REVISTA ESTADISTICA

Vol. 54 - Nros. 162-163 - 2002
Special Issue on Robust Statistics

Contents

Statistical Procedures and Robust Statistics

P.L.Davies
Fachbereich Mathematik und Informatik
Universit (at Essen 45117 Essen, Germany}

Abstract

It is argued that a main aim of statistics is to produce statistical procedures which in this article are defined as algorithms with inputs and outputs. The structure and properties of such procedures are investigated with special reference to topological and testing considerations. Procedures which work well in a large variety of situations are often based on robust statistical functionals. In the final section some aspects of robust statistics are discussed again with special reference to topology and continuity.

Robust Tests in Statistics - a Review of the Past Decade

Xuming He
Department of Statistics
University of Illinois
Champaign, IL 61820, USA

Abstract

The purpose of this article is to provide an up-to-date review of the literature on robust hypothesis testing. While much of the activities in robust statistics have been concerned with estimation, the work on testing seems to deserve wider attention and better promotion. Although this review is by all means partial and at times subjective, the author hopes that it will encourage and facilitate new minds to this area of research.

Robust parametric means of asymmetric distributions: estimation and testing

A. Marazzi and Giulia Barbati
Institut de mèdecine sociale et préventive, Université de Lausanne

Abstract

A robust parametric mean is the mean of a robustly estimated parametric model. Here we review a family of robust means for location-scale and shape-scale univariate models such as Lognormal, Weibull, Gamma, and Pareto distributions. The procedures are based on three steps: first the model is adjusted using a high breakdown-point estimator; second, outliers with respect to the initial estimate are rejected; third, an efficient estimate of mean is computed with the remaining data. In addition, the final estimate is corrected to make it Fisher consistent. The procedures include the truncated mean and the truncated maximum likelihood estimate. We also discuss the use of the bootstrap for the computation of the finite sample distribution of a robust mean and consider the “robust bootstrap" as a tool for improving the robustness of the approximation. For the problem of testing hypotheses concerning robust means we describe methods to estimate the models under the null hypotheses. We detail the two-sample case and provide examples with real data.

Robustness and Heavy-Tailed Distributions

Stephan Morgenthaler
FSB -- Institut de mathématiques
Ecole polytechnique fédérale de Lausanne (EPFL)
1015 Lausanne
Switzerland

Abstract

The paper describes equivariant estimator having efficiencies that are poly-optimal within finite sets of distributional shapes. For simplicity and because robustness theory is most highly developed in this situation, the pure location case is considered. Both point estimation and interval estimation are discussed and computational ideas involved in “configural polysampling" are also mentioned.

On Smoothing the Corrected Content of Tolerance Intervals

Luisa Turrin Fernholz
Temple University
Philadelphia, PA 19122

Abstract

This article first reviews the content-corrected method for tolerance limits proposed by Fernholz and Gillespie (2001) and extended by Fernholz (2002) to include robust statistics for the end-points. The content-corrected method for *k*-factor tolerance limits consists of obtaining a bootstrap corrected value *p** that is robust in the sense of preserving the confidence coefficient for a variety of distributions. The article also addresses the issue of obtaining a more accurate content-corrected value by using a method based on the smoothed bootstrap. The smoothed version $\tilde{p}^*$ of the corrected content is obtained by using a kernel smoothed empirical distribution function and smoothed bootstrap samples to approximate quantile values. By smoothing, the Hadamard differentiability and the bootstrap consistency of the smoothed corrected content are preserved for large samples. The robustness of the method is studied through the influence function and examples and simulations showing the advantage of using smoothed and robust alternatives are included.

Globally Robust Inference

Jorge Adrover
Univ. Nacional de Cordoba and CIEM, Argentina

Jose Ramon Berrendero
Universidad Autonoma de Madrid, Spain

Matias Salibian-Barrera
Carleton University, Canada

Ruben H. Zamar
University of British Columbia, Canada

Abstract

The robust approach to data analysis uses models that do not completely specify the distribution of the data, but rather assume that this distribution belongs to a certain neighborhood of a parametric model.
Consequently, robust inference should be valid under all the distributions in these neighborhoods. Regarding robust inference, there are two important sources of uncertainty: (i) sampling variability and (ii) bias caused by outlier and other contamination of the data. The estimates of the sampling variability provided by standard asymptotic theory generally require assumptions of symmetric error distribution or alternatively known scale.
None of these assumptions are met in most practical problems where robust methods are needed. One alternative approach for estimating the sampling variability is to bootstrap a robust estimate. However, the classical bootstrap has two shortcomings in robust applications. First, it is computationally very expensive (in some cases unfeasible). Second, the bootstrap quantiles are not robust. An alternative bootstrap procedure overcoming these problems is presented. The bias uncertainty is usually ignored even by robust inference procedures. The consequence of ignoring the bias can result in true probability coverage for confidence intervals much lower that the nominal ones. Correspondingly, the true significance levels of tests may be much higher than the nominal ones. We will show how the bias uncertainty can be dealt with by using maximum bias curves, obtaining confidence interval and test valid for the entire neighborhood. Applications of these ideas to location and regression models will be given.

Computing LTS regression for large data sets

Peter J. Rousseeuw
Department of Mathematics and Computer Science, Universitaire Instelling Antwerpen (UIA), Universiteitsplein 1, B-2610 Wilrijk, Belgium
and
Katrien Van Driessen
Faculty of Applied Economics UFSIA-RUCA, University of Antwerp, Prinsstraat 13, B-2000 Antwerp, Belgium

Abstract

Data mining aims to extract previously unknown patterns or substructures from large databases. In statistics, this is what methods of robust estimation and outlier detection were constructed for (see, e.g., Rousseeuw and Leroy, 1987). Here we will focus on Least Trimmed Squares (LTS) regression, which is based on the subset of *h* cases (out of *n*) whose least squares fit possesses the smallest sum of squared residuals. The coverage h may be set between *n*/2 and *n*. The computation time of existing LTS algorithms grows too much with the size of the data set, precluding their use for data mining. In this paper we develop a new algorithm called FAST-LTS. The basic ideas are an inequality involving order statistics and sums of squared residuals, and techniques which we call ‘selective iteration’ and ‘nested extensions'. We also use an intercept adjustment technique to improve the precision. For small data sets FAST-LTS typically finds the exact LTS, whereas for larger data sets it gives more accurate results than existing algorithms for LTS and is faster by orders of magnitude. This allows us to apply FAST-LTS to large databases.

Robust estimates for high dimensional data based on projections

Ricardo A. Maronna
University of La Plata and C.I.C.P.B.A

Abstract

High breakdown point estimates of multivariate location and scatter are reviewed. Most of them can be classified into those based on the minimization of a scale of the Mahalanobis distances. and those based on the projections of the data points. The projection approach has yielded the first high breakdown point equivariant estimate, estimates with a maximum bias not depending on dimension, and fast estimates for high dimensional data.

High Breakdown Point Multivariate M-Estimation

David E. Tyler
Rutgers University

Abstract

In this paper, a general study of the existence and breakdown points of the M-estimates of multivariate location and scatter with auxiliary scale is presented.
The multivariate M-estimates with auxiliary scale include as special cases the minimum volume ellipsoid estimates, the multivariate S-estimates, the multivariate constrained M-estimates, and the recently introduced multivariate MM-estimates. A related estimate, the multivariate τ-estimate, is also treated. The results presented here for the multivariate S-estimates, the multivariate constrained M-estimates and the multivariate τ-estimates are mainly intended to serve as a unifying review.
The results presented for the multivariate MM-estimates, though, are entirely new.
The breakdown points of the redesecending M-estimates of multivariate location and scatter for fixed scale are also derived. This result generalizes the results on the breakdown points of the univariate redescending M-estimates of location with fixed scale given by Huber (1984) to the multivariate setting. As in the univariate setting,
The breakdown point of a redescending multivariate M-estimate depends on the specific distributional model, and can be arbitrary close to 1/2 regardless of the dimension of the data.

A robust approach to partly linear autoregressive models

Ana Bianco
Universidad de Buenos Aires, Argentina
Graciela Boente
Universidad de Buenos Aires and CONICET, Argentina

Abstract

This paper first reviews existing procedures to estimate the autoregression function through a kernel *p*-dimensional smoother and recalls their properties under different mixing conditions. Both linear and M-smoothers are considered. In the last years, to solve the curse of dimensionality, there has been an increasing interest in the area of partly linear models. This article also provides an up-to-date presentation of the existing literature on partly linear autoregression. The sensitivity to outliers of the classical estimates for these models is good evidence that robust methods are needed. The problem of obtaining a family of robust estimates, in a partly linear autoregression model, is then addressed introducing a three-step robust procedure. Through a Monte Carlo study, the performance of the proposed estimates is compared with the classical ones. This study shows the advantage of considering the three-step robust estimates based on nearest neighbor with kernel M-smoothers.

The Identification Of Multiple Outliers In Online Monitoring Data

Ursula Gather, Marcus Bauer, Roland Fried
Department of Statistics
University of Dortmund
44221 Dortmund, Germany

Abstract

We present a new procedure for routine detection of isolated and patchy outliers in univariate time series. This procedure is especially designed for online identification of outliers, but of course it can be used retrospectively, too. It is based on an embedding of the time series which allows to regard the time series as a multivariate sample with identically distributed but non independent observations. Thus, multivariate outlier identifiers can be transferred into the context of time series. This gives interesting insights in some features of outliers in time dependent data, which are not recognizable by other methods.
Some applications to online monitoring data from intensive care are included.