Preface

It has now been eleven years since the publication of the first edition of "Robust Statistics: Theory and Methods" in 2006. During that time period, there have been two developments prompting the need for a second edition. The first development is that since 2006 a number of new robust statistics theory and methods research results have been developed and published by several researchers, in particular the book´s authors. The second development is that S-PLUS has been supplanted by open source R, rendering our original focus on use of the S-PLUS robust package in support of the book no longer tenable. So for this second edition we have created a new R based package RobStatTM in support of the book, and at the publisher web site provide scripts for computing all the examples in the book. We now discuss the main research advances included in this second edition.

Finite-sample robustness

Asymptotically normal robust estimators have tuning constants that allow users to control their normal distribution variance efficiency, in a trade-off with robustness toward non-normal distributions. The resulting finite-sample performance in terms of mean-squared error (MSE), that takes into account bias as well as variance, can be considerably worse than their asymptotic performance implies. This second edition contains useful new results concerning finite-sample MSE performance of robust linear regression and robust covariance estimators that is briefly described below.

Linear regression estimators

Section 5.9.3 focuses on finite sample efficiency and robustness and introduces a new "distance-constrained maximum-likelihood" (DCML) estimator. The DCML estimator is shown to provide the best trade-off between finite-sample robustness and normal distribution efficiency in comparison with an MM estimator that is asymptotically 85% efficient, and an adaptive estimator described in Section 5.9.2 that is asymptotically fully efficient for normal distributions.

Multivariate location and scatter

There exist a number of proposed robust covariance matrix estimators that were discussed in the first edition, and some comments of choice of estimator were made. In

this second edition, a new Section 6.10 "Choosing a Location/Scatter Estimator" replaces the previous Section 6.8, and this new section provides new recommendations for choosing a robust covariance matrix estimator, based on extensive finite-sample performance simulation studies.

Fast and reliable starting points for initial estimators

The standard starting point for computing initial S-estimators for linear regression and covariance matrix estimators is based on a subsampling algorithm. Subsampling algorithms have two disadvantages: the first is that their computation time increases exponentially in the number of variables. The second disadvantage of the subsampling method is that the method is stochastic, which means that different final S-estimator and MM-estimator can occur when the computation is repeated.

Linear regression

Section 5.7.4 describes a deterministic algorithm due to Peña and Yohai (1999) for obtaining a starting point for robust regression. Since this algorithm is deterministic it always yields the same final MM-estimator. This is particularly important in some applications, e.g., in financial risk calculations. Furthermore, it is shown in Section 5.7.6 that the Peña-Yohai starting value algorithm is much faster than the subsampling method, and has smaller maximum MSE than the subsampling algorithm, sometimes substantially so.

Multivariate location and scatter

Subsampling methods have also been used to get starting values for robust estimators of location and dispersion (scatter), but they have a similar difficulty as in linear regression, namely they will be too slow when the number of variables is large. Fortunately, there is an improved algorithm for computing starting values due to Peña and Prieto (2007) that makes use of finding projection directions of maximum and minimum kurtosis plus a set of random directions obtained by a "stratified sampling" procedure. This method, which is referred to as the KSD method, is described in Section 6.9.2. While the KSD method is still stochastic in nature, it provides fast reliable starting values, and is more stable than ordinary subsampling, as is discussed in Sections 6.10.2 and 6.10.3.

Robust regularized regression

Penalized regression estimators to obtain good results for high-dimensional but sparse predictor variables has been a hot topic in the "Machine Learning" literature in the last decade or so. These estimators add L_1 and L_2 penalties to the least squares objective function, and the leading estimators of this type are Lasso regression, Least Angle Regression, and Elastic Net regression, among others. A new section on robust regularized regression describes how to extend robust linear model regression to obtain robust versions of the above type of non-robust least-squares based regularized regression estimators.

Multivariate location and scatter estimation with missing data

A new Section 6.12 provides a method for solving the problem of robust estimation of scatter and location with missing data. The method contains two main components. The first is the introduction of a generalized S-estimator of scatter and location that depends on Mahalanobis distances for the non-missing data in each observation. The second component is a weighted version of the well-known expectation-minimization (EM) algorithm for missing data.

Robust estimation with independent outliers in variables

The Tukey-Huber outlier-generating family of distribution models has been a commonly accepted standard model for robust statistics research and associated empirical studies for independent and identically distributed data. In the case of multivariate data, the Tukey-Huber model describes the distribution of the rows, i.e., "cases", of a data matrix whose columns represent variables, and outliers generated by this model are "case outliers". However, there are important problems where outliers occur independently across cells, i.e., across variables, in each row of a data matrix. For example with portfolios of stock returns, where the columns represent different stocks and the rows represent observations at different times, outlier returns in different stocks (representing idiosyncratic risk) occur independently across stocks, i.e., across cells/variables.

Section 6.13 discusses an important and relatively new outlier generating model for independent outliers across cells (across variables), called the *independent contamination* (IC) model. It turns out that estimators that have good robustness properties under the Tukey-Huber model are shown to have very poor robustness properties for the IC model. For example, estimators that have high breakdown points under the Tukey-Huber model can have very low breakdown point under the IC model. This section surveys the current state of research on robust methods for IC models, and on robust methods for simultaneously dealing with outliers from both Tukey-Huber and IC models. The problem of obtaining robust estimators that work well for both Tukey-Huber and IC models is an important ongoing area of research.

Mixed linear models

Section 6.15 discusses robust methods for mixed linear models. Two primary methods are discussed, the first of which is an S-estimator method that has good robustness properties for Tukey-Huber model case-wise outliers, but does not perform well for cell-wise independent outliers. The second method is designed to do well for both types of outliers, and achieves a breakdown point of 50% for Tukey-Huber models and 29% for IC models.

Generalized linear models

New material on a family of robust estimators has been added to the chapter on generalized linear models (GLM's). These estimators are based on using M-estimators after a variance-stabilizing transformation is applied to the response variable.

Regularized robust estimators of the inverse covariance matrix

In the Multivariate Analysis Chapter 6, a short Section 6.14 on regularizing robust estimators of inverse covariance matrices in cases where is close to or larger than one.

A note on software

The section "Recommendations and software" at the end of each chapter indicates the procedures recommended by the authors and the R functions that implement them and other procedures described in the chapter. These functions are located in several libraries, in particular the library `RobStatTM`, which was especially developed for this book; and are all available in the CRAN network.

The R scripts and datasets that enable the reader to reproduce the book's examples are available at the book's web site. Each dataset has the same name as the respective script. The reader is adviced to download all scripts and datasets in the same folder.