Data-driven efficient score tests for deconvolution hypotheses

M A Langovoy

Institute for Mathematical Stochastics, Georg-August-University at Göttingen,
Maschmuellenweg 8/10, 37077 Göttingen, Germany

E-mail: langovoy@math.uni-goettingen.de

Received 25 August 2007, in final form 22 February 2008
Published 25 March 2008
Online at stacks.iop.org/IP/24/025028

Abstract
We consider testing statistical hypotheses about densities of signals in deconvolution models. A new approach to this problem is proposed. We constructed score tests for the deconvolution density testing with the known noise density and efficient score tests for the case of unknown density. The tests are incorporated with model selection rules to choose reasonable model dimensions automatically by the data. Consistency of the tests is proved.

1. Introduction to statistical testing

Suppose that one wishes to decide whether some hypothesis that has been formulated is correct. The decision has to be based on the value of some random variable $X$, the distribution $P$ of which is known to belong to a family of distributions $\mathcal{P}$. We assume that the distributions of $P$ can be classified into those for which the hypothesis is true and those for which it is false. A decision procedure for such a problem is called a statistical test for the hypothesis. For any values $x$ of $X$, such a test chooses between the two decisions, rejection or acceptance, with certain probabilities that depend on $x$. We refer to Lehmann and Romano (2005) for further details and a classic presentation of the subject.

Constructing good tests for statistical hypotheses is an essential problem of statistics. There are two main approaches to constructing test statistics. In the first approach, roughly speaking, some measure of distance between the theoretical and the corresponding empirical distributions is proposed as the test statistic. Classical examples of this approach are the Cramer–von Mises and the Kolmogorov–Smirnov statistics. More general examples of tests from this class are given by $L^2$-distance based tests and by tests based on confidence bands. Although, these tests work and are capable of giving very good results, each of these tests is asymptotically optimal only in a finite number of directions of alternatives to a null hypothesis (see Nikitin (1995)).

Nowadays, there is an increasing interest in the second approach of constructing test statistics. The idea of this approach is to construct tests in such a way that the tests would
be asymptotically optimal. Test statistics constructed following this approach are often called (efficient) score test statistics. The pioneer of this approach was Neyman (1937) and then many other works followed: Neyman (1959), Cox and Hinkley (1974), Bickel and Ritov (1992), Le Cam (1956), and Ledwina (1994). This approach is also closely related to the theory of efficient (adaptive) estimation—Bickel et al (1993), Ibragimov and Has’minskii (1981). Score tests are asymptotically optimal in the sense of intermediate efficiency in an infinite number of directions of alternatives (see Inglot and Ledwina (1996)) and show good overall performance in practice (Kallenberg and Ledwina (1995), Kallenberg and Ledwina (1997)).

Another important line of development in the area of optimal testing concerns minimax testing, see Ingster (1993)), and adaptive minimax testing, see Spokoiny (1996). Those tests are optimal in a certain minimax sense against wide classes of nonparametric alternatives. We do not discuss minimax testing theory in this paper. See Bickel et al (2006) for a recent general overview of this and other existing theories of statistical testing, and a discussion of some advantages and disadvantages of different classes of testing methods.

2. Inference in statistical inverse problems

We describe the situation in classical hypothesis testing, i.e., testing hypotheses about random variables $X_1, \ldots, X_n$, whose values are directly observable. But, it is important from a practical point of view to be able to construct tests or estimates for situations where $X_1, \ldots, X_n$ are corrupted or can only be observed with an additional noise term. These kinds of problems are termed statistical inverse problems. Statistical inference in inverse problems is more difficult than classical inference for directly available observations.

One of the most well-known examples of statistical inverse problems is the deconvolution density testing problem. This problem appears when one has noisy signals or measurements: in physics, seismology, optics and imaging, and engineering. It is a building block for many complicated statistical inverse problems. Here we briefly highlight two specific examples from physics and engineering. In Butucea (2007), other examples are given and it is shown that the deconvolution density testing problem is of importance for many applications in biology, medicine and astronomy.

Example 1. Frequency combs recently received a lot of attention in physics. They are used in frequency metrology, high precision spectroscopy, laser noise characterization, and telecommunications. In fundamental physics, the frequency combs technique provides a possibility for the determination of fundamental constants such as the Rydberg constant (see Niering et al (2000)). In the core of this technique there are high precision measurements of certain frequencies that are often subject to quantum uncertainties and noise (see Paschotta et al (2006)). These measurements provide powerful tools for new tests of fundamental physics laws (see Hänisch (2006)). Statistical tests of deconvolution density hypotheses can serve as a mathematical foundation for tests of these laws.

Example 2. Statistical tests for densities of noisy signals can be used in various signal detection problems. In these problems, one awaits a certain signal of interest to come through a channel, but there is always a random noise in the channel. The statistical hypothesis of interest here would be that the signal of interest is constantly zero (or that the signal intensity does not exceed certain threshold). One encounters such problems in the seismic monitoring of nuclear explosions (see Picard and Bryson (1992)) or in the underwater sound detection (see Bailey et al (1998)).
Due to the importance of the deconvolution problem, testing statistical hypotheses related to this problem have been widely studied in the literature. But, to our knowledge, all the proposed tests were based on some kind of distance (usually a $L_2$-type distance) between the theoretical density function and the empirical estimate of the density (see, for example, Bickel and Rosenblatt (1973), Delaigle and Gijbels (2002)). Graphical tests based on confidence bands give an example of tests built on $L_\infty$-type distance. Thus, only the first approach described above was implemented for the deconvolution density testing problem.

In this paper, we treat the deconvolution density testing problem with the second approach. We construct efficient score tests for the problem. From classical hypothesis testing, it was shown that for applications of efficient score tests, it is important to select the right number of components in the test statistic (see Bickel and Ritov (1992), Eubank et al. (1993), Kallenberg and Ledwina (1995), Fan (1996)). Thus, we provide corresponding refinement of our tests. Following the solution proposed in Kallenberg (2002), we make our tests data-driven, i.e., the tests are capable of choosing a reasonable number of components in the test statistics automatically by the data.

In section 3, we formulate the simple deconvolution density testing problem. In section 4, we construct the score tests for the parametric deconvolution hypothesis. In section 6, we prove consistency of our tests against nonparametric alternatives. In section 7, we turn to the deconvolution density testing with an unknown error density. We derive the efficient scores for the composite parametric deconvolution hypothesis in section 8. In section 9, we construct the efficient score tests for this case. In section 10, we make our tests data-driven. In section 11, we prove consistency of the tests against nonparametric alternatives. Additionally, in sections 6 and 11, we explicitly characterize the class of nonparametric alternatives such that our tests are inconsistent and therefore should not be used for testing against the alternatives from this class. Some simple examples of applications of the theory are also presented in this paper.

3. Notation and basic assumptions

The problem of testing whether independent identically distributed (i.i.d.) real-valued random variables $X_1, \ldots, X_n$ are distributed according to a given density $f$ is classical in statistics. We consider a more difficult problem, namely the case when $X_i$ can only be observed with an additional noise term, i.e., instead of $X_i$ one observes $Y_i$, where

$$Y_i = X_i + \varepsilon_i,$$

and $\varepsilon_i$'s are i.i.d. with a known density $h$ with respect to the Lebesgue measure $\lambda$; also $X_i$ and $\varepsilon_i$ are independent for each pair of $i$ and $j$, and $E\varepsilon_i = 0, 0 < E\varepsilon_i^2 < \infty$. For brevity of notation say that $X_i, Y_i, \varepsilon_i$ have the same distribution as random variables $X, Y, \varepsilon$ correspondingly. Assume that $X$ has a generalized density $f$ with respect to $\lambda$ (i.e., discrete distributions are included).

Our null hypothesis $H_0$ is the simple hypothesis that $X$ has a known density $f_0$ with respect to $\lambda$. The most general possible nonparametric alternative hypothesis $H_A$ is that $f \neq f_0$. Since this class of alternatives is too broad, first we would be concerned with a special class of submodels of the model described above. In this paper we will first assume that all possible alternatives from $H_A$ belong to some parametric family. Then we will propose a test that is expected to be asymptotically optimal (in some sense) against the alternatives from this parametric family. However, we will prove that our test is also consistent against other alternatives even if they do not belong to the initial parametric family. The test is therefore applicable in many nonparametric problems. Moreover, the test is expected to be asymptotically optimal (in some sense) for testing against an infinite number of directions of
nonparametric alternatives (see Inglot and Ledwina (1996)). This is the general plan for our construction.

4. Score test for simple deconvolution density testing

Suppose that all possible densities of $X$ belong to some parametric family $\{f_\theta\}$, where $\theta$ is a $k$-dimensional Euclidean parameter, $\Theta \in \mathbb{R}^k$ is a parameter set. Then all the possible densities $q(y; \theta)$ of $Y$ have in a such model the form

$$q(y; \theta) = \int_{\mathbb{R}} f_\theta(s) h(y - s) \, ds.$$  \hfill (1)

The score function $\dot{l}$ is defined as

$$\dot{l}(y; \theta) = \frac{(q(\theta))'}{q(\theta)} l_{q(\theta) > 0},$$  \hfill (2)

where $q(\theta) := q(y; \theta)$ and $l(\theta) := l(y; \theta)$ for brevity. The score function is just the derivative of statistical log-likelihood and is therefore commonly used to maximize the likelihood function in order, for example, to find the most probable value of unknown parameter $\theta$. The Fisher information matrix $I(\theta)$ is defined as

$$I(\theta) = \int_{\mathbb{R}} \dot{l}(y; \theta) \dot{l}^T(y; \theta) \, dQ_\theta(y).$$  \hfill (3)

The symbol ‘$T$’ denotes the transposition and all vectors are supposed to be row ones. The information matrix is essentially the variance of the score function and is therefore commonly used, for example, to find bounds on the mean-squared error in estimates for $\theta$. We refer to Cox and Hinkley (1974) for an extensive introduction into the field of maximum likelihood methods and score tests.

**Definition 1.** Call our problem a regular deconvolution density testing problem if

$(B1)$ for all $\theta \in \Theta$ $q(y; \theta)$ is continuously differentiable in $\theta$

for $\lambda$ – almost all $y$ with gradient $q(\theta)$

$(B2)$ $|\dot{l}(\theta)| \in L_2(\mathbb{R}, Q_\theta)$ for all $\theta \in \Theta$

$(B3)$ $I(\theta)$ is nonsingular for all $\theta \in \Theta$ and continuous in $\theta$.

The model $M_k(\theta)$ is the set of all pairs of densities $(q(\cdot; \theta), h)$ such that in each pair $q(\theta)$ and $h$ satisfy the above assumptions. If $\theta$ is a true parameter value, then we denote by $Q_\theta$ the probability distribution function, and by $E_\theta$ the expectation, corresponding to the density $q(\cdot; \theta)$.

If conditions $(B1)$–$(B3)$ hold, then by proposition 1, p 13 of Bickel et al (1993) we calculate for all $y \in \text{supp} q(\cdot; \theta)$

$$\dot{l}(\theta) = \dot{l}(y; \theta) = \frac{(q(y; \theta))'}{q(y; \theta)} = \frac{\frac{\partial}{\partial \theta} \int_{\mathbb{R}} f_\theta(s) h(y - s) \, ds}{\int_{\mathbb{R}} f_\theta(s) h(y - s) \, ds}. $$  \hfill (4)

Then for $y \in \text{supp} q(\cdot; \theta)$ the efficient score vector for testing $H_0 : \theta = 0$ is

$$l^*(y) := \dot{l}(y; 0) = \frac{\partial}{\partial \theta} \left( \int_{\mathbb{R}} f_\theta(s) h(y - s) \, ds \right) \big|_{\theta = 0}. $$  \hfill (5)

Set

$$L = \{E_0[l^*(Y)]^T l^*(Y)\}^{-1}.$$  \hfill (6)
Theorem 1. For the regular deconvolution density testing problem the efficient score vector \( l^* \) for testing \( \theta = 0 \) in \( M_k(\theta) \) is given for all \( x \in \mathbb{R} \) by (5). Moreover, under \( H_0 : \theta = 0 \) we have \( U_k \rightarrow_d \chi^2_k \) as \( n \rightarrow \infty \), where \( \chi^2_k \) denotes a random variable with central \( \chi^2 \) distribution with \( k \) degrees of freedom.

We construct the test based on the test statistic \( U_k \) as follows: the null hypothesis \( H_0 \) is rejected if the value of \( U_k \) exceeds standard critical points for \( \chi^2_k \)-distribution. Note that we do not need to estimate the scores \( l^* \).

Remark 2. The test constructed on the basis of theorem 1 is expected to perform very well asymptotically, i.e., when the number of observations \( n \) is sufficiently large. For moderate sample sizes, the distribution of \( U_k \) can be far from the \( \chi^2_k \)-distribution and one has to use an approximate version of the test. An approximate test can be obtained by referring the test statistic to simulated critical values rather than to the standard critical points for the \( \chi^2_k \)-distribution. One of the possible methods to realize this strategy is bootstrap. This is a resampling method commonly used by statisticians. See Hinkley (1988) for a good general description of the bootstrap methodology.

Example 3. Consider one important special case. Assume that each submodel of interest is given by the following restriction: all possible densities \( f \) of \( X \) belong to a parametric exponential family, i.e., \( f = f_\theta \) for some \( \theta \), where

\[
f_\theta(x) = f_0(x)b(\theta)\exp(\theta \cdot u(x)),
\]

where the symbol \( \cdot \) denotes the inner product in \( \mathbb{R}^k \), \( u(x) = (u_1(x), \ldots, u_k(x)) \) is a vector of known Lebesgue measurable functions, \( b(\theta) \) is the normalizing factor and \( \theta \in \Theta \subseteq \mathbb{R}^k \).

We assume that the standard regularity assumptions on exponential families (see Barndorff-Nielsen (1978)) are satisfied. All the possible densities \( q(y; \theta) \) of \( Y \) have in such a model the form

\[
q(y; \theta) = \int_\mathbb{R} f_0(s)b(\theta)\exp(\theta \cdot u(s))h(y - s)\, ds.
\]

These densities no longer need to form an exponential family. If we assume, for example, that \( h > 0 \) \( \lambda \)-almost everywhere on \( \mathbb{R} \) and the functions \( f_0, h, u_1, \ldots, u_k \) are bounded and \( \lambda \)-measurable and that there exists an open subset \( \Theta_1 \subseteq \Theta \) such that \( [y; \theta] \in L_2(Q_\theta) \) and the Fisher information matrix \( I(\Theta) \) is nonsingular and continuous in \( \theta \), then conditions (B1)–(B3) are satisfied for this problem and the previous results are applicable. The score vector for the problem is

\[
l^*(y) = \frac{\int_\mathbb{R} u(s)f_0(s)h(y - s)\, ds}{\int_\mathbb{R} f_0(s)h(y - s)\, ds} - \int_\mathbb{R} u(s)f_0(s)\, ds.
\]

In other words, if we denote by \( \ast \) the standard convolution of functions,

\[
l^*(y) = (u\ast f_0) \ast h(y) - E_0u(X).
\]
Let $L$ be defined by (6) and

$$
V_k = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l^*(Y_j) \right\} L \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l^*(Y_j) \right\}^T.
$$

(12)

This is the score test statistic designed to be asymptotically optimal for testing $H_0$ against the alternatives from the exponential family (8). Its asymptotic distribution under the null hypothesis $H_0$ is given by theorem 1.

5. Selection rule

For the use of score tests in classical hypotheses testing it was shown (see the introduction for the references) that it is important to select the right dimension $k$ of the space of possible alternatives. Incorrect choice of the model dimension can substantially decrease the power of a test. In section 6, we give a theoretical explanation of this fact for the case of deconvolution density testing. The possible solution of this problem is to incorporate the test statistic of interest by some procedure (called a selection rule) that chooses a reasonable dimension of the model automatically by the data. See Kallenberg (2002) for an extensive discussion and practical examples. In this section we implement this idea for testing the deconvolution hypothesis. First we give a definition of selection rule, generalizing ideas from Inglot and Ledwina (2006).

Let $M_k(\theta)$ be the model described in section 4 and let the true parameter $\theta$ belong to the parameter set, say $\Theta_k$, and dim $\Theta_k = k$. The family of submodels $M_k(\theta)$ for $k = 1, 2, \ldots$ is nested in the sense that for their parameter sets of interest it holds that $\Theta_1 \subset \Theta_2 \subset \cdots$.

**Definition 2.** Consider a nested family of submodels $M_k(\theta)$ for $k = 1, \ldots, d$, where $d$ is fixed but otherwise arbitrary. Choose a function $\pi(\cdot, \cdot) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$, where $\mathbb{N}$ is the set of natural numbers. Assume that $\pi(1, n) < \pi(2, n) < \cdots < \pi(d, n)$ for all $n$ and $\pi(j, n) - \pi(1, n) \to \infty$ as $n \to \infty$ for every $j = 2, \ldots, d$. Call $\pi(j, n)$ a penalty attributed to the $j$th model $M_j(\theta)$ and the sample size $n$. Then a selection rule $S$ for the test statistic $U_k$ is an integer-valued random variable satisfying the condition

$$
S = \min\{k : 1 \leq k \leq d; U_k - \pi(k, n) \geq U_j - \pi(j, n), j = 1, \ldots, d\}.
$$

(13)

We call $U_S$ a data-driven efficient score test statistic for testing validity of the initial model.

From theorem 3 it follows that for our problem (as well as in the classical case, see Kallenberg (2002)) many possible penalties lead to consistent tests. So the choice of the penalty should be dictated by practical reasons. Our simulation study is not so vast to recommend the most practically suitable penalty for the deconvolution density testing problem. Possible choices are, for example, Schwarz’s penalty $\pi(j, n) = j \log n$, or minimum description length penalties (see Rissanen (1983)).

Denote by $P^n_0$ the probability measure corresponding to the case when $X_1, \ldots, X_n$ all have the density $f_0$. For simplicity of notation we will further sometimes omit index ’n’ and write simply $P_0$. The main result about the asymptotic null distribution of $U_S$ is the following.

**Theorem 3.** Suppose that assumptions (B1)–(B3) hold. Then under the null hypothesis $H_0$ it holds that $P^n_0(S > 1) \to 0$ and $U_S \to_d \chi^2_d$ as $n \to \infty$.

**Remark 4.** The selection rule $S$ can be modified in order to make it possible to choose not only models of dimension less than some fixed $d$ but to allow arbitrary large dimensions of
M_k(\theta) as n grows to infinity. In this case an analogue of theorem 3 still holds, but the proof becomes more technical and one should take care about the possible rates of growth of the model dimension. Though, one can argue that even \( d = 10 \) is often enough for practical purposes (see Kallenberg and Ledwina (1995)).

6. Consistency of tests

Let \( F \) be a true distribution function of \( X \). Here \( F \) is not necessarily parametric and possibly does not have a density with respect to \( \lambda \). Let us choose for every \( k \leq d \) an auxiliary parametric family \( \{f_\theta\}, \theta \in \Theta \subseteq \mathbb{R}^k \) such that \( f_0 \) from this family coincides with \( f_0 \) from the null hypothesis \( H_0 \). Suppose that the chosen family \( \{f_\theta\} \) gives us the regular deconvolution density testing problem in the sense of definition 1. Then one is able to construct the score test statistic \( U_k \) defined by (7) despite the fact that the true \( F \) possibly has no relation to the chosen \( \{f_\theta\} \). One can use the exponential family from example 3 as \( \{f_\theta\} \), or some other parametric family, whatever is convenient. This is our goal in this section to determine under what conditions thus built \( U_k \) will be consistent for testing against \( F \).

Suppose that the following condition holds

\[ \langle D_1 \rangle \quad \text{there exists an integer } K \geq 1 \text{ such that } K \leq d \text{ and } E_F l^*_1 = 0, \ldots, E_F l^*_{K-1} = 0, \quad E_F l^*_K = C_K \neq 0, \]

where \( l^*_i \) is the \( i \)-th coordinate function of \( l^* \) and \( l^* \) is defined by (5), \( d \) is the maximal possible dimension of our model as in definition 2 of section 5 and \( E_F \) denotes the mathematical expectation with respect to \( F \). h.

We prove in this section various consistency results for our tests. It is assumed that we test the simple null hypothesis \( H_0 \) from section 3 against an alternative hypothesis \( H_A \), where as \( H_A \) one can take any set of (fixed) alternatives satisfying condition \( \langle D_1 \rangle \).

Condition \( \langle D_1 \rangle \) is a weak analogue of nondegeneracy: if for all \( k \langle D_1 \rangle \) fails, then \( F \) is orthogonal to the whole system \( \{l^*_i\}_{i=1}^\infty \), and if this system is complete, then \( F \) is degenerate. Also \( \langle D_1 \rangle \) is related to the identifiability of the model (see the beginning of section 11 for more details).

We start with investigation of consistency of \( U_k \), where \( k \) is some fixed number, \( 1 \leq k \leq d \). The following result shows why it is important to choose the right dimension of the model.

**Proposition 5.** Let \( \langle D_1 \rangle \) hold. Then for all \( 1 \leq k \leq K - 1 \), if \( F \) is the true distribution function of \( X \), then \( U_k \rightarrow_d \chi^2_k \) as \( n \rightarrow \infty \).

This result and theorem 1 show that if the dimension of the model is too small, then the test does not work since it does not distinguish between \( F \) and \( f_0 \).

**Proposition 6.** Let \( \langle D_1 \rangle \) hold. Then for \( k \geq K \), if \( F \) is the true distribution function of \( X \), then \( U_k \rightarrow \infty \) in probability as \( n \rightarrow \infty \).

Now we turn to the data-driven statistic \( U_S \). Suppose that the selection rule \( S \) is defined as in section 5. Assume that

\[ \langle S_1 \rangle \quad \text{for every fixed } k \geq 1 \text{ it holds that } \pi(k, n) = o(n) \text{ as } n \rightarrow \infty. \]

Denote by \( P_F \) the probability measure corresponding to the case when \( X_1, \ldots, X_n \) all have the distribution \( F \). Consider consistency of the ‘adaptive’ test based on \( U_S \).

**Proposition 7.** Let \( \langle D_1 \rangle \) and \( \langle S_1 \rangle \) hold. If \( F \) is the true distribution function of \( X \), then \( P_F(S \geq K) \rightarrow 1 \) and \( U_S \rightarrow \infty \) as \( n \rightarrow \infty \).
The main result of this section is the following.

**Theorem 8.**

(i) The test based on $U_k$ is consistent for testing against all alternative distributions $F$ such that $\langle D_1 \rangle$ is satisfied with $K \leq k$.

(ii) The test based on $U_k$ is inconsistent for testing against all alternative distributions $F$ such that $\langle D_1 \rangle$ is satisfied with $K > k$.

(iii) If the selection rule $S$ satisfies $\langle S_1 \rangle$, then test based on $U_S$ is consistent against all alternative distributions $F$ such that $\langle D_1 \rangle$ is satisfied with some $K$.

This theorem provides an explicit rule allowing us to decide whether our tests based on $U_k$ or $U_S$ could be used for testing a certain hypothesis or not. Having some specific testing problem, one has to check if the alternatives to the null hypothesis in this problem satisfy condition $\langle D_1 \rangle$. As theorem 8 shows, our tests should be used only if the consistency condition holds.

**Remark 9.** Theorem 8 is an asymptotic result, i.e., the tests based on $U_k$ or $U_S$ are guaranteed to work when the number of observations $n$ is sufficiently large. For moderate sample sizes, one has to use approximate versions of the tests. Approximate tests can be based on simulated critical values obtained by bootstrap, see remark 2.

**Remark 10.** Note that in many classical statistical problems $n = 50$ is already large enough for asymptotic theory to work very well without using bootstrapping or other approximation methods, but in statistical inverse problems the situation is much more difficult and unpredictable. In the deconvolution density testing problem, accuracy of the asymptotic predictions heavily depends, for example, on the smoothness of the noise density. The dependence is nontrivial. Sometimes 100–200 observations are enough, and sometimes one has to use 800 or more observations.

7. Composite deconvolution

In the previous sections, we treated the simplest case of the deconvolution density testing problem. The following sections are devoted to the more realistic case of unknown error density. Our main ideas and constructions will be similar to those for the simple case. Our goal is to modify the technics and constructions from the simple hypothesis case in order to apply them in the new situation. In order to do this we will have to impose on our new model additional regularity assumptions of uniformity. These assumptions are quite standard in statistics. They are a necessary payment for our ability to keep simple and general constructions for the more complicated problem. We will have to modify the scores we used in the simple case. Such a modification is called efficient scores.

Despite all the changes, we will be able to build a selection rule for the new problem. We will need a slightly modified definition of the selection rule, but the rest of our construction is essentially the same as for the simple hypothesis case.

Consider the situation described in the first paragraph of section 3, but with the following complication introduced. Suppose further on that the density $h$ of $\varepsilon$ is unknown.

Then the most general possible null hypothesis $H_0$ in this setup is that $f = f_0$ and the error $\varepsilon$ has expectation 0 and finite variance. The most general alternative hypothesis $H_A$ is that $f \neq f_0$. Since both $H_0$ and $H_A$ are in this case too broad, we would first consider a special class of submodels of the model described above. First we assume that all possible densities $f$ of $X$ belong to some specific and preassigned parametric family $\{f_\theta\}$, i.e., $f = f_\theta$ for some $\theta$ and $\theta$ is a $k$-dimensional Euclidian parameter and $\Theta \subseteq R^k$ is a parameter set for $\theta$. 

Our starting assumption about the density of the error \( \varepsilon \) will be that \( \varepsilon \) belongs to some specific parametric family \( \{h_{\eta}\} \), where \( \eta \in \Lambda \) and \( \Lambda \subseteq \mathbb{R}^m \) is a parameter set. Thus, \( \eta \) is a nuisance parameter.

If \( (\theta, \eta) \) is a true parameter value, denote in this case the density of \( Y \) by \( g(\cdot; (\theta, \eta)) \) and the corresponding expectation by \( E(\theta, \eta) \). The model \( M_{k,m}(\theta, \eta) \) for the parameter \( (\theta, \eta) \) is defined analogously to section 4. Let the null hypothesis \( H_0 \) be \( \theta = \theta_0 \), where it is assumed that \( \theta_0 \in \Theta \). We prove below that the alternative hypothesis \( H_A \) can be any collection of parametric or nonparametric (fixed) alternatives satisfying consistency condition (C1) below.

8. Efficient scores

All possible densities \( g(y; (\theta, \eta)) \) of \( Y \) have in our model the form

\[
g(y; (\theta, \eta)) = \int_{\mathbb{R}} f_\theta(s)h_\eta(y-s) \, ds.
\] (14)

It is not always possible to identify \( \theta \) or/and \( \eta \) in this model. Since we are concerned with testing hypotheses and not with estimation of parameters, it is not necessary for us to impose a restrictive assumption of identifiability on the model. We will need only a (weaker) consistency condition to build a sensible test (see section 11).

The score function for \( (\theta, \eta) \) at \((\theta_0, \eta_0)\) is defined as (see Bickel et al 1993, p 28):

\[
\dot{i}_{\theta_0, \eta_0}(y) = (\dot{i}_{\theta_0}(y), \dot{i}_{\eta_0}(y)),
\] (15)

where \( \dot{i}_{\theta_0} \) is the score function for \( \theta \) at \( \theta_0 \) and \( \dot{i}_{\eta_0} \) is the score function for \( \eta \) at \( \eta_0 \), i.e.,

\[
\dot{i}_{\theta_0}(y) = \frac{d}{d\theta} \left( \int_{\mathbb{R}} f_\theta(s)h_{\eta_0}(y-s) \, ds \right) \bigg|_{\theta=\theta_0} 1_{\{y:g(y; (\theta_0, \eta_0)) > 0\}}.
\] (16)

\[
\dot{i}_{\eta_0}(y) = \frac{d}{d\eta} \left( \int_{\mathbb{R}} f_{\theta_0}(s)h_{\eta_0}(y-s) \, ds \right) \bigg|_{\eta=\eta_0} 1_{\{y:g(y; (\theta_0, \eta_0)) > 0\}}.
\] (17)

The Fisher information matrix of parameter \( (\theta, \eta) \) is defined as

\[
I(\theta, \eta) = \int_{\Theta} \dot{i}_{\theta, \eta}^T(y)I_{\theta, \eta}(y) \, dG_{\theta, \eta}(y),
\] (18)

where \( G_{\theta, \eta}(y) \) is the probability measure corresponding to the density \( g(y; (\theta, \eta)) \). We assume that \( M_{k,m}(\theta, \eta) \) is a regular composite deconvolution density testing problem (see Langovoy 2007).

Let us write \( I(\theta_0, \eta_0) \) in the block matrix form

\[
I(\theta_0, \eta_0) = \begin{pmatrix} I_{11}(\theta_0, \eta_0) & I_{12}(\theta_0, \eta_0) \\ I_{21}(\theta_0, \eta_0) & I_{22}(\theta_0, \eta_0) \end{pmatrix},
\] (19)

where \( I_{11}(\theta_0, \eta_0) \) is \( k \times k \), \( I_{12}(\theta_0, \eta_0) \) is \( k \times m \) and so on. Thus, denoting for simplicity of formulae \( \Omega := \{y: g(y; (\theta_0, \eta_0)) > 0\} \) we can write explicitly

\[
I_{11}(\theta_0, \eta_0) = E_{\theta_0, \eta_0} \dot{i}_{\theta_0}^T \dot{i}_{\theta_0} = \int_{\Omega} \dot{i}_{\theta_0}^T(y)\dot{i}_{\theta_0}(y) \, dG_{\theta_0, \eta_0}(y)
\]

\[
= \int_{\Omega} \frac{d}{d\theta_0} \left( \int_{\mathbb{R}} f_{\theta_0}(s)h_{\eta_0}(y-s) \, ds \right) \bigg|_{\theta=\theta_0} \frac{d}{d\theta_0} \left( \int_{\mathbb{R}} f_{\theta_0}(s)h_{\eta_0}(y-s) \, ds \right) \bigg|_{\theta=\theta_0} \, dy,
\] (20)

\[
I_{12}(\theta_0, \eta_0) = E_{\theta_0, \eta_0} \dot{i}_{\eta_0}^T \dot{i}_{\eta_0} = \int_{\Omega} \dot{i}_{\eta_0}^T(y)\dot{i}_{\eta_0}(y) \, dG_{\theta_0, \eta_0}(y),
\] (21)
and analogously for $I_{21}(\theta_0, \eta_0)$ and $I_{22}(\theta_0, \eta_0)$. The efficient score function for $\theta$ in $M_{k,m}(\theta, \eta)$ is defined as (see Bickel et al. (1993), p 28):

$$I^*_\theta(y) = \frac{\partial}{\partial \theta} \log f(y; \theta, \eta) = - \frac{\partial}{\partial \theta} \log f_{\theta,\eta}(y),$$

and the efficient Fisher information matrix for $\theta$ in $M_{k,m}(\theta, \eta)$ is defined as

$$I^* = E_{\theta,\eta} I^*_\theta \left( I^*_\theta \right)^T = \int I^*_\theta(y)^2 \, dG_{\theta,\eta}(y).$$

Before closing this section we consider two simple examples.

**Example 4.** Suppose that we are interested in the parameter $\theta$ and testing for each parameter, the problem is symmetric in the sense that it is possible to consider estimating $\theta$ and $\eta$ equally well whether we know the true $\eta_0$ or not. Though, we will not be concerned with estimation here. From (20) we get

$$I_{10}(\theta, \eta) = \int_\mathbb{R} \frac{y - \theta_0}{\eta_0^2 + 1} \frac{(y - \theta_0)^2 \eta_0}{(\eta^2 + 1)^2} \, dN(\theta, \eta^2 + 1)(y),$$

for all $\theta, \eta$. This means that adaptive estimation of $\theta$ is possible in this model, i.e., we can estimate $\theta$ equally well whether we know the true $\eta_0$ or not. Though, we will not be concerned with estimation here. From (20) we get

$$I_{10}^{-1}(\theta, \eta) = \int_{\mathbb{R}} (y - \theta)^2 dN(\theta, \eta^2 + 1)(y) = \frac{1}{\eta_0^2 + 1}.$$  

**Example 5.** Suppose that we are interested in the parameter $\eta$ from example 4 and the null hypothesis is $H_0 : \eta = \eta_0$. Since we proved for this example $I_{12} = I_{21} = 0$, the efficient score function $l^*_\eta$ for $\eta$ at $\eta_0$ is given by (24) as well. We calculate now

$$(I_{10}^*)^{-1} = \int_{\mathbb{R}} \left( \frac{(y - \theta)^2 \eta_0}{(\eta_0^2 + 1)^2} - \frac{\eta_0}{\eta_0^2 + 1} \right)^2 \, dN(\theta, \eta_0^2 + 1)(y) = \frac{1}{C(\eta_0)}.$$  

The constant $C(\eta_0)$ in (26) can be expressed explicitly in terms of $\eta_0$. By the symmetry of $\theta$ and $\eta$ we have $I^*_\eta(y) = I_{10}(y) - I_{12}(\theta_0, \eta_0) I_{11}^{-1}(\theta_0, \eta_0) l_{10}(y) = l_{10}(y)$.

**Remark 11.** The problem is symmetric in the sense that it is possible to consider estimating and testing for each parameter, $\theta$ or $\eta$. From the noisy signal one can partially recover some ‘information’ not only about the pure signal but also about the noise.

### 9. Efficient score test

Let $l^*_\theta$ be defined by (22) and $l^*_\eta$ by (23). Note that both $l^*_\theta$ and $l^*_\eta$ depend (at least in principle) on the unknown nuisance parameter $\eta_0$. Let $I^*_\eta$ and $L$ be some estimators of $l^*_\eta(Y_j)$ and $(I^*_\eta)^{-1}$ correspondingly. These estimators are supposed to depend only on the observable $Y_1, \ldots, Y_n$, but not on the $X_1, \ldots, X_n$.

**Definition 3.** We say that $l^*_\eta$ is a sufficiently good estimator of $l^*_\eta(Y_j)$ if the average

$$\frac{1}{n} \sum_{j=1}^n l^*_\eta(Y_j) \approx E_{\theta_0,\eta_0} l^*_\eta \text{ is } \sqrt{n}\text{-consistently estimated.}$$
We refer to Langovoy (2007) for more details and here only illustrate this definition by some examples.

**Example 4** (continued). Define

\[ l_j^* := \frac{Y_j - \theta_0}{\hat{\sigma}_n^2}, \]

where \( \hat{\sigma}_n^2 \) is any \( \sqrt{n} \)-consistent estimator of the variance of \( Y \). One can take, for example, the sample variance \( \hat{s}_n^2 = \frac{1}{n} \sum_{j=1}^{n} (Y_j - \bar{Y})^2 \) as such an estimate. Then \( l_j^* \) satisfies definition 3. See Langovoy (2007) for the proof.

**Example 5** (continued). For simplicity of notations we write \( l_{\eta_0}^* = C_1(\eta_0)(Y_j - \theta_0)^2 - C_2(\eta_0) \). Let \( \hat{\theta}_n \) be any \( \sqrt{n} \)-consistent estimate of \( \theta_0 \) and put \( l_j^* := C_1(\eta_0)(Y_j - \hat{\theta}_n)^2 - C_2(\eta_0) \). Then definition 3 is satisfied, see Langovoy (2007).

Definition 3 reflects the basic idea of the method of estimated scores. This method is widely used in statistics (see Bickel et al (1993), Schick (1986), Ibragimov and Has’minskiĭ (1981), Inglot and Ledwina (2006) and others). These authors show that for different problems it is possible to construct nontrivial parametric, semi- and nonparametric estimators of scores such that these estimators will satisfy definition 3. See also Langovoy (2007) for many examples and a discussion of different methods that can be used to estimate efficient score functions.

**Definition 4.** Define

\[
W_k = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l_j^* \right\} \hat{L} \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l_j^* \right\}^T, \tag{27}
\]

where \( \hat{L} \) is an estimate of \( (I_n^*)^{-1} \) depending only on \( Y_1, \ldots, Y_n \). Note that \( l_j^* \) is a \( k \)-dimensional vector and \( \hat{L} \) is a \( k \times k \) matrix. We call \( W_k \) the efficient score test statistic for testing \( H_0 : \theta = \theta_0 \) in \( M_{k,2}(\theta, \eta) \). It is assumed that the null hypothesis is rejected for large values of \( W_k \).

Normally it should be possible to construct reasonably good estimators \( \hat{\eta}_n \) of \( \eta \) by standard methods since at this point our construction is parametric. After that it would be enough to plug in these estimates in (22) and get the desired \( l_j^* \)'s. As for practical implementation of this test in the case of moderate sample sizes, everything can go along the lines of remarks 2 and 9, although the amount of computations needed for estimating critical values increases due to appearance of several estimators in definition 4.

**Example 4** (continued). Let \( \hat{\sigma}^2(Y) \) be any \( \sqrt{n} \)-consistent estimate of \( \eta^2 + 1 \) such that this estimate is based on \( Y_1, \ldots, Y_n \). Then by (25), (24) and definition (27) the efficient score test statistic for testing \( H_0 : \theta = \theta_0 \) (in the model \( M_{1,1}(\theta, \eta) \)) is

\[
W_1 = \left( \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{Y_j - \theta_0}{\hat{\sigma}^2_n(Y)} \right)^2 \hat{\sigma}^2_n(Y) = \frac{1}{\hat{\sigma}^2_n(Y)} \left( \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (Y_j - \bar{Y}) \right)^2. \tag{28}
\]

**Example 5** (continued). Using any \( \sqrt{n} \)-consistent estimate \( \hat{\theta} \) of \( \theta \), we get the efficient score test statistic

\[
W_1 = \left( \frac{1}{\sqrt{n}} \frac{\eta_0}{\eta_0^2 + 1} \sum_{j=1}^{n} (Y_j - \hat{\theta}_n)^2 - \frac{\eta_0}{\eta_0^2 + 1} \right)^2 C(\eta_0). \tag{29}
\]
Remark 12. For the simple hypothesis we had the score test statistics and now we have the efficient score test statistics. Here the word ‘efficient’ has a technical meaning and the statistics for simple hypotheses are not ‘inefficient’. Because of the presence of the nuisance parameter, we have to extract information about the parameter of interest from the score. We want to do this efficiently in some sense. This is the explanation of the term. In this paper, the terminology of Bickel et al (1993) is used. Sometimes, efficient score tests are also called Neyman C(α)-tests, conditional moment restriction tests or Lagrange multiplier tests (in the econometrics literature).

The following theorem describes asymptotic behavior of $W_k$ under the null hypothesis (see Langovoy (2007)).

Theorem 13. Assume the null hypothesis $H_0 : \theta = \theta_0$ holds true, $\langle A1 \rangle$–$\langle A3 \rangle$ are fulfilled, estimators of efficient scores are sufficiently good, and $\hat{L}$ is any consistent estimate of $(I_{\theta_0}^*)^{-1}$. Then $W_k \to_d \chi^2_k$ as $n \to \infty$.

10. Selection rule

In this section we extend the construction of section 5 to the case of composite hypotheses. First we modify the definition of a selection rule.

Let $M_{k,m}(\theta, \eta)$ be the model described in section 7 and such that the true parameter $(\theta, \eta)$ belongs to a parameter set, say $\Theta_k \times \Lambda$, and $\text{dim } \Theta_k = k$. The family $M_{k,m}(\theta, \eta)$ for $k = 1, \ldots, d$ is nested in a sense that $\Theta_1 \times \Lambda \subset \Theta_2 \times \Lambda \subset \cdots$.

Definition 5. Consider a nested family of submodels $M_{k,m}(\theta, \eta)$ for $k = 1, \ldots, d$, where $d$ is fixed but otherwise arbitrary, and $m$ is fixed. Let $\pi(k, n)$ be a penalty, attributed to the $k$th model $M_{k,m}(\theta, \eta)$ and the sample size $n$, and satisfying definition 2. Then a selection rule $S(l^*)$ for the test statistic $W_k$ is an integer-valued random variable satisfying the condition $S(l^*) = \min\{k : 1 \leq k \leq d; W_k - \pi(k, n) \geq W_j - \pi(j, n), j = 1, \ldots, d\}$. (30)

We call the random variable $W_S$ a data-driven efficient score test statistic for testing validity of the initial model. In this case we also assume that condition $\langle S1 \rangle$ holds.

Unlike the case of the simple null hypothesis, in the case of the composite hypotheses the selection rule depends on the estimator $\hat{l}_j^*$ of the unknown values $l_{0j}^*(Y_j)$ of the efficient score function. This means that we need to estimate the nuisance parameter $\eta$, or corresponding scores, or their sum. Surprising result follows from theorem 14: for our problem many possible penalties and, moreover, essentially all sensible estimators plugged in $W_k$ give consistent selection rules.

The main result about the asymptotic null distribution of $W_S$ is the following (see Langovoy (2007)).

Theorem 14. Under the conditions of theorem 13, it holds that $W_S \to_d \chi^2_1$ as $n \to \infty$.

Remark 15. The selection rule $S(l^*)$ can be modified in order to allow arbitrary large dimensions of $M_{k,m}(\theta, \eta)$ as the number of observations grows. See remark 4.

Remark 16. It is possible to modify the definition of selection rule so that both dimensions $k$ and $m$ would be selected by the test from the data. A corresponding test statistic will be of the form $W_S$, where this time $S = (S_1, S_2)$. Proofs of the asymptotic properties for this statistic are analogous to those presented in this paper.
11. Consistency of tests

Let $F$ be a true distribution function of $X$ and $H$ a true distribution of $\varepsilon$. Here $F$ and $H$ are not necessarily parametric and possibly these distribution functions do not have densities with respect to the Lebesgue measure $\lambda$. Let us choose for every $k \leq d$ an auxiliary parametric family $\{f_0\}, \theta \in \Theta \subseteq \mathbb{R}^d$ such that $f_0$ from this family coincides with $f_0$ from the null hypothesis $H_0$. Correspondingly, let us fix an integer $m$ and choose an auxiliary parametric family $\{h_0\}, \eta \in \Lambda \subseteq \mathbb{R}^m$. Suppose that the chosen families $\{f_0\}$ and $\{h_0\}$ give us the regular composite deconvolution density testing problem. Then one is able to construct the score test statistic $W_k$ defined by (27) despite the fact that the true $F$ and $H$ possibly do not have any relation to the chosen $\{f_0\}$ and $\{h_0\}$. This is our goal in this section to determine under what conditions thus built $W_k$ will be consistent for testing against $H_A$.

Suppose that the following condition holds

$$(C1) \quad \text{there exists integer } K \geq 1 \text{ such that } K \leq d \text{ and }$$

$$E_{F \ast H} l^*_0(1) = 0, \ldots, E_{F \ast H} l^*_0(K-1) = 0, \quad E_{F \ast H} l^*_0(K) = C_K \neq 0,$$

where $l^*_0(i)$ is the $i$th coordinate function of $l^*_0$ and $l^*_0$ is defined by (22), $d$ is the maximal possible dimension of our model as in section 10, and $E_{F \ast H}$ denotes the mathematical expectation with respect to $F \ast H$.

Condition $(C1)$ is a weak analogue of nondegeneracy: if for all $k(C1)$ fails, then $F$ is orthogonal to the whole system $l^*_0(1) = \ldots = l^*_0(K) = 0$ and if this system is complete, then $F \ast H$ is degenerate. Also $(C1)$ is related to the identifiability of the model: if the model is not identifiable, then $F \ast H = F_0 \ast H$ can happen and $(C1)$ fails. Establishing identifiability for the parametric deconvolution is not trivial (see Sclove and Van Ryzin (1969)). It is important to note also that although $(C1)$ has something common with both nondegeneracy and identifiability, it is in general pretty far from both these notions.

The main result of this section is the following.

**Theorem 17.** If the estimators of efficient scores are sufficiently good and $\hat{\Lambda}$ is a consistent estimate of $(l^*_0)^{-1}$, then

(i) the test based on $W_k$ is consistent for testing against all alternative distributions $F, H$ such that $(C1)$ is satisfied with $K \leq k$;

(ii) the test based on $W_k$ is inconsistent for testing against alternative distributions $F, H$ such that $(C1)$ is satisfied with $K > k$;

(iii) if the selection rule $S(l^*)$ satisfies (S1), then test based on $W_k$ is consistent against all alternative distributions $F \ast H$ such that $(C1)$ is satisfied with some $K$.

This is an analogue of theorem 8 for the case of unknown error density. The proof can be found in Langovoy (2007). Practical implementation of this test for small sample sizes goes along the lines of remarks 2, 9 and 10. Now we give two illustrative theoretical examples.

**Example 4** (continued). By theorem 17 the test based on $W_l$ is consistent if and only if for true $F$ and $H$ it holds that

$$\frac{1}{\eta^2 + 1} E_{F \ast H} (Y - \theta_0) \neq 0, \quad \text{i.e. } E_{F \ast H} (Y) \neq \theta_0.$$  \hspace{1cm} (31)

For example, $W_l$ does not work when the true $H$ is symmetric about 0 and the true $F \neq F_0$ has the mean equal to $\theta_0$. 
Example 5 (continued). By theorem 17 $W_1$ is consistent if and only if for true $F$ and $H$ it holds that
\begin{equation}
\begin{aligned}
E_{F+H} \left( \frac{(y - \theta)^2 \eta_0}{(\eta_0^2 + 1)^2} - \eta_0 \eta_0 + 1 \right) \neq 0, \quad \text{i.e.}
\end{aligned}
\end{equation}
\begin{equation}
E_{F+H} (y - \theta)^2 \eta_0^2 + 1, \quad \text{or equivalently} \quad \text{Var}_{F+H} Y \neq \text{Var}_{F+H} Y_0.
\end{equation}

One cannot expect more from such a simple test as $W_1$. In contrast, in most of the cases when $W_1$ fails, the data-driven test statistic $W_S$ provides a consistent testing procedure for any $d \geq 2$ and all the reasonable penalties.

12. Conclusions

Score tests and data-driven score tests can be used in statistical inverse problems. Data-driven score tests are preferable from the asymptotic point of view because they usually have a wider range of consistency than ordinary score tests. It is useful to have an explicit consistency condition (like $\langle C_1 \rangle$ or $\langle D_1 \rangle$) when a score test is used. The consistency condition allows us to check whether it is reasonable to use the score test for any particular problem.

Acknowledgments

The author would like to thank Axel Munk for suggesting this topic of research, Fadoua Balabdaoui, Andrei Volotka, Ta-Chao Kao, Wilbert Kallenberg, Janis Valeinis for helpful discussions, and Luis Tenorio and two referees for valuable suggestions that helped to greatly improve the presentation of the paper. The financial support of the Deutsche Forschungsgemeinschaft GK 1023 ‘Identifikation in Mathematischen Modellen’ is gratefully acknowledged.

Appendix

Proof (theorem 1). We calculated the efficient score vector in (4) and (5). By proposition 1, p 13 of Bickel et al (1993) and our regularity assumptions matrix $L$ exists and is positive definite and nondegenerate of rank $k$. Under $\langle B1 \rangle$--$\langle B3 \rangle$ $E_0 l^*(y) = 0$ (see Bickel et al (1993), p 15) and our statement follows. □

Proof (proposition 5). Follows by the multivariate central limit theorem. □

Proof (theorem 3). Denote $\Delta(k, n) := \pi(k, n) - \pi(1, n)$. For any $k = 2, \ldots, d$
\begin{equation}
P_0^n (S = k) \leq P_0^n (U_k - \pi(k, n) \geq U_1 - \pi(1, n)) \leq P_0^n (U_k \geq \pi(k, n) - \pi(1, n)) = P_0^n (U_k \geq \Delta(k, n)).
\end{equation}

By theorem 1 $U_k \rightarrow_d \chi_1^2$ as $n \rightarrow \infty$, thus for $\Delta(k, n) \uparrow \infty$ as $n \rightarrow \infty$ we have $P_0^n (U_k \geq \Delta(k, n)) \rightarrow 0$ as $n \rightarrow \infty$, so for any $k = 2, \ldots, d$ we have $P_0^n (S = k) \rightarrow 0$ as $n \rightarrow \infty$. This proves that
\begin{equation}
P_0^n (S \geq 2) = \sum_{k=2}^d P_0^n (S = k) \rightarrow 0, \quad n \rightarrow \infty.
\end{equation}
and so $P^n_0(S = 1) \to 1$. Now write for arbitrary real $t > 0$

$$P^n_0(|US - U_1| \geq t) = P^n_0(|U_1 - U_1| \geq t; S = 1) + \sum_{m=2}^{d} P^n_0(|U_m - U_1| \geq t; S = m)$$

$$= \sum_{m=2}^{d} P^n_0(|U_m - U_1| \geq t; S = m). \quad (A.1)$$

For $m = 2, \ldots, d$ we have $P^n_0(S = m) \to 0$, so

$$0 \leq \sum_{m=2}^{d} P^n_0(|U_m - U_1| \geq t; S = m) \leq \sum_{m=2}^{d} P^n_0(S = m) \to 0$$

as $n \to \infty$ and thus by (A.1) it follows that $U_S$ tends to $U_1$ in probability as $n \to \infty$. But $U_1 \to_d \chi^2_1$ by theorem 1, so $U_S \to_d \chi^2_1$ as $n \to \infty$. \hfill \Box

**Proof** (proposition 6). From $\langle D_1 \rangle$ by the law of large numbers we get

$$\frac{1}{n} \sum_{j=1}^{n} l^*_i(Y_j) \to_P 0 \quad \text{for} \quad 1 \leq i \leq K - 1 \quad (A.2)$$

$$\frac{1}{n} \sum_{j=1}^{n} l^*_K(Y_j) \to_P C_K \neq 0. \quad (A.3)$$

Since all the eigenvalues of the matrix $L$ defined in (6) are positive, one can choose a positive number $\delta$ less than the smallest eigenvalue of $L$. We then obtain the following inequality:

$$U_k = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l^*_i(Y_j) \right\} L \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l^*_i(Y_j) \right\}^T$$

$$\geq \delta \left\| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l^*_i(Y_j) \right\|^2 = \delta n \sum_{i=1}^{k} \left( \frac{1}{n} \sum_{j=1}^{n} l^*_i(Y_j) \right)^2$$

$$\geq \delta n \left( \frac{1}{n} \sum_{j=1}^{n} l^*_K(Y_j) \right)^2. \quad (A.4)$$

Now by (A.2) and (A.3) we get for all $s \in \mathbb{R}$

$$P(U_k \leq s) \leq P \left( \delta n \left( \frac{1}{n} \sum_{j=1}^{n} l^*_K(Y_j) \right)^2 \right) \leq s$$

$$= P \left( \left| \frac{1}{n} \sum_{j=1}^{n} l^*_K(Y_j) \right| \leq \frac{s}{\delta n} \right) \to 0 \quad \text{as} \quad n \to \infty,$$

and this proves the proposition. \hfill \Box

**Proof** (proposition 7). Let $\pi(k, n)$ and $\Delta(k, n)$ be defined as in section 5. For any $i = 1, \ldots, K - 1$ we have

$$P_F(S = i) \leq P_F(U_i - \pi(i, n) \geq U_K - \pi(K, n))$$

$$= P_F(U_i \geq U_K - (\pi(K, n) - \pi(i, n))). \quad (A.5)$$
By (A.3) and (A.4) we get
\[
P_{F}\left(U_{K} \geq \frac{\delta C_{K}}{2n}\right) \to 1 \quad \text{as} \quad n \to \infty. \tag{A.6}
\]

Note that
\[
P_{F}\left(U_{i} \geq U_{K} - (\pi(K, n) - \pi(i, n))\right) \leq P_{F}\left(U_{i} \geq \frac{\delta C_{K}}{2n} - (\pi(K, n) - \pi(i, n)); U_{K} \geq \frac{\delta C_{K}}{2n}\right)
+ P_{F}\left(U_{K} \leq \frac{\delta C_{K}}{2n}\right). \tag{A.7}
\]

Since by (S1) it holds that \(\pi(K, n) - \pi(i, n) = o(n)\), we get
\[
P_{F}\left(U_{i} \geq \frac{\delta C_{K}}{2n} - (\pi(K, n) - \pi(i, n)); U_{K} \geq \frac{\delta C_{K}}{2n}\right) \leq P_{F}\left(U_{i} \geq \frac{\delta C_{K}}{2n} - (\pi(K, n) - \pi(i, n))\right) \leq P_{F}\left(U_{i} \geq \frac{\delta C_{K}}{2n}\right) \to 0 \tag{A.8}
\]

as \(n \to \infty\) by Chebyshev’s inequality since by proposition 5 we have \(U_{i} \to d \chi_{i}^{2}\) as \(n \to \infty\), for all \(i = 1, \ldots, K - 1\). Substituting (A.6) and (A.8) into (A.7) we get \(P_{F}(S = i) \to 0\) as \(n \to \infty\) for all \(i = 1, \ldots, K - 1\). This means that \(P_{F}(S \geq K) \to 1\) as \(n \to \infty\).

Now write for \(t \in \mathbb{R}\)
\[
P_{F}(U_{S} \leq t) = P_{F}(U_{S} \leq t; S \leq K - 1) + P_{F}(U_{S} \leq t; S \geq K) =: R_{1} + R_{2}.
\]

But \(R_{1} \to 0\) since \(P_{F}(S = i) \to 0\) for \(i = 1, \ldots, K - 1\) and \(K \leq d < \infty\). Since \(U_{l_{1}} \geq U_{l_{2}}\) for \(l_{1} \geq l_{2}\), we get
\[
R_{2} \leq \sum_{l=K}^{d} P_{F}(U_{K} \leq t) \to 0
\]
as \(n \to \infty\) by proposition 6. Thus \(P_{F}(U_{S} \leq t) \to 0\) as \(n \to \infty\) for all \(t \in \mathbb{R}\). □

**Proof** (theorem 8). Part 1 follows from theorem 1 and proposition 6, part 2 from theorem 1 and proposition 5, part 3 from theorem 3 and proposition 7. □

**References**

Bailey T C, Sapatinas T, Powell K J and Krzanowski W J 1998 Signal detection in underwater using wavelets J. Am. Stat. Assoc. 93 73–83

Barndorff-Nielsen O 1978 Information and Exponential Families in Statistical Theory (Wiley Series in Probability and Mathematical Statistics) (Chichester: Wiley) (ISBN 0-471-99545-2)

Bickel P J, Klaassen C A J, Ritov Y and Wellner J A 1993 Efficient and Adaptive Estimation for Semiparametric Models (Johns Hopkins Series in the Mathematical Sciences) (Baltimore, MD: Johns Hopkins University Press) (ISBN 0-8018-4541-6)

Bickel P J and Rosenblatt M 1973 On some global measures of the deviations of density function estimates Ann. Stat. 1 1071–95 (ISSN 0090-5364)
Butucea C 2007 Goodness-of-fit testing and quadratic functional estimation from indirect observations Ann. Stat. 35 1907

Cox D R and Hinkley D V 1974 Theoretical Statistics (London: Chapman and Hall)

Delaigle A and Gijbels I 2002 Estimation of integrated squared density derivatives from a contaminated sample J. R. Stat. Soc. Ser. B Stat. Methodol. 64 869–86 (ISSN 1369-7412)

Eubank R L, Hart J D and LaRiccia V N 1993 Testing goodness of fit via nonparametric function estimation techniques Commun. Stat. Theory Methods 22 3327–54 (ISSN 0361-0926)

Fan J 1996 Test of significance based on wavelet thresholding and [N]eyman’s truncation J. Am. Stat. Assoc. 91 674–88 (ISSN 0162-1459)

Hänsch T W 2006 Nobel lecture: passion for precision Rev. Mod. Phys. 78 1297 doi:10.1103/RevModPhys.78.1297

Hinkley D V 1988 Bootstrap methods J. R. Stat. Soc. Ser. B 50 321–37, 355 (ISSN 0035-9246. With discussion and a reply by the author)

Ibragimov I A and Has’minski˘ı R Z 1981 Statistical estimation, Applications of Mathematics vol 16 (New York: Springer) (ISBN 0-387-90523-5. Asymptotic theory, translated from the Russian by Samuel Kotz.)

Inglot T and Ledwina T 1996 Asymptotic optimality of data-driven Neyman’s tests for uniformity Ann. Stat. 24 1982–2019 (ISSN 0090-5364)

Inglot T and Ledwina T 2006 Asymptotic optimality of new adaptive test in regression model Ann. Inst. Henri Poincaré Probab. Stat. 42 579–90 (ISSN 0246-0203)

Ingster Yu I 1993 Asymptotically minimax hypothesis testing for nonparametric alternatives. I, II, III Math. Methods Stat. 2 85–114, 171 (ISSN 1066-5307)

Kallenberg W C M 2002 The penalty in data driven Neyman’s tests Math. Methods Stat. 11 323–40 (2003) (ISSN 1066-5307)

Kallenberg W C M and Ledwina T 1995 Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests Ann. Stat. 23 1594–608 (ISSN 0990-5364)

Kallenberg W C M and Ledwina T 1997 Data driven smooth tests for composite hypotheses: comparison of powers J. Stat. Comput. Simul. 59 101–21 (ISSN 0094-9655)

Langovoy M 2007 Data-driven goodness-of-fit tests PhD Thesis, University of Göttingen, Göttingen

Le Cam L 1956 On the asymptotic theory of estimation and testing hypotheses Proc. Third Berkeley Symp. on Mathematical Statistics and Probability, 1954–1955, vol. I (Berkeley, CA: University of California Press) pp 129–56

Ledwina T 1994 Data-driven version of Neyman’s smooth test of fit J. Am. Stat. Assoc. 89 1000–5 (ISSN 0162-1459)

Lehmann E L and Romano J P 2005 Testing statistical hypotheses Springer Texts in Statistics 3rd edn (New York: Springer) (ISBN 0-387-98864-5)

Neyman J 1937 Smooth test for goodness of fit Skand. Aktuarieidskr. 20 150–99

Neyman J 1959 Optimal asymptotic tests of composite statistical hypotheses Probability and Statistics: the Harald Cramér Volume ed U Grenander (Stockholm: Almqvist) pp 213–34

Niering M et al 2000 Measurement of the hydrogen 1s–2s transition frequency by phase coherent comparison with a microwave cesium fountain clock Phys. Rev. Lett. 84 5496–9 doi:10.1103/PhysRevLett.84.5496

Nikitin Y 1995 Asymptotic Efficiency of Nonparametric Tests (Cambridge: Cambridge University Press) (ISBN 0-521-47029-3)

Paschotta R, Schlatter A, Zeller S C, Telle H R and Keller U 2006 Optical phase noise and carrier-envelope offset noise of mode-locked lasers Appl. Phys. B 82 265–73 doi:10.1007/s00304-005-2041-9

Picard R and Bryson M 1992 Calibrated seismic verification of the threshold test ban treaty J. Am. Stat. Assoc. 87 293–9

Rissanen J 1983 A universal prior for integers and estimation by minimum description length Ann. Stat. 11 416–31 (ISSN 0090-5364)

Schick A 1986 On asymptotically efficient estimation in semiparametric models Ann. Stat. 14 1139–51 (ISSN 0090-5364)

Sclove S L and Van Ryzin J 1969 Estimating the parameters of a convolution J. R. Stat. Soc. Ser. B 31 181–91 (ISSN 0035-9246)

Spokoiny V G 1996 Adaptive hypothesis testing using wavelets Ann. Stat. 24 2477–98 (ISSN 0090-5364)