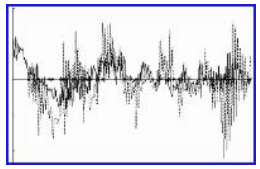
Lehrstuhl
Mathematik VII



Mathematics VII

R-packages for infinitesimal robustness

ICORS 2005

Jyväskylä

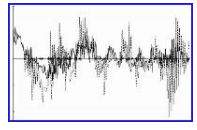
June 13th 2005

Peter Ruckdeschel

Matthias Kohl

E-mail: peter.ruckdeschel@uni-bayreuth.de

matthias.kohl@uni-bayreuth.de



1 "RobASt" - Concept and Organization

1 (a) Motivation

Why R?

- platform independent, open source, easily available
- huge developer & user community
- easily extensible, in particular: OO-capability

Need for OO in providing robust software

- state of the art:
  - lots of neat implementations of robustness concepts—
    in R: 22 packages on CRAN (probably more in the mean time)

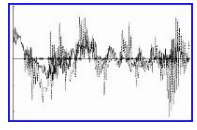
    Design, Hmisc, MASS, PTAk, covRobust, edci, epsi, fdim, forward, fpc, ftnonpar, geoR, multinomRob, ncomplete, noverlap, pheno, quantreg, rqmcmb2, rrcov, sfsmisc, stats, wle

but somewhat disparate and to some extent lack of a unified approach...

UNIVERSITÄT BAYREUTH
Mathematics VII
Peter Ruckdeschel
Matthias Kohl

R-packages for infinitesimal robustness

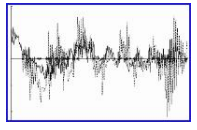
06/13/2005



- advantages of OO in this context
  - general interfaces (c.f. **lm**, **glm**, rlm,) possible
  - by dispatching mechanism on run-time: general code using particularized methods
  - code (may / will) be:
    less redundant, better maintainable, better readable, better extensible

OO concept in S

- S3– and S4-class concept: — confer Chambers[92,98]; in R from 1.7.0 on
- Terminology of Bengtson[03]: *FOOP* vs *COOP*:
  - *function-OOP*: methods belong to *generic functions*
  - *class-OOP*: methods belong to *classes*
- our approach will use both paradigms



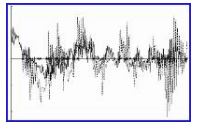
Why an R-bundle for infinitesimal robustness?

- there is no such thing up to now
- infinitesimal robustness provides unified treatment on high abstraction level $\rightsquigarrow$ particularly well suited for OO

1 (b) Organization

(Co-)Authors

- Matthias Kohl: matthias.kohl@uni-bayreuth.de
- [P.R.: peter.ruckdeschel@uni-bayreuth.de]
- Thomas Stabla: statho3@web.de
- Florian Camphausen: fcampi@gmx.de



Organization in packages within RobASt

- distr: provides classes for distributions
- distrEx: provides extensions to distr such as expectation etc.
- RandVar: provides classes for random variables (also for dim>1)
- ROptEst: provides classes for optimally-robust estimation in infinitesimal robustness setup
- RobLox (not here): optimally robust ICs for location and scale
- RobRex (not here): optimally robust ICs for regression and scale
- ROptRegTS (not here): infinitesimal robustness for regression and time series models

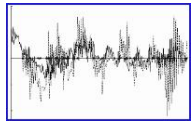
Availability

- distr: published on CRAN; current version 1.5; extensive documentation available (see references)
- RobASt: preliminary version available on
  http://www.staff.uni-bayreuth.de/~btm722/diss/diss.html
  user name: MKohl, password: Diss05

UNIVERSITÄT BAYREUTH
Mathematics VII
Peter Ruckdeschel
Matthias Kohl

R-packages for infinitesimal robustness

06/13/2005



2 distr

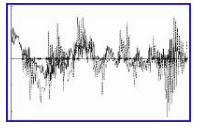
2 (a) Motivation

- realization of distributions in S:
  - highly-skilled implementations for virtually any useful (univariate) distribution
  - naming convention: [prefix]<name>
    [prefix] $\hat{=}$ r, d, p, or q and <name> is the name of the distribution, e.g. norm
- limitation: how to formulate an algorithm once for all distributions?
- works, but. . . : **eval**(**parse** (....))
- possible with distr: — see R-example

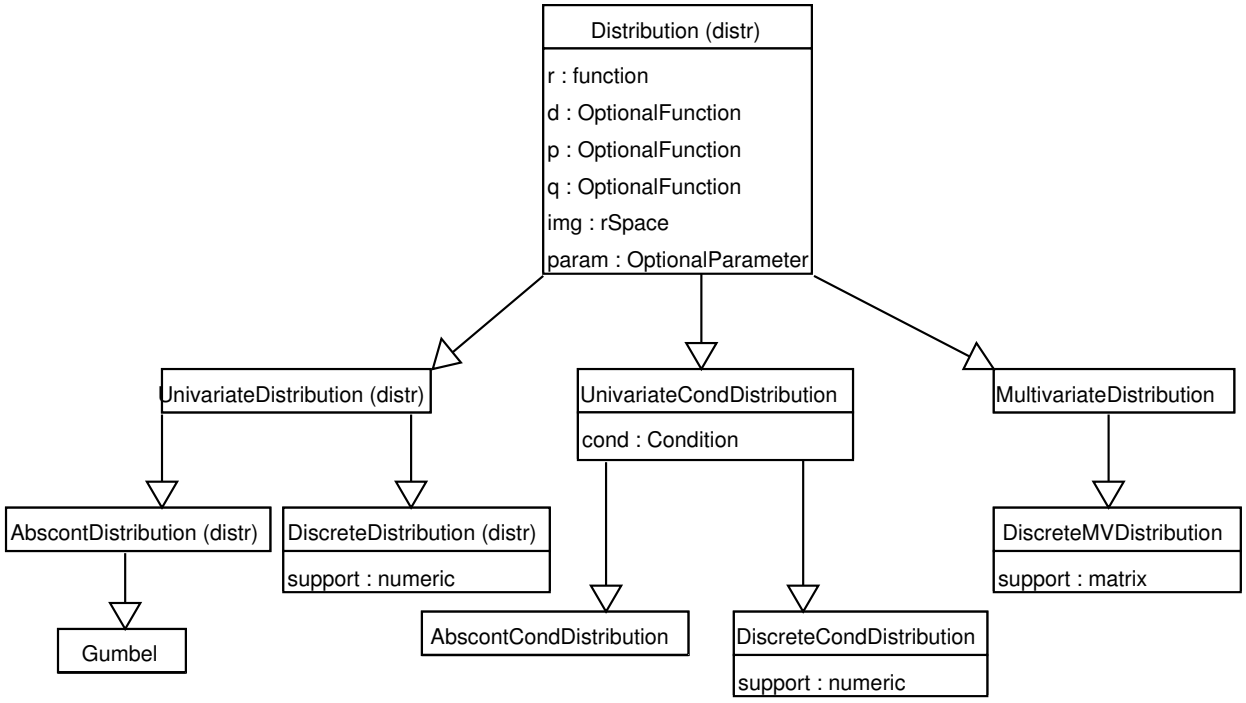
UNIVERSITÄT BAYREUTH
Mathematics VII
Peter Ruckdeschel
Matthias Kohl

R-packages for infinitesimal robustness

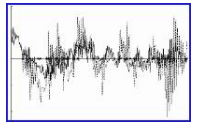
06/13/2005



2 (b) Organization in classes

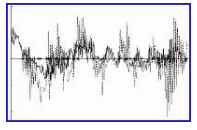


- further subclasses of `AbscontDistribtion`: `Beta`, `Cauchy`, `Chisq`, `Exp`, `Fd`, `Gammad`, `Logis`, `Lnorm`, `Norm`, `Td`, `Unif`, `Weibull` (all from base package)
- further subclasses of `DiscreteDistribtion`: `Binom`, `Dirac`, `Geom`, `Hyper`, `Nbinom`, `Pois` (all from base package)



2 (c) Methods

- overloaded: operators `"+"`, `"-"`, `"*"`, `"/"` — e.g. Y←(3∗X+5)/4
- group `math` of unary mathematical operations is available for objects of class `Distribution` e.g. **exp**(**sin**(3∗X+5)/4)
- `RtoDPQ`: default method for filling slots `d`, `p`, `q` on basis of simulations
- a default convolution method for two independent r.v.'s by means of FFT; c.f. K., R., & Stabla[04]
- particular methods for **plot**, **summary**,...



3 distrEx

3 (a) Functionals for distributions

for a distribution F on $\mathbb{B}^k$, e.g. D1 $\leftarrow$ **Norm**(**mean**=2)

- expectation E of a distribution (with or without transformation); see R-example
- for robust statistics: truncated moments

$$\texttt{m1df(F,t)} := \int \mathrm{I}_{(-\infty;t]}(x)x\,F(dx),$$
$$\texttt{m2df(F,t)} := \int \mathrm{I}_{(-\infty;t]}(x)x^2\,F(dx)$$

- easy but yet to be done: variance, median, MAD, IQR, ...

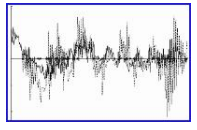
3 (b) Distances between distributions

- ContaminationSize —pseudo-distance:
  $d_c(P,Q) := \inf\{r > 0 \mid \exists\ \text{p.m. } H\colon Q = (1-r)P + rH\}$
- Hellinger-, Kolmogorov-, total-variation-distance
- compare R-example

UNIVERSITÄT BAYREUTH
Mathematics VII
Peter Ruckdeschel
Matthias Kohl

R-packages for infinitesimal robustness

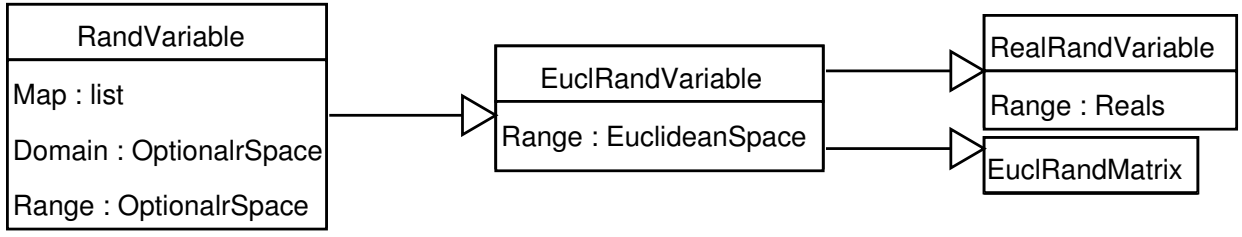
06/13/2005



4 RandVar

4 (a) Random variable as a class concept

- Definition



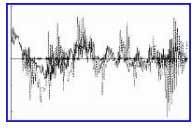
4 (b) Mathematical operations

- there are **many**...
- for `"RealRandVector"`, `"EuclRandVector"` and `"EuclRandMatrix"` all arithmetic und matrix operations are available and are used as with numeric vectors and matrices
- also: group `math` may be applied to it
- further methods: `compatibleDomains`, `dimension`, `E`, `evalRandVar`, `imageDistr`, `length`,...

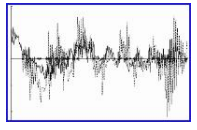
Peter Ruckdeschel
Matthias Kohl

R-packages for infinitesimal robustness

06/13/2005



- accessor und replacement functions for `Map`, `Domain`, `Range`, `Dim`
- e.g. for random variable objects `X` and `Y`, a numerical vector `v` and a matrix `M` (with compatible dimensions), we can generate `exp(X - v)`, `X %*% Y` or `M %*% Y` where "`%*%`" stands for matrix multiplication.



5 ROptEst

5 (a) "Main ingredients" of infinitesimal robustness

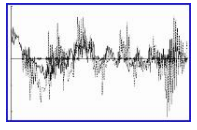
compare Rieder[94], Rieder, K., R.[2K]

- L_2-differentiable model $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$ open
- differentiable parameter transformation $\tau\colon \mathbb{R}^k \to \mathbb{R}^p$, $\tau'(\theta) = D = D(\theta)$
- influence curves (IC): $\eta_\theta \in L_2^p(P_\theta)$ s.t. $\quad \mathrm{E}_{P_\theta}\, \eta_\theta = 0, \quad \mathrm{E}_{P_\theta}\, \eta_\theta \Lambda_\theta^\tau = D.$
- asymptotically linear estimators (ALEs): with expansion
  $$\sqrt{n}\,(S_n - \theta) = \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n \eta_\theta(X_i) + \mathrm{o}_{P_\theta^n}(n^0)$$
  for some IC η_θ
- one-step-estimators to starting estimate θ_0 and IC η
  $$S_n^{(1)} := \tau(\theta_0) + \frac{1}{n}\sum_{i=1}^n \eta_{\theta_0}(X_i)$$
- (shrinking) neighborhood system to radius r given or unknown
- risk

Peter Ruckdeschel
Matthias Kohl

R-packages for infinitesimal robustness

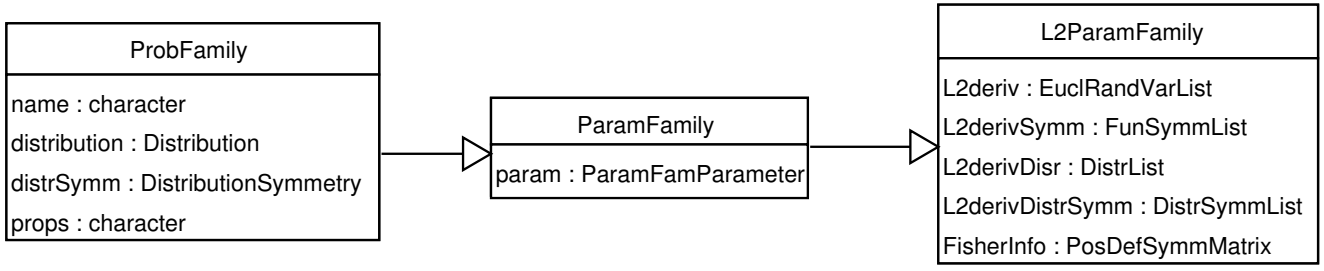
06/13/2005



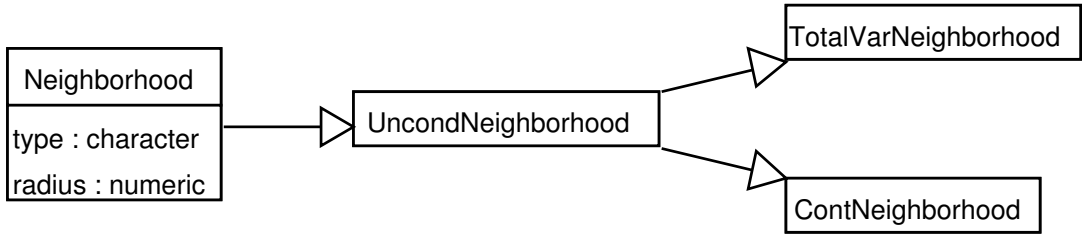
5 (b) Implementation in S4-Classes

Classes

- L_2-differentiable model:



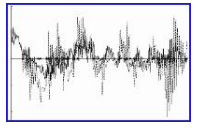
- neighborhood system to some given radius r



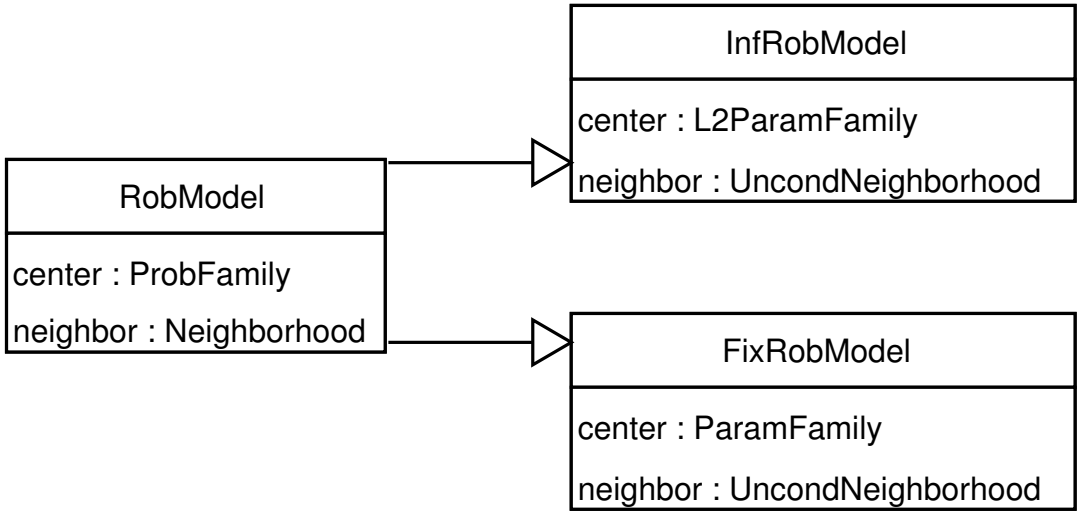
Peter Ruckdeschel
Matthias Kohl

R-packages for infinitesimal robustness

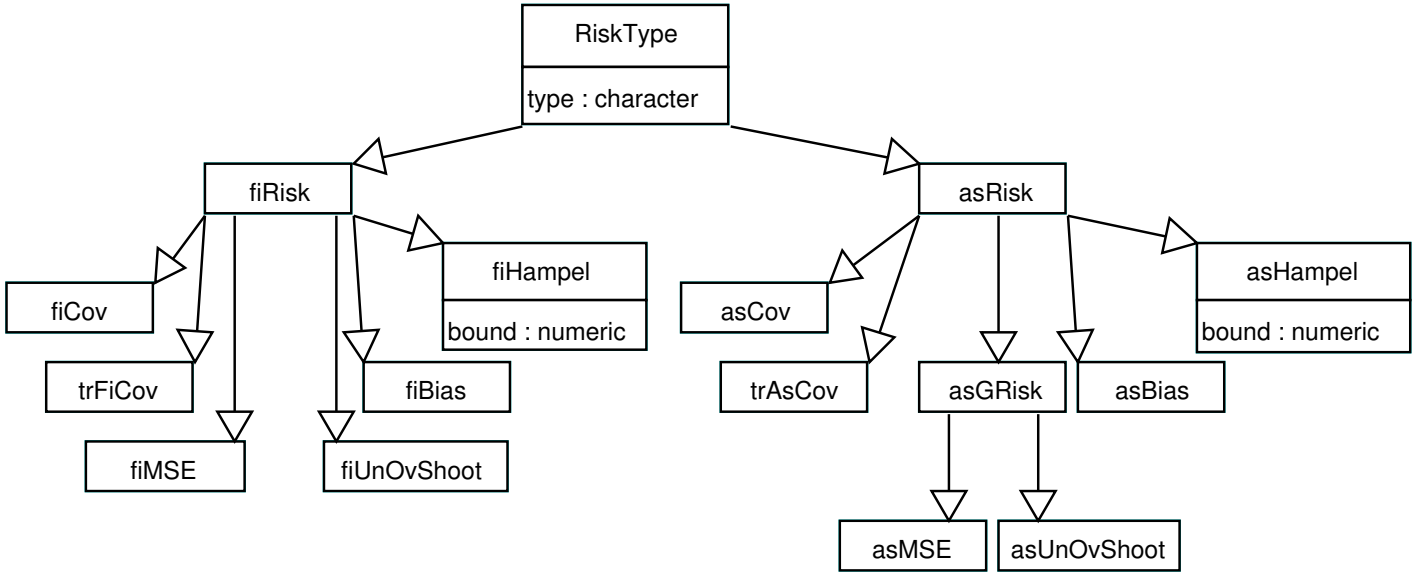
06/13/2005

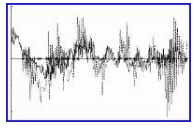


- robust model

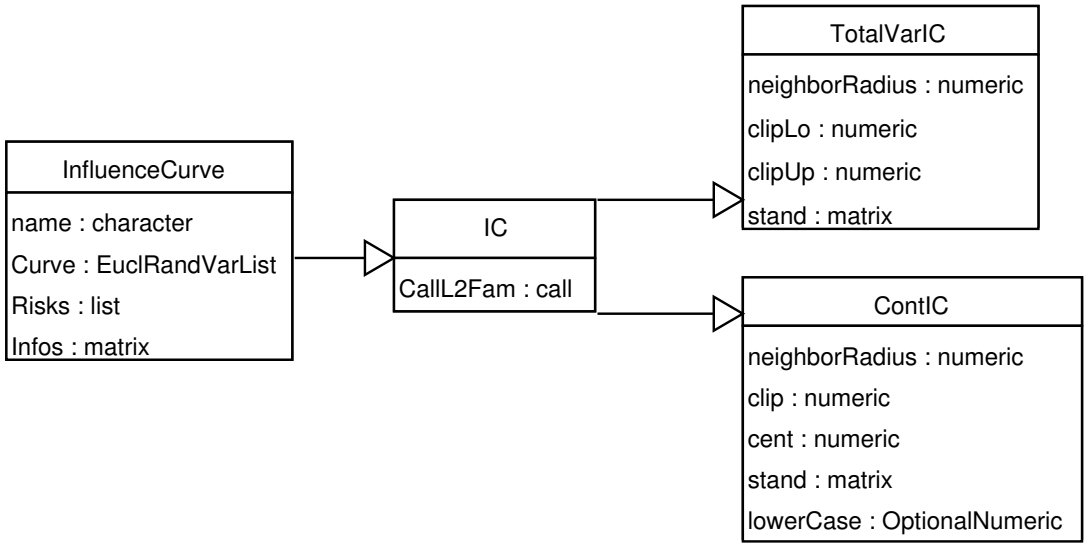


- risk



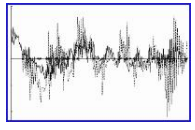


- IC



Methods

- accessor and replacement functions, **show**, **plot**
- `addInfo`, `addProp`, `addRisk`
- `checkL2deriv`, `checkIC`, `evalIC`, `getRiskIC`, `infoPlot`, `ksEstimator`, `leastFavorableRadius`, `locMEstimator`, `oneStepEstimator`, `optIC`, `optRisk`, `radiusMinimaxIC`
- easy generating functions for implemented L_2-families like `NormLocationScaleFamily`, `BinomFamily`



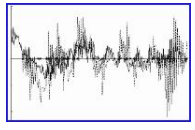
Special meta-information slots

- information gathered during generation of objects is stored in information slots, e.g.

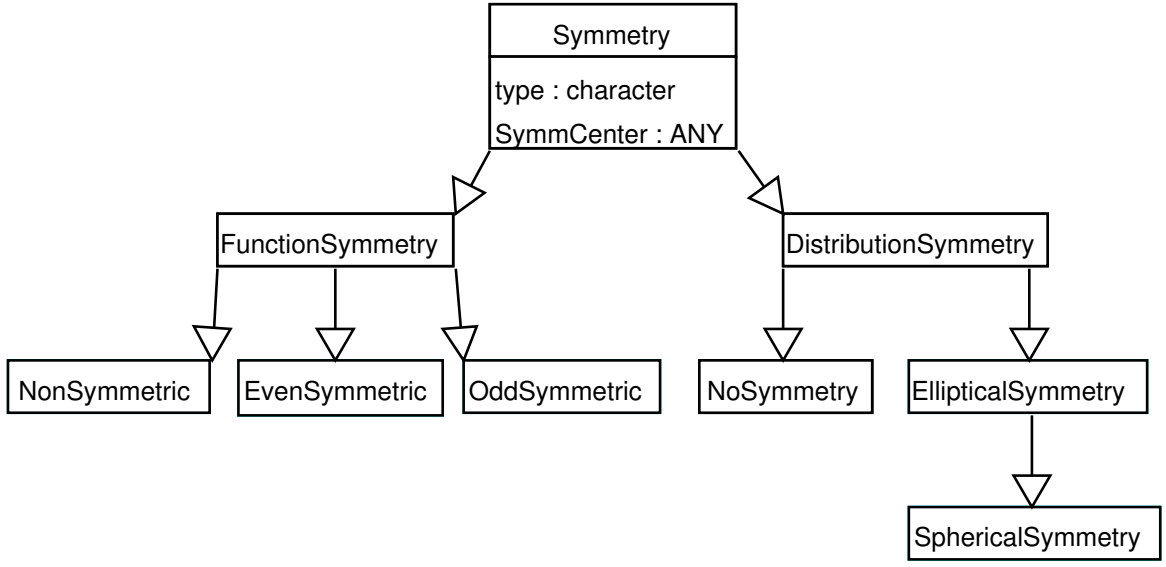
```
### props:
[1] "The normal location and scale family is invariant under"
[2] "the group of transformations 'g(x) = sd*x + mean'"
[3] "with location parameter 'mean' and scale parameter 'sd'"
```

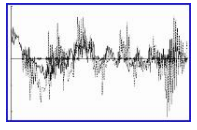
Semi-symbolic calculus

- Situation:
  - we have a certain abstract property for our model (e.g. symmetry)
  - whether this property holds or not cannot be decided (exactly) on basis of numeric evaluations (e.g. convergence?)
  - as a logical statement we can "calculate" with this property and even deduce further properties
  - important for evaluation of high dimensional integrals



- Approach
  - in classical (linear) hierarchical inheritance relations of objects: not clear in which order we should inherit abstract properties...
  - introduce symbolic/logical flags as members(slots) of objects and interfere into dispatching mechanism...
- Realization





1. Estimation of location and scale
   - `X` a contaminated sample from $\mathcal{N}(\texttt{mean}, \texttt{sd}^2)$
   - want to estimate location and scale
   - next steps see `R`-Example
2. Generation of a new L_2-differentiable family:
   - censored Poisson distribution with parameter $\lambda > 0$, i.e. we only observe realizations > 0
   - next steps see `R`-Example

6 Summary/Outlook

6 (a) Summary

covered so far:

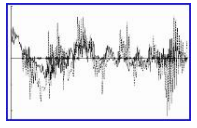
- computation of optimal ICs for all(!) L_2-diff'ble models based on univariate distributions
- Kolmogorov minimum distance estimator as starting estimator

UNIVERSITÄT BAYREUTH

Mathematics VII

Peter Ruckdeschel

Matthias Kohl

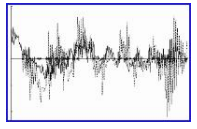


- provide optimally robust estimators by means of one-step constructions

6 (b) Open issues

1. reporting: use of `XML` for the storage of meta-information about generated objects
2. extension of class `RiskType`: getRiskIC
3. mStepEstimator `m = Inf`: $\hat{=}$ iteration until "convergence"
4. further distances
5. use of package `Matrix`
6. Lower case for Dimension > 1
7. one generic method for `ksEstimator`
8. use of S-classes for model formula $\rightsquigarrow$ `rlm` extending `lm` also available for infinitesimal robustness
9. better use of symmetry and group invariances
10. special group generic for invertible operators for the exact determination of image distributions
11. `liesInSupport`: allow for logical operations for slot `'img'` of distributions
12. further functionals for `distrEx`

. . . many more



References

[1] H. Bengtsson. The R.oo package - object-oriented programming with references using standard R code. In K. Hornik, F. Leisch, and A. Zeileis, editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Vienna, Austria, March 2003. Published as http://www.ci.tuwien.ac.at/Conferences/DSC-2003/ .

[2] J. M. Chambers. *Programming with data. A guide to the S language*. Springer, 1998. http://cm.bell-labs.com/stat/Sbook/index.html.

[3] —— and T. J. Hastie. *Statistical Models in S.*. Chapman & Hall, 1992.

[4] M. Kohl. *Numerical contributions to the asymptotic theory of robustness*. Dissertation, Universität Bayreuth, Bayreuth, 2005. preliminary version available under http://www.staff.uni-bayreuth.de/~btm722/diss/diss.html, user name: MKohl, password: Diss05

[5] ——, P. Ruckdeschel, and T. Stabla. General Purpose Convolution Algorithm for Distributions in S4-Classes by means of FFT. Technical report, Feb. 2005. Also available under http://www.uni-bayreuth.de/departments/math/org/ /mathe7/RUCKDESCHEL/pubs/comp.pdf.

[6] H. Rieder. *Robust asymptotic statistics*. Springer Series in Statistics. Springer, 1994.

[7] ——, M. Kohl, and P. Ruckdeschel. The Costs of not Knowing the Radius. Submitted. Appeared as discussion paper Nr. 81. SFB 373 (Quantification and Simulation of Economic Processes), Humboldt University, Berlin; http://www.uni-bayreuth.de/departments/math/org/mathe7/RIEDER/pubs/RR.pdf, September 2001.

[8] P. Ruckdeschel, M. Kohl, T. Stabla, and F. Camphausen. S4 *Classes for Distributions— a manual for package* *distr* version 1.5, March 2005. http://www.uni-bayreuth.de/departments/math/org/mathe7/DISTR/distr.pdf.

[9] ——. S4 Classes for Distributions. Submitted. March 2005. Also available in http://www.uni-bayreuth.de/departments/math/org/mathe7/RUCKDESCHEL/pubs/distr-rnews.pdf.

[10] R Development Core Team. R*: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. (2005) http://www.R-project.org.