

Robust tests for testing the parameters of a normal population

A. Mahdavi^{*,1}, M. Towhidi²

¹ Department of Statistics, Faculty of Mathematical Sciences, Vali-e-Asr University of Rafsanjan, Rafsanjan, Islamic Republic of Iran

² Department of Statistics, College of Sciences, Shiraz University, Shiraz, Islamic Republic of Iran

Received: 20 September 2013 / Revised: 11 May 2014 / Accepted: 10 August 2014

Abstract

This article aims to provide a simple robust method to test the parameters of a normal population by using the new diagnostic tool called the “Forward Search” (FS) method. The most commonly used procedures to test the mean and variance of a normal distribution are Student’s t test and Chi-square test, respectively. These tests suffer from the presence of outliers. We introduce the FS version of these tests that is not affected by the outliers. The performances of these procedures are illustrated by some simulation studies and a real data example.

Keywords: Forward search procedure; Robust approach; Student’s t test; Chi-square test; Outlier.

Introduction

All statistical methods require a number of assumptions explicitly or implicitly. These assumptions are used in data analysis or statistical modeling problem. It often happens in practice that an assumption such as normality approximately holds true for the majority of observations, but some observations follow a different pattern or no pattern at all. Such atypical data are called outliers. Regarding the disturbing effect of outliers, it should be mentioned that even a single outlier can have a large disturbing effect on a classical statistical method that is optimal under the classical assumption. The robust approach to statistical modeling and data analysis provides methods that are not unduly affected by outliers or other small departures from model assumptions. The Forward Search (FS) approach is a powerful method for investigating the effect of outliers.

The purpose of this article is to adopt the FS method proposed by Atkinson and Riani [4] to test hypotheses

regarding the mean and the variance of a normal distribution when these parameters are unknown. The most frequently used tests under classical assumptions for these hypotheses are Student’s t test and Chi-square test, respectively. Although these tests are uniformly most powerful unbiased, but the presence of outliers influences these tests strongly since the sample mean and variance are not robust statistics.

The FS approach is a powerful general method that provides diagnostic plots for finding outliers and determining their effect on the fitted models. The FS method starts from a small, robustly chosen subset of the data and increases the subset size until all the data are fitted. The outliers enter the model in the last steps and their entrance point can be determined by monitoring some statistics of interest during the process. Initially, Hadi [15] and Atkinson [2] presented the method of fitting a model to subsets of an increasing size for multivariate data analysis. The FS was used in regression by Hadi and Simonoff [16], and its

* Corresponding author: Tel: +983913202365; Fax: +983913202294; Email: a.mahdavi@vru.ac.ir

development for regression was introduced by Atkinson and Riani [3,4]. Also, Atkinson et al. [5] used the FS method for multivariate procedure. Other scholars such as Bertaccini and Varriale [8], Coin [13], and Mavridis and Moustaki [18] implemented the FS procedure for the ANOVA framework, to test normality and to identify atypical subjects in factor analysis models, respectively. Cheng and Biswas [11] developed this method for the analysis of mixed continuous and categorical data. Solaro and Pagani [20] used the FS in the context of multidimensional scaling and Bellini [7] in detecting atypical observations in financial data. Regarding the use of FS procedure for estimation proposes, Grossi and Laurini [14] use it in robust estimation of efficient mean–variance frontiers, Bastero and Barrios [6] in robust estimation of a spatiotemporal model with structural change and Campano and Barrios [10] in robust estimation of a time series model with structural change.

This paper aims to introduce an original procedure to investigate the effect of outliers on testing hypotheses with regard to the mean and the variance of a normal distribution. Our aim is to determine how many and which observations agree with the null hypothesis. In this procedure we try not only to identify the outlying observations but also to analyse their effect on the inferences about the parameters of a normal population.

The paper is organized as follows. In the next section, we briefly review the most commonly used tests for testing the parameters of a normal distribution. After that, the forward search algorithm is presented for these tests, and then the performance of the method is illustrated with some simulated data and the behavior of our procedure is analysed. Finally, the application of the proposed approach to real data is shown using the blood clotting data set.

Testing hypotheses for unknown parameters of a normal distribution

Because of their wide applicability, the problems of testing the mean μ and variance σ^2 of a normal distribution are of particular importance. Let X_1, X_2, \dots, X_n be a random sample from $N(\mu, \sigma^2)$ and consider the hypotheses $H_1: \mu = \mu_0$ and $H_2: \sigma^2 = \sigma_0^2$ against the alternatives $K_1: \mu \neq \mu_0$ and $K_2: \sigma^2 \neq \sigma_0^2$, respectively. The test statistics of uniformly most powerful unbiased (UMPU) tests for testing H_1 and H_2 are

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S}, \tag{1}$$

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}, \tag{2}$$

respectively, where $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is the sample mean and $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is the sample variance. The UMPU test of H_1 has the acceptance region ($|T| \leq c$) and the one for H_2 has the acceptance region ($c_1 \leq \chi^2 \leq c_2$).

Forward search in testing hypotheses about the parameters of a normal distribution

In this section the FS method has been used not only to detect outliers, but also to investigate the effect of outliers on the statistics T and χ^2 in testing the mean and variance of a normal distribution. The FS method has three steps. These steps include choosing outlier free subset of all observations, presenting a plan to progressing in FS, and monitoring statistics during the search, respectively. An exhaustive explanation of the adoption of these three steps in our work has been presented.

1. Step 1: Choice of the initial subset

The starting point of the FS procedure is to choose outlier free subset of observations robustly. If the vector of observations $\mathbf{x} = (x_1, x_2, \dots, x_n)$ comes from a normal distribution with unknown parameters μ and σ^2 , we can write a regression model for each ordered observation as

$$x_{(i)} = \mu + \sigma \alpha_i + \varepsilon_i, \tag{3}$$

where $x_{(i)}$ denotes i -th ordered observation and α_i is the expected value of the i -th standard ordered statistics.

The two parameters of (3) can be estimated using the Generalized Least Squares estimator (GLS) (see [1] or [12]). To find an outlier free subset to be used as starting point of the forward procedure, we use robust regression estimation, Least Median of Squares (LMS) proposed by Rousseeuw [19] instead of GLS. If $\boldsymbol{\beta} \in \mathbb{R}^p$ denotes the vector of parameters in the classical linear regression model

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i \quad i = 1, 2, \dots, n, \tag{4}$$

where $y_i \in \mathbb{R}, \mathbf{x}_i \in \mathbb{R}^p$ and ε_i is the error term, then the LMS estimator for $\boldsymbol{\beta}$ is defined as

$$\hat{\boldsymbol{\beta}}_{\text{LMS}} = \min_{\boldsymbol{\beta}} \text{med } e_i^2 \tag{5}$$

where e_i denotes the i -th residual

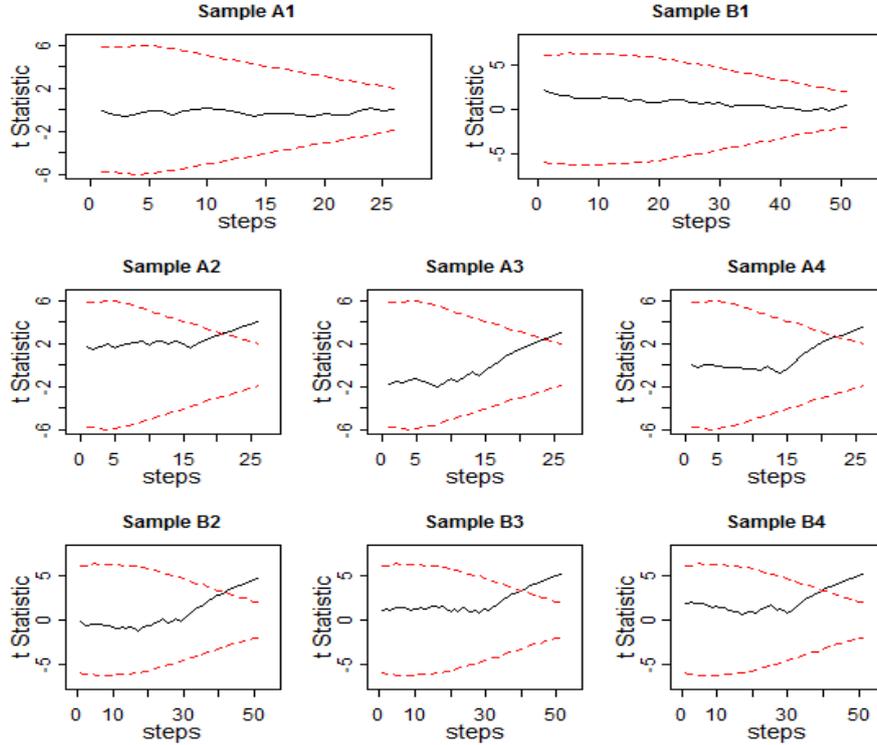


Figure 1. Forward plots of \mathbf{T}_{FS} during the search for samples A1-A4 and B1-B4. The dashed lines refer to 2.5% and 97.5% empirical quantiles of \mathbf{T}_{FS} .

$$e_i = y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}} \quad i = 1, 2, \dots, n. \quad (6)$$

After estimating the parameters of model (3) by LMS estimation method, the estimated expected value for $x_{(i)}$ is of the form

$$\hat{x}_{(i)} = \hat{\mu}_{LMS} + \hat{\sigma}_{LMS} \alpha_i \quad ; i = 1, 2, \dots, n. \quad (7)$$

Let $r_i = |x_{(i)} - \hat{x}_{(i)}|$ be, the i -th absolute residual resulting from (7). The elements of $\mathbf{x}_{(i)} = (x_{(1)}, x_{(2)}, \dots, x_{(n)})$ are reordered based on the values of r_i , and this new vector of reordered observations is denoted by $\mathbf{x}_{(LMS)}$.

To start the FS approach, the size of initial subset must be specified. The breakdown point of (5) is 50%. Besides, a larger initial subset will give more stable estimates and smoother forward plots, therefore, we set the size of initial subset $[(n + 1)/2]$. We start the process with the first $[(n + 1)/2]$ observations of $\mathbf{x}_{(LMS)}$ and denote this subset by $S^{(*)}$.

2. Step 2: Adding observations during the FS

After choosing the initial subset $S^{(*)}$, in the $n - [(n + 1)/2]$ remaining steps of FS, all observations must be added to $S^{(*)}$. At each step, the observations

closer to the previously fitted model are added to the subset $S^{(*)}$. Since we use a robust method to estimate the parameters of (3), it is not necessary to reorder the observations $\mathbf{x}_{(LMS)}$ at each step of the search. Therefore, in each step of the search we add the next observation of $\mathbf{x}_{(LMS)}$ to the previously chosen subset.

3. Step 3: Monitoring the search

For detecting and determining the effect of outliers some statistics of interest must be monitored during the search. Let $S^{(m)}$ be the subset of the first m observations of $\mathbf{x}_{(LMS)}$. The collections of T statistics and χ^2 statistics computed for different subsets $S^{(m)}$ during the FS procedure are defined as follows

$$\mathbf{T}_{FS} = (T_{S^{(*)}}, \dots, T_{S^{(m)}}, \dots, T_{S^{(n)}}). \quad (8)$$

$$\boldsymbol{\chi}_{FS}^2 = (\chi_{S^{(*)}}^2, \dots, \chi_{S^{(m)}}^2, \dots, \chi_{S^{(n)}}^2). \quad (9)$$

For comparing the values of \mathbf{T}_{FS} and $\boldsymbol{\chi}_{FS}^2$ with corresponding bounds of the rejection region, the empirical quantiles of these statistics must be estimated by a simulation study in each step of the search.

Simulation study

Simulation studies are conducted to consider the behavior of the proposed statistics (8) and (9) in the presence of outliers and the ability of FS to detect them. We generate some samples in different ways and consider the behavior of the test statistics T and χ^2 in testing $H_1: \mu = 0$ and $H_2: \sigma^2 = 1$ during the search. In this study the effect of the outliers is clearly revealed.

1. Example 1

Consider eight samples generated in the following way:

- Sample A1: 50 observations are generated from a standard normal distribution.
- Sample A2: 40 observations are generated from standard normal distribution and 10 observations are generated from normal distribution with parameters $\mu = 5, \sigma^2 = 1$ for contamination.
- Sample A3: 40 observations are generated from a standard normal distribution and 10 observations are generated from normal distribution with parameters $\mu = 5, \sigma^2 = 0.5$ for contamination.
- Sample A4: 40 observations are generated from standard normal distribution and 10 observations are generated from uniform distribution with parameters $a = 10, b = 11$ for contamination.
- Sample B1: 100 observations are generated from a

standard normal distribution.

- Sample B2: 80 observations are generated from standard normal distribution and 20 observations are generated from normal distribution with parameters $\mu = 5, \sigma^2 = 1$ for contamination.
- Sample B3: 80 observations are generated from a standard normal distribution and 20 observations are generated from normal distribution with parameters $\mu = 5, \sigma^2 = 0.5$ for contamination.
- Sample B4: 80 observations are generated from standard normal distribution and 20 observations are generated from uniform distribution with parameters $a = 10, b = 11$ for contamination.

Fig.1 shows how the T statistic of subsets $S^{(m)}$ changes during the search for samples A1-A4 and B1-B4. For clean samples A1 and B1 these values lie between the 2.5% and 97.5%, empirical quantiles (dashed lines) of this statistic, and there is no jump in their plots. The effect of outliers on T_{FS} for strongly contaminated samples and the extreme change of T value after steps 15 and 30 corresponding to the samples A2-A4 and B2-B4, respectively, can be observed. Also, Fig.1 is indicative of the cases A2-A4 where the null hypothesis H_0 is rejected from step 20 onwards and cases B2-B4 where null hypothesis is rejected from step 40 onwards.

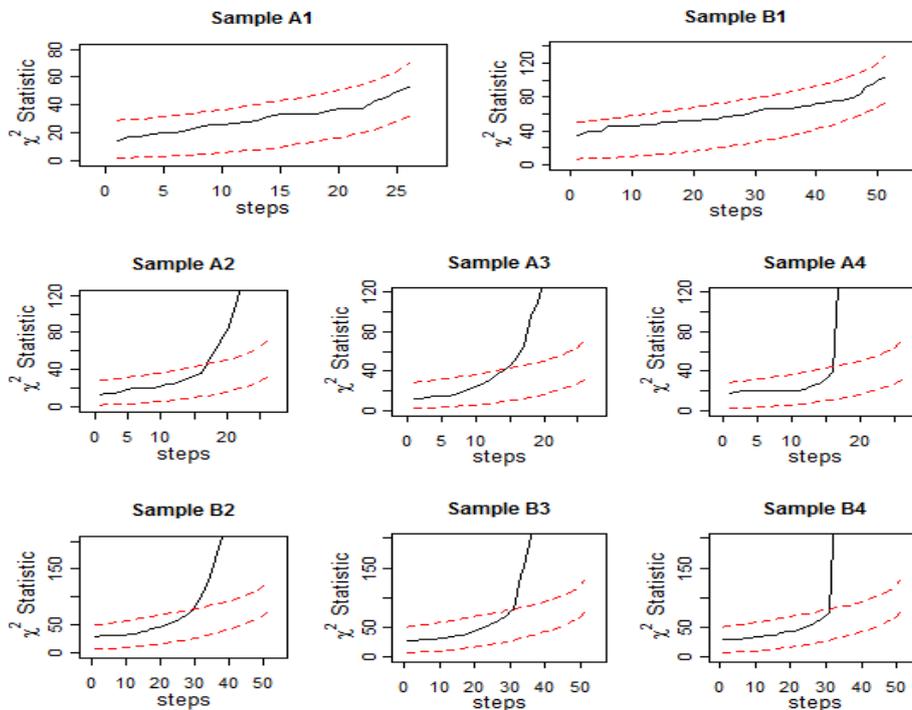


Figure 2. Forward plots of χ^2_{FS} during the search for samples A1-A4 and B1-B4. The dashed lines refer to 2.5% and 97.5% empirical quantiles of χ^2_{FS} .

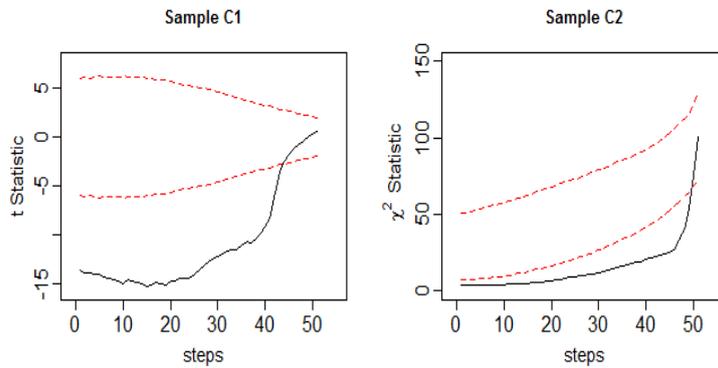


Figure 3. Forward search results for sample C1 and C2. The dashed lines refer to 2.5% and 97.5% empirical quantiles of statistics.

The behavior of χ_{FS}^2 during the FS for samples A1-A4 and B1-B4 has been plotted in Fig 2. Since the search proceeds by adding observations which have small residuals, the value of χ^2 statistics increases by adding each further unit. The fact that the plots of these statistics for clean samples A1 and B1 are smooth indicates that there are no important outliers in these samples. But, there is a big jump in samples A2-A4 and B2-B4 in steps 15 and 30, respectively, indicating the entrance points of outliers in these steps. The null hypothesis about the variance is accepted in all steps of the search for clean samples A1 and B1, and it's strictly rejected after steps 15 and 30 for strongly contaminated samples A2-A4 and B2-B4, respectively. This study shows that the statistic χ^2 is more sensitive to outliers than the statistic T .

2. Example 2

The outliers can increase type II error of the test and lead to an incorrect decision. The FS approach can show this effect of outliers. Here we simulate two samples to emphasize this issue.

- Sample C1: 90 observations are generated from a normal distribution with parameters $\mu = -1, \sigma^2 = 1$ and 10 observations are generated from normal distribution with parameters $\mu = 10, \sigma^2 = 1$ for contamination.

- Sample C2: 95 observations are generated from a normal distribution with parameters $\mu = 0, \sigma^2 = 0.5$ and 5 observations are generated from normal distribution with parameters $\mu = 4, \sigma^2 = 1$ for contamination.

Supposing that we want to test the null hypothesis $\mu = 0$ against the alternative hypothesis $\mu \neq 0$ for sample C1 and test the null hypothesis $\sigma^2 = 1$ against the alternative hypothesis $\sigma^2 \neq 1$ for sample C2. The FS results plotted in Fig.3 show that the false null

hypothesis is accepted in the last steps of the search by the entrance of outliers.

3. Empirical power of T_{FS} and χ_{FS}^2

In this subsection we generate 10000 samples of deferent sizes (50 and 100 units) from normal distribution with various parameters to evaluate the empirical power of our approach. These alternative normal distributions are: $N(\mu = 0.5, \sigma^2 = 1)$, $N(\mu = 1, \sigma^2 = 1)$, $N(\mu = 1, \sigma^2 = 0.5)$ and $N(\mu = 1, \sigma^2 = 2)$ to test the null hypothesis $\mu = 0$ against the alternative hypothesis $\mu \neq 0$ and $N(\mu = 0, \sigma^2 = 0.5)$ and $N(\mu = 0, \sigma^2 = 2)$ to test the null hypothesis $\sigma^2 = 1$ against the alternative hypothesis $\sigma^2 \neq 1$. In each step of the search the values of T_{FS} and χ_{FS}^2 are compared with corresponding empirical quantiles of these tests to determine whether the null hypothesis is rejected or not, the results are plotted in Fig.4 and Fig.5, respectively.

By increasing the size of sample, the estimated powers during the search reach more quickly to the estimated powers in the last step. Therefore, when the sample size is large enough, the procedure can be safely used and the FS method will have enough efficiency.

The blood clotting data

To investigate the performance of the proposed approach to real-world data, we use the blood clotting data set, created by Heiberg [17], related to percent blood clotting activity (PCA) that was measured for 158 Norway rats before (baseline) and four days after injection of an anticoagulant. This data set contains 67 male instances. For our purposes, we used the percent blood clotting activity on day 4 (PCA4) for the male gender. Also, we use the Box-Cox transformation [9] with $\lambda = 0.5$ to transform the data to normal distribution.

The histogram and boxplot for transformed PCA4 (TPCA4), plotted in Fig.6, show some observations that

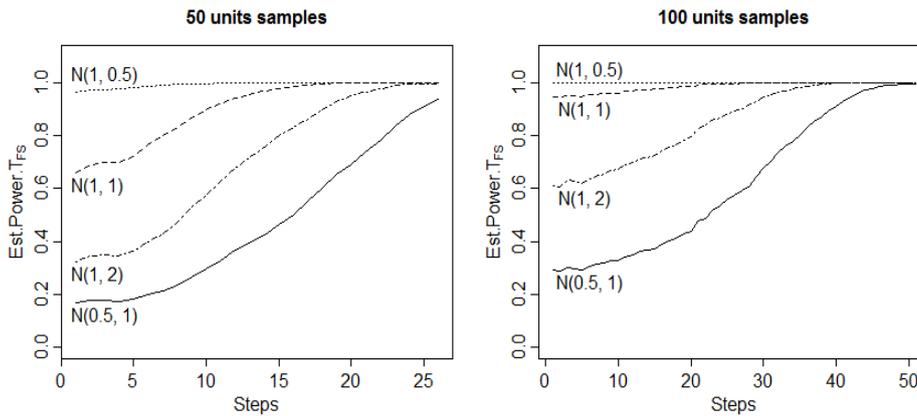


Figure 4. Empirical power of T_{FS} versus alternative normal distributions.

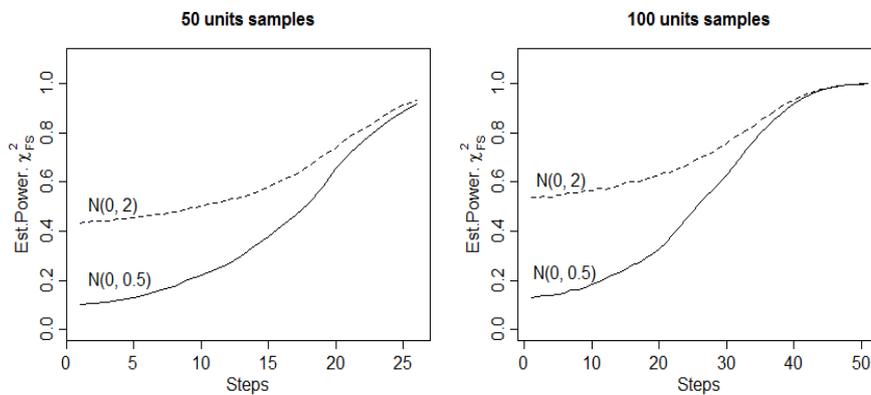


Figure 5. Empirical power of χ^2_{FS} versus alternative normal distributions.

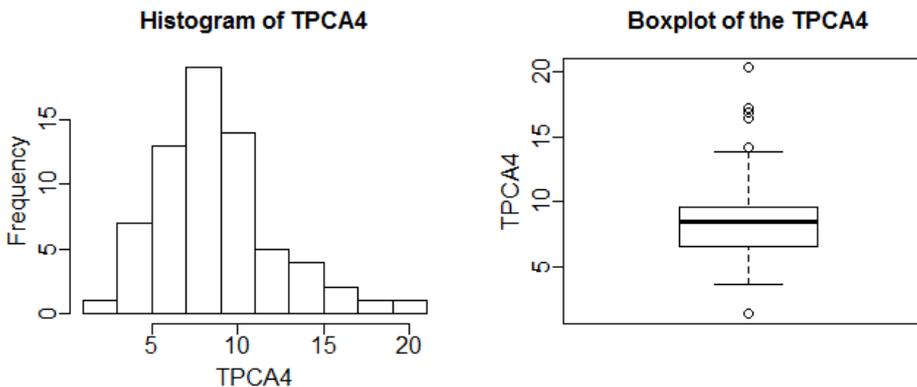


Figure 6. Histogram and boxplot for TPCA4.

could be regarded as outliers. The normality assumption may be affected by these observations and tend to depart in the estimation of normal distribution parameters. We consider two hypotheses to be tested in terms of the mean and the variance of TPCA4 that are defined as follow: Test 1: the null hypothesis $\mu = 9$ against the alternative hypothesis $\mu \neq 9$ and Test 2: the

null hypothesis $\sigma^2 = 15$ against the alternative hypothesis $\sigma^2 \neq 15$.

After transforming these tests to the standard ones, the results of forward search were plotted in Fig.7. The null hypothesis for Test 1 and Test 2 are accepted in the last steps due to the entrance of outliers after steps 22 onwards indicating 12 observations are outliers. The

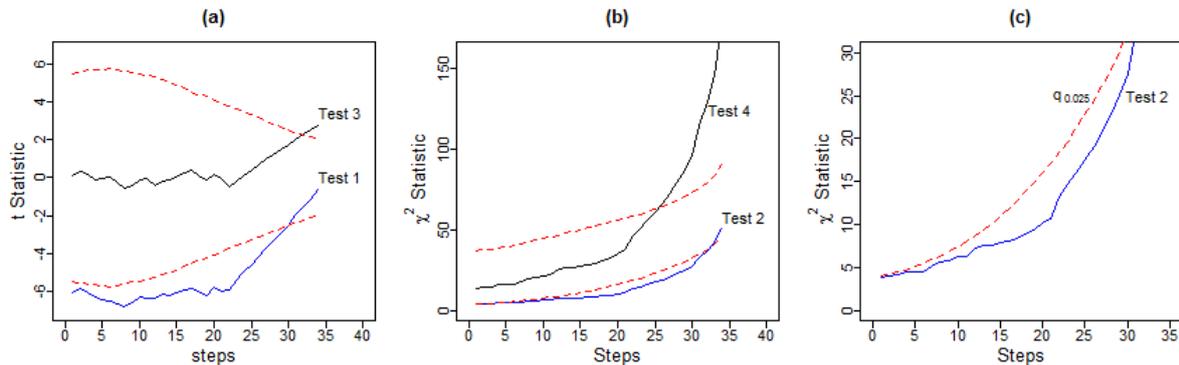


Figure 7. (a) Forward plots of T_{FS} during the search for Test 1 and Test 3. (b) Forward plots of χ^2_{FS} during the search for Test 2 and Test 4. The dashed lines refer to 2.5% and 97.5% empirical quantiles. (c) Zoom of forward plots for Test 2.

zoom plot of Test 2 reveals the jumping point well (Fig.7.(c)). We adjust mean and variance of TPCA4 by removing observations that are included to the process after step 22. Therefore, we repeat the tests for mean and variance of TPCA4 equal to adjusted values as follow: Test 3: the null hypothesis $\mu = 7.6$ against the alternative hypothesis $\mu \neq 7.6$ and Test 4: the null hypothesis $\sigma^2 = 4.3$ against the alternative hypothesis $\sigma^2 \neq 4.3$.

Forward search results for Test 3 and Test 4 are also plotted on Fig.7. (a) and Fig.7. (b). The null hypothesis for Test 3 is accepted in each step except the last steps, but the rejection point for Test 4 is the step 25, indicating the test statistics χ^2_{FS} , for testing variance is very sensitive to the outliers with respect to the test statistics T_{FS} , for testing the mean.

Results

In this paper a new robust method has been proposed to test the parameters of a normal population. This study is concerned with the effect of outliers in testing hypotheses in terms of the mean and the variance of a normal distribution when these parameters are unknown. The robust FS method was implemented to individuate the outliers. In each step of the FS, the proposed robust statistics were computed. This search method separates the group of outliers from the other observations graphically. Simulation studies were conducted to illustrate the application and the advantage of the FS approach. Furthermore, we showed an application of the proposed approach to a real data.

References

1. Aitken A.C. On least squares and linear combination of observations. *J. R. Statist. Soc.*, **55**:42–48(1935).

2. Atkinson A.C. Fast very robust methods for the detection of multiple outliers. *J. Am. Statistic. Assoc.*, **89**: 1329-1339 (1994).

3. Atkinson A.C. and Riani M. Forward search added-variable t-tests and the effect of masked outliers on model selection. *Biometrika*, **89**(4): 939–946 (2002).

4. Atkinson A.C. and Riani M. *Robust Diagnostic Regression Analysis*. Springer, New York (2000).

5. Atkinson A.C., Riani M., and Cerioli A. *Exploring Multivariate Data with the Forward Search*. Springer, New York (2004).

6. Bastero R.C. and Barrios E.B. Robust estimation of a spatiotemporal model with structural change. *Commun. Stat. Simulat.*, **40**: 448-468 (2011).

7. Bellini T. Detecting atypical observations in financial data: the forward search for elliptical copulas. *Adv. Data Anal. Classif.*, **4**: 287-299 (2010).

8. Bertaccini B. and Varriale R. Robust analysis of variance: An approach based on the forward search. *Comp. Statist. Data. Anal.*, **51**: 5172-5183 (2007).

9. Box G.E. and Cox D.R. An analysis of transformations. *J. R. Statist. Soc., Series B*, **26**(2): 211-252 (1964).

10. Campano W.Q. and Barrios E.B. Robust estimation of a time series model with structural change. *J. Statist. Comput. Simulation.*, **81**: 909-927 (2011).

11. Cheng T.-C. and Biswas A. Maximum trimmed likelihood estimator for multivariate mixed continuous and categorical data. *Comp. Statist. Data. Anal.*, **52**: 2042-2065 (2008).

12. Cohen C. and Balakrishnan N. *Order statistics and Inference*. Academic, New York(1991).

13. Coin D. Testing normality in the presence of outliers. *Statist. Meth. Appl.*, **17**: 3-12 (2008).

14. Grossi L. and Laurini F. Robust estimation of efficient mean-variance frontiers. *Adv. Data Anal. Classif.*, **5**: 3-22 (2011).

15. Hadi A.S. Identifying multiple outliers in multivariate data. *J. R. Statist. Soc., Series B*, **54**: 761-771 (1992).

16. Hadi A.S. and Simonoff J.S. Procedures for the identification of multiple outliers in linear models. *J. Am. Statistic. Assoc.*, **88**: 1264-1272 (1993).

17. Heiberg, A.-C. Project at The Royal Veterinary and Agricultural University (1999).

18. Mavridis D. and Moustaki I. Detecting outliers in factor analysis using the forward search algorithm. *Multivariate Behavioral Research.*, **43**: 453–475 (2008).
19. Rousseeuw P.J. Least median of squares regression. *J. Am. Statistic. Assoc.*, **79**: 871–880 (1984).
20. Solaro N. and Pagani M. The forward search for classical multidimensional scaling when the starting data matrix is known, in: Lauro C., Palumbo F., and Greenacre M. (Eds.), *Data analysis and classification: From the exploratory to the confirmatory approach*. Springer, Berlin, pp. 101-109 (2010).