Nonparametric Comparison of Signals Based on Statistical Bootstrap Techniques

De Brabanter, J.[*,**], Pelckmans, K.[**], Suykens, J.A.K.[**],
Saey, P.[*], Verhiest, G.[*], De Moor, B. [**]

January 18, 2006

[*] KaHo Sint Lieven, Department Industrieel Ingenieur
G.Desmetstraat.1, B–9000 GENT
jos.debrabanter@kahosl.be

[**] K.U.Leuven, ESAT-SCD/SISTA
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
jos.debrabanter@esat.kuleuven.be

Abstract

In this paper the problem of detecting the equality of signals is studied. It is common practice to check the equality of the known original signal and the received signal by parametric methods in the Gaussian case. Several method based on nonparametric estimators of the signal are described. In this setting, the distribution of the test statistic is difficult to compute or cannot be determined analytically. For this reason, we propose the use of statistical bootstrap algorithms as an alternative approach. A simulation study is conducted to investigate the finite sample properties of the test.
Keywords: *Hypothesis test, Nonparametric estimator, LS-SVM, External Bootstrap.*

1 Introduction

Detection of signals is a key area in signal processing applications such as radar, sonar and telecommunications. The theory of signal detection has been extensively covered in the literature. Many textbooks exists, including the series by Van Trees [1,2,3,4], the text on radar detection by DiFranco and Rubin [5], and several texts on estimation and detection (Scharf [6], Poor [7], Kay [8,9]).

The main role of a communication system is to transmit signals (information) from the source of information (system input) to the user or destination (system output). The transmission is done over a communication channel using a transmitter and a receiver. A simplified basic communication system is presented in Figure 1. The original signal, usually called the baseband

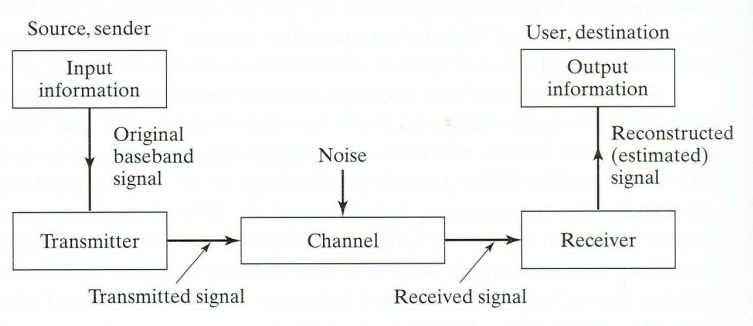


Figure 1: Basic communication system

signal, is first transformed into a signal convenient for transmission by the transmitter. The transmitter sends the signal which may be electrical or optical (electromagnetic), over a communication channel, which represents a physical medium convenient for propagation of electromagnetic waves. Communication channels can be guided media (such as copper wire or optical fiber cable), or free-space channels (such as satellite or wireless (radio)). The role of the receiver is to convert received signals, theoretically, into baseband signals, and pass them to the user. Due to channel attenuation, distortion, and noise, the receiver produces a signal that is similar but not identical to the baseband signal. Such a signal is called an estimated signal. The estimated signal can be slightly different than the original (baseband) signal, especially for voice and video transmissions. However, in the case of data, signal transmission must be error free.

Signal detection theory is well established when the interference is Gaussian. However, methods for detection in the non-Gaussian case are often cumbersome and in many cases non-optimal. Signal detection is formulated as a test of a hypothesis [10]. In this paper, an hypothesis test based on external statistical bootstrap methods will be proposed.

The bootstrap is a computer-intensive method that provides answers to a large class of statistical inference problems without stringent structural assumptions on the underlying random process generating the data. Since its introduction by Efron [11], the bootstrap has found its application to a number of statistical problems, including many standard ones, where it has outperformed the existing methodology as well as to many complex problems involving independent data where conventional approaches failed to provide

satisfactory answers. However, the general perception that the bootstrap is a generally applicable method, giving accurate results in all problems automatically, is misleading. An example of this appears in Singh [12], which points out the inadequacy of this resampling scheme under dependence. A breakthrough was achieved with block resampling, an idea that was put forward by Hall [13] and others in various forms and in different inference problems. The most popular bootstrap methods for dependent data are block, sieve [14], local and external bootstrap [15].

The goal of this paper is to construct Monte Carlo tests, based on test statistics and bootstrap techniques, for testing the equality of signals.

This paper is organized as follows. In Section 2 we discuss methods for testing the Equality of Signals by nonparametric methods. We briefly discuss nonparametric estimators of the signal. Next we give several test statistics for checking the hypothesis of equality of signals. Finally, we apply bootstrap techniques in the case of hypothesis testing. Section 4 reports results on an artificial data set.

2 Testing the Equality of Signals by nonparametric Methods

In many cases of practical interest we have a sample of $N = n_1 + n_2$ observations,

$$Y_{i,t} = g_i\left(x_{i,t}\right) + e_{i,t}, \qquad t = 1, ..., n_i, \quad i = 1, 2 \tag{1}$$

where $\{e_{i,t}\}_{t=1}^{n_i}$ are random errors with probability distribution functions G_i, $E\left[e_{1,t}\right] = E\left[e_{2,t}\right] = 0$ and $Var\left[e_{i,t}^2\right] = \sigma_i^2 < \infty$. The design (sampled) points $\{x_{i,t}\}$ are fixed and usually rescaled into the unit interval, so $0 \leq x_{i,1} \leq ... \leq x_{i,n_i} \leq 1$. In this context we are interested in the problem of testing the equality of the mean functions, that is,

$$H_0 : g_1 = g_2 \qquad \text{versus} \qquad H_1 : g_1 \neq g_2 \tag{2}$$

2.1 Nonparametric Estimation of Signals

In this subsection we review some nonparametric methods for estimating the function g in (1). Model (1) has the format of a nonlinear regression problem for which many smoothing methods exist when the observations are independent. The description concerning the four paradigms in nonparametric regression is based on [16]. The kernel estimate (local averaging) is due to [17] and [18]. The principle of least squares (global modeling) is much older.

For historical details we refer to [19]. The principle of penalized modeling, in particular, smoothing splines, goes back to [20]. Final, The principle of complexity regularization is due to [21], in particular for least squares estimates, see [22]. Hart [23] demonstrates that these methods can be "borrowed" for time series analysis where the observations are correlated by making use of the "whitening by windowing principle".

Given a time series $\{Y_t, t = 1, ..., n\}$, we can assume that

$$Y_t = g(X_t) + e_t \tag{3}$$

where $X_t = (Y_{t-1}, ..., Y_{t-p})$, g is an unknown function, and $\{e_t\} \sim i.i.d.\left(0, \sigma^2\right)$. Instead of imposing a concrete form on g, we only make some qualitative assumptions, such as that the function $g \in C^{\infty}(\mathbb{R})$. Model (3) is called a nonparametric autoregressive model. The structure in (3) is very general, making very few assumptions on how the data were generated.

Nadaraya-Watson kernel estimate .

A typical situation for an application to a time series $\{Y_t, t = 1, ..., n\}$ is that the regressor vector X consists of past time series values $X_t = (Y_{t-1}, ..., Y_{t-p})^T$. Let $K : \mathbb{R}^p \to \mathbb{R}_+$ be a function called the kernel function, and let $h > 0$ be a bandwidth. For $X \in \mathbb{R}^p$, $X_t = (Y_{t-1}, ..., Y_{t-p})^T$ and weights

$$w_{n,i}(x) = \frac{K\left(\frac{x - X_i}{h}\right)}{\sum_{t=p+1}^{n} K\left(\frac{x - X_t}{h}\right)} \tag{4}$$

where $w_{n,i} : \mathbb{R}^p \to \mathbb{R}$. The Nadaraya-Watson kernel estimator in model (3) is given by

$$\hat{g}_{n,NW}(x) = \sum_{i=p+1}^{n} w_{n,i}(x) Y_i. \tag{5}$$

LS-SVM regression .

Consider the model

$$g_{n,LS}(x) = w^T \varphi(x) + b \tag{6}$$

with so-called feature map $\varphi : \mathbb{R}^p \to \mathbb{R}^{D_\varphi}$, $w \in \mathbb{R}^{D_\varphi}$ and $b \in \mathbb{R}$. Consider the regularized least squares cost function [22]

$$\min_{w,e,b} \mathcal{J} = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{t=p}^{n} e_t^2$$
$$\text{s.t.} \quad w^T \varphi(x_t) + b + e_t = y_t \quad \forall t = p+1, \ldots, n \tag{7}$$

Then the dual solution is characterized by the following linear system

$$\left[\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + \frac{1}{\gamma} I_n \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right], \tag{8}$$

where $\Omega \in \mathbb{R}^{N \times N}$ with $\Omega_{ij} = K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$, e.g. $\Omega_{ij} = K\left(\frac{x_i - x_j}{h}\right)$ for all $i, j = 1, \ldots, N$ and $y = (y_1, \ldots, y_n)^T$. The estimated model can be evaluated at a new point $x_* \in \mathbb{R}^p$ as follows

$$\hat{g}_{n,LS}(x) = \sum_{t=p+1}^{n} \alpha_t K\left(\frac{x - X_t}{h}\right) + b. \tag{9}$$

2.2 Testing the Equality of Signals

Several test statistics for checking the hypothesis of equality of signals given in (2) are next summarized.

An ANOVA-type statistic The following method for testing the equality of the signals was introduced by Young and Bowman [24]. The test statistic is defined by

$$\hat{Q}_n^{(1)} = \frac{1}{N} \sum_{i=1}^{2} \sum_{t=1}^{n_i} \left(\hat{g}_p\left(x_{i,t}\right) - \hat{g}_i\left(x_{i,t}\right)\right)^2 \tag{10}$$

where $\hat{g}_p(x)$ is e.g., the Nadaraya-Watson kernel estimator or the LS-SVM estimator of the signal obtained on the basis of the total combined sample.

Pairwise comparison of signals Following Härdle and Mammen [15], a test of the hypothesis (2) could be obtained from a pairwise comparison of the estimators of the signals. We consider the statistic

$$\hat{Q}_n^{(2)} = \int \left(\hat{g}_1(x) - \hat{g}_2(x)\right)^2 w_{12}(x)\, dx \tag{11}$$

where $w_{12}(x)$ are positive weight functions.

An Empirical process Neumeyer and Dette [25] proposed a new test based on the difference of two empirical processes, which are constructed

from the residuals obtained under the assumption of equality of the two signals. The residuals are generated as

$$\hat{e}_{i,t} = Y_{i,t} - \hat{g}_p\left(x_{i,t}\right), \quad t = 1, ..., n_i \quad \text{and} \quad i = 1, 2 \tag{12}$$

and the difference between the corresponding empirical processes would be given by

$$\hat{R}_n\left(s\right) = \frac{1}{n_1}\sum_{t=1}^{n_1} \hat{e}_{1,t} I\left(x_{1,t} \leq s\right) - \frac{1}{n_2}\sum_{t=1}^{n_2} \hat{e}_{2,t} I\left(x_{2,t} \leq s\right) \tag{13}$$

where $s \in [0,1]$ and $I\left(\cdot\right)$ denotes the indicator function. Neumeyer and Dette [25] suggested to test the hypothesis of equal signals on the basis of real valued functionals of the process $\hat{R}_n\left(s\right)$. In particular, the next two test statistics were proposed

$$\hat{Q}_n^{(3)} = \int \hat{R}_n^2\left(s\right) ds \tag{14}$$

$$\hat{Q}_n^{(4)} = \sup_{s\in[0,1]} \left|\hat{R}_n\left(s\right)\right|. \tag{15}$$

2.3 External Bootstrap Algorithm

In what follows, the regression model of the fixed design given in (1), with $n = n_1 = n_2$, are considered. It is assumed that $g(x)$ is defined in $[0,1]$. The sample points are taken equidistantly spaced, $x_t = t/n$, for $t = 1, \ldots, n$.

In this context, our interest is focused on checking the null hypothesis H_0 : $g_1 = g_2$. For this, any of the test statistics introduced in Subsection 2.2, $\hat{Q}_n^{(\cdot)}$ can be used. In all the cases, the criterion is to reject H_0 for large values of $\hat{Q}_n^{(\cdot)}$. In practice, it is obviously necessary to know the distribution of $\hat{Q}_n^{(\cdot)}$ in order to compute the critical values of the test. It is in general complicated to determine the distribution of these statistics. In this paper, an alternative way of approximating the distribution of the test statistic $\hat{Q}_n^{(\cdot)}$ by means of statistical bootstrap techniques is proposed.

Algorithm 1 *External bootstrap.*

1. The test statistic $\hat{Q}_n^{(\cdot)}$ is computed from the initial sample given by $\{(x_t, Y_{1,t}, Y_{2,t}\}_{t=1}^n$.
2. Under the null hypothesis, residuals $\hat{e}_{i,t}$ are obtained by

$$\hat{e}_{i,t} = Y_{i,t} - \hat{g}_p(x_t), t = 1, \ldots, n \quad \text{and} \quad i = 1, 2 \tag{16}$$

with $\hat{g}_p(x_t)$ the LS-SVM estimator computed from the total combined sample.

3. Draw the bootstrap residuals $\hat{e}^*_{i,t}$ from a two-point centered distribution in order that its second and third moment fit the square and the cubic power of the residual $\hat{e}_{i,t}$. For instance, one can choose $\hat{e}^*_{i,t} \sim \hat{e}_{i,t}\left(\frac{Z_1^2}{\sqrt{2}} + \frac{Z_2^2-1}{2}\right)$, with Z_1 and Z_2 being two independent standard Normal random variables, also independent of $\hat{e}_{i,t}$.

4. A bootstrap sample $\{(x_t, Y^*_{1,t}, Y^*_{2,t}\}_{t=1}^n$ is obtained, making

$$Y^*_{i,t} = \hat{g}_p(x_t) + \hat{e}^*_{i,t}, t = 1, \ldots, n \quad \text{and} \quad i = 1, 2 \tag{17}$$

   The test statistic $\hat{Q}_n^{(\cdot)}$ is now computed with this bootstrap sample.

5. This whole proces must be repeated B times, so that the sequence $\{\hat{Q}^{(\cdot)*}_{n,1}, \ldots, \hat{Q}^{(\cdot)*}_{n,B}\}$ is obtained. A bootstrap critical region of a significance level α is then given as

$$\hat{Q}_n^{(\cdot)} > \hat{Q}^{(\cdot)*}_{n,(\lfloor(1-\alpha)B\rfloor)}, \tag{18}$$

   where $\lfloor\cdot\rfloor$ represents the integer part and $\{\hat{Q}^{(\cdot)*}_{n,(k)}\}_{k=1}^B$ is the sample $\{\hat{Q}^{(\cdot)*}_{n,k}\}_{k=1}^B$ arranged in increasing order of magnitude.

The number of bootstrap replications B, depends on the statistical inference problem. $B = 100$ is often enough to give a good estimate of a standard error. Much larger values of B are required for bootstrap confidence (prediction) intervals, for example $B = 1000$ [26].

3 Illustrative Examples

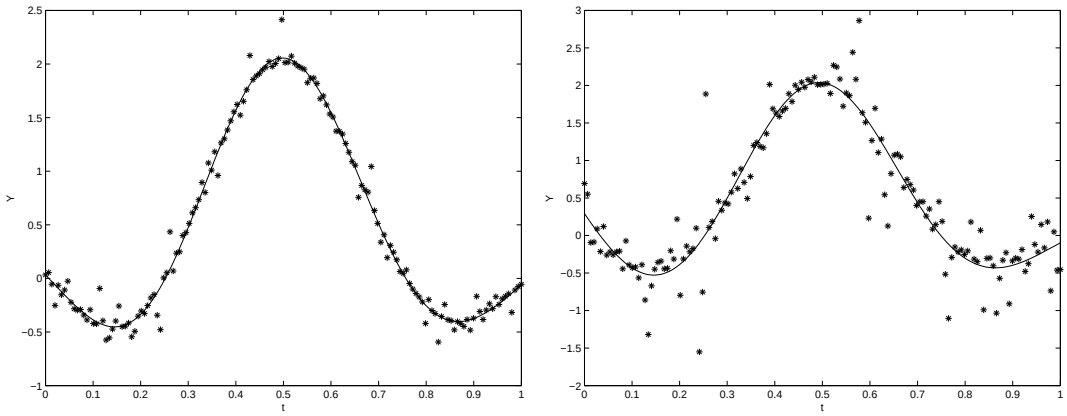
To illustrate the Monte Carlo test, we present two examples using the following data sets. Consider the nonlinear signals defined on

$$Y_{i,t} = g_i\left(x_{i,t}\right) + e_{i,t}, \qquad t = 1, ..., 150, \quad i = 1, 2 \tag{19}$$

where the noise is uniform distributed, denoted by $e_{1,t} \sim U[-0.1, 0.1]$ and $e_{2,t} \sim U[-0.5, 0.5]$. The LS-SVM with RBF kernel was used to construct the nonparametric estimation. By using 1000 bootstrap samples, we obtain 1000 test $\hat{Q}_n^{(\cdot)*}$.

3.1 Example 1

Figure 2 shows the transmitted signal and the received signal. The histogram of the test statistics $\hat{Q}_n^{(\cdot)*}$ is illustrated in Figure 3. Let $\alpha = 0.05$, based on $\hat{Q}_n^{(\cdot)} < \hat{Q}^{(\cdot)*}_{n,(\lfloor(1-\alpha)B\rfloor)}$ we can not reject the null hypothesis $H_0 : g_1 = g_2$.



(a) (b)

Figure 2: (a) Transmitted signal: (solid line) true function (sinc function), * data points; (b) Received signal: (solid line) true function (sinc function), * data points

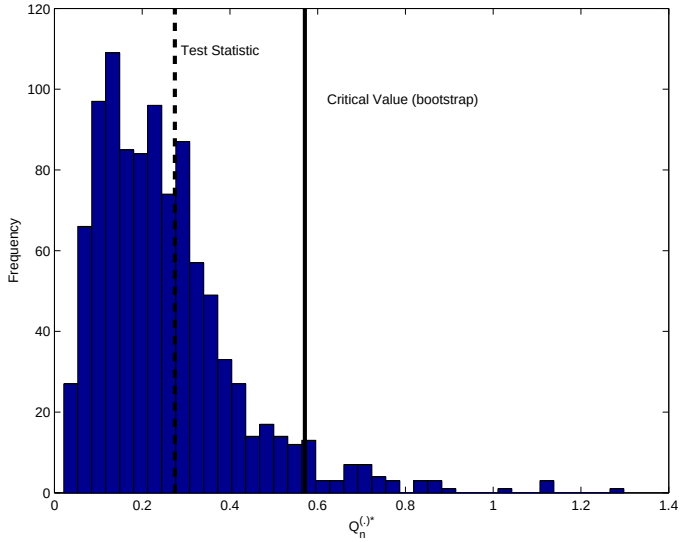
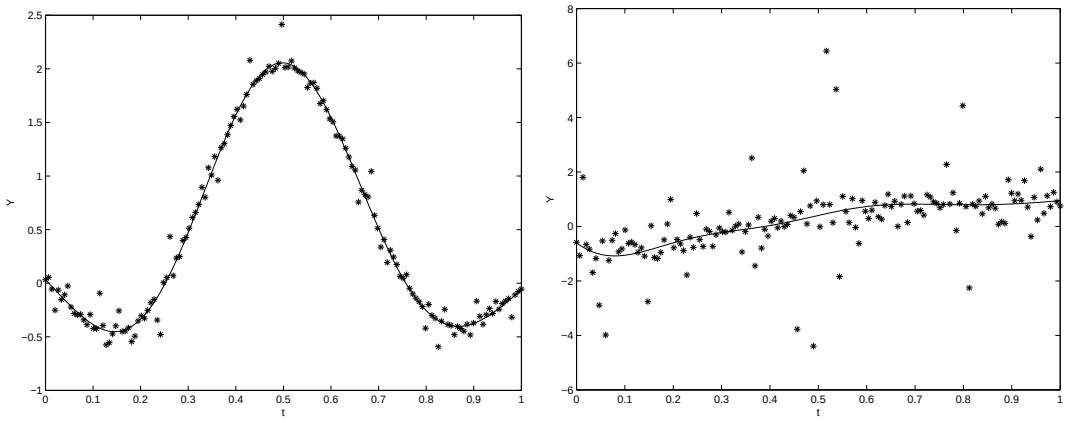


Figure 3: Null distribution of the test statistic. The doted line indicates the value of $\hat{Q}_n^{(\cdot)}$ and the solid line indicates the critical value based on bootstrap samples $\hat{Q}_n^{(\cdot)*}$

3.2 Example 2

Figure 4 shows the transmitted signal and the received signal. The histogram of the test statistics $\hat{Q}_n^{(\cdot)*}$ is illustrated in Figure 5. Let $\alpha = 0.05$, based on $\hat{Q}_n^{(\cdot)} > \hat{Q}_{n,(\lfloor(1-\alpha)B\rfloor)}^{(\cdot)*}$ we reject the null hypothesis $H_0 : g_1 = g_2$.



(a) (b)

Figure 4: (a) Transmitted signal: (solid line) true function (sinc function), * data points; (b) Received signal: (solid line) true function (tanh function), * data points

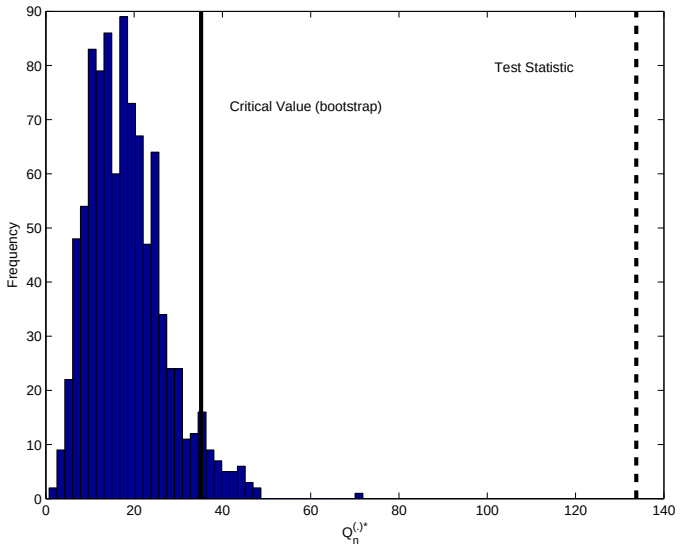


Figure 5: Null distribution of the test statistic. The doted line indicates the value of $\hat{Q}_n^{(\cdot)}$ and the solid line indicates the critical value based on bootstrap samples $\hat{Q}_n^{(\cdot)*}$

Acknowledgements Research supported by

(Research Council KUL): GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering, several PhD/postdoc & fellow grants; (Flemish Government): (FWO): PhD/postdoc grants, projects, G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0553.06, research com-

munities (ICCoS, ANMMM, MLDM); (IWT): PhD Grants,GBOU (McKnow), Eureka-Flite2 - Belgian Federal Science Policy Office: IUAP P5/22,PODO-II,- EU: FP5-Quprodis; ERNSI; - Contract Research/agreements: ISMC/IPCOS, Data4s, TML, Elia, LMS, Mastercard JS is an associate professor and BDM is a full professor at K.U.Leuven Belgium, respectively.

References

[1] Van Trees, H.L. (2001) *Detection, Estimation, and Modulation Theory. Part I.* John Wiley and Sons.

[2] Van Trees, H.L. (2001) *Detection, Estimation, and Modulation Theory. Part III. Radar-Sonar Signal Processing and Gaussian Signals in Noise.* John Wiley and Sons.

[3] Van Trees, H.L. (2002) *Detection, Estimation, and Modulation Theory. Part II. Nonlinear Modulation Theory.* John Wiley and Sons.

[4] Van Trees, H.L. (2002) *Detection, Estimation, and Modulation Theory. Part IV. Optimum Array Signal Processing.* John Wiley and Sons.

[5] DiFranco, J.V. and Rubin, W.L. (1980) *Radar Detection.* Artech House, Dedham, Mass.

[6] Scharf, L.L.(1991) *Statistical Signal Processing, Detection, Estimation, and Time Series Analysis.* Addison Wesley, Reading, Mass.

[7] Poor, H.V.(1986) *Robustness in Signal Detection.* In Communications and Networks: A Survey of Recent Advances, Blake, I.F. and Poor, H.V., Editors, 131-156, Springer-Verlag.

[8] Kay, S.M. (1993) *Fundamentals of Statistical Signal Processing. Estimation Theory, Volume I.* Prentice Hall, Englewood Cliffs, New Jersey.

[9] Kay, S.M. (1998) *Fundamentals of Statistical Signal Processing. Detection Theory, Volume II.* Prentice Hall, Englewood Cliffs, New Jersey.

[10] Lehmann, E.L. (1991) *Testing Statistical Hypotheses.* Wadsworth and Brooks, Pacific Grove, Calif.

[11] Efron, B., Bootstrap methods: Another look at the jackknife, *Annals of Statistics*, 7: 1-26 (1979)

[12] Singh, K., On the asymptotic accuracy of the Efron's bootstrap, *Annals of Statistics*, 9: 1187-1195 (1981)

[13] Hall, P., Resampling a coverage pattern, *Stochastic Processes and Their Applications* 20, 231-246 (1985)

[14] Bühlmann, P., Sieve bootstrap for time series, *Bernoulli*, 3, 123-148 (1997)

[15] Härdle, W. and Mammen, E., Comparing nonparametric versus parametric regression fits, *The Annals of Statistics*, 21, 1926-1947, (1993)

[16] Györfi, L., Kohler, M., Krzyżak, A., Walk, H., *A distribution-free theory of nonparametric regression*, Springer-Verlag, New York, (2002)

[17] Nadaraya, E.A., On nonparametric estimates of density functions and regression curves, *Theory of Applied Probability*, 10, 186-190 (1965)

[18] Watson, G.S., Smooth regression analysis, *Sankhya*, 26:15 175-184 (2002)

[19] Stigler, S.M., *Statistics on the table: The History of Statistical Concepts and Methods*, Harvard University Press, Cambrigde, MA. (1999)

[20] Reinsch, C., Smoothing by spline functions, *Numerische Mathematik*, 10: 177-183 (1967)

[21] Vapnik, V.N., *Statistical Learning Theory*, John Wiley and Sons, INC. (1999)

[22] Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J. *Least Squares Support Vector Machines*, World Scientific, Singapore (2002)

[23] Hart, D.J., *Nonparametric Smoothing and Lack-of-fit Tests*, Springer-Verlag, New York, (1997)

[24] Young, S.G. and Bowman, A.W., An Introduction to the bootstrap, *Biometrics*, 51, 920-931, (1995)

[25] Neumeyer, N. and Dette, H., Nonparametric comparision of regression curves: an empirical process approach, *The Annals of Statistics*, 31 , 880-920, (2003)

[26] Efron, B. and Tibshirani, R.J., *An Introduction to the bootstrap*, Chapman and Hall, London (1993)