Model selection, large deviations and consistency of data-driven tests.

Mikhail Langovoy[*]

Mikhail Langovoy
EURANDOM, Technische Universiteit Eindhoven,
Postbus 513,
Den Dolech 2,
5600 MB Eindhoven, The Netherlands
e-mail: langovoy@eurandom.tue.nl
Phone: (+31) (40) 247 - 8113
Fax: (+31) (40) 247 - 8190

Abstract: We consider three general classes of data-driven statistical tests. Neyman's smooth tests, data-driven score tests and data-driven score tests for statistical inverse problems serve as important special examples for the classes of tests under consideration. Our tests are additionally incorporated with model selection rules. The rules are based on the penalization idea. Most of the optimal penalties, derived in statistical literature, can be used in our tests. We prove general consistency theorems for the tests from those classes. Our proofs make use of large deviations inequalities for deterministic and random quadratic forms. The paper shows that the tests can be applied for simple and composite parametric, semi- and nonparametric hypotheses. Applications to testing in statistical inverse problems and statistics for stochastic processes are also presented.

Primary 62G10; secondary 60F10, 62H10, 62E20.
Keywords and phrases: Data-driven test, model selection, large deviations, asymptotic consistency, generalized likelihood, random quadratic form, statistical inverse problems, stochastic processes.

1. Introduction

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 generally, L^p-distance based tests, as well as graphical tests based on confidence bands, usually belong to this type. Although, these tests works and are capable of giving very good results, but each of these tests is asymptotically optimal only in a finite number of directions of alternatives to a null hypothesis (see [1]).

Nowadays, there is an increasing interest to 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 in some sense, or most powerful, at least in a reach enough set of directions. Test statistics constructed

[*] Financial support of the Deutsche Forschungsgemeinschaft GK 1023 "Identifikation in Mathematischen Modellen" and of the Hausdorff Institute for Mathematics is gratefully acknowledged.

following this approach are often called score test statistics. The pioneer of this approach was [2]. See also, for example, [3], [4], [5], [6], [7], [8] for subsequent developments and improvements, and [9], [10] and [11] for recent results in the field. This approach is also closely related to the theory of efficient (adaptive) estimation - [12], [13]. Additionally, it was shown, at least in some basic situations, that data-driven score tests are asymptotically optimal in the sense of intermediate efficiency in an infinite number of directions of alternatives (see [14]) and show good overall performance in practice ([15], [16]).

Another important line of development in the area of optimal testing concerns with minimax testing, see [17]), and adaptive minimax testing, see [18]. 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, except for Remark 21 below. See [10] 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.

Classical score tests have been substantially generalized in recent literature: see, for example, the generalized likelihood ratio statistics for nonparametric models in [9], tailor-made tests in [10] and the semiparametric generalized likelihood ratio statistics in [11]. The situation is similar to the one in estimation theory: there is a classical estimation method based on the use of maximum likelihood equations, and there is a more general method of M-estimation.

In this paper we propose a new development of the theory of *data-driven* score tests. We introduce the notions of NT- and GNT-tests, generalizing the concepts of Neyman's smooth test statistics and data-driven score tests, for both simple and composite hypotheses. The main goal of this paper is to give a unified approach for proving consistency of NT- and GNT-tests.

When proving consistency of a data-driven score test for any specific problem, one has to:

1. establish large deviation inequalities for the test statistic
2. derive consistency of the test from those inequalities.

Usually both steps require lengthy and problem-specific computations. In some cases the model is so complicated that direct calculations are hardly possible to carry out. But consistency theorems of this paper allow to pass through step 2 almost automatically. This helps to substantially reduce the length of the proofs in many specific cases, and to obtain consistency results in some semiparametric problems where direct computations are ineffective. Additionally, the method of this paper gives a lot of freedom in the choice of penalties, dimension growth rates, and flexibility in model regularity assumptions.

The method is applicable to dependent data and statistical inverse problems. Additionally, we conjecture (and provide some support for this claim) that both semi- and nonparametric generalized likelihood ratio statistics from [9] and [11], score processes from [10], and empirical likelihood from [19], could be used to build consistent data-driven NT- and GNT-tests.

Moreover, for any NT- or GNT-test, we have an explicit rule to determine, for every particular alternative, whether the test will be consistent against this alternative. This rule allows us to describe, in a closed form, the set of "bad" alternatives for every NT- and GNT-test.

In Section 2, we describe the framework and introduce a class of SNT-statistics. In Section 3, we consider model selection rules that will be used in our tests. Section 4 is devoted to the definition of NT-statistics. This is the main concept of this paper. In Section 5, we study behaviour of NT-statistics for the case when the alternative hypothesis is true, while in Section 6 we investigate what happens under the null hypothesis. In the end of Section 6, a consistency theorem for NT-statistics is given. Section 7 is devoted to some direct applications of our method. In Section 8, a new notion concerning the use of quadratic forms in statistics is introduced. This section is somewhat auxiliary for this paper. In Section 9, we introduce a notion of GNT-statistics. This notion generalizes the notion of score tests for composite hypotheses. We prove consistency theorems for GNT-statistics.

2. Notation and basic assumptions

Let $X_1, X_2, \ldots$ be a sequence of random variables with values in an arbitrary measurable space $\mathbb{X}$. Suppose that for every m the random variables $X_1, \ldots, X_m$ have the joint distribution P_m from the family of distributions $\mathbb{P}_m$. Suppose there is a given functional $\mathcal{F}$ acting from the direct product of the families $\otimes_{m=1}^{\infty} \mathbb{P}_m = (\mathbb{P}_1, \mathbb{P}_2, \ldots)$ to a known set Θ, and that $\mathcal{F}(P_1, P_2, \ldots) = \theta$. We consider the problem of testing the hypothesis

$$H_0 : \quad \theta \in \Theta_0 \subset \Theta$$

against the alternative

$$H_A : \quad \theta \in \Theta_1 = \Theta \setminus \Theta_0$$

on the basis of observations $Y_1, \ldots, Y_n$ having their values in an arbitrary measurable space $\mathbb{Y}$ (i.e. not necessarily on the basis of $X_1, \ldots, X_m$).

Here Θ can be *any* set, for example, a functional space; correspondingly, parameter θ can be infinite dimensional. The situation is, therefore, nonparametric. It is not assumed that $Y_1, \ldots, Y_n$ are independent or identically distributed. The measurable space $\mathbb{Y}$ can be, for example, infinite dimensional. This allows to apply the results of this paper in statistics for stochastic processes. Additional assumptions on Y_i's will be imposed below, when it would be necessary.

The exact form of the null hypothesis H_0 is not important for us at this moment: H_0 can be composite or simple, H_0 can be about Y's densities or expectations, or it can be of any other form. The important feature of our approach is that we are able to consider the case when H_0 is not about observable Y_i's, but about some other random variables $X_1, \ldots, X_m$. This makes it possible to use our method in the case of statistical inverse problems. Under some conditions (see Theorem 9) it would be still possible to extract from Y_i's some information about X_i's and build a consistent test.

Definition 1. Consider the following statistic of the form

$$T_k = \sum_{j=1}^{k} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_j(Y_i) \right\}^2 . \tag{1}$$

where n is the number of available observations $Y_1, \ldots, Y_n$ and $l_1, \ldots, l_k$, $l_i : \mathbb{Y} \to \mathbb{R}$, are some known Lebesgue measurable functions. We call T_k the *simplified statistic of Neyman's type* (or SNT-statistic).

Here $l_1, \ldots, l_k$ can be some score functions, as was the case for the classical Neyman's test, but it is possible to use any other functions, depending on the problem under consideration. We prove below that under additional assumptions it is possible to construct consistent tests of such form without using scores in (1). We will discuss different possible sets of meaningful additional assumptions on $l_1, \ldots, l_k$ below (see Sections 5 - 9).

Scores (and efficient scores) are based on the notion of maximum likelihood. In our constructions it possible to use, for example, truncated, penalized or partial likelihood to build a test. In this sense, our theory generalizes the score tests theory, like M-estimation generalizes classical likelihood estimation. It is even possible to use functions $l_1, \ldots, l_k$ such that they are unrelated to any kind of a likelihood. Under conditions imposed in Sections 5 and 6, these tests will still be consistent.

Example 1. Basic example of an SNT-statistic is the Neyman's smooth test statistic for simple hypotheses (see [2] or [8]). Let $X_1, \ldots, X_n$ be i.i.d. random variables. Consider the problem of testing the simple null hypothesis H_0 that the X_i's have the uniform distribution on $[0,1]$. Let $\{\phi_j\}$ denote the family of orthonormal Legendre polynomials on $[0,1]$. Then for every k one has the test statistic

$$T_k = \sum_{j=1}^{k} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_j(X_i) \right\}^2 .$$

We see that Neyman's classical test statistic is an SNT-statistics.

Example 2. Partial likelihood. [20] proposed the notion of partial likelihood, generalizing the ideas of conditional and marginal likelihood. Applications of partial likelihood are numerous, including inference in stochastic processes. Below we give Cox's definition of partial likelihood and then construct SNT-statistics based on this notion.

Consider a random variable Y having the density $f_Y(y;\theta)$. Let Y be transformed into the sequence

$$(X_1, S_1, X_2, S_2, \ldots, X_m, S_m), \tag{2}$$

where the components may themselves be vectors. The full likelihood of the sequence (2) is

$$\prod_{j=1}^{m} f_{X_j|X^{(j-1)},S^{(j-1)}}(x_j|x^{(j-1)}, s^{(j-1)};\theta) \prod_{j=1}^{m} f_{S_j|X^{(j)},S^{(j-1)}}(s_j|x^{(j)}, s^{(j-1)};\theta), \tag{3}$$

where $x^{(j)} = (x_1, \ldots, x_j)$ and $s^{(j)} = (s_1, \ldots, s_j)$. The second product is called the partial likelihood based on S in the sequence $\{X_j, S_j\}$. The partial likelihood is useful especially when it is substantially simpler than the full likelihood, for example when it involves only the parameters of interest and not nuisance parameters. In [20] some specific examples are given.

Assume now, for simplicity of notation, that θ is just a real parameter and that we want to test the simple hypothesis $H_0 : \theta = \theta_0$ against some class of alternatives. Define for $j = 1, \ldots, m$ functions

$$t_j = \frac{\partial \log f_{S_j | X^{(j)}, S^{(j-1)}}(s_j | x^{(j)}, s^{(j-1)}; \theta)}{\partial \theta} \bigg|_{\theta = \theta_0}, \qquad (4)$$

and $\sigma_j^2 := var(t_j)$. If we define $l_j := t_j / \sigma_j$, we can form the SNT-test statistic

$$PL_m = \sum_{j=1}^{m} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_j \right\}^2 . \qquad (5)$$

□

Consistency theorems for SNT-statistics will follow from consistency theorems for more general NT-statistics that are introduced in Section 4. See, for example, Theorem 10.

Remark 1. There is a direct method that makes it possible to find asymptotic distributions of SNT-statistics, both under the null hypothesis and under alternatives. The idea of the method is as follows: one approximates the quadratic form T_k (that has the form $Z_1^2 + \ldots + Z_k^2$) by the quadratic form $N_1^2 + \ldots + N_k^2$, where N_i is the Gaussian random variable with the same mean and covariance structure as Z_i, i.e. the $i-$th component of T_k. This approximation is possible, for example, if $l(Y_j)'$s are i.i.d. random vectors with nondegenerate covariance operators and finite third absolute moments. Then the error of approximation is of order $n^{-1/2}$ and depends on the smallest eigenvalue of the covariance of $l(Y_1)$. See [21], p. 1078 for more details. And the asymptotic distribution and large deviations of the quadratic form $N_1^2 + \ldots + N_k^2$ has been studied extensively.

3. Selection rule

Since 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 [7], [22], [15], [23], [24]), it is desirable to provide a corresponding refinement for our construction. Using the idea of a penalized likelihood, we propose a general mathematical framework for constructing a rule to find reasonable model dimensions. We make our tests data-driven, i.e., the tests are capable to choose a reasonable number of components in the test statistics automatically by the data. Our construction offers a lot of freedom in the choice of penalties and building blocks for the statistics. A statistician could take into account specific features of his particular problem and choose among all the theoretical possibilities the most suitable penalty and the most suitable structure of the test statistic to build a test with desired properties.

We will not restrict a possible number of components in test statistics by some fixed number, but instead we allow the number of components to grow unlimitedly as the number of observations grows. This is important because the more observations $Y_1, \ldots, Y_n$ we have, the more information is available about the problem. This makes it possible to give a more detailed description of the phenomena under investigation. In our case this means that the complexity of the model and the possible number of components in the corresponding test statistic grow with n at a controlled rate.

Denote by M_k a statistical model designed for a specific statistical problem satisfying assumptions of Section 2. Assume that the true parameter value θ belongs to the parameter set of M_k, call it Θ_k. We say that the family of models M_k for $k = 1, 2, \dots$ is nested if for their parameter sets it holds that $\Theta_1 \subseteq \Theta_2 \subseteq \dots$. We do not require Θ_k's to be finite dimensional. We also do not require that all Θ_k's are different (this has a meaning in statistics: see the first remark on the page 221 of [25]).
Let T_k be an *arbitrary* statistic for testing validity of the model M_k on the basis of observations $Y_1, \dots, Y_n$. The following definition applies for the sequence of statistics $\{T_k\}$.

Definition 2. Consider a nested family of models M_k for $k = 1, \dots, d(n)$, where $d(n)$ is a control sequence, giving the largest possible model dimension for the case of n observations. 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 $\pi(1, n) < \pi(2, n) < \dots < \pi(d(n), n)$ for all n and $\pi(j, n) - \pi(1, n) \to \infty$ as $n \to \infty$ for every $j = 2, \dots, d(n)$. Call $\pi(j, n)$ a *penalty attributed to the j-th model M_j and the sample size n.* Then a *selection rule S* for the sequence of statistics $\{T_k\}$ is an integer-valued random variable satisfying the condition

$$S = \min\big\{k : 1 \le k \le d(n);\, T_k - \pi(k, n) \ge T_j - \pi(j, n),\, j = 1, \dots, d(n)\big\}\,. \quad (6)$$

We call T_S a *data-driven test statistic* for testing validity of the initial model. The definition is meaningful, of course, only if the sequence $\{T_k\}$ is increasing in the sense that $T_1(Y_1, \dots, Y_n) \le T_2(Y_1, \dots, Y_n) \le \dots$.

In statistical literature, one usually tries to choose penalties such that they possess some sort of minimax or Bayesian optimality. Classical examples of the penalties constructed via this approach are Schwarz's penalty $\pi(j, n) = j \log n$ (see [26]), and minimum description length penalties, see [27]. For more examples of optimal penalties and recent developments, see [28], [25] or [29]. In this paper, we do not aim for optimality of the penalization; our goal is to be able to build consistent data-driven tests based on different choices of penalties. The penalization technic that we use in this paper allows for many possible choices of penalties. It seems that in our framework it is possible to use most of the penalties from the abovementioned papers. As an illustration, see Example 3 below.

Example 2 (continued). We have an interesting possibility concerning the statistic PL_m. This statistic depends on the number m of components in the sequence (2). Suppose now that Y can be transformed into sequences (X_1, S_1), or (X_1, S_1, X_2, S_2), or even $(X_1, S_1, X_2, S_2, \dots, X_m, S_m)$ for any natural m. If we are free to choose the partition number m, then which m is the best choice? If m is too small, one can loose a lot of information about the problem; and if m is too big, then the resulting partial likelihood can be as complicated as the full one. Definition 2 proposes a solution to this problem. The adaptive statistic PL_S will choose a reasonable number of components in the transformed sequence automatically by the data. □

Example 3 (Gaussian model selection). Birgé and Massart in [25] proposed a method of model selection in a framework of Gaussian linear processes. This framework is quite general and includes as special cases a Gaussian regression with fixed design, Gaussian sequences and the model of Ibragimov and Has'minskii. In this example we briefly describe the construction (for more details see the original paper) and then discuss the relations with our results.

Given a linear subspace $\mathbb{S}$ of some Hilbert space $\mathbb{H}$, we call Gaussian linear process on $\mathbb{S}$, with mean $s \in \mathbb{H}$ and variance ε^2, any process Y indexed by $\mathbb{S}$ of the form

$$Y(t) = \langle s, t\rangle + \varepsilon Z(t),$$

for all $t \in \mathbb{S}$, and where Z denotes a linear isonormal process indexed by $\mathbb{S}$ (i.e. Z is a centered and linear Gaussian process with covariance structure $E[Z(t)Z(u)] = \langle t, u\rangle$). Birgé and Massart considered estimation of s in this model.

Let S be a finite dimensional subspace of $\mathbb{S}$ and set $\gamma(t) = \|t\|^2 - 2Y(t)$. One defines the projection estimator on S to be the minimizer of $\gamma(t)$ with respect to $t \in S$. Given a finite or countable family $\{S_m\}_{m\in\mathcal{M}}$ of finite dimensional linear subspaces of S, the corresponding family of projection estimators $\widehat{s}_m$, built for the same realization of the process Y, and given a nonnegative function pen defined on $\mathcal{M}$, Birgé and Massart estimated s by a penalized projection estimator $\widetilde{s} = \widehat{s}_{\widehat{m}}$, where $\widehat{m}$ is any minimizer with respect to $m \in \mathcal{M}$ of the penalized criterion

$$crit(m) = -\|\widehat{s}_m\|^2 + pen(m) = \gamma(\widehat{s}_m) + pen(m).$$

They proposed some specific penalties pen such that the penalized projection estimator has the optimal order risk with respect to a wide class of loss functions. The method of model selection of this paper has a relation with the one of [25].

In the model of Birgé and Massart $\gamma(t)$ is the least squares criterion and $\widehat{s}_m$ is the least squares estimator of s, which is in this case the maximum likelihood estimator. Therefore $\|\widehat{s}_m\|^2$ is the Neyman score for testing the hypothesis $s = 0$ within this model. Risk-optimizing penalties pen proposed in [25] satisfy the conditions of Definition 2 (after the change of notations $pen(m) = \pi(m, n)$; for the explicit expressions of pen's see the original paper). Therefore, $\|\widehat{s}_{\widehat{m}}\|^2$ is, in our terminology, the data-driven SNT-statistic. As follows from the consistency Theorem 9 below, $\|\widehat{s}_{\widehat{m}}\|^2$ can be used for testing $s = 0$ and has a good range of consistency.

Of course, there are differences in the two approaches as well. We consider the case when possible models form a growing and countable sequence, while Birgé and Massart allow also sets of models which cannot be ordered with respect to inclusion, and finite sequences of models. On the other hand, Birgé and Massart work in a Hilbert space and all submodels S_m are finite dimensional, whereas we are able to work with a much more general case of topological sets Θ_k's. These differences in two setups are due to the fact that we consider consistent *testing* of general hypotheses and Birgé and Massart do *estimation* for less general models. □

4. NT-statistics

Now we introduce the main concept of this paper. Suppose that we are under the general setup of Section 2.

Definition 3. Suppose we have n random observations $Y_1, \ldots, Y_n$ with values in a measurable space $\mathbb{Y}$. Let k be a fixed number and $l = (l_1, \ldots, l_k)$ be a vector-function, where $l_i : \mathbb{Y} \to \mathbb{R}$ for $i = 1, \ldots, k$ are some known Lebesgue measurable functions. We assume that $Y_i's$ and $l_i's$ are as general as in Definition 1. Set

$$L = \left\{E_0[l(Y)]^T l(Y)\right\}^{-1}, \tag{7}$$

where the mathematical expectation E_0 is taken with respect to P_0, and P_0 is the (fixed and known in advance) distribution function of some auxilliary random variable Y, where Y is assuming its values in the space $\mathbb{Y}$. Assume that $E_0\, l(Y) = 0$ and L is well defined in the sense that all its elements are finite. Put

$$T_k = \left\{\frac{1}{\sqrt{n}} \sum_{j=1}^{n} l(Y_j)\right\} L \left\{\frac{1}{\sqrt{n}} \sum_{j=1}^{n} l(Y_j)\right\}^T . \tag{8}$$

We call T_k a *statistic of Neyman's type* (or NT-statistic).

If, for example, Y_i's are equally distributed, then the natural choice for P_0 is their distribution function under the null hypothesis. Thus, L will be the inverse to the covariance matrix of the vector $l(Y)$. Such a constraction is often used in score tests for simple hypothesis. But our definitions allow to use a reasonable substitution instead of the covariance matrix. This possibility can help for testing in a semi- or nonparametric case, where instead of finding a complicated covariance in a nonparametric situation one could use P_0 from a much simpler parametric family, thus getting a reasonably working test and avoiding a considerable amount of technicalities. Of course, this P_0 will have to satisfy consistency conditions, but after that we get the consistent test regardless of the unusual choice of P_0. Consistency conditions put a serious restriction on possible P_0; they are a mathematical formalization of the idea of how P_0 should be connected to Y_i's.

If we choose P_0 absolutely irrespectively of Y_i's, we will simply not be able to satisfy consistency conditions of Theorem 9, and our test will not be consistent against meaningful alternatives.

Note that, even if consistency conditions are satisfied, we cannot expect to get an efficient test by using something different than the covariance matrix. This is a drawback for using unusual L's.

Example 2 (continued). It is possible to define by the formula (8) a version of the partial likelihood statistic PL_m for the case when θ is multidimensional or even infinite dimensional. In [20] it is shown that under additional regularity assumptions $E(t_j) = 0$. In this case PL_m will be an NT-statistic (but not an

SNT-statistic).

Example 4 (trivial). If for the SNT-statistic T_k defined by (1) additionally $E_0\, l(Y) = 0$, then T_k is obviously an NT-statistic. Therefore, in most situations of interest the notion of NT-statistics is more general than the one of SNT-statistics. The first reason for introducing SNT-statistics as a special class is that for this special case there is a well-developed theory for finding asymptotic distributions of corresponding quadratic forms, and therefore there could be some asymptotic results and rates for SNT-statistics such that they are stronger than the corresponding results for NT-statistics (see Remark 1). The second reason is that there exist SNT-statistics of interest such that they are not NT-statistics. Though, they will not be studied in this paper.

Example 5. Statistical inverse problems. The most well-known example here is the deconvolution problem. This problem appears when one has noisy signals or measurements: in physics, seismology, optics and imaging, engineering. It is a building block for many complicated statistical inverse problems. Due to importance of the deconvolution, testing statistical hypotheses related to this problem has been widely studied in a literature. But, to our knowledge, all the proposed tests were based on some kind of distance (usually an L_2-type distance) between the theoretical density function and the empirical estimate of the density. See [30], [31], [32]. But, as was shown in [33], it is also possible to construct score tests for the problem.

In [33] we went even further and constructed *data-driven* score tests for the problem. However, our method allowed only to choose among dimensions not exceeding some fixed finite number, regardless of the number of available observations. In this paper we permit the model dimension to grow unlimitedly.

The problem is formulated as follows. Suppose that instead of X_i one observes Y_i, where

$$Y_i = X_i + \varepsilon_i,$$

and ε_i's are i.i.d. with a known density h with respect to the Lebesgue measure λ; also X_i and ε_i are independent for each i and $E\,\varepsilon_i = 0$, $0 < E\,\varepsilon^2 < \infty$. Assume that X has a density with respect to λ. Our null hypothesis H_0 is the simple hypothesis that X has a known density f_0 with respect to λ. Let us choose for every $k \leq d(n)$ 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. The true F possibly has no relation to the chosen $\{f_\theta\}$. Set

$$l(y) = \frac{\frac{\partial}{\partial\theta}\left(\int_{\mathbb{R}} f_\theta(s)\, h(\,y-s)\, ds\right)\Big|_{\theta=0}}{\int_{\mathbb{R}} f_0(s)\, h(\,y-s)\, ds} \tag{9}$$

and define the corresponding test statistic U_k by the formula (8). Under appropriate regularity assumptions, U_k is an NT-statistic (see [34]).

Example 6. Rank Tests for Independence. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. random variables with the distribution function D and the marginal distribution functions F and G for X_1 and Y_1. Assume that F and G are continuous, but unknown. It is the aim to test the null hypothesis of independence

$$H_0 : \quad D(x,y) = F(x)G(y), \quad x, y \in \mathbb{R}, \tag{10}$$

against a wide class of alternatives. The following construction was proposed in [35].
Let b_j denote the $j-$th orthonormal Legendre polynomial (i.e., $b_1(x) = \sqrt{3}(2x-1)$, $b_2(x) = \sqrt{5}(6x^2 - 6x + 1)$, etc.). The score test statistic from [35] is

$$T_k = \sum_{j=1}^{k} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_j\left(\frac{R_i - 1/2}{n}\right) b_j\left(\frac{S_i - 1/2}{n}\right) \right\}^2 , \tag{11}$$

where R_i stands for the rank of X_i among $X_1, \ldots, X_n$ and S_i for the rank of Y_i among $Y_1, \ldots, Y_n$. Thus defined T_k satisfies Definition 3 of NT-statistics: put

$$Z_i = (Z_i^{(1)}, Z_i^{(2)}) := \left(\frac{R_i - 1/2}{n}, \frac{S_i - 1/2}{n}\right)$$

and $l_j(Z_i) := b_j(Z_i^{(1)})\, b_j(Z_i^{(2)})$. Under the null hypothesis $L_k = E_{k\times k}$, and $E_0 l(Z) = 0$. Thus, T_k is an NT-statistic. New Z_i depends on the original $(X_i, Y_i)'$s in a nontrivial way, but still contains some information about the pair of interest.

This example shows why we needed so much generality in the definition of NT-statistics.
The selection rule proposed in [35] to choose the number of components k in T_k was

$$S = \min\bigl\{k : 1 \le k \le d(n);\, T_k - k\log n \ge T_j - j\log n,\, j = 1, 2, \ldots, d(n)\bigr\}. \tag{12}$$

This selection rule satisfies Definition 2, and so the data-driven statistic T_S from [35] is a data-driven NT-statistic. □

5. Alternatives

Now we shall investigate consistency of tests based on data-driven NT-statistics. In this section we study the behavior of NT-statistics under alternatives.

We impose additional assumptions on the abstract model of Section 2. First, we assume that $Y_1, Y_2, \ldots$ are identically distributed. We do *not* assume that $Y_1, Y_2, \ldots$ are independent. It is possible that the sequence of interest $X_1, X_2, \ldots$ consists of dependent and nonidentically distributed random variables. It is only important that the new (possibly obtained by a complicated transformation) sequence $Y_1, Y_2, \ldots$ obeys the consistency conditions. Then it is possible to build consistent tests of hypotheses about X_i's. The reason for this is that, even after a complicated transformation, the transformed sequence still can contain some part of the information about the sequence of interest. However, if the transformed sequence $Y_1, Y_2, \ldots$ is not chosen reasonably, then test can be meaningless: it can be formally consistent, but against an empty or almost empty set of alternatives.

Let P denote an alternative distribution of Y_i's. Suppose that $E_P\, l(Y)$ exists. Another assumption we impose is that $l(Y_i)'$s satisfy both the law of large numbers and the multivariate central limit theorem, i.e. that for the vectors $l(Y_1), \ldots, l(Y_n)$ it holds that

$$\frac{1}{n}\sum_{j=1}^{n} l(Y_j) \to E_P\, l(Y) \quad \text{in } P-\text{probability as } n \to \infty,$$

$$n^{-1/2}\sum_{j=1}^{n}(l(Y_j) - E_P\, l(Y)) \to_d \mathcal{N}(0, L^{-1})\,, \tag{13}$$

where L is defined by (7) and $\mathcal{N}(0, L^{-1})$ denotes the $k-$dimensional normal distribution with mean 0 and covariance matrix L^{-1}.
These assumptions put a serious restriction on the choice of the function l and leave us with a uniquely determined P_0. In this paper we are not using the full generality of Definition 3. Nonetheless, random variables of interest $X_1, \ldots, X_n$ are still allowed to be arbitrarily dependent and nonidentically distributed, and their transformed counterparts $Y_1, \ldots, Y_n$ are still allowed to be dependent.
Now we formulate the following *consistency condition*:

$\langle \mathbf{C} \rangle$ there exists integer $K = K(P) \geq 1$ such that
$E_P\, l_1(Y) = 0, \ldots, E_P\, l_{K-1}(Y) = 0,\; E_P\, l_K = C_P \neq 0\,,$

where $l_1, \ldots, l_k$ are as in Definition 3.
We assume additionally (without loss of generality) that

$$\lim_{n\to\infty} d(n) \,=\, \infty\,. \tag{14}$$

Remark 2. Assumption (14) describes the most interesting case. It is not very important from statistical point of view to include the possibility that $d(n)$ is non-monotone. And the case when $d(n)$ is nondecreasing and bounded from above by some constant D can be handled analogously to the method of this paper, only the proofs will be shorter.

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ be the ordered eigenvalues of L, where L is as in Definition 3. To avoid possible confusion in the statement of the next theorem, we have to modify our notations a little bit. We remind that in Definition 3 L is a $k \times k-$matrix. Below we will sometimes need to denote it L_k in order to stress the model dimension. Accordingly, ordered eigenvalues of L_k will be denoted $\lambda_1^{(k)} \geq \lambda_2^{(k)} \geq \ldots \geq \lambda_k^{(k)}$. We have the sequence of matrices $\{L_k\}_{k=1}^{\infty}$ and each matrix has its own eigenvalues. When it will be possible, we will use the simplified notation from Definition 3.

Theorem 3. *Let $\langle C \rangle$ and (14) holds and*

$$\lim_{n\to\infty}\sup_{k\leq d(n)} \frac{\pi(k,n)}{n\lambda_k^{(k)}} = 0\,. \tag{15}$$

Then

$$\lim_{n\to\infty} P(S \geq K) = 1\,.$$

Remark 4. Condition (15) means that not only n tends to infinity, but that it is also possible for k to grow infinitely, but at the controlled rate.

Now suppose that the alternative distribution P is such that $\langle C\rangle$ is satisfied and that there exists a sequence $\{r_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} r_n = \infty$ and

$$\langle \mathbf{A}\rangle \qquad P\left(\frac{1}{n}\left|\sum_{i=1}^{n}\bigl[l_K(Y_i) - E_P\, l_K(Y_i)\bigr]\right| \geq y\right) = O\biggl(\frac{1}{r_n}\biggr)\,.$$

Note that in $\langle A\rangle$ we do not require uniformity in y, i.e. r_n gives us the rate, but the exact bound can depend on y. In some sense condition $\langle A\rangle$ is a way to make the weak law of large numbers for $l_K(Y_i)'$s more precise. As an illustration, we prove the next lemma.

Lemma 5. *Let $l_K(Y_i)'s$ be bounded i.i.d. random variables with finite expectation and variance σ^2. Then condition $\langle A\rangle$ is satisfied with $r_n = \exp(ny^2/2\sigma)$.*

Therefore, one can often expect exponential rates in $\langle A\rangle$, but even a much slower rate is not a problem. The main theorem of this section is

Theorem 6. *Let $\langle A\rangle$, $\langle C\rangle$, (14) and (15) holds and*

$$d(n) = o(r_n) \quad \textit{as} \quad n\to\infty\,. \tag{16}$$

Then $T_S \ \to_P\ \infty$ as $n\to\infty$.

6. The null hypothesis

Now we study the asymptotic behavior of data-driven NT-statistics under the null hypothesis. We need one more definition first.

Definition 4. Let $\{T_k\}$ be a sequence of NT-statistics and S be a selection rule for it. Suppose that $\lambda_1 \geq \lambda_2 \geq \ldots$ are ordered eigenvalues of L, where L is defined by (7). We say that the penalty $\pi(k,n)$ in S is of *proper weight*, if the following conditions holds:

1. there exists sequences of real numbers $\{s(k,n)\}_{k,n=1}^{\infty}$, $\{t(k,n)\}_{k,n=1}^{\infty}$, such that
   (a)
   $$\lim_{n\to\infty}\sup_{k\leq u_n}\frac{s(k,n)}{n\lambda_k^{(k)}} = 0\,,$$
   where $\{u_n\}_{n=1}^{\infty}$ is some real sequence such that $\lim_{n\to\infty} u_n = \infty$.
   (b) $\lim_{n\to\infty} t(k,n) = \infty$ for every $k\geq 2$
   $\lim_{k\to\infty} t(k,n) = \infty$ for every fixed n.
2. $s(k,n) \leq \pi(k,n) - \pi(1,n) \leq t(k,n)$ for all k, n

3.

$$\lim_{n\to\infty} \sup_{k\le m_n} \frac{\pi(k,n)}{n\lambda_k^{(k)}} = 0\,,$$

where $\{m_n\}_{n=1}^{\infty}$ is some real sequence such that $\lim_{n\to\infty} m_n = \infty$.

For notational convenience, we define for $l = (l_1, \dots, l_k)$ from Definition 3

$$\bar{l}_j := \frac{1}{n} \sum_{i=1}^{n} l_j(Y_i)\,, \tag{17}$$

$$\bar{l} := (\bar{l}_1, \bar{l}_2, \dots, \bar{l}_k) \tag{18}$$

and, using the notation L from Definition 3, a quadratic form

$$Q_k(\bar{l}) = (\bar{l}_1, \bar{l}_2, \dots, \bar{l}_k)\, L\, (\bar{l}_1, \bar{l}_2, \dots, \bar{l}_k)^T\,. \tag{19}$$

The first reason for the new notation is that $T_k = Q_k(\bar{l})$, where T_k is the statistic from Definition 3. It is more convenient to formulate and prove Theorem 7 below using the quadratic form Q_k rather than T_k itself. And the main value of introducing Q_k will be seen in Section 8, where Q_k is the central object.

Below we use the notation of Definitions 3 and 4.

Definition 5. Let S be with a penalty of proper weight. Assume that there exists a Lebesgue measurable function $\varphi(\cdot,\cdot) : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$, such that φ is monotonically decreasing in the second argument and monotonically nondecreasing in the first one, and assume that

1. (B2) for every $\varepsilon > 0$ there exists $K = K_\varepsilon$ such that for every $n > n(\varepsilon)$
$$\sum_{k=K_\varepsilon}^{u_n} \varphi(k; s(k,n)) \,<\, \varepsilon\,,$$
where $\{u_n\}_{n=1}^{\infty}$ is as in Definition 4.
2. (B)
$$P_0\left(n\, Q_k(\bar{l}) \ge y\right) \,\le\, \varphi(k; y)$$
for all $k \ge 1$ and $y \in [s(k,n); t(k,n)]$, where P_0 is as in Definition 4.

We call φ a *proper majorant* for (large deviations of) the statistic T_S. Equivalently, we say that (large deviations of) the statistic T_S are *properly majorated* by φ.

To prove consistency of a test based on some test statistic, usually it is required to use some large deviations inequality for this test statistic. NT-statistics are no exception from this. In order to prove consistency of an NT-test, one has to choose some specific large deviations inequality to use in the proof. Part of the model regularity assumptions and the rate $d(n)$ will be determined by this choice. Without a general consistency theorem, if one would like to use another inequality, the proof of consistency should be started anew.

In our method it is easier to prove different types of consistency theorems. Sometimes, it is desirable to have a better rate $d(n)$ by the cost of more restrictive regularity assumptions, arising from the use of a strong probabilistic inequality; sometimes, it is better to use a simple inequality that requires less regularity assumptions, but gives worse rate $d(n)$. The meaning of Definitions 4 and 5 and Theorem 9 below is that one can be sure in advance that whatever inequality he chooses, he will succeed in proving a consistency theorem, provided that the chosen inequality satisfies conditions (B) and (B2). Moreover, once an inequality is chosen, the rate of $d(n)$ is obtained from Theorem 9.

Some of the previously published proofs of consistency of data-driven tests relied heavily on the use of Prohorov's inequality. For many test statistics this inequality can't be used to estimate the large deviations. This is usually the case for more complicated models where the matrix L is not diagonal. This is typical for statistical inverse problems and even for such a basic problem as the deconvolution. Our method helps to surpass this difficulty. It is possible to use, for example, Dvoretzky-Kiefer-Wolfowitz inequality for dependent data, and many kinds of non-standard large deviations inequalities.

Theorem 7. *Let $\{T_k\}$ be a sequence of NT-statistics and S be a selection rule for it. Assume that the penalty in S is of proper weight and that large deviations of statistics T_k are properly majorated. Suppose that*

$$d(n) \,\leq\, \min\{u_n\,, m_n\}\,. \tag{20}$$

Then $S = O_{P_0}(1)$ and $T_S = O_{P_0}(1)$.

Remark 8. In Definition 5 we need $s(k,n)$ to be sure that the penalty π is not "too light", i.e. that the penalty somehow affects the choice of the model dimension and protects us from choosing a "too complicated" model. In nontrivial cases, it follows from (B2) that $s(k,n) \to \infty$ as $k \to \infty$. But $t(k,n)$ is introduced for the reason of statistical sense. Practically, the choice of $t(k,n)$ is dictated by the form of inequality (B) established for the problem. Additionally, one can drop assumptions 1 and 3 in Definition 4 and still prove a modified version of Theorem 7. But usually it happens that if the penalty does not satisfy all the conditions of Definitions 4 and 5, then T_S has the same distribution under both alternative and null hypotheses and the test is inconsistent. Then, formally, the conclusions of Theorem 7 holds, but this has no statistical meaning.

Now we formulate the general consistency theorem for NT-statistics. We understand consistency of the test based on T_S in the sense that under the null hypothesis T_S is bounded in probability, while under fixed alternatives $T_S \to \infty$ in probability.

Theorem 9. *Let $\{T_k\}$ be a sequence of NT-statistics and S be a selection rule for it. Assume that the penalty in S is of proper weight. Assume that conditions (A), (14) and (15) are satisfied and that $d(n) = o(r_n)$, $d(n) \leq \min\{u_n, m_n\}$. Then the test based on T_S is consistent against any (fixed) alternative distribution P satisfying condition (C).*

7. Applications

As the first application, we have the following result.

Theorem 10. *Let $\{T_k\}$ be a family of SNT-statistics and S a selection rule for the family. Assume that $Y_1, \ldots, Y_n$ are i.i.d.. Let $E\, l(Y_1) = 0$ and assume that for every k the vector $(l_1(Y_i), \ldots, l_k(Y_i))$ has the unit covariance matrix. Suppose that $\|(l_1(Y_1), \ldots, l_k(Y_1))\|_k \leq M(k)$ a.e., where $\|\cdot\|_k$ is the norm of the $k-$dimensional Euclidean space. Assume $\pi(k,n) - \pi(1,n) \geq 2k$ for all $k \geq 2$ and*

$$\lim_{n\to\infty} \frac{M(d(n))\,\pi(d(n),n)}{\sqrt{n}} = 0. \tag{21}$$

Then $S = O_{P_0}(1)$ and $T_S = O_{P_0}(1)$.

Example 1 (continued). As a simple corollary, we derive the following theorem that slightly generalizes Theorem 3.2 from [24].

Theorem 11. *Let T_S be the Neyman's smooth data-driven test statistic for the case of simple hypothesis of uniformity. Assume that $\pi(k,n) - \pi(1,n) \geq 2k$ for all $k \geq 2$ and that for all $k \leq d(n)$*

$$\lim_{n\to\infty} \frac{d(n)\pi(d(n),n)}{\sqrt{n}} = 0.$$

Then $S = O_{P_0}(1)$ and $T_S = O_{P_0}(1)$.

Proof. It is enough to note that in this case $M(k) = \sqrt{(k-1)(k+3)}$ and apply Theorem 10. □

Remark 12. In my point of view, the precise rate at which $d(n)$ tends to infinity is not crucial for many practical applications. Typical rates such as $d(n) = o(\log n)$ or $d(n) = o(n^{1/3})$ are not better for applications with $n = 50$ than, say, just $d(n) \equiv 10$. It seems that in practical applications one should not put in much effort to increase $d(n)$ as much as possible for each n.

Example 5 (continued). In [35] the following consistency result was established.

Theorem 13. *Suppose that $d(n) = o\big(\{\frac{n}{\log n}\}^{1/10}\big)$. Let $\mathbb{P}$ be an alternative and let F and G be the marginal distribution functions of X and Y under $\mathbb{P}$. Let*

$$E_{\mathbb{P}}\, b_j(F(X))b_j(G(Y)) \neq 0 \tag{22}$$

for some j. If $d(n) \to \infty$, then $T_S \to \infty$ as $n \to \infty$ when $\mathbb{P}$ applies (i.e. T_S is consistent against $\mathbb{P}$).

For this problem, our condition $\langle C \rangle$ is equivalent to the following one: there exists $K = K_{\mathbb{P}}$ such that $E_{\mathbb{P}}\, l_K \neq 0$, i.e.

$$E_{\mathbb{P}}\, b_K\left(\frac{R_1 - 1/2}{n}\right) b_K\left(\frac{S_1 - 1/2}{n}\right) \neq 0. \tag{23}$$

For continuous F and G (23) is asymptotically equivalent to (22) since both $F(X)$ and $G(Y)$ are distributed as $U[0,1]$ and

$$\frac{R_i - 1/2}{n} \to U[0,1], \quad \frac{S_i - 1/2}{n} \to U[0,1].$$

We see that Theorem 6 is applicable to get a result similar to Theorem 13. We do not go into technical details here. □

8. Quadratic forms of P-type

Now we introduce another notion, concerning quadratic forms.

Definition 6. Let $Z_1, Z_2, \dots, Z_n$ be identically distributed (not necessarily independent) random vectors with k components each. Denote their common distribution function by F. Let Q be a $k \times k$ symmetric matrix. Then $Q(x) := x\,Q\,x^T$ defines a quadratic form, for $x \in \mathbb{R}^k$. We say that $Q(x)$ is a *quadratic form of Prohorov's type* (or just $P-$type) *for the distribution* F, if for some $\{s(k,n)\}_{k,n=1}^{\infty}$, $\{t(k,n)\}_{k,n=1}^{\infty}$ satisfying (B1) it holds that for all k, and for all $y \in [s(k,n); t(k,n)]$

$$P_F\left(n\,Q\bigg(\frac{Z_1 + Z_2 + \ldots + Z_n}{n} - E_F Z_1\bigg) \geq y\right) \;\leq\; \varphi(k;y)\,, \tag{24}$$

with φ being a proper majorant for P_F and of the form

$$\varphi(k;y) \;=\; C_1\,\varphi_1(k)\,\varphi_2(\lambda_1, \lambda_2, \dots, \lambda_k)\,y^{k-1}\,\exp\bigg\{-\frac{y^2}{C_2}\bigg\}\,, \tag{25}$$

where $\lambda_1, \lambda_2, \dots, \lambda_k$ are the eigenvalues of matrix Q, and C_1, C_2 are uniform in the sense that they do not depend on y, k, n. We will shortly say that $Q(x)$ is of $P-$type for Z_i's.

We have the following direct consequence of Theorem 9 .

Corollary 14. *Suppose that for T_S condition $\langle A \rangle$ holds, L is of P-type for the distribution function of the vector $\{(l_1(Y_1), l_2(Y_1), \dots, l_k(Y_1))\}_{i=1}^n$ and that the penalty in S is of proper weight. Then the test based on T_S is consistent against any alternative P satisfying (C).*

If $Z_1, Z_2, \dots, Z_n$ are i.i.d. and Q is a diagonal positive definite matrix, then $Q(x)$ is of P-type because of the Prohorov inequality. Definition 6 is meant to incorporate all the cases when Prohorov's inequality or some of its variations holds. Thus, Definition 6 is just some specification of the general condition (B) from Theorem 7. The definition is useful in the sense that it shows which kind of majorating functions φ could (and typically would) occur in condition (B) when.

The simple sufficient condition for L to be of P-type is not known. But there is a method that makes it possible to establish P-type property in many particular situations. This method consists of two steps. On the first step, one approximates the quadratic form $Q(\bar{l}(Y))$ by the simpler quadratic form $Q(N)$, where N is the Gaussian random variable with the same mean and covariance structure as $l(Y)$. This approximation is possible, for example, under conditions given in [36] or

[21]. These authors gave the rate of convergence for such approximation. Then the second step is to establish a large deviation result for the quadratic form $Q(N)$; this form has a more predictable distribution. For strongly dependent random variables, one can hope to use some technics from [37].

On the side note, many of the conditions for the existence of such approximation of $Q(\bar{l}(Y))$ are rather technical and specific on the structure of L. For example, sometimes assumptions on the 5 largest eigenvalues of L can be required. See the above papers by Gotze, Bentkus, Tikhomirov and references therein.

9. GNT-statistics

The notion of NT-statistics is helpful if the null hypothesis is simple. However, for composite hypotheses it is not always possible to find a suitable L from Definition 3. Therefore the concept of NT-statistics needs to be modified to be applicable for composite hypotheses. As a possible solution, we introduce a notion of GNT-statistics.

Definition 7. Suppose we have n random observations $Y_1, \ldots, Y_n$ assuming values in a measurable space $\mathbb{Y}$. For simplicity of presentation, assume they are identically distributed. Let k be a fixed number and $l = (l_1, \ldots, l_k)$ be a vector-function, where $l_i : \mathbb{Y} \to \mathbb{R}$ for $i = 1, \ldots, k$ are some (maybe unknown) Lebesgue measurable functions. Set

$$L^{(0)} = \left\{E_0[l(Y)]^T l(Y)\right\}^{-1} . \tag{26}$$

where the expectation E_0 is taken w.r.t. P_0, and P_0 is (possibly unknown) distribution function of $Y's$ under the null hypothesis. Assume that $E_0 l(Y) = 0$ and that $L^{(0)}$ is well-defined in the sense that all of its elements are finite. Let L_k denote, for every k, a $k \times k$ symmetric positive definite (known) matrix with finite elements such that for the sequence $\{L_k\}$ it holds that

$$\left\|L_k - L^{(0)}\right\| = o_{P_0}(1) . \tag{27}$$

Let $l_1^*, \ldots, l_n^*$ be *sufficiently good* estimators of $l(Y_1), \ldots, l(Y_n)$ with respect to P_0 in the sense that for every $\varepsilon > 0$

$$P_0^n\left(\frac{1}{\sqrt{n}}\left\|\sum_{i=1}^{n}(l_j^* - l(Y_j))\right\| \geq \varepsilon\right) \to 0 \quad \text{as} \quad n \to \infty , \tag{28}$$

where $\|\cdot\|$ denotes the Euclidian $k-$norm of a given vector. Set

$$GT_k = \left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} l_j^*\right\} L_k \left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} l_j^*\right\}^T . \tag{29}$$

We call GT_k a *generalized statistic of Neyman's type* (or a GNT-statistic). Let the selection rule S satisfy Definition 3. We call GT_S a data-driven GNT-statistic.

Remark 15. Now it is not obligatory to know functions $l_1, \dots, l_k$ explicitly (in Definition 3 we assumed that we know those functions). It is only important that we should be able to choose reasonably good L and l_j^*'s. Definition 7 generalizes the idea of efficient score test statistics with estimated scores.

Remark 16. Establishing (28) in parametric problems is usually not difficult and can be done if a $\sqrt{n}-$consistent estimate of the nuisance parameter is available (see [34]). In semiparametric models, finding estimators for the score function that satisfy (28) is more difficult and not always possible, but there exist effective methods for constructing such estimates. Often the sample splitting technic is helpful. See [38], [39], [40] for general results related to the topic. See also Example 10 below.

Example 7 (trivial). If $Y_1, \dots, Y_n$ are equally distributed and T_k is an NT-statistic, then T_k is also a GNT-statistic. Indeed, put in Definition 7 $L := L^{(0)}$ and $l_j^*(Y_1, \dots, Y_n) := l_j(Y_1)$.

Example 8. Let $X_1, \dots X_n$ be i.i.d. random variables with density $f(x)$. Consider testing the composite hypothesis

$$H_0 : \quad f(x) \in \{f(x;\beta), \beta \in \mathcal{B}\},$$

where $\mathcal{B} \subset \mathbb{R}^q$ and $\{f(x;\beta), \beta \in \mathcal{B}\}$ is a given family of densities. In [41], the data-driven score test for testing H_0 was constructed using score test for composite hypotheses from [6]. Here we briefly describe the construction from [41]. Let F be the distribution function corresponding to f and set

$$Y_n(\beta) = n^{-1} \sum_{i=1}^{n} (\phi_1(F(X_i;\beta)), \dots, \phi_j(F(X_i;\beta)))^T$$

with j depending on the context. Let I be the $k \times k$ identity matrix. Define

$$I_\beta = \Big\{ - E_\beta \frac{\partial}{\partial \beta_t} \phi_j(F(X_i;\beta)) \Big\}_{t=1,\dots,q;\, j=1,\dots,k},$$

$$I_{\beta\beta} = \Big\{ - E_\beta \frac{\partial^2}{\partial \beta_t \partial \beta_u} \log f(X;\beta) \Big\}_{t=1,\dots,q;\, u=1,\dots,q},$$

$$R(\beta) = I_\beta^T (I_{\beta\beta} - I_\beta I_\beta^T) I_\beta \,.$$

Let $\widehat{\beta}$ denotes the maximum likelihood estimator of β under H_0. Then the score statistic is given by

$$W_k(\widehat{\beta}) = n\, Y_n^T(\widehat{\beta}) \{I + R(\widehat{\beta})\}\, Y_n(\widehat{\beta}). \tag{30}$$

As follows from the results of [6], Section 9.3, pp.323-324, in a regular enough situation $W_k(\widehat{\beta})$ satisfies Definition 7 and is a GNT-statistic. Practically useful sets of such regularity assumptions are given in [41].

Example 9. Consider the problem described in Example 5, but with the following complication introduced. Suppose that the density h of ε is *unknown.* The score function for (θ, η) at (θ_0, η_0) is

$$\dot{l}_{\theta_0,\eta_0}(y) \, = \big(\, \dot{l}_{\theta_0}(y),\, \dot{l}_{\eta_0}(y) \,\big)\,, \tag{31}$$

where $\dot{l}_{\theta_0}$ is the score function for θ at θ_0 and $\dot{l}_{\eta_0}$ is the score function for η at η_0, i.e.

$$\dot{l}_{\theta_0}(y) = \frac{\frac{\partial}{\partial\theta}\left(\int_{\mathbb{R}} f_\theta(s)\, h_{\eta_0}(\, y - s)\, ds \right)\Big|_{\theta=\theta_0}}{\int_{\mathbb{R}} f_{\theta_0}(s)\, h_{\eta_0}(\, y - s)\, ds}\, 1_{[y:\, g\,(y\,;(\theta_0,\eta_0))>0]}\,, \tag{32}$$

$$\dot{l}_{\eta_0}(y) = \frac{\frac{\partial}{\partial\eta}\left(\int_{\mathbb{R}} f_{\theta_0}(s)\, h_{\eta}(\, y - s)\, ds \right)\Big|_{\eta=\eta_0}}{\int_{\mathbb{R}} f_{\theta_0}(s)\, h_{\eta_0}(\, y - s)\, ds}\, 1_{[y:\, g\,(y\,;(\theta_0,\eta_0))>0]}\,. \tag{33}$$

The *Fisher information matrix* of parameter (θ, η) is

$$I(\theta, \eta) \, = \, \int_{\mathbb{R}} \dot{l}^T_{\theta,\eta}(y)\, \dot{l}_{\theta,\eta}(y)\ dG_{\theta,\eta}(y)\,, \tag{34}$$

where $G_{\theta,\eta}(y)$ is the probability measure corresponding to the density $g\,(y\,; (\theta, \eta))$. Let us write $I(\theta_0, \eta_0)$ in the block matrix form:

$$I(\theta_0, \eta_0) \, = \, \begin{pmatrix} I_{11}(\theta_0, \eta_0) & I_{12}(\theta_0, \eta_0) \\ I_{21}(\theta_0, \eta_0) & I_{22}(\theta_0, \eta_0) \end{pmatrix}\,, \tag{35}$$

where $I_{11}(\theta_0, \eta_0) \, = \, E_{\theta_0,\eta_0} \dot{l}^T_{\theta_0}\, \dot{l}_{\theta_0}, \quad I_{12}(\theta_0, \eta_0) \, = \, E_{\theta_0,\eta_0} \dot{l}^T_{\theta_0}\, \dot{l}_{\eta_0}$, and analogously for $I_{21}(\theta_0, \eta_0)$ and $I_{22}(\theta_0, \eta_0)$. The efficient score function for θ in this model is

$$l^*_{\theta_0}(y) \, = \, \dot{l}_{\theta_0}(y) \, - \, I_{12}(\theta_0, \eta_0)\, I^{-1}_{22}(\theta_0, \eta_0)\, \dot{l}_{\eta_0}(y)\,, \tag{36}$$

and the efficient Fisher information matrix for θ is

$$I^*_{\theta_0} \, = \, E_{\theta_0,\eta_0} {l^*_{\theta_0}}^T\, l^*_{\theta_0} \, = \, \int_{\mathbb{R}} l^*_{\theta_0}(y)^T\, l^*_{\theta_0}(y)\, dG_{\theta_0,\eta_0}(y)\,. \tag{37}$$

The efficient score test statistics for composite deconvolution problem is

$$W_k = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n \widehat{l^*_{\theta_0}(Y_i)} \right\} (\widehat{I^*_{\theta_0}})^{-1} \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n \widehat{l^*_{\theta_0}(Y_i)} \right\}^T .$$

This is a GNT-statistics if the estimators satisfy (27) and (28). See [33] for more details.

Example 10. The following semiparametric example belongs to [42]. Let $Z = (X, Y)$ denote a random vector in $I \times \mathbb{R}$, $I = [0, 1]$. We would like to test the null hypothesis

$$H_0 : \quad Y = \beta[v(X)]^T + \varepsilon,$$

where X and ε are independent, $E\,\varepsilon = 0$, $E\,\varepsilon^2 < \infty$, $\beta \in \mathbb{R}^q$ a vector of unknown real valued parameters, $v(x) = (v_1(x), \dots, v_q(x))$ is a vector of known functions. Suppose X has an unknown density f, and ε an unknown density f with respect to Lebesgue measure λ.
Choose some real functions $u_1(x), u_2(x), \dots$. Set

$$l^*(z) = l^*(x, y) := -\left[\frac{f'}{f}(y - v(x)\beta^T)\right][\widetilde{u}(x) - \widetilde{v}(x)V^{-1}M] +$$

$$+\frac{1}{\tau}[y - v(x)\beta^T][m_1 - m_2 V^{-1}M],$$

where

$$m_1 = E_g u(X), \quad m_2 = E_g v(X), \quad m = (m_1, m_2),$$

$$\widetilde{w}(x) = (\widetilde{u}(x), \widetilde{v}(x)), \quad \widetilde{u}(x) = u(x) - m_1, \quad \widetilde{v}(x) = v(x) - m_2,$$

while M and V are blocks in

$$W = \begin{pmatrix} U & M^T \\ M & V \end{pmatrix} = \frac{1}{4}\{J \cdot E_g[\widetilde{w}(X)]^T[\widetilde{w}(X)] + \frac{1}{\tau}m^T m\},$$

where $J = J(f) = \int_{\mathbb{R}} \frac{[f'(y)]^2}{f(y)} d\lambda(y)$. Finally set

$$W^{11} = (U - M^T V^{-1} M)^{-1}, \quad L = \frac{1}{4}W^{11},$$

then the efficient score statistic is

$$W_k = \left\{\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \widehat{l^*}(Z_i)\right\} \widehat{L} \left\{\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \widehat{l^*}(Z_i)\right\}^T,$$

where $\widehat{l^*}(\cdot)$ is an estimator of l^*, while $\widehat{L}$ is an estimator of L. Inglot and Ledwina proposed, under additional regularity assumptions on the model, certain estimators for these quantities such that conditions (27) and (28) are satisfied. Therefore, W_k becomes a GNT-statistic and its asymptotic properties can be studied by the method of this paper. □

Remark 17. In general, it seems to be possible to use the idea of a score process and some other technics from [10] in order to construct and analyze NT- and GNT-statistics. This can be seen by the fact that such applications as in Examples 6 and 8 naturally appear in both papers. The difference with the above paper would be that we prefer to use test statistics of the form (29) rather than integrals or supremums of score processes.

In semi- and nonparametric models, generalized likelihood ratios from [9] and [11], as well as different modifications of empirical likelihood, could also be a powerful tool for constructing NT- and GNT-statistics.

A general consistency theorem for GNT-statistics is required. Without a general consistency theorem, one has to perform the whole proof of consistency anew for every particular problem. This becomes difficult in cases where sample splitting, complicated estimators and infinitedimensional parameters are involved. For this reason, sometimes in statistical literature authors do not prove consistency of their tests. Therefore, in my opinion, for most of the semi- and nonparametric problems general consistency theorems are the most convenient tool for proving consistency of data-driven NT- and GNT-tests. If one has a general consistency theorem analogous to Theorem 9 for data-driven NT-statistics, then at least some consistency result will follow automatically.

Now we prove consistency theorems for data-driven GNT-statistics. First, note that Definitions 4 and 5 are also meaningful for a sequence of GNT-statistics $\{GT_k\}$, if only instead of L we use in Definition 4 and in (19) the matrix $L^{(0)}$ from Definition 7. To be technically correct in the statement of the next theorem, we introduce the auxiliary random variable R_k that approximates the statistic of interest GT_k :

$$R_k = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l(Y_j) \right\} L^{(0)} \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l(Y_j) \right\}^T .$$

Definition 8. We would say that the penalty in the data-driven test statistic GT_S is of *proper weight*, if this penalty is of proper weight for R_S in the sense of Definition 4. We say that GT_S is *properly majorated*, if R_S is properly majorated in the sense of Definition 5.

Due to conditions (27) and (28) from the definition of GNT-statistics, this definition serves just for purposes of formal technical correctness.

Theorem 18. *Let $\{GT_k\}$ be a sequence of GNT-statistics and S be a selection rule for it. Assume that the penalty in S is of proper weight and that large deviations of GT_k are properly majorated. Suppose that $d(n) \le \min\{u_n, m_n\}$. Then under the null hypothesis it holds that $S = O_{P_0}(1)$ and $GT_S = O_{P_0}(1)$.*

To ensure consistency of GT_S against some alternative distribution P, it is necessary and sufficient to show that under P it holds that $GT_S \to \infty$ in $P-$probability as $n \to \infty$. There are different possible additional sets of assumptions on the construction that make it possible to prove consistency against different sets of alternatives. For example, suppose that

$$\langle \mathbf{C1} \rangle \qquad \left\| L - L^{(0)} \right\| = o_P(1) \tag{38}$$

and that $l_1^*, \ldots, l_n^*$ are sufficiently good estimators of $l(Y_1), \ldots, l(Y_n)$ with respect to P, i.e. that for every $\varepsilon > 0$

$$P^n \left(\frac{1}{\sqrt{n}} \left\| \sum_{i=1}^{n} (l_j^* - l(Y_j)) \right\| \ge \varepsilon \right) \to 0 \quad \text{as} \quad n \to \infty . \tag{39}$$

These assumptions are very strong: they mean that the estimators, plugged in GT_k, are not only good at one point P_0, but that the estimators also possess some globally good quality.

Theorem 19. *Let $\{GT_k\}$ be a sequence of GNT-statistics and S be a selection rule for it. Assume that the penalty in S is of proper weight. Assume that conditions $\langle A \rangle$, (14) and (15) are satisfied and that $d(n) = o(r_n)$, $d(n) \leq \min\{u_n, m_n\}$. Then the test based on T_S is consistent against any (fixed) alternative distribution P satisfying $\langle C \rangle$, $\langle C1 \rangle$ and (39).*

Remark 20. Substantial relaxation of assumptions (38) and (39) should be possible. Indeed, these assumptions ensure us not only that $GT_S \to \infty$, but also that $GT_S \to R_S$ under P, where R_S is as in Definition 8. This is much stronger than required for our purposes, since for us $GT_S \to \infty$ is enough and the order of growth is not important for proving consistency.

Remark 21. In the literature on nonparametric testing, some authors consider the number of observations n tending to infinity and alternatives (of specific form) that tend to the null hypothesis at some speed. For such alternatives, some kind of minimax rate for testing can be established. The hardness of the testing problem, and the efficiency of the test, can be measured by this rate. See, for example, [43], [44], [45]. We do not consider rates for testing in this paper, but it is possible to consider local alternatives in this general setup as well. For example, minimax optimality of the penalized likelihood estimators, in a rather general setting of l_0-type penalties, was studied in [28]. In [9], it was shown that, for certain class of statistical problems, the generalized likelihood ratio statistics achieve optimal rates of convergence given in [44]. In our case, this remains to be investigated.

Acknowledgments. Author would like to thank Fadoua Balabdaoui for helpful discussions and Axel Munk for several literature references. Large parts of this research were done at Georg-August-University of Göttingen, Germany and at the University of Bonn, Germany.

References

[1] Ya. Nikitin. *Asymptotic efficiency of nonparametric tests.* Cambridge University Press, Cambridge, 1995.

[2] J. Neyman. Smooth test for goodness of fit. *Skand. Aktuarietidskr.*, 20:150–199, 1937.

[3] S.S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.*, 9:60–62, 1938.

[4] L. Le Cam. On the asymptotic theory of estimation and testing hypotheses. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pages 129–156, Berkeley and Los Angeles, 1956. University of California Press.

[5] J. Neyman. Optimal asymptotic tests of composite statistical hypotheses. In *Probability and statistics: The Harald Cramér volume (edited by Ulf Grenander)*, pages 213–234. Almqvist & Wiksell, Stockholm, 1959.

[6] D. R. Cox and D. V. Hinkley. *Theoretical statistics.* Chapman and Hall, London, 1974.

[7] P. J. Bickel and Y. Ritov. Testing for goodness of fit: a new approach. In *Nonparametric statistics and related topics (Ottawa, ON, 1991)*, pages 51–57. North-Holland, Amsterdam, 1992.

[8] T. Ledwina. Data-driven version of Neyman's smooth test of fit. *J. Amer. Statist. Assoc.*, 89(427):1000–1005, 1994.

[9] J. Fan, C. Zhang, and J. Zhang. Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Stat.*, 29(1):153–193, 2001.

[10] P. J. Bickel, Y. Ritov, and T. M. Stoker. Tailor-made tests for goodness of fit to semiparametric hypotheses. *Ann. Statist.*, 34(2):721–741, 2006.

[11] R. Li and H. Liang. Variable selection in semiparametric regression modeling. *Annals of Statistics, to appear*, 2007.

[12] P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, 1993.

[13] I. A. Ibragimov and R. Z. Has′minskiĭ. *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York, 1981. Asymptotic theory, Translated from the Russian by Samuel Kotz.

[14] T. Inglot and T. Ledwina. Asymptotic optimality of data-driven Neyman's tests for uniformity. *Ann. Statist.*, 24(5):1982–2019, 1996.

[15] W. C. M. Kallenberg and T. Ledwina. Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests. *Ann. Statist.*, 23(5):1594–1608, 1995.

[16] W. C. M. Kallenberg and T. Ledwina. Data driven smooth tests for composite hypotheses: comparison of powers. *J. Statist. Comput. Simulation*, 59(2):101–121, 1997.

[17] Yu. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. I, II, III. *Math. Methods Statist.*, 2:85–114, 171–189, 249–268, 1993.

[18] V. G. Spokoiny. Adaptive hypothesis testing using wavelets. *Ann. Statist.*, 24(6):2477–2498, 1996.

[19] A. B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.

[20] D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[21] F. Götze and A. N. Tikhomirov. Asymptotic distribution of quadratic forms. *Ann. Probab.*, 27(2):1072–1098, 1999.

[22] R. L. Eubank, J. D. Hart, and V. N. LaRiccia. Testing goodness of fit via nonparametric function estimation techniques. *Comm. Statist. Theory Methods*, 22(12):3327–3354, 1993.

[23] J. Fan. Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.*, 91(434):674–688, 1996.

[24] W. C. M. Kallenberg. The penalty in data driven Neyman's tests. *Math. Methods Statist.*, 11(3):323–340 (2003), 2002.

[25] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[26] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[27] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Ann. Statist.*, 11(2):416–431, 1983.

[28] F. Abramovich, V. Grinshtein, and M. Pensky. On optimality of bayesian

testimation in the normal means problem. *Annals of Statistics (to appear)*, 2007.

[29] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for gaussian regression. *Annals of Statistics (to appear)*, 2007.

[30] P. J. Bickel and M. Rosenblatt. On some global measures of the deviations of density function estimates. *Ann. Statist.*, 1:1071–1095, 1973.

[31] A. Delaigle and I. Gijbels. Estimation of integrated squared density derivatives from a contaminated sample. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(4):869–886, 2002.

[32] H. Holzmann, N. Bissantz, and A. Munk. Density testing in a contaminated sample. *J. Multivariate Anal.*, 98(1):57–75, 2007.

[33] M. Langovoy. Data-driven efficient score tests for deconvolution hypotheses. *Inverse Problems*, 24(2):025028 (17pp), 2008.

[34] M. Langovoy. *Data-driven goodness-of-fit tests.* University of Göttingen, Göttingen, 2007. Ph.D. thesis.

[35] W. C. M. Kallenberg and T. Ledwina. Data-driven rank tests for independence. *J. Amer. Statist. Assoc.*, 94(445):285–301, 1999.

[36] V. Bentkus and F. Götze. Optimal rates of convergence in the CLT for quadratic forms. *Ann. Probab.*, 24(1):466–490, 1996.

[37] L. Horváth and Q.-M. Shao. Limit theorems for quadratic forms with applications to Whittle's estimate. *Ann. Appl. Probab.*, 9(1):146–187, 1999.

[38] A. Schick. On asymptotically efficient estimation in semiparametric models. *Ann. Statist.*, 14(3):1139–1151, 1986.

[39] A. Schick. A note on the construction of asymptotically linear estimators. *J. Statist. Plann. Inference*, 16(1):89–105, 1987.

[40] C. A. J. Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.*, 15(4):1548–1562, 1987.

[41] T. Inglot, W. C. M. Kallenberg, and T. Ledwina. Data driven smooth tests for composite hypotheses. *Ann. Statist.*, 25(3):1222–1250, 1997.

[42] T. Inglot and T. Ledwina. Asymptotic optimality of new adaptive test in regression model. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(5):579–590, 2006.

[43] F. Abramovich and R. Heller. Local functional hypothesis testing. *Math. Methods Statist.*, 14(3):253–266, 2005.

[44] Yu. I. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2003.

[45] V. G. Spokoiny. Adaptive and spatially adaptive testing of a nonparametric hypothesis. *Math. Methods Statist.*, 7(3):245–273, 1998.

[46] A. V. Prohorov. Sums of random vectors. *Teor. Verojatnost. i Primenen.*, 18:193–195, 1973.

Appendix.

We use the following standard lemmas from linear algebra.

Lemma 22. *Let $x \in \mathbb{R}^k$, and let A be a $k \times k$ positive definite matrix; if for some real number $\delta > 0$ we have $A > \delta$ (in the sense that matrix $(A - \delta\, I_{k\times k})$ is positive definite, where $I_{k\times k}$ is $k \times k$ identity matrix), then for all $x \in \mathbb{R}^k$ it holds that $xAx^T > \delta \|x\|^2$.*

Lemma 23. *Let A be a $k \times k$ positive definite matrix and $\{A_n\}_{n=1}^{\infty}$ be sequence of $k \times k$ matrices such that $A_n \to A$ in the Euclidian matrix norm. Suppose that for some real number $\delta > 0$ we have $A > \delta$ in the sense that matrix $(A - \delta I_{k\times k})$ is positive definite, where $I_{k\times k}$ is the $k \times k$ identity matrix. Then for all sufficiently large n it holds that $A_n > \delta$.*

Proof. (Theorem 3). By the law of large numbers, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} l_K(Y_i) \ \to_P \ C_P \neq 0. \tag{40}$$

We get by Lemma 22

$$\begin{aligned} T_K \ &= \ \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\overrightarrow{l}(Y_i)\right\} L_k \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\overrightarrow{l}(Y_i)\right\}^T \\ &\geq \ \lambda_K^{(k)}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\overrightarrow{l}(Y_i)\right\|^2 \\ &\geq \ \lambda_K^{(k)}\cdot\frac{1}{n}\left(\sum_{i=1}^{n} l_K(Y_i)\right)^2 . \end{aligned} \tag{41}$$

By (40)

$$\begin{aligned} T_K - \pi(K,n) \ &\geq \ n\lambda_K^{(k)}\cdot\Big(\frac{1}{n}\sum_{i=1}^{n} l_K(Y_i)\Big)^2 - \pi(K,n) \\ &= \ n\lambda_K^{(k)}\big(C_K^2 + o_P(1)C_K\big) \ - \pi(K,n) \\ &= \ n\lambda_K^{(k)}C_K^2 + o_P\big(n\lambda_K^{(k)}\big) \ - \pi(K,n)\,, \end{aligned}$$

and, because K and C_K are constants determined by fixed P, condition (15) yields

$$T_K - \pi(K,n) \ \to_P \ \infty \quad \text{as} \quad n \to \infty\,. \tag{42}$$

On the other hand, by (13)

$$\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l_1(Y_i), \ldots, \frac{1}{\sqrt{n}}\sum_{i=1}^{n} l_{K-1}(Y_i),\Big) \ \to_P \ \mathcal{N}\,,$$

where $\mathcal{N}$ is a $(K-1)-$dimensional multivariate normal distribution with the expectation vector equal to zero. This implies that $T_k = O_P(1)$ for all $k = 1, 2, \ldots, K-1$, because

$$T_k \leq \lambda_1^{(k)}\left\|\frac{1}{n}\sum_{i=1}^{n} l(Y_i)\right\|^2 \ = \ \lambda_1^{(k)} O_P(1) \ = \ O_P(1)$$

and $\lambda_1^{(1)}$, $\lambda_1^{(2)}, \ldots,$ $\lambda_1^{(K-1)}$ are constants and $K < \infty$. Now by (42)

$$\lim_{n\to\infty} \sum_{k=1}^{K-1} P\left(T_k - \pi(k,n) \geq T_K - \pi(K,n)\right) \,=\, 0\,.$$

But for $d(n) \geq K$

$$P(S < K) \leq \sum_{k=1}^{K-1} P\left(T_k - \pi(k,n) \geq T_K - \pi(K,n)\right),$$

and the theorem follows. □

Because of assumption $\langle A\rangle$ we can prove the following lemma.

Lemma 24.

$$P\biggl(\biggl|\frac{1}{n}\sum_{i=1}^{n} l_K(Y_i)\biggr| \leq \sqrt{\frac{x}{\lambda_K n}}\,\biggr) \,=\, O\biggl(\frac{1}{r_n}\biggr)\,.$$

Proof. Denote $x_n := \sqrt{\frac{x}{\lambda_K n}}$, and remember that by $\langle C\rangle$ we have $E_P\, l_K(Y_i) = C_K$. Obviously, $x_n \to 0$ as $n \to \infty$. We have

$$P\biggl(\biggl|\frac{1}{n}\sum_{i=1}^{n} l_K(Y_i)\biggr| \leq x_n\biggr) \,=\, P\biggl(-x_n \,\leq\, \frac{1}{n}\sum_{i=1}^{n} l_K(Y_i) \,\leq x_n\biggr)$$

$$= P\biggl(-x_n - C_K \,\leq\, \frac{1}{n}\sum_{i=1}^{n}\bigl(l_K(Y_i) - E_P l_K(Y_i)\bigr) \,\leq x_n - C_K\biggr)\,.$$

Here we get two cases. First, suppose $C_K > 0$. Then we continue as follows:

$$P\biggl(-x_n - C_K \,\leq\, \frac{1}{n}\sum_{i=1}^{n}\bigl(l_K(Y_i) - E_P l_K(Y_i)\bigr) \,\leq x_n - C_K\biggr)$$

$$\leq \quad P\biggl(\frac{1}{n}\sum_{i=1}^{n}\bigl(l_K(Y_i) - E_P l_K(Y_i)\bigr) \,\leq x_n - C_K\biggr)$$

$$\leq \quad P\biggl(\biggl|\frac{1}{n}\sum_{i=1}^{n}\bigl(l_K(Y_i) - E_P l_K(Y_i)\bigr)\biggr| \,\geq \bigl|x_n - C_K\bigr|\biggr)$$

(for all $n \geq$ some n_K)

$$\leq \quad P\biggl(\biggl|\frac{1}{n}\sum_{i=1}^{n}\bigl(l_K(Y_i) - E_P l_K(Y_i)\bigr)\biggr| \,\geq\, \frac{C_K}{2}\biggr) \,=\, O\biggl(\frac{1}{r_n}\biggr)$$

by $\langle A\rangle$, and so we proved the lemma for the case $C_K > 0$. In case if $C_K < 0$, we write

$$P\biggl(-x_n - C_K \,\leq\, \frac{1}{n}\sum_{i=1}^{n}\bigl(l_K(Y_i) - E_P l_K(Y_i)\bigr) \,\leq x_n - C_K\biggr)$$

$$\leq \quad P\Big(\frac{1}{n}\sum_{i=1}^{n}\big(l_K(Y_i) - E_P l_K(Y_i)\big) \,\geq -x_n - C_K\Big)$$

and then we proceed analogously to the previous case. □

Proof. (Lemma 5) We will use Sloane's asymptotic expansion for the standard normal distribution function Φ : for $x \to \infty$

$$\Phi(x) = 1 - (2\pi)^{-1/2}\exp(-x^2/2)(x^{-1} + o(x^{-1})).$$

From this expansion and the CLT it follows that

$$\begin{aligned}
&P\left(\frac{1}{n}\sum_{i=1}^{n}\big[l_K(Y_i) - E_P\, l_K(Y_i)\big] \geq y\right)\\
&= \quad P\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{l_K(Y_i) - E_P\, l_K(Y_i)}{\sigma} \geq \frac{y\sqrt{n}}{\sigma}\right)\\
&= \quad 1 - P\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{l_K(Y_i) - E_P\, l_K(Y_i)}{\sigma} < \frac{y\sqrt{n}}{\sigma}\right)\\
&= \quad 1 - \Phi(y\sqrt{n}/\sigma)\\
&\sim \quad (2\pi)^{-1/2}\frac{\sigma}{y\sqrt{n}}\exp\Big(-\frac{1}{2}\frac{ny^2}{\sigma^2}\Big),
\end{aligned}$$

and we see that $r_n = \exp(ny^2/2\sigma)$ is even more than enough to guarantee $\langle A\rangle$. □

Proof. (Theorem 6). Let $x > 0$. Since $T_j > T_K$ if $j > K$ and (14) holds, we get by Theorem 3 that

$$\begin{aligned}
P(T_S \leq x) \quad &= \quad \sum_{j=K}^{d(n)} P(T_j \leq x,\, S = j) + o(1)\\
&\leq \quad d(n)\, P(T_K \leq x) + o(1)\\
&\leq \quad d(n)\, P\Big(\lambda_K\,\frac{1}{n}\left(\sum_{i=1}^{n} l_K(Y_i)\right)^2 \leq x\Big) + o(1)\\
&= \quad d(n)\, P\Big(\left|\frac{1}{n}\sum_{i=1}^{n} l_K(Y_i)\right| \leq \sqrt{\frac{x}{\lambda_K n}}\Big) + o(1)\,.
\end{aligned}$$

Now by Lemma 24 and (16) we get

$$P(T_S \leq x) \,=\, O\Big(\frac{d(n)}{r_n}\Big) + o(1) \,=\, o(1)\,.$$

□

Proof. (Theorem 7). If $S \geq K$, then $T_k - T_1 \geq \pi(k,n) - \pi(1,n)$ for some $K \leq k \leq d(n)$ and so, equivalently,

$$\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l(Y_i)\right\} L \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l(Y_i)\right\}^T$$

$$-\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l_1(Y_i)\right\}^2 \left\{E_0[l_1(Y)]^T l_1(Y)\right\}^{-1} \geq \pi(k,n) - \pi(1,n) \qquad (43)$$

for some $K \leq k \leq d(n)$, where $l = (l_1, l_2, \dots, l_k)$. We can rewrite (43) in terms of the notation (17)-(19) as follows:

$$\begin{aligned}&(\sqrt{n}\;\bar{l}_1, \dots, \sqrt{n}\;\bar{l}_k)\, L\, (\sqrt{n}\;\bar{l}_1, \dots, \sqrt{n}\;\bar{l}_k)^T \qquad (44)\\ &\quad = n\,(\bar{l}_1, \dots, \bar{l}_k)\, L\, (\bar{l}_1, \dots, \bar{l}_k)^T \geq \frac{n\,\bar{l}_1^{\,2}}{E_0\, l_1^{\,2}} + \big(\pi(k,n) - \pi(1,n)\big),\end{aligned}$$

for some $K \leq k \leq d(n)$. Denote $\Delta(k,n) := \pi(k,n) - \pi(1,n)$; then with the help of (19) we rewrite (44) as

$$n\, Q_k(\bar{l}) \geq \Delta(k,n) + \frac{n\,\bar{l}_1^{\,2}}{E_0\, l_1^{\,2}}\,, \qquad (45)$$

for some $K \leq k \leq d(n)$. Clearly,

$$\begin{aligned} P_0(S \geq K) &\leq P_0\big((43) \text{ holds for some } K \leq k \leq d(n)\big)\\ &= P_0\big((45) \text{ holds for some } K \leq k \leq d(n)\big)\\ &\leq P_0\big(n\, Q_k(\bar{l}) \geq \Delta(k,n) \text{ for some } K \leq k \leq d(n)\big)\,. \end{aligned}$$

But now by condition (B) we have

$$\begin{aligned} P_0(S \geq K) &\leq P_0\big(n\, Q_k(\bar{l}) \geq \Delta(k,n) \text{ for some } K \leq k \leq d(n)\big)\\ &\leq \sum_{k=K}^{d(n)} P_0\Big(n\, Q_k(\bar{l}) \geq \Delta(k,n)\Big)\\ &\leq \sum_{k=K}^{d(n)} \varphi\big(k; \Delta(k,n)\big)\,, \qquad (46) \end{aligned}$$

if only $d(n) \leq \min\{u_n, m_n\}$ (see Definition 4). Thus, because of the Condition (B), for each $\varepsilon > 0$ there exists $K = K_\varepsilon$ such that for all $n > n(\varepsilon)$ we have $P_0(S \geq K) \leq \varepsilon$, i.e. $S = O_{P_0}(1)$.

Now, by standard inequalities, it is possible to show that $T_S = O_{P_0}(1)$. Let us write for an arbitrary real $t > 0$

$$
\begin{aligned}
P_0(|T_S| \ge t) &= \sum_{m=1}^{K_\varepsilon} P_0(|T_m| \ge t;\, S = m) \\
&\quad + \sum_{m=K_\varepsilon+1}^{d(n)} P_0(|T_m| \ge t;\, S = m) \\
&\le \sum_{m=1}^{K_\varepsilon} P_0(|T_m| \ge t) + \sum_{m=K_\varepsilon+1}^{d(n)} P_0(S = m) \\
&= \sum_{m=1}^{K_\varepsilon} P_0(|T_m| \ge t) + P_0(S \ge K_\varepsilon + 1) \\
&\le \sum_{m=1}^{K_\varepsilon} P_0(|T_m| \ge t) + \varepsilon \\
&=: R(t) + \varepsilon.
\end{aligned}
$$

For $t \to \infty$ we have $P_0(|T_m| \ge t) \to 0$ for every fixed m, so $R(t) \to 0$ as $t \to \infty$. Now it follows that for arbitrary $\varepsilon > 0$

$$\overline{\lim_{t\to\infty}} P_0(|T_S| \ge t) \le \varepsilon,$$

therefore

$$\overline{\lim_{t\to\infty}} P_0(|T_S| \ge t) = 0$$

and

$$\lim_{t\to\infty} P_0(|T_S| \ge t) = 0.$$

This completes the proof. □

Proof. (Theorem 9). Follows from Theorems 3, 6 and 7 and our definition of consistency. □

In the next proof we will need the following theorem from [46].

Theorem 25. *Let $Z_1, \dots, Z_n$ be i.i.d. random vectors with values in $\mathbb{R}^k$. Let $EZ_i = 0$ and let the covariance matrix of Z_i be equal to the identity matrix. Assume $\|Z_1\|_k \le L$ a.e. Then, for $2k \le y^2 \le nL^{-2}$, we have*

$$Pr\bigg(\|n^{-1/2}\sum_{i=1}^{n} Z_i\|_k \ge y \bigg) \le \frac{150210}{\Gamma(k/2)} \left(\frac{y^2}{2}\right)^{\frac{k-1}{2}} \exp\bigg\{ -\frac{y^2}{2}\bigg(1-\eta_n\bigg)\bigg\},$$

where $0 \le \eta_n \le Lyn^{-1/2}$.

Proof. (Theorem 10) The SNT-statistic T_S is an NT-statistic with $L_k = E_{k\times k}$ and $\lambda_1^{(k)} = \dots = \lambda_k^{(k)} = 1$. Therefore Theorem 7 is applicable. Put (in Theorem 7) $s(k,n) = \sqrt{2k}$, $t(k,n) = \sqrt{n}M(k)^{-1}$. The Prohorov inequality is applicable if $M(k)\,\pi(k,n) \le \sqrt{n}$ and $M^2(k)\,\pi(k,n) \le n$ for all $k \le d(n)$; therefore assumption (21) guarantees that the Prohorov inequality is applicable and, moreover, that (B) holds with

$$\varphi(k;y) \;=\; \frac{150210}{\Gamma(k/2)}\left(\frac{y^2}{2}\right)^{\frac{k-1}{2}} \exp\Biggl\{ -\frac{y^2}{2}\left(1-\frac{M(k)\,y}{\sqrt{n}}\right)\Biggr\}. \tag{47}$$

Since φ is exponentially decreasing in y under (21), it is a matter of simple calculations to prove that (B2) is satisfied with $u_n = d(n)$ for any sequence $\{d(n)\}$ such that (21) holds. Thus Theorem 10 follows from Theorem 7. □

Proof. (Theorem 18). Consider the auxiliary random variable

$$R_k = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l(Y_j) \right\} L^{(0)} \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l(Y_j) \right\}^T . \tag{48}$$

This is not a test statistic, but formally this random variable satisfies Definition 3. Therefore Theorem 7 is applicable for R_k. Since under the null hypothesis $GT_k \to R_k$ and $GT_S \to R_S$ in P_0-probability by Definition 7, we get the statement of the theorem by the Slutsky lemma. □

Proof. (Theorem 19). Consider the random variable R_k defined in the proof of Theorem 18. Theorems 3, 6 and 7 are valid for the random variable R_S. Under the assumptions of the theorem, $GT_S \to R_S$ in $P-$probability, and we get the statement of the theorem by the Slutsky lemma. □