Romanazzi, S. Sirkiä and M. Templ.

Last updated December 13, 2005.

What topics should be discussed here?

- Multivariate location and scatter
- MANOVA (estimates and tests)
- Multivariate distributions
- Multivariate regression
- Canonical correlation, independence studies (redundant anlysis, PLS)
- Dimensional reduction (PCA, ICA, FA, Ordinal PCA)
- Discriminant analysis
- Cluster analysis CA

Let us have a look at what is already done an what should be done in the future. People mentioned by certain topics/packages indicate who have already done or might be asked for writing the packages.

Multivariate location and scatter

Already in R

Available in a package:

package-name if existing	functionname/method	author/maintainer
rrcov	`FastMCD`	V. Todorov
cov.Robust	`cov.nnve`	Wang
mvoutlier	`arw`	P. Filzmoser
MASS	`cov.mve`	
MASS	`cov.mcd`	
MASS	`cov.trob`	

Available as R-functions:

- David Olive implemented the median ball algorithm (MBA) estimator. The function implemented in R is called `covmba()` and information on the MBA estimator can be found on the website of his book "Applied Robust Statistics": `http://www.math.siu.edu/olive/`

ol-bookp.htm
Function covmba in http://www.math.siu.edu/olive/rpack.txt
Information can be found in Chapter 10 http://www.math.siu.edu/olive/run.pdf or in the paper: A Resistant Estimator of Multivariate Location and Dispersion, Computational Statistics and Data Analysis, 46, 99-102.

- Kjell Konis implemented the OGK estimator completely in R. It is an alpha-version and so, at the moment, it should be used with a bit of caution!
  Available at: http://www.stats.ox.ac.uk/~konis/pairwise.q
- The FORTRAN code of D. Hawkins can be found at ftp://www.stat.umn.edu/pub/fsa/readme.ftp.
  Valentin Todorov has an R interface to some of the FORTRAN programs of Hawkins e.g MWMCD (linear discriminant analysis) and will release it in the new rrcov.

Coming soon in R

package-name if existing	functionname/method	author/maintainer
pcaPP	L1median	P. Filzmoser
...	OGK	V. Todorov
...	FastS	K. Joossens
...	M-estimates	S. Sirkia

Other

Other estimators of which we should think

- D. Peña en F. Prieto
  Matlab code can be found at http://halweb.uc3m.es/fjp/kur.zip
  Paper: Multivariate outlier detection and robust covariance matrix estimation, Technometrics, 43, 18-21.
- Y. Zuo, Projection Pursuit based. No software available yet.
  Paper: Projection-based affine equivariant multivariate location estimators with the best possible finite sample breakdown point, Statistica Sinica 14(2004), 1199-1208.
- Juan-Prieto, Projection Pursuit
  Paper: A subsampling method for the computation of multivariate estimators with high breakdown point, Journal of Computational and Graphical Statistics, 4(1995), 319-334.
- J. Agulló
  Paper: Exact Iterative Computation of the Multivariate Minimum Volume Ellipsoid Estimator With a Branch and Bound Algorithm, in Proceedings in Computational Statistics 1996, ed. A. Prat, Heidelberg: Physica-Verlag, pp. 175-180.
- Agulló. Least Quartile Difference estimator.
  Paper: An exchange algorithm for computing the least quartile difference estimator, Metrika, 55(2002), 3-16.
  Originally introduced by C. Croux, P. J. Rousseeuw O. and Hössjer in the paper: Generalised S-estimators, Journal of the American Statistical Society, 79(1994), 871-880.

Input and Output

How should the in and output look like for the location and scatter estimator. We propose to use S4-classes!

INPUT

- `x`: data object
- `wt`: weights/frequency vector
- `cor=FALSE` indicates whether the correlation derived from the estimated scatter should be added to the output arguments
- method=???? method used. e.g. c("covMcd","ogk","S",...)
  Here the question rises: Which robust estimator will be used as default?
- `function.control(...)`,e.g. `covrob.control(list(MCD=list(...),OGK,...))`

OUTPUT

- method parameters
- center
- cov
- wt if called in input or explicitly given as output by the method
- cor if input `cor=TRUE`
- summary(md) summary of the Mahalanobis distances
  e.g. [0 25 50 75 100]

INSTRUCTIONS: We should make an instruction-file for new authors.
If you are writing new functions, think about:

- missing values
- correction factor which depends on the method

EXTRA

- show/print: (center, covariance, ...)
- summary: e.g. method, parameters, dimension of data matrix, possible warnings, condition numbers; (Maronna & Zamar, Technometrics 2002, Vol 44-4)
- plots: e.g. density plots, ellipses ...[Stefano Pagnotta]
- a function compare.plots: If only one input is given, compare the robust to the classical estimator; if two robust estimators are given, compare those two.
  Comparison can be done by DD-plot screeplot biplot .... plenty of choices [Stefano Pagnotta]

MANOVA (estimates and tests)

Is theory in robust manova existing? Maybe our centerCov can be considered as input so that within groups the specific estimator is used., [Maybe Seija Sirkiä and Hannu Oja]

Multivariate distributions

mdistr multivariate distributions [Peter Ruckdeschel] (like random generator, en distributions, p and especially q are much more difficult.)

Multivariate regression

regression fastS [Matias Salibian-Barrera and Victor Yohai?]
..... Use centerCov [Kristel Joossens]

Canonical correlation, independence studies (redundant analysis, PLS)

- Dimensional reduction (PCA, ICA, FA, Ordinal PCA)
- Already in R or coming soon pcaPP PCA [Peter Filzmoser]
- FA,PCA, CC, (DA) [Student Peter Filzmoser]
- ICA [Seija Sirkia and Hannu Oja]
- ICA: uses 2 scatter matrices. So centerCov would be very handy!
- PCA: Can the use of a superfunction be intersting?

Discriminant analysis

use `centerCov` extend the da [Valentin Todorov]

Cluster analysis CA

Usually not covariance based so, here centerCov is not The solution Much work still needs to be done First new theoretical robust methods should be developed! hclust, within-groups

Main remarks

Superfunction centerCov is needed.
We should make comparing robust methods possible and easy. It is maybe one of the most interesting things, that you are able to compare two methods.
You can always contact Seijia and Kristel, who will soon concern about work, when, what, how much ....
It might be interesting to try to organise an extra day on future meetings to spend on evaluating the developments.

Document source: `http://www.econ.kuleuven.ac.be/public/ndbae49/R/`
Contact for comments and suggestions: `kristel.joossens@econ.kuleuven.be`

Report of the Working Group "Time Series" from RsR

State of the Art

- unfortunately not much has been done in R in robust times series context
  e.g. google search on CRAN: 'robust autocovariance'
- we collected important contributions as to papers / not necessarily implementations
  in the following domains:

I. robust tests in time series
II. robust prediction/robust filtering
III. robust change-point analysis
IV. outlier detection
V. model choice
VI. specific parametric models
VII. non-parametric models
VIII. robust autocovariances
IX. robust spectra
X. general robust techniques/methods

(we came up with a not-so-small list of references too long to be displayed in a minute :-)
If anyone is interested, we will send them our results off-line on request.

Open Questions / Wish list

We came up with some open questions / a wish list
for the whole rsr--list

General questions:

1. Is there a robust counterpart of classical "ts" in stats package (which we have overseen?)?
2. How "open" are time series routines in S-plus? What are the legal restrictions as to Insightful?
3. If known, should concepts like breakdown point, influence function of a procedure be accesible by S functions or at least be mentioned in the rd-File?

Request for references to papers & implementations

4. Is there any robust decomposition (season/trend/...) in time series?
5. Are there diagnostic tools for robust autocovariance?
6. Are there freely available implementations of robust inference for serial correlation with GM-estimation?
7. What is known on robust tests for stationarity (e.g. unit root-type- tests) even for more general models, other than autoregressive models?
8. Extending robust regression to kalman filtering and state space ---which implementations do exist in S?
9. What is known in robust change point analysis experiences with sac library, kza library?

Could you help us with (additional) references to robust

10. model selection
11. estimation in GARCH/TAR/BILINEAR models
12. estimation/testing in ARIMA, ARFIMA, ARCH, GARCH families
13. estimation/testing in TAR - models

General design principles

- problem: there is no such universal modeling structures / formula interfaces as e.g. in lm (compare definition of an ARMA-model...)
- -> discussion within whole rsr-list and probably beyond (people working on classical time series)
- WHICH data structures -> S3/S4 (e.g. state space models exist in several implementations...)
- tentatively: preferable reference class: ts (other suggestions?)
- should be: a commonly accepted user interface
- should be: commonly accepted output methods (plot, print, diagnostics)

to come to solutions at short term:

- *compromise:* take one of the existing implementations of corresp. classical procedure (preferably from stats and tseries) and
  then use that user-interface (that is "classical arguments + extra robust control strucure")

(tentative) Proposals for R packages on robust time series:

Robust Kalman Filtering

responsible: B. Spangl & P. Ruckdeschel

envisaged contents: (for point II above)

- (S4) class definitions for state space models
  [contact with Paul Gilbert (author dse) for coherent definitions]
- rLS filter (P.R.[01])
- acm filter (-> if code is available)
- other filters are welcome to be contributed

design principles:

- multivariate concept
- S4 based time-varying State space models should be covered
- will implement the rLS filter and integrate the Kalman filter;
- would appreciate contribution of code for different robust filters -- in particular acm;
  if no code contributed for acm: will try and implement it
  (-> questions "Howto" for Victor Yohai and Doug Martin)

next steps:
Preparation

- discussion with P. Gilbert (dse) and R-Core (KalmanLike)

goals:

- universaly acceptable (S4?) class definition of a mother class state space model (ssm)
- classes for simulated time series / ssm's
- reasonable hereditary structure
- which output methods are convenient for ssm's?
- is there a need for C-code in this field?

deadline: beginning of Jan 2006

Implementation

- implementation of the class definitions
- implementation of the accessor / replacement functions
- implementation of the (abstract) methods for filtering, smoothing, prediction as well as simulation of state-space-models
- implementation of the (abstract) methods for output: summary, plot, print methods for filtered/smoothed/predicted time series
- implementation of the algorithms / collection of existing code
- documentation --- rd-files
- vignette
- package building
- first release .....
- at long term: (robust) algorithms for the estimation of the hyper-parameters

deadline: first prototypes: May 2006

Robust Time Series Estimation

responsibles: Giuliana Cortese & Antonio Parisi & Luca Greco & Moreno Bevilacqua (& (?) further Italian coworkers)

evisaged contents:

- collection of existing (!) robust code to points I, VI, VIII above in one package
- perhaps integration of Claudio's methods for V

design principles:

- no change of interface w.r.t classical routine, additional argument for robustness
- possibly provide unified interfaces for different methods (e.g. autocovariances)
- robust procedures are preformed in extra functions
- result/value of "robust" function is then converted to convenient form to make available generic output methods plot print summary

work to be done:

- contact with researchers / authors of papers on that field
- ask them for code (preferably in S, in C, in C++, or in FORTRAN)
- write rd-files for these routines
- possibly write a vignette?
- package building ...

- at longer term: also porting from other languages

next steps:

- contact with researchers / authors of papers on that field

deadlines: Jan 15th for the first step
(contact with researchers, ask them for code, write rd-files for these routines).

Robust Signal Extraction

responsibles: Roland Fried
evisaged contents: collection of R functions for Robust Signal Extraction
possibly also methods for points III, IV above

work to be done:

- write rd-files for these routines
- possibly write a vignette?
- package building

next steps:

- write rd-files for these routines

deadlines: tentatively june 2006

Robust Outlier Detection

responsibles: Roland Fried
evisaged contents:

- collection of R functions for Robust outlier detection

depends:

- robust time series estimation

next steps:

- wait for packages on Robust Signal Extraction and robust time series estimation

deadlines: after release of package on Robust Signal Extraction

Robust Nonparametric Smoothing [cancelled]

Reason:
Arne Kovac pointed out that it was too early for a package particularly designed for robust methods in non-parametric time series,
regarding the lack of a unified approach in this area and the non-standardized situation on the implementation side in R.

More urgent would be setting up a task-view with all material available in R as to smoothing.

Minutes from Working Group "Large Data Sets"

Friedrich Leisch

27.10.2005

Participants: Ruggero Bellio, Kjell Konis, Friedrich Leisch, Giovanni Pace, Silvia Salini, Francesca Santello, Valentin Todorov, Stefan Van Aelst, Mark Werner

- We first discussed, what a "large data set" actually is and agreed on the following definition: Everything that does not fit into the main memory of the computer due to the number of observations and/or variables. This also includes data which themselves fit into main memory, but leave not enough free memory for computations on them.
- We also agreed that none of us considers himself an expert for the topic, most ended up in this group because they either merely expressed interest in the topic or had dealt with larger amounts of data to some extent in the past.
- We then brainstormed some ideas that may be important for the topic or could offer solutions:

  Pairwise calculations: If a data set has many variables (=columns), then pairwise algorithms could help, because at no point in time the complete data set needs to be loaded to memory. For pairwise algorithms only 2 columns of the data need to be loaded, this could even be distributed over a cluster of workstations with access to a central database. Each node of the cluster receives the information which pair(s) of data columns it shall use and only loads the corresponding subset of the data.

  Fast MCD & LTS: Algorithms that perform only a single pass over the observations (=rows) of the data can be parallelized by creating blocks of rows and performing operations on these blocks, e.g., using different nodes of a workstation cluster for each block of data.

  Online algorithms: Another solutions may be online algorithms which use only one observation (or small blocks of observations) at a time. Updating formulas and incemental algorithms are needed.

  Robust ensemble methods: Another solution may be to again divide the data into subsets, fit a model to each subset and then use this ensemble of models similar to bagging or boosting. Of course this would need robust ways of model averaging (none of us was aware of any existing work in this direction).

  Outliers: In high-dimensional spaces every observation is an "outlier" even if the number of observations is very large. New defintions of what an oulier is are probably necessary if outliers shall be downweighted in the model fitting process.

  Algebra: Parallel algorithms and sparse linear algebra may help and could offer valuable building blocks.

  Graphics: Traditional scatterplots quickly become useless if too many points overplot each other. Binning and alpha shading are possible solutions.

- We concluded that we are probably not experts enough to write down solutions or specification lists what needs to be done in R. Instead, we decided to experiment a little bit on three of the problems above by writing some actual R code and running some simulations:

pairwise: We implemented a version of the quadrant correlation algorithm that runs the computations for each pair of variables on a different node in a workstation cluster using packages `Rmpi` and `snow`.

ensemble; Fitting robust regression lines to subsets of a large data sets shows that the performace of the different models do vary a lot, and that there are "outlier" models where the model fitting went wrong, i.e., models that are unusually bad. If the single models shall be combined into an ensemble, identification of this outliers will be helpful: We need a definition of outliers in the space of models.

fast-mcd: To compute the minimum covariance determinant estimator, the FAST-MCD algorithm can be modified to accommodate large data sets using virtual sampling.

alpha-shading: Alpha-shading of points in a plot makes the symbols transparent, such that only regions with higher data density are visible. Inverting the schemeon the other hand highlights outliers.

Please have a look at the corresponding R code examples and/or PDF files for details of our findings.

The descriptive statistics/EDA group - minutes

We felt that we were in a difficult position with our group's topic because the field of descriptive statistics and EDA is very broad and though some of us have written code that could be seen to belong to this topic, nobody of the Treviso group members is currently involved in any kind of project concerning the implementation or unification of implementations of any existing methods - we just implemented our own stuff to make it available.
Therefore our group didn't start any continuing projects. Instead, we discussed some basic issues and tried to work through some existing functions and documentation.
First we tried to demarcate our topic "Descriptive Statistics and EDA" somehow. Generally, more or less every technique in statistics can be seen as descriptive or explorative as long as it is not used with a model-based background. This comprises more or less simple summary statistics, smoothing methods, and the entire field of data visualization. In respect to robust statistics, model-free outlier identification rules and sensitivity curves can also be included.
According to Tukey's distinction, the term "robust" is used only in connection with model-based settings while the term "resistant" is preferred for descriptive/EDA techniques, which refers to the change of the results of a method caused by changes in the data. (See e.g. Mosteller & Tukey (1978): Data Analysis und Regression, Addison-Wesley, Chapter 10 on "Robust and Resistant Measures".) Concepts such as minimax asymptotic variance are of course not of interest in non-model-based settings. It is quite difficult to decide whether some methods (especially graphical ones) are qualified to be called "resistant" or "robust".
We then excluded projection techniques such as PCA, which are usually considered under the title "multivariate analysis" from what we tried to discuss.
Quite few resistant techniques are implemented, and they are difficult to find. Descriptive statistics and/or EDA are seemingly not official keywords (though "robust", "smooth" and something like "graphics" are). For example, only a very small part of the techniques suggested in Tukey and co-author's legendary books seem to be implemented.
There is no unification of input/output procedures at all, and standard techniques like the boxplot do not use plot/summary-methods and the like. However, given the variety in the whole field, we agreed that full unification is not desirable.
The best thing to do is perhaps to give some non-binding recommendations such as

- read documentation and code of existing related methods first (this refers to methods serving the same purpose such as smoothing methods, projection methods, visualization of univariate data etc.) and try to make input and output similar,
- use print/plot/summary-methods following the standard conventions whenever possible,
- if your method has an in-built outlier identification, export an outlier identifier or score vector,
- use "robust" in the keywords (and hope that eda or descriptive will be introduced as official keyword, if it applies).