Data-driven goodness-of-fit tests

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch–Naturwissenschaftlichen Fakultäten
der Georg–August–Universität zu Göttingen

vorgelegt von
Mikhail Langovoy
aus
St Petersburg, Russland

Göttingen 2007

D7
Referent: Prof. Dr. Axel Munk
Koreferent: Prof. Dr. Manfred Denker
Tag der Mündlichen Prüfung: 09. Juli 2007

Contents

Chapter 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. 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 Appendix III for related definitions and [32] for the general theory).

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. Test statistics constructed following this approach are often called (efficient) score test statistics. The pioneer of this approach was Neyman [30] and then many other works followed: [31], [10], [4], [28], [29]. This approach is also closely related to the theory of efficient (adaptive) estimation - [3], [16]. Score tests are asymptotically optimal in the sense of intermediate efficiency in an infinite number of directions of alternatives (see Appendices I, II and IV for related definitions and [18] for some theoretical results) and show good overall performance in practice (see [23], [24]).

This thesis attempts to generalize the theory of score tests. 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. Our theory offers, in particular, an analogous generalization of the theory of data-driven score tests. We introduce the notions of NT- and GNT-tests, and other abstract concepts generalizing the concepts of Neyman's smooth test statistics, score tests and data-driven score tests.

The main goal of this thesis is to propose an unified theory to automatize the

process of building NT- and GNT-tests for different statistical problems, and to give an unified approach for proving consistency of such tests. We propose a general method for constructing consistent data-driven tests for parametric, semi- and nonparametric problems.

Examples in this thesis tries to show that the method is applicable also to dependent data and statistical inverse problems. Moreover, for any test constructed, we have an explicit rule to determine, for every particular alternative, whether the test will be consistent against it. This rule allows us to describe, in a closed form, the set of "bad" alternatives for every NT- and GNT-test. This is an important feature of the approach of this thesis.

The new theory generalizes some constructions and results of Cox, Choi, Hall, Inglot, Kallenberg, Ledwina, Neyman, Schick, van der Vaart and others.

These general results are presented in Chapter 3. But before going into the mathematical theory, we start in Chapter 2 with an important special example.

Classical hypothesis testing is concerned with testing hypotheses about random variables $X_1, \ldots, X_n$, whose values are directly observable. But, it is important from practical point of view to be able to construct tests for situations where $X_1, \ldots, X_n$ are corrupted or can only be observed with an additional noise term. These kind of problems are termed *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.

Due to importance of the deconvolution problem, testing statistical hypotheses related to this problem has been widely studied in the literature. But, to our knowledge, only the first approach described above was implemented for the problem.

In this thesis, we treat the deconvolution problem with the second approach. In Chapter 2, score tests and data-driven score tests for both simple and composite deconvolution problems are constructed. This Chapter is mostly orientated towards applied statisticians. Material in this Chapter is presented in such a way that the tests will be easy to use, even if one do not read proofs of consistency theorems. We tried to indicate situations when the tests are consistent and working fine, and also those situations where the theory predicts these tests to be not very useful. Simple and clear criterions are provided for how one can decide whether the test should be (or should not) applied in any particular situation.

In Appendices I - IV, some auxiliary definitions, lemmas and theorems are collected for the convenience of the reader. Appendices are mostly suited to provide technical references while one reads the thesis. Section 7.2 of Appendix IV, however, contains a discussion of some results on intermediate optimality.

Acknowledgements

I am grateful to my advisor, Prof. Dr. Axel Munk, for proposing the topic of my dissertation and for helpful discussions. I wish to thank Prof. Dr. Manfred Denker for taking the Koreferat, for organising many interesting seminars and for his encouraging support. I also thank Prof. Dr. Andrei Borodin for teaching me many important things about mathematical research and Prof. Dr. Mikhail Gordin for his helpful suggestions.

During my time as a Ph.D. student I was a member of the Graduiertenkolleg "Identifikation in mathematischen Modellen: Synergie stochastischer und numerischer Methoden", and I would like to thank them for their financial support. I am grateful to all the people from the Institute for Mathematical Stochastics and Graduiertenkolleg 1023 for providing me with an excellent working environment.

Many thanks go to Dr. Andrei Volotka and Dr.Dr. Elena Sivukhina, Dr. Janis Valeinis and Daina Valeina, Dr. Fadoua Balabdaoui, Dr. Dmitry Zaporozhets and Elena Tsoi, Dr. Mikhail Danilov, Dr. Ivan Yudin, Dr. Sachar Kabluchko, Dr. Marina Schnellen, Dr. Leif Boysen, Dr. Rada Dakovic (Matic), Dr. Natalia Kan-Dobrosky, Ta-Chao Kao, Mihaela Manescu, Razmig Dijekenjan, Olha Ivanyshin, Anna Levina, Michael Scheuerer, Vladislav Vysotsky, Achim Wübker, Krzysztof Mieloch, Anna Solska-Mieloch, Yuriy Botnikov and Dmitry Matveev.

I thank Vladimir Shirokov, Vitaly Burylev, Alexander Alexeev, Victor Rovsky, Mikhail Rzhevskiy, Vladimir Putin and all my other friends for their support throughout my life.

I especially thank my wife Anna, my sister Stanislava and my parents Valentina and Anatoly for everything.

Chapter 2

Deconvolution Problems

2.1 Introduction

Classical hypothesis testing deals with hypotheses about random variables $X_1, \ldots, X_n$, whose values are directly observable. But it is important from practical point of view to be able to construct tests for situations where $X_1, \ldots, X_n$ are corrupted or can only be observed with an additional noise term. We call this kind of problems *statistical inverse problems*. The most well known example here is the deconvolution. It 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 the importance of the deconvolution problem, testing statistical hypotheses related to this problem has 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, [5], [11], [15]). Thus, only the first approach described above was implemented for the deconvolution problem.

In this thesis, we treat the deconvolution 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 [4], [12], [23], [13]). Thus, we provide corresponding refinement of our tests. Following the solution proposed in [22], 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.

In Section 2.2, we formulate the simple deconvolution problem. In Section 2.3,

we construct the score tests for the parametric deconvolution hypothesis. In Section 2.5, we prove consistency of our tests against nonparametric alternatives. In Section 2.6, we turn to the deconvolution with an unknown error density. We derive the efficient scores for the composite parametric deconvolution hypothesis in Section 2.7. In Section 2.8, we construct the efficient score tests for this case. In Section 2.9, we make our tests data-driven. In Section 2.10, we prove consistency of the tests against nonparametric alternatives. Additionally, in Sections 2.5 and 2.10, we explicitly characterize the class of nonparametric alternatives such that our tests are inconsistent and therefore shouldn't be used for testing against the alternatives from this class. Some simple examples of applications of the theory are also presented in this Chapter.

2.2 Notation and basic assumptions

The problem of testing whether 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 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 $E\,\varepsilon_i = 0$, $0 < E\,\varepsilon^2 < \infty$. For brevity of notation say that X_i, Y_i, ε_i have the same distribution as random variables X, Y, ε correspondingly. 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. 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 Chapter we will at 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 consistent also 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 [18]). This is the general plan for our construction.

2.3 Score test for simple deconvolution

Suppose that all possible densities of X belong to some parametric family $\{f_\theta\}$, where θ 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 such model the form

$$q\,(y\,;\theta) = \int_{\mathbb{R}} f_\theta(s)\, h(\,y - s)\, ds\,. \tag{2.1}$$

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

$$\dot{l}\,(y\,;\theta) = \frac{\big(q\,(\theta)\big)'_\theta}{q\,(\theta)}\, 1_{[q\,(\theta)>0]}\,, \tag{2.2}$$

where $q\,(\theta) := q\,(y;\theta)$ and $l\,(\theta) := l\,(y;\theta)$ for brevity. The *Fisher information matrix* of parameter θ is defined as

$$I(\theta) = \int_{\mathbb{R}} \dot{l}\,(y\,;\theta)\dot{l}^T\,(y\,;\theta)\, dQ_\theta(y)\,. \tag{2.3}$$

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

$\langle$**B1**$\rangle$ for all $\theta \in \Theta$ $\quad q\,(y;\theta)$ is continuously differentiable in θ
for $\lambda -$ almost all y with gradient $\dot{q}\,(\theta)$

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

$\langle$**B3**$\rangle$ $\quad I(\theta)$ is nonsingular for all $\theta \in \Theta$ and continuous in $\theta\,.$

If θ is a true parameter value, call such model $GM_k(\theta)$ and denote by Q_θ the probability distribution function and by E_θ the expectation corresponding to the density $q\,(\cdot;\theta)$.

If conditions $\langle B1\rangle - \langle B3\rangle$ holds, then by Proposition 1, p.13 of [3] we calculate

for all $y \in supp\ q(\cdot;\theta)$

$$\dot{l}(\theta) = \dot{l}(y;\theta) = \frac{\left(q(y;\theta)\right)'_\theta}{q(y;\theta)} = \frac{\frac{\partial}{\partial\theta}\int_{\mathbb{R}} f_\theta(s)\,h(y-s)\,ds}{\int_{\mathbb{R}} f_\theta(s)\,h(y-s)\,ds}. \tag{2.4}$$

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

$$l^*(y) := \dot{l}(y;0) = \frac{\frac{\partial}{\partial\theta}\left(\int_{\mathbb{R}} f_\theta(s)\,h(y-s)\,ds\right)\Big|_{\theta=0}}{\int_{\mathbb{R}} f_0(s)\,h(y-s)\,ds}. \tag{2.5}$$

Set

$$L = \{E_0[l^*(Y)]^T l^*(Y)\}^{-1} \tag{2.6}$$

and

$$U_k = \left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} l^*(Y_j)\right\} L \left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} l^*(Y_j)\right\}^T. \tag{2.7}$$

Theorem 2.1. *For the regular deconvolution problem the efficient score vector l^* for testing $\theta = 0$ in $GM_k(\theta)$ is given for all $x \in \mathbb{R}$ by (2.5). Moreover, under $H_0 : \theta = 0$ we have $U_k \to_d \chi^2_k$ as $n \to \infty$.*

Proof. (Theorem 2.1). We calculated the efficient score vector in (2.4)-(2.5). By Proposition 1, p.13 of [3] 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 [3], p.15) and our statement follows. □

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 χ^2_k-distribution. Note that we do not need to estimate the scores l^*.

Corollary 2.2. *If the deconvolution problem is regular and $f_\theta(\cdot)$ is differentiable in θ for all $\theta \in \Theta$, then the conclusions of Theorem 2.1 are valid and the efficient score vector for testing $H_0 : \theta = 0$ can be calculated by the formula*

$$l^*(y) = \frac{\int_{\mathbb{R}} \left(\frac{\partial}{\partial\theta} f_\theta(s)\right)\Big|_{\theta=0} h(y-s)\,ds}{\int_{\mathbb{R}} f_\theta(s)\,h(y-s)\,ds}. \tag{2.8}$$

Example 1. 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 θ, where

$$f_\theta(x) \;=\; f_0(x)\, b(\theta) \exp(\theta \circ u(x))\,, \tag{2.9}$$

where the symbol $\circ$ 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 [1]) are satisfied. All the possible densities $q\,(y\,;\theta)$ of Y have in such model the form

$$q\,(y\,;\theta) = \int_{\mathbb{R}} f_0(s)\, b(\theta) \exp(\theta \circ u(s))\; h(\,y - s)\, ds\,. \tag{2.10}$$

These densities no longer need to form an exponential family. If we assume, for example, that $h > 0 \quad \lambda-$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 $\left|\dot{l}\,(y\,;\theta)\right| \in L_2(Q_\theta)$ and the Fisher information matrix $I(\Theta)$ is nonsingular and continuous in θ, then conditions $\langle B1\rangle - \langle B3\rangle$ are satisfied for this problem and the previous results are applicable. The score vector for the problem is

$$l^*(y) = \frac{\int_{\mathbb{R}} u(s)\, f_0(s)\, h(\,y - s)\, ds}{\int_{\mathbb{R}} f_0(s)\, h(\,y - s)\, ds} \;-\; \int_{\mathbb{R}} u(s)\, f_0(s)\, ds\,. \tag{2.11}$$

In other words, if we denote by $*$ the standard convolution of functions,

$$l^*(y) = \frac{(uf_0) * h}{f_0 * h}(y) \;-\; E_0 u(X)\,. \tag{2.12}$$

Let L be defined by (2.6) and

$$V_k = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l^*(Y_j) \right\} L \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l^*(Y_j) \right\}^T . \tag{2.13}$$

This is the score test statistic designed to be asymptotically optimal for testing H_0 against the alternatives from the exponential family (2.9). Its asymptotic distribution under the null hypothesis H_0 is given by Theorem 2.1.

2.4 Selection rule

For the use of score tests in classical hypotheses testing it was shown (see the Introduction) 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 2.5 we give a theoretical explanation of this fact for the case of deconvolution. 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 [22] 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 [19].

Denote by $M_k(\theta)$ the model described in Section 2.3 such that the true parameter θ belongs to the parameter set, say Θ_k, and $\dim \Theta_k = k$. By a *nested family* of submodels $M_k(\theta)$ for $k = 1, 2, \ldots$ we mean a sequence of these models such that for their parameter sets it holds that $\Theta_1 \subset \Theta_2 \subset \ldots$.

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 $\pi(1, n) < \pi(2, n) < \ldots < \pi(d, n)$ 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

$$S = \min\left\{k : \, 1 \le k \le d;\, U_k - \pi(k, n) \ge U_j - \pi(j, n),\, j = 1, \ldots, d\right\}. \tag{2.14}$$

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

From Theorem 2.3 below it follows that for our problem (as well as in the classical case, see [22]) many possible penalties lead to consistent tests. So the choice of the penalty should be dictated by external practical considerations. Our simulation study is not so vast to recommend the most practically suitable penalty for the deconvolution problem. Possible choices are, for example, Schwarz's penalty $\pi(j, n) = j \log n$, or Akaike's penalty $\pi(j, n) = j$.

Denote by P_0^n 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 2.3. *Suppose that assumptions* $\langle B1\rangle - \langle B3\rangle$ *holds. Then under the null hypothesis* H_0 *it holds that* $P_0^n(S > 1) \to 0$ *and* $U_S \to_d \chi_1^2$ *as* $n \to \infty$.

Proof. (Theorem 2.3). Denote $\Delta(k, n) := \pi(k, n) - \pi(1, n)$. For any $k = 2, \ldots, d$

$$\begin{aligned} P_0^n(S = k) &\leq P_0^n\big(U_k - \pi(k, n) \geq U_1 - \pi(1, n)\big) \\ &\leq P_0^n\big(U_k \geq \pi(k, n) - \pi(1, n)\big) \\ &= P_0^n\big(U_k \geq \Delta(k, n)\big)\,. \end{aligned}$$

By Theorem 2.1 $U_k \to_d \chi_k^2$ as $n \to \infty$, thus for $\Delta(k, n) \uparrow \infty$ as $n \to \infty$ we have $P_0^n\big(U_k \geq \Delta(k, n)\big) \to 0$ as $n \to \infty$, so for any $k = 2, \ldots, d$ we have $P_0^n(S = k) \to 0$ as $n \to \infty$. This proves that

$$P_0^n(S \geq 2) = \sum_{k=2}^{d} P_0^n(S = k) \to 0, \quad n \to \infty,$$

and so $P_0^n(S = 1) \to 1$. Now write for arbitrary real $t > 0$

$$\begin{aligned} P_0^n(|U_S - U_1| \geq t) &= P_0^n(|U_1 - U_1| \geq t;\, S = 1) \\ &\quad + \sum_{m=2}^{d} P_0^n(|U_m - U_1| \geq t;\, S = m) \\ &= \sum_{m=2}^{d} P_0^n(|U_m - U_1| \geq t;\, S = m). \end{aligned} \tag{2.15}$$

For $m = 2, \ldots, d$ we have $P_0^n(S = m) \to 0$, so

$$0 \leq \sum_{m=2}^{d} P_0^n(|U_m - U_1| \geq t;\, S = m) \leq \sum_{m=2}^{d} P_0^n(S = m) \to 0$$

as $n \to \infty$ and thus by (2.15) it follows that U_S tends to U_1 in probability as $n \to \infty$. But $U_1 \to_d \chi_1^2$ by Theorem 2.1, so $U_S \to_d \chi_1^2$ as $n \to \infty$. □

Remark 2.4. 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 2.3 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 [23]).

2.5 Consistency of tests

Let F be a true distribution function of X. Here F is *not* necessarily parametric and possibly doesn't have a density with respect to λ. 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 problem in the sense of Definition 1. Then one is able to construct the score test statistic U_k defined by (2.7) despite the fact that the true F possibly has no relation to the chosen $\{f_\theta\}$. One can use the exponential family from Example 1 as $\{f_\theta\}$, or some other parametric family whatever is convenient. This is our goal in this section to determine under what conditions thus build U_k will be consistent for testing against F.

Suppose that the following condition holds

$\langle$**D1**$\rangle$ there exists an integer $K \geq 1$ such that $K \leq d$ and

$$E_F\, l_1^* = 0, \ldots, E_F\, l_{K-1}^* = 0,\ E_F\, l_K^* = C_K \neq 0\,,$$

where l_i^* is the $i-$th coordinate function of l^* and l^* is defined by (2.5), d is the maximal possible dimension of our model as in Definition 2 of Section 2.4, and E_F denotes the mathematical expectation with respect to $F * h$.

Condition $\langle D1\rangle$ is a weak analog of nondegeneracy: if for all k $\langle D1\rangle$ 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 $\langle D1\rangle$ is related to the identifiability of the model (see the beginning of Section 2.10 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 2.5. *Let* $\langle D1\rangle$ *holds. Then for all* $1 \leq k \leq K-1$, *if* F *is the true distribution function of* X, *then* $U_k \to_d \chi_k^2$ *as* $n \to \infty$.

Proof. (Proposition 2.5). Follows by the multivariate Central Limit Theorem. □

This result and Theorem 2.1 show that if the dimension of the model is too small, then the test doesn't work since it doesn't distinguish between F and f_0.

Proposition 2.6. *Let $\langle D1\rangle$ holds. Then for $k \geq K$, if F is the true distribution function of X, then $U_k \to \infty$ in probability as $n \to \infty$.*

Proof. (Proposition 2.6). We shall use the following standard lemma from linear algebra.

Lemma 2.7. *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 the 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 $x \in \mathbb{R}^k$ it holds that $xAx^T > \delta\|x\|^2$.*

From $\langle D1\rangle$ by the law of large numbers we get

$$\frac{1}{n}\sum_{j=1}^{n} l_i^*(Y_j) \;\to_P\; 0 \quad \text{for} \quad 1 \leq i \leq K-1 \tag{2.16}$$

$$\frac{1}{n}\sum_{j=1}^{n} l_i^*(Y_j) \;\to_P\; C_K \neq 0. \tag{2.17}$$

We apply Lemma 2.7 to the matrix L defined in (2.6); since all the eigenvalues of L are positive we can choose δ to be any fixed positive number less than the smallest eigenvalue of L. We obtain the following inequality

$$\begin{aligned} U_k &= \left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} l^*(Y_j)\right\} L \left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} l^*(Y_j)\right\}^T \\ &> \delta\left\|\frac{1}{\sqrt{n}}\sum_{j=1}^{n} l^*(Y_j)\right\|^2 = \delta\, n \sum_{i=1}^{k}\left(\frac{1}{n}\sum_{j=1}^{n} l_i^*(Y_j)\right)^2 \\ &\geq \delta\, n\left(\frac{1}{n}\sum_{j=1}^{n} l_K^*(Y_j)\right)^2. \end{aligned} \tag{2.18}$$

Now by (2.16) and (2.17) we get for all $s \in \mathbb{R}$

$$\begin{aligned} P(U_k \le s) &\le P\Big(\delta\, n \Big(\frac{1}{n}\sum_{j=1}^{n} l_K^*(Y_j)\Big)^2 \le s\Big) \\ &= P\Big(\Big(\frac{1}{n}\sum_{j=1}^{n} l_K^*(Y_j)\Big)^2 \le \frac{s}{\delta\, n}\Big) \\ &= P\Big(\Big|\frac{1}{n}\sum_{j=1}^{n} l_K^*(Y_j)\Big| \le \sqrt{\frac{s}{\delta\, n}}\Big) \to 0 \quad \text{as} \quad n \to \infty\,, \end{aligned}$$

and this proves the Proposition.

□

Now we turn to the data-driven statistic U_S. Suppose that the selection rule S is defined as in Section 2.4. Assume that

⟨**S1**⟩ for every fixed $k \ge 1$ it holds that $\pi(k,n) = o(n)$ as $n \to \infty$.

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 2.8. *Let* $\langle D1\rangle$ *and* $\langle S1\rangle$ *holds. If* F *is the true distribution function of* X, *then* $P_F(S \ge K) \to 1$ *and* $U_S \to \infty$ *as* $n \to \infty$.

Proof. (Proposition 2.8). Let $\pi(k,n)$ and $\Delta(k,n)$ be defined as in Section 2.4. For any $i = 1, \ldots, K-1$ we have

$$\begin{aligned} P_F(S = i) &\le P_F\big(U_i - \pi(i,n) \ge U_K - \pi(K,n)\big) \\ &= P_F\big(U_i \ge U_K - (\pi(K,n) - \pi(i,n))\big). \end{aligned} \tag{2.19}$$

By (2.17) and (2.18) we get

$$P_F\left(U_K \ge \delta\, \frac{C_K}{2}\, n\right) \to 1 \quad \text{as} \quad n \to \infty\,. \tag{2.20}$$

Note that

$$P_F\left(U_i \geq U_K - (\pi(K,n) - \pi(i,n))\right) \tag{2.21}$$

$$\begin{aligned} \leq \quad & P_F\left(U_i \geq \delta \frac{C_K}{2} n - (\pi(K,n) - \pi(i,n));\, U_K \geq \delta \frac{C_K}{2} n\right) \\ & + P_F\left(U_K \leq \delta \frac{C_K}{2} n\right). \end{aligned}$$

Since by $\langle S1 \rangle$ it holds that $\pi(K,n) - \pi(i,n) = o(n)$, we get

$$P_F\left(U_i \geq \delta \frac{C_K}{2} n - (\pi(K,n) - \pi(i,n));\, U_K \geq \delta \frac{C_K}{2} n\right) \tag{2.22}$$

$$\begin{aligned} \leq \quad & P_F\left(U_i \geq \delta \frac{C_K}{2} n - (\pi(K,n) - \pi(i,n))\right) \\ \leq \quad & P_F\left(U_i \geq \delta \frac{C_K}{2} n\right) \to 0 \end{aligned}$$

as $n \to \infty$ by Chebyshev's inequality since by Proposition 2.5 we have $U_i \to_d \chi^2_i$ as $n \to \infty$ for all $i = 1, \ldots, K-1$. Substituting (2.20) and (2.22) to (2.21) 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 2.6. Thus $P_F(U_S \leq t) \to 0$ as $n \to \infty$ for all $t \in \mathbb{R}$. $\square$

The main result of this section is the following

Theorem 2.9.

1. *The test based on U_k is consistent for testing against all alternative distributions F such that $\langle D1 \rangle$ is satisfied with $K \leq k$*
2. *The test based on U_k is inconsistent for testing against all alternative distributions F such that $\langle D1 \rangle$ is satisfied with $K > k$*
3. *If the selection rule S satisfies $\langle S1 \rangle$, then test based on U_S is consistent against all alternative distributions F such that $\langle D1 \rangle$ is satisfied with some K.*

Proof. (Theorem 2.9). Part 1 follows from Theorem 2.1 and Proposition 2.6, part 2 from Theorem 2.1 and Proposition 2.5, part 3 from Theorem 2.3 and Proposition 2.8. □

2.6 Composite deconvolution

In the previous sections we treated the simplest case of the deconvolution problem. The next sections are devoted to the more realistic case of unknown error density. Our main ideas and constructions will be similar to the ones 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 concerning 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. The modification we will use is called efficient scores.

Despite of all the changes, we will still be able to build a selection rule for the new problem. We will need a new and modified definition of the selection rule. Big part of the new model uniformity assumptions will be needed not to build an efficient score test, but to make such test data-driven (see section 2.9).

Consider the situation described in the first paragraph of Section 2.2, but with the following complication introduced. Suppose further on that the density h of ε is *unknown*.

Then the most general possible null hypothesis H_0 in this setup is that $f = f_0$ and the error ε 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. At 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 θ and θ is a $k-$dimensional Euclidian parameter and $\Theta \subseteq R^k$ is a parameter set for θ. Our starting assumption about the density of the error ε will be that h belongs to some specific parametric family $\{h_\eta\}$, where $\eta \in \Lambda$ and $\Lambda \subseteq R^m$ is a parameter set. Thus, η is a nuisance parameter. The null hypothesis H_0 is the following composite hypothesis: X has particular density f_0 with respect to λ.

Then we will propose a test that is expected to be asymptotically optimal (in some sense) for testing in this parametric situation. After that we will prove that our test is consistent also against a wide class of nonparametric alternatives. Moreover, the test is expected to be asymptotically optimal (in some sense) for testing against an infinite number of directions of nonparametric alternatives. This is essentially the same plan as for the simple case.

If (θ, η) is a true parameter value, we call such submodel $M_{k,m}(\theta, \eta)$. Denote in this case the density of Y by $g(\cdot; (\theta, \eta))$ and the corresponding expectation by $E_{(\theta,\eta)}$. Let the null hypothesis H_0 be $\theta = \theta_0$, where it is assumed that $\theta_0 \in \Theta$. Then the alternative hypothesis $\theta \neq \theta_0$ is a parametric subset of the original general and nonparametric alternative hypothesis H_A.

2.7 Efficient scores

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

$$g\,(y\,;(\theta,\eta)) = \int_{\mathbb{R}} f_\theta(s)\, h_\eta(\,y - s)\, ds\,. \tag{2.23}$$

It is not always possible to identify θ or/and η 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 2.10).

The *score function* for (θ, η) at (θ_0, η_0) is defined as (see [3], p.28):

$$\dot{l}_{\theta_0,\eta_0}(y) \;= \big(\,\dot{l}_{\theta_0}(y),\, \dot{l}_{\eta_0}(y)\,\big)\,, \tag{2.24}$$

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.

$$\dot{l}_{\theta_0}(y) = \frac{\frac{\partial}{\partial\theta}\left(g\left(y\,;(\theta,\eta_0)\right)\right)\Big|_{\theta=\theta_0}}{g\left(y\,;(\theta_0,\eta_0)\right)}\,1_{[y:\,g\,(y\,;(\theta_0,\eta_0))>0]} \tag{2.25}$$

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

$$\dot{l}_{\eta_0}(y) = \frac{\frac{\partial}{\partial\eta}\left(g\left(y\,;(\theta_0,\eta)\right)\right)\Big|_{\eta=\eta_0}}{g\left(y\,;(\theta_0,\eta_0)\right)}\,1_{[y:\,g\,(y\,;(\theta_0,\eta_0))>0]} \tag{2.26}$$

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

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

$$I(\theta,\eta) \;=\; \int_{\mathbb{R}} \dot{l}^{T}_{\theta,\eta}(y)\,\dot{l}_{\theta,\eta}(y)\;dG_{\theta,\eta}(y)\,, \tag{2.27}$$

where $G_{\theta,\eta}(y)$ is the probability measure corresponding to the density $g\left(y\,;(\theta,\eta)\right)$. The symbol 'T' denotes the transposition and all vectors are supposed to be row ones.

We assume that $M_{k,m}(\theta,\eta)$ is a regular parametric model in the sense of the following definition.

Definition 3. Call our problem a *regular deconvolution problem* if

$\langle$**A1**$\rangle$ for all $(\theta,\eta)\in\Theta\times\Lambda$ $\quad g\left(y\,;(\theta,\eta)\right)$ is continuously differentiable in (θ,η) for $\lambda-$almost all y

$\langle$**A2**$\rangle$ $\quad\left|\dot{l}\,(\theta,\eta)\right|\in L_2(\mathbb{R},G_{\theta,\eta})$ for all $(\theta,\eta)\in\Theta\times\Lambda$

$\langle$**A3**$\rangle$ $\quad I(\theta,\eta)$ is nonsingular for all $(\theta,\eta)\in\Theta\times\Lambda$ and continuous

in (θ, η) .

This is a joint regularity condition and it is stronger than the assumption that the model is regular in θ and η separately. Let us write $I(\theta_0, \eta_0)$ in the block matrix form:

$$I(\theta_0, \eta_0) \;=\; \begin{pmatrix} I_{11}(\theta_0, \eta_0) & I_{12}(\theta_0, \eta_0) \\ I_{21}(\theta_0, \eta_0) & I_{22}(\theta_0, \eta_0) \end{pmatrix} , \tag{2.28}$$

where $I_{11}(\theta_0, \eta_0)$ is $k \times k$, $I_{12}(\theta_0, \eta_0)$ is $k \times m$, $I_{21}(\theta_0, \eta_0)$ is $m \times k$, $I_{11}(\theta_0, \eta_0)$ is $m \times m$. Thus, denoting for simplicity of formulas $\Omega := [y : \; g\,(y\,; (\theta_0, \eta_0)) > 0]$ we can write explicitly

$$I_{11}(\theta_0, \eta_0) \;=\; E_{\theta_0,\eta_0} \dot{l}^T_{\theta_0}\, \dot{l}_{\theta_0} \;=\; \int_{\mathbb{R}} \dot{l}^T_{\theta_0}(y)\, \dot{l}_{\theta_0}(y)\; dG_{\theta_0,\eta_0}(y) \tag{2.29}$$

$$= \int_{\Omega} \frac{\frac{\partial}{\partial\theta} \left(\int_{\mathbb{R}} \, f_\theta(s)\, h_{\eta_0}(\, y - s)\, ds \right)^T \Big|_{\theta=\theta_0} \frac{\partial}{\partial\theta} \left(\int_{\mathbb{R}} \, f_\theta(s)\, h_{\eta_0}(\, y - s)\, ds \right) \Big|_{\theta=\theta_0}}{\int_{\mathbb{R}} \, f_{\theta_0}(s)\, h_{\eta_0}(\, y - s)\, ds}\, dy \,,$$

$$I_{12}(\theta_0, \eta_0) \;=\; E_{\theta_0,\eta_0} \dot{l}^T_{\theta_0}\, \dot{l}_{\eta_0} \;=\; \int_{\mathbb{R}} \dot{l}^T_{\theta_0}(y)\, \dot{l}_{\eta_0}(y)\; dG_{\theta_0,\eta_0}(y) \tag{2.30}$$

$$= \int_{\Omega} \frac{\frac{\partial}{\partial\theta} \left(\int_{\mathbb{R}} \, f_\theta(s)\, h_{\eta_0}(\, y - s)\, ds \right)^T \Big|_{\theta=\theta_0} \frac{\partial}{\partial\eta} \left(\int_{\mathbb{R}} \, f_{\theta_0}(s)\, h_{\eta}(\, y - s)\, ds \right) \Big|_{\eta=\eta_0}}{\int_{\mathbb{R}} \, f_{\theta_0}(s)\, h_{\eta_0}(\, y - s)\, ds}\, dy \,,$$

and analogously for $I_{21}(\theta_0, \eta_0)$ and $I_{22}(\theta_0, \eta_0)$. The *efficient score function* for θ in $M_{k,m}(\theta, \eta)$ is defined as (see [3], p.28):

$$l^*_{\theta_0}(y) \;=\; \dot{l}_{\theta_0}(y) \;-\; I_{12}(\theta_0, \eta_0)\, I^{-1}_{22}(\theta_0, \eta_0)\, \dot{l}_{\eta_0}(y) \,, \tag{2.31}$$

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

$$I^*_{\theta_0} \;=\; E_{\theta_0,\eta_0} l^{*T}_{\theta_0}\, l^*_{\theta_0} \;=\; \int_{\mathbb{R}} l^*_{\theta_0}(y)^T\, l^*_{\theta_0}(y)\, dG_{\theta_0,\eta_0}(y) \,. \tag{2.32}$$

Before closing this section we consider two simple examples.

Example 2. Suppose $\theta \in \mathbb{R}$, $\eta \in \mathbb{R}^+$ and, moreover, $\{f_\theta\}$ is a family $\{N(\theta,1)\}$ of normal densities with mean θ and variance 1, and $\{h_\eta\}$ is a family $\{N(0,\eta^2)\}$. Then $g(\theta,\eta) = f_\theta * h_\eta \sim N(\theta,\eta^2+1)$. Let θ be the parameter of interest and η the nuisance one. Let H_0 be $\theta = \theta_0$. By (2.25) and (2.26) for all y

$$\dot{l}_{\theta_0}(y) = \frac{y-\theta_0}{\eta_0^2+1}\,, \quad \dot{l}_{\eta_0}(y) = \frac{(y-\theta_0)^2\,\eta_0}{(\eta_0^2+1)^2} - \frac{\eta_0}{\eta_0^2+1}\,. \tag{2.33}$$

By (2.30)

$$I_{12}(\theta,\eta) \;=\; \int_{\mathbb{R}} \frac{y-\theta}{\eta^2+1}\left[\frac{(y-\theta)^2\eta}{(\eta^2+1)^2} - \frac{\eta}{\eta^2+1}\right] dN(\theta,\eta^2+1)(y) \;=\; 0\,,$$

for all θ,η. This means that adaptive estimation of θ is possible in this model, i.e., we can estimate θ equally well whether we know the true η_0 or not. Though, we will not be concerned with estimation here. From (2.29) we get

$$(I_\theta^*)^{-1} \;=\; \int_{\mathbb{R}} \frac{(y-\theta)^2}{(\eta^2+1)^2}\, dN(\theta,\eta^2+1)(y) \;=\; \frac{1}{\eta^2+1}\,. \tag{2.34}$$

Example 3. Suppose now that we are interested in the parameter η in the situation of Example 2 and the null hypothesis is $H_0 : \eta = \eta_0$. There is a sort of symmetry between signal and noise: "what is a signal for one person is a noise for the other" (see also Remark 2.10). From Example 2 we know that the score function $\dot{l}_{\eta_0}$ for η at η_0 is given by (2.33). Since we proved for this example $I_{12} = I_{21} = 0$, the efficient score function $l^*_{\eta_0}$ for η at η_0 is given by (2.33) as well. We calculate now

$$(I^*_{\eta_0})^{-1} \;=\; \int_{\mathbb{R}} \left(\frac{(y-\theta)^2\,\eta_0}{(\eta_0^2+1)^2} - \frac{\eta_0}{\eta_0^2+1}\right)^2 dN(\theta,\eta_0^2+1)(y) \;=:\; \frac{1}{C(\eta_0)}\,. \tag{2.35}$$

The constant $C(\eta_0)$ in (2.35) can be expressed explicitly in terms of η_0, but this is not the point of this example. By the symmetry of θ and η we have $l^*_{\eta_0}(y) = \dot{l}_{\eta_0}(y) - I_{21}(\theta,\eta_0)\, I_{11}^{-1}(\theta,\eta_0)\, \dot{l}_{\theta_0}(y) = \dot{l}_{\eta_0}(y)\,.$

Remark 2.10. Note that the problem is symmetric in θ and η in the sense that it is possible to consider estimating and testing for each parameter, θ or η. Physically this means that from the noisy signal one can recover some "information" not only

about the pure signal but also about the noise. This is actually natural since a noise is in fact also a signal. We are observing two signals at once. The payment for this possibility is that except for some trivial cases one can't recover full information about both the signal of interest as well as about the noise.

2.8 Efficient score test

Let $l^*_{\theta_0}$ be defined by (2.31) and $I^*_{\theta_0}$ by (2.32). Note that both $l^*_{\theta_0}$ and $I^*_{\theta_0}$ depends (at least in principle) on the unknown nuisance parameter η_0. Let l^*_j and L be some estimators of $l^*_{\theta_0}(Y_j)$ and $(I^*_{\theta_0})^{-1}$ correspondingly. These estimators are supposed to depend only on the observable $Y_1, \dots, Y_n$, but not on the $X_1, \dots, X_n$.

Definition 4. We say that l^*_j is a *sufficiently good* estimator of $l^*_{\theta_0}(Y_j)$ if for each $(\theta_0, \eta_0) \in \Theta \times \Lambda$ it holds that for every $\varepsilon > 0$

$$G^n_{\theta_0,\eta_0}\left(\frac{1}{\sqrt{n}}\left\|\sum_{i=1}^{n}(l^*_j - l^*_{\theta_0}(Y_j))\right\| \geq \varepsilon\right) \to 0 \quad \text{as} \quad n \to \infty\,, \tag{2.36}$$

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

In other words, condition (2.36) means that the average $\frac{1}{n}\sum_{i=1}^{n} l^*_{\theta_0}(Y_j) \approx E_{\theta_0,\eta_0} l^*_{\theta_0}$ is $\sqrt{n}-$consistently estimated. We illustrate this definition by some examples.

Example 2 (continued). We have (denoting variance of Y by $\sigma^2(Y)$):

$$l^*_{\theta_0}(Y_j) = \frac{Y_j - \theta_0}{\sigma^2(Y)}.$$

Define

$$l^*_j := \frac{Y_j - \theta_0}{\widehat{\sigma}^2_n},$$

where $\widehat{\sigma}^2_n$ is any $\sqrt{n}-$consistent estimator of the variance of Y. One can take, for example, the sample variance $s^2_n = s^2_n(Y_1, \dots, Y_n)$ as such an estimate. Then, since by the model assumptions $\sigma^2(Y) > 0$, thus constructed l^*_j satisfies Definition 4. See Appendix for the proof. □

Example 3 (continued). We have in this case

$$l^*_{\eta_0}(Y_j) = \frac{\eta_0}{\eta_0^2 + 1}(Y_j - \theta_0)^2 - \frac{\eta_0}{\eta_0^2 + 1}.$$

For simplicity of notations we write $l^*_{\eta_0}(Y_j) = C_1(\eta_0)(Y_j - \theta_0)^2 - C_2(\eta_0)$. Let $\widehat{\theta}_n$ be any $\sqrt{n}-$consistent estimate of θ_0 and put $l^*_j := C_1(\eta_0)(Y_j - \widehat{\theta}_n)^2 - C_2(\eta_0)$. Then Definition 4 is satisfied in this Example also. This is proved in Appendix. □

Definition 4 reflects the basic idea of the method of estimated scores. This method is widely used in statistics (see [3], [35], [16], [19] 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 (2.36).

Definition 5. Define

$$W_k = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l^*_j \right\} \widehat{L} \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l^*_j \right\}^T, \tag{2.37}$$

where $\widehat{L}$ is an estimate of $(I^*_{\theta_0})^{-1}$ depending only on $Y_1, \ldots, Y_n$. Note that l^*_j is a $k-$dimensional vector and $\widehat{L}$ is a $k \times k$ matrix. We call W_k the *efficient score test statistic* for testing $H_0 : \theta = \theta_0$ in $M_{k,m}(\theta, \eta)$. 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 $\widehat{\eta}_n$ of η by standard methods since at this point our construction is parametric. After that it would be enough to plug in these estimates in (2.31) and get the desired $l^{*\prime}_j$s satisfying (2.36).

Example 2 (continued). Let $\widehat{\sigma}^2(Y)$ be any $\sqrt{n}-$consistent estimate of η^2+1 such that this estimate is based on $Y_1, \ldots, Y_n$. Then by (2.34), (2.33) and definition (2.37) the efficient score test statistic for testing $H_0 : \theta = \theta_0$ (in the model $M_{1,1}(\theta, \eta)$) is

$$W_1 = \left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{Y_j - \theta_0}{\widehat{\sigma}^2_n(Y)} \right)^2 \widehat{\sigma}^2_n(Y) = \frac{1}{\widehat{\sigma}^2_n(Y)} \left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (Y_j - \theta_0) \right)^2. \tag{2.38}$$

Example 3 (continued). Using any $\sqrt{n}-$ consistent estimate $\widehat{\theta}$ of θ, we get the efficient score test statistic

$$W_1 = \left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left[\frac{(Y_j - \widehat{\theta}_n)^2 \eta_0}{(\eta_0^2+1)^2} - \frac{\eta_0}{\eta_0^2+1} \right] \right)^2 C(\eta_0)$$

$$= \left(\frac{1}{\sqrt{n}} \frac{\eta_0}{(\eta_0^2+1)^2} \sum_{j=1}^{n} (Y_j - \widehat{\theta}_n)^2 - \sqrt{n} \frac{\eta_0}{\eta_0^2+1} \right)^2 C(\eta_0)\,. \qquad (2.39)$$

Remark 2.11. We make now the following remark to avoid possible confusions. For the simple deconvolution we had the score test statistics and now we have the *efficient* score test statistics. This does not mean that the statistics for simple deconvolution is "inefficient". Here the word "efficient" has a strictly technical meaning. Because of the presence of the nuisance parameter we have to extract information about the parameter of interest. We want to do this efficiently in some sense. This is the explanation of the terminology.

The following theorem describes asymptotic behavior of W_k under the null hypothesis.

Theorem 2.12. *Assume the null hypothesis $H_0 : \theta = \theta_0$ holds true, $\langle A1\rangle$-$\langle A3\rangle$ are fulfilled, (2.36) is satisfied, and $\widehat{L}$ is any consistent estimate of $(I^*_{\theta_0})^{-1}$. Then*

$$W_k \to_d \chi^2_k \quad as \quad n \to \infty\,,$$

where χ^2_k denotes a random variable with central chi-square distribution with k degrees of freedom.

Proof. (Theorem 2.12). Put

$$V_k = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l^*_{\theta_0}(Y_j) \right\} (I^*_{\theta_0})^{-1} \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l^*_{\theta_0}(Y_j) \right\}^T, \qquad (2.40)$$

where $l^*_{\theta_0}$ is defined by (2.31) and $I^*_{\theta_0}$ by (2.32). Of course, V_k is *not* a statistic since it depends on the unknown η_0. But if the true η_0 is known, then because of $\langle B1\rangle$-$\langle B3\rangle$ we can apply the multivariate Central Limit Theorem and obtain $V_k \to_d \chi^2_k$ as $n \to \infty$. Condition (2.36) implies that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} l^*_j \;\to\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l^*_{\theta_0}(Y_j) \quad \text{in } G_{\theta_0,\eta_0}\text{−probability as } n \to \infty$$

and by consistency of $\widehat{L}$ we get the statement of the theorem by Slutsky's Lemma. □

2.9 Selection rule

In this section we extend the construction of Section 2.4 to the case of composite hypotheses. First we give a general definition of a selection rule.

Denote by $M_{k,m}(\theta,\eta)$ the model described in Section 2.6 and such that the true parameter (θ,η) belongs to a parameter set, say $\Theta_k \times \Lambda$, and $\dim \Theta_k = k$. By a *nested family* of submodels $M_{k,m}(\theta,\eta)$ for $k = 1,\dots$ we would mean a sequence of these models such that for their parameter sets it holds that $\Theta_1 \times \Lambda \subset \Theta_2 \times \Lambda \subset \dots$.

Definition 6. Consider a nested family of submodels $M_{k,m}(\theta,\eta)$ for $k = 1, \dots, d$, where d is fixed but otherwise arbitrary, and m is fixed. 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 $\pi(1,n) < \pi(2,n) < \dots < \pi(d,n)$ for all n and $\pi(j,n) - \pi(1,n) \to \infty$ as $n \to \infty$ for every $j = 2,\dots,d$. Call $\pi(j,n)$ a *penalty attributed to the j-th model $M_j(\theta)$ and the sample size n.* Then a *selection rule* $S(l^*)$ for the test statistic W_k is an integer-valued random variable satisfying the condition

$$S(l^*) = \min\big\{k : 1 \le k \le d;\, W_k - \pi(k,n) \ge W_j - \pi(j,n),\, j = 1,\dots,d\big\}. \quad (2.41)$$

We call the random variable W_S a *data-driven efficient score test statistic* for testing validity of the initial model. We also assume that the following condition holds.

⟨**S1**⟩ for every fixed $k \ge 1$ it holds that $\pi(k,n) = o(n)$ as $n \to \infty$.

Unlike the case of the simple null hypothesis, in the case of the composite hypotheses the selection rule depends on the estimator l_j^* of the unknown values $l_{\theta_0}^*(Y_j)$ of the efficient score function. This means that we need to estimate the nuisance parameter η, or corresponding scores, or their sum. Surprising result follows from Theorem 2.13 below: for our problem many possible penalties and, moreover, essentially all sensible estimators plugged in W_k, give consistent selection rules. Possible choices of penalties are, for instance, Shwarz's penalty $\pi(j,n) = j \log n$, or Akaike's penalty $\pi(j,n) = j$.

Denote by $P^n_{\theta_0,\eta_0}$ the probability measure corresponding to the case when $X_1,\dots,X_n$ all have the density $f(\theta_0,\eta_0)$. The main result about the asymptotic null distribution of W_S is the following theorem (it is proved analogously to Theorem 2.3).

Theorem 2.13. *Under the conditions of Theorem 2.12, as $n \to \infty$ it holds that*

$$P^n_{\theta_0,\eta_0}(S(l^*) > 1) \to 0 \quad \textit{and} \quad W_S \to_d \chi^2_1.$$

Condition (2.36) is what makes this direct reference to the case of the simple hypothesis possible. Estimation of the efficient score function $l^*_{\theta_0}$ can be done by different ways. First way is to estimate the whole expression from the right side of (2.31). For this method of estimation condition (2.36) is natural. The second and probably more convenient method of estimating $l^*_{\theta_0}$ is via estimation of the nuisance parameter η by some estimator $\widehat{\eta}$. But for this approach condition (2.36) becomes something that have to be proved for each particular estimator. We hope that this inconvenience is excused by the fact that we are only introducing the new test here. It is possible to reformulate condition (2.36) explicitly in terms of conditions on $\widehat{\eta}$, $\{f_\theta\}$, and $\{h_\eta\}$ (see an analogue in [17]).

Remark 2.14. The selection rule $S(l^*)$ 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 $M_{k,m}(\theta, \eta)$ as the number of observations grows. See Remark 2.4.

Remark 2.15. 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 Chapter. Possibly this statistic could be useful since the situation with the noise of an unknown dimension often seems to be more realistic. On the other hand, this statistic will also have some disadvantages. One will have to impose more strict assumptions on both signal and noise (including an analogue of the double-identifiability assumption). Also the final result will be weaker than the result of this section. This will be a payment for an attempt to extract information about a larger number of parameters from the same amount of observations $Y_1, \ldots, Y_n$.

2.10 Consistency of tests

Let F be a true distribution function of X and H a true distribution of ε. Here F and H are *not* necessarily parametric and possibly these distribution functions do not have densities with respect to the Lebesgue measure λ. Let us choose for every $k \le d$ 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. Correspondingly, let us fix an integer m and choose an auxiliary parametric family $\{h_\eta\}$, $\eta \in \Lambda \subseteq \mathbb{R}^m$. Suppose that the chosen families $\{f_\theta\}$ and $\{h_\eta\}$ give us the regular deconvolution problem in

the sense of Definition 3. Then one is able to construct the score test statistic W_k defined by (2.37) despite the fact that the true F and H possibly do not have any relation to the chosen $\{f_\theta\}$ and $\{h_\eta\}$. This is our goal in this section to determine under what conditions thus build W_k will be consistent for testing against H_A.

Suppose that the following condition holds

$\langle$**C1**$\rangle$ there exists integer $K \geq 1$ such that $K \leq d$ and

$$E_{F*H}\, l^*_{\theta_0(1)} = 0, \ldots, E_{F*H}\, l^*_{\theta_0(K-1)} = 0,\ E_{F*H}\, l^*_{\theta_0(K)} = C_K \neq 0\,,$$

where $l^*_{\theta_0(i)}$ is the $i-$th coordinate function of $l^*_{\theta_0}$ and $l^*_{\theta_0}$ is defined by (2.31), d is the maximal possible dimension of our model as in Definition 3 of Section 2.9, and E_{F*H} denotes the mathematical expectation with respect to $F * H$.

Condition $\langle C1\rangle$ is a weak analog of nondegeneracy: if for all k $\langle C1\rangle$ fails, then F is orthogonal to the whole system $l^*_{\theta_0(i)}{}_{i=1}^{\infty}$ and if this system is complete, then $F * H$ is degenerate. Also $\langle C1\rangle$ is related to the identifiability of the model: if the model is not identifiable, then $F * H = F_0 * H$ can happen and $\langle C1\rangle$ fails. Establishing identifiability for the parametric deconvolution is not trivial (see [37], e.g.). It is important to note also that although $\langle C1\rangle$ 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 2.16. *If (2.36) is satisfied and $\widehat{L}$ is any consistent estimate of $(I^*_{\theta_0})^{-1}$, then*

1. *the test based on W_k is consistent for testing against all alternative distributions F, H such that $\langle C1\rangle$ is satisfied with $K \leq k$*
2. *the test based on W_k is inconsistent for testing against alternative distributions F, H such that $\langle C1\rangle$ is satisfied with $K > k$*
3. *if the selection rule $S(l^*)$ satisfies $\langle S1\rangle$, then test based on W_S is consistent against all alternative distributions $F * H$ such that $\langle C1\rangle$ is satisfied with some K.*

Part 2 of Theorem 2.16 shows why it is important to choose the suitable model dimension. Now we give two specific examples.

Proof. (Theorem 2.16). Because of condition (2.36) the proof is analogous to the proof of Theorem 2.9. Indeed, after obvious change of notations Propositions 2.5, 2.6, and 2.8 are true for W_k, $W_{S(l^*)}$, $S(l^*)$ instead of U_k, U_S, S. Proofs of the new versions of propositions are analogous to the proofs of the previous versions. The only difference is that the proof of the key inequality analogous to (2.18) requires the use of the following lemma.

Lemma 2.17. *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 the 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 $A_n > \delta$.*

□

Example 2 (continued). By Theorem 2.16 the test based on W_1 is consistent if and only if for true F and H it holds that

$$\frac{1}{\eta^2+1} E_{F*H}(Y-\theta_0) \neq 0\,, \quad \text{i.e.} \quad E_{F*H}(Y) \neq \theta_0\,. \tag{2.42}$$

For example, W_1 doesn't work when the true H is symmetric about 0 and the true $F \neq F_0$ has the mean equal to θ_0.

Example 3 (continued). By Theorem 2.16 W_1 is consistent if and only if for true F and H it holds that

$$E_{F*H}\left[\frac{(y-\theta)^2\,\eta_0}{(\eta_0^2+1)^2} - \frac{\eta_0}{\eta_0^2+1}\right] \neq 0\,, \quad \text{i.e.}$$

$$E_{F*H}\,(y-\theta)^2 \neq \eta_0^2+1, \quad \text{or equivalently} \quad Var_{F*H}\,Y \neq Var_{F*H_0}\,Y\,. \tag{2.43}$$

Note that condition (2.42) can be interpreted as "W_1 is consistent for testing the hypothesis about the mean in this model iff the expectation of Y under alternative is different from the expectation under the null hypothesis" and (2.43) as "W_1 is consistent for testing the hypothesis about the variance in this model iff the variance of Y under alternative is different from the variance under the null hypothesis". One cannot expect more from such a simple test as W_1. On contrary, the data-driven test statistic W_S provides a consistent testing procedure.

2.11 Appendix

Proof. (The statement about l_j^* from Example 2). Indeed,

$$\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}(l_j^*-l_{\theta_0}^*(Y_j))\right| = \frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}(Y_j-\theta_0)\Big(\frac{1}{\sigma^2(Y)}-\frac{1}{\widehat{\sigma}_n^2}\Big)\right|$$
$$= \sqrt{n}\left|\frac{1}{\sigma^2(Y)}-\frac{1}{\widehat{\sigma}_n^2}\right|\cdot\frac{1}{n}\left|\sum_{i=1}^{n}(Y_j-\theta_0)\right|.$$

But

$$\frac{1}{n}\left|\sum_{i=1}^{n}(Y_j-\theta_0)\right| = \left|\overline{Y}-\theta_0\right| = \left|\overline{Y}-E_Y\right| \to 0$$

in $G_{\theta_0,\eta_0}-$probability, therefore Definition 4 is satisfied if $\sqrt{n}\left|\frac{1}{\sigma^2(Y)}-\frac{1}{\widehat{\sigma}_n^2}\right|$ is bounded in $G_{\theta_0,\eta_0}-$probability, and this holds if $\widehat{\sigma}_n^2$ is a $\sqrt{n}-$consistent estimate of $\sigma^2(Y)$. Here $\overline{Y}$ denotes the sample mean $\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_j$. □

Proof. (The statement about l_j^* from Example 3).

$$\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}(l_j^*-l_{\eta_0}^*(Y_j))\right|$$

$$= \frac{1}{\sqrt{n}}\left|C_1(\eta_0)\right|\left|\sum_{i=1}^{n}\big((Y_j-\widehat{\theta}_n)^2-(Y_j-\theta_0)^2\big)\right|$$
$$= \frac{1}{\sqrt{n}}\left|C_1(\eta_0)\right|\left|\sum_{i=1}^{n}(\widehat{\theta}_n-\theta_0)(-2Y_j+\widehat{\theta}_n+\theta_0)\right|$$

$$
\begin{aligned}
&= \left|C_1(\eta_0)\right| \sqrt{n}\left|\widehat{\theta}_n - \theta_0\right| \; \frac{1}{n}\left|\sum_{i=1}^{n}(Y_j - \widehat{\theta}_n) + \sum_{i=1}^{n}(Y_j - \theta_0)\right| \\
&= \left|C_1(\eta_0)\right| \sqrt{n}\left|\widehat{\theta}_n - \theta_0\right| \; \left|(\overline{Y} - \widehat{\theta}_n) + (\overline{Y} - \theta_0)\right| \\
&\leq \left|C_1(\eta_0)\right| \sqrt{n}\left|\widehat{\theta}_n - \theta_0\right| \; \left(\left|\overline{Y} - \widehat{\theta}_n\right| + \left|\overline{Y} - \theta_0\right|\right) \to 0
\end{aligned}
$$

in $G_{\theta_0,\eta_0}-$probability since for $n \to \infty$ it holds that $\left|\overline{Y} - \widehat{\theta}_n\right| \to 0$ and $\left|\overline{Y} - \theta_0\right| \to 0$, both in $G_{\theta_0,\eta_0}-$probability, and $\sqrt{n}\left|\widehat{\theta}_n - \theta_0\right|$ is bounded in $G_{\theta_0,\eta_0}-$probability. $\square$

Chapter 3

General Theory

3.1 Introduction.

In the previous Chapter, we constructed data-driven score goodness-of-fit tests for the deconvolution problem. This shows that score tests can be built for statistical inverse problems as well.

In this Chapter, we attempt to generalize the theory of score tests. 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. Our theory offers, in particular, an analogous generalization of the theory of data-driven score tests. We introduce abstract concepts generalizing the concepts of Neyman's smooth test statistics, score tests and data-driven score tests.

The main goal of this thesis is to propose an unified theory to automatize the process of building NT-tests for different statistical problems, and to give an unified method for proving consistency of such tests. We propose a general method for constructing consistent data-driven tests for parametric, semi- and nonparametric problems. Usually, proofs of consistency for data-driven tests consisted of two parts:

1) establishing large deviation inequalities for the test statistic

2) deriving consistency of the test from these inequalities.

Our method gives the tool to pass through step 2 automatically. Additionally, in our theorems, we allow a lot of freedom in the choice of penalties, dimension growth rates and model regularity assumptions.

Examples in this Chapter tries to show that the method is applicable also to

dependent data and statistical inverse problems. Moreover, for any test constructed, we have an explicit rule to determine, for every particular alternative, whether the test will be consistent against it. This rule allows us to describe, in a closed form, the set of "bad" alternatives for every NT-, SNT- and GNT-test.

In Section 3.2, we describe the framework and introduce an abstract notion of SNT-statistic. In Section 3.3, we propose a general definition of a model selection rule. Section 3.4 is devoted to the definition of NT-statistics. This is the main concept of this Chapter. In Section 3.5, we study behaviour of NT-statistics for the case when the alternative hypothesis is true, while in Section 3.6 we investigate what happens under the null hypothesis. In the end of Section 3.6, a consistency theorem for NT-statistics is given. Section 3.7 is devoted to some direct applications of our method. In Section 3.8, a new notion concerning the use of quadratic forms in statistics is introduced. In Section 3.9, we introduce a notion of GNT-statistics. This notion generalizes the notion of score tests for composite hypotheses. We prove a general consistency theorem for GNT-statistics.

3.2 Notation and basic assumptions.

Let X_1, X_2, ... 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, ..., X_m have the 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, ..., X_m).

Here Θ can be *any* set, for example, a functional space; correspondingly, parameter θ can be infinite dimensional. 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 means that results of this Chapter are applicable, in principle, in statistics for stochastic processes. Additional assumptions on Y_i's will be imposed below, when it would be necessary.

Assume that we have some hypothesis H_0 to test. H_0 will be specified below, when necessary. The exact form of 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 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 3.9) it would be still possible to extract from Y_i's some information about X_i's and build a consistent test.

Definition 7. Consider the following (abstract) statistic of the form

$$T_k = \sum_{j=1}^{k} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_j(Y_i) \right\}^2 . \tag{3.1}$$

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

Here $l_1, \ldots, l_k$ can be some score functions, but they can also be 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 (3.1). We will discuss different possible sets of meaningful additional assumptions on $l_1, \ldots, l_k$ below (see Sections 3.5 - 3.9).

Scores (and efficient scores) are based on the notion of maximum likelihood. Our constructions below will make it possible to use, for example, truncated, penalized or partial likelihood to build a test. In this sense our theory generalizes 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 totally unrelated to any kind of a likelihood.

Example 1. Basic example of SNT-statistic is the following (see, e.g., [30] or [29]). It is known as Neyman's smooth test statistic for simple hypotheses. 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 smooth test statistic is an SNT-statistics.

Example 2. Partial likelihood. Cox in [9] 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 random variable Y having density $f_Y(y;\theta)$. Let Y be transformed into the sequence

$$(X_1, S_1, X_2, S_2, \ldots, X_m, S_m), \tag{3.2}$$

where the components may themselves be vectors. The full likelihood of the sequence (3.2) is

$$\prod_{j=1}^{m} f_{X_j|X^{(j-1)},S^{(j-1)}}(x_j|x^{(j-1)}, s^{(j-1)};\theta) \prod_{j=1}^{m} f_{S_j|X^{(j)},S^{(j-1)}}(s_j|x^{(j)}, s^{(j-1)};\theta), \tag{3.3}$$

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. Cox in [9] gives some specific examples.

Assume now for simplicity of notations 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

$$t_j = \left. \frac{\partial \log f_{S_j|X^{(j)},S^{(j-1)}}(s_j|x^{(j)}, s^{(j-1)};\theta)}{\partial\theta} \right|_{\theta=\theta_0}, \tag{3.4}$$

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

$$PL_m = \sum_{j=1}^{m} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_j \right\}^2 . \tag{3.5}$$

□

Consistency theorems for SNT-statistics will follow from consistency theorems for more general NT-statistics (they are introduced in Section 3.4). See Theorem 3.10.

Remark 3.1. There is a direct method that makes it possible to find the asymptotic distribution of SNT-statistic, both under the null hypothesis and under alternatives. The idea of the method is as follows. First, 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 [14], 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.

3.3 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 [4], [12], [23], [13]), it is desirable to provide a corresponding refinement of our construction. Using the ideas from [22], we propose a general mathematical framework for constructing a rule to find a reasonable model dimension. We make our tests data-driven, i.e., tests are capable to choose a reasonable number of components in test statistics automatically by the data. Our construction offers a lot of freedom in the choice of penalties and building blocks for statistics. A statistician could take into account specific features of his particular problem and choose among all 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 possible number of components in test statistics by some fixed number, but instead we allow this number 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 is attainable by allowing the complexity of the model and the number of components in test statistics to grow with n at a controlled rate.

Denote by M_k a statistical model designed for a specific statistical problem satisfying assumptions of Section 3.2. 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 $k = 1, 2, \ldots$ is nested we if for their parameter sets it holds that $\Theta_1 \subseteq \Theta_2 \subseteq \ldots$. 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 some statistical meaning: see the first remark on the page 221 of [6]).

Let T_k be an *arbitrary* statistic for testing validity of the model M_k on the basis of observations $Y_1, \ldots, Y_n$. The following definition applies for the sequence of statistics $\{T_k\}$.

Definition 8. Consider a nested family of models M_k for $k = 1, \ldots, d(n)$, 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 $\pi(1, n) < \pi(2, n) < \ldots < \pi(d(n), n)$ for all n and $\pi(j, n) - \pi(1, n) \to \infty$ as $n \to \infty$ for every $j = 2, \ldots, d(n)$. Call $\pi(j, n)$ a *penalty attributed to jth model M_j and sample size n.* Then a *selection rule* S for the sequence of statistics $\{T_k\}$ is an integer-valued random variable satisfying the condition

$$S = \min\big\{k : 1 \le k \le d(n);\, T_k - \pi(k, n) \ge T_j - \pi(j, n),\, j = 1, \ldots, d(n)\big\}. \quad (3.6)$$

We call T_S a *data-driven test statistic* for testing validity of the initial model.

Possible choices of penalties are, for example, Schwarz's penalty $\pi(j, n) = j \log n$, or Akaike's penalty $\pi(j, n) = j$. The definition is statistically meaningful, of course, only if the sequence $\{T_k\}$ is increasing in the sense that $T_1(Y_1, \ldots, Y_n) \le T_2(Y_1, \ldots, Y_n) \le \ldots$.

Example 2 (continued). We have an interesting possibility concerning statistics PL_m. This statistic depends on the number m of components in the sequence (3.2). Suppose now that Y can be transformed into sequences (X_1, S_1), or (X_1, S_1, X_2, S_2), or even $(X_1, S_1, X_2, S_2, \ldots, X_m, S_m)$ 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 8 gives

some solution of this problem. The adaptive statistic PL_S is capable to choose the reasonable number of components in partial likelihood automatically by the data. □

Example 3 (Gaussian model selection). Birge and Massart in [6] 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 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 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 I use has close relations with the one of [6].

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 [6] satisfy the conditions of Definition 8 (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-statistics. As follows from the consistency Theorem 3.9 below, $\|\widehat{s}_{\widehat{m}}\|^2$ can be used for testing $s = 0$ and has a good range of consistency.

3.4 NT-statistics.

Now we introduce the main concept of this Chapter. Suppose that we are under the general setup of Section 3.2.

Definition 9. 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 7. Set

$$L = \{E_0[l(Y)]^T l(Y)\}^{-1}, \tag{3.7}$$

where the mathematical expectation E_0 is taken with respect to P_0, and P_0 is the distribution function of some (fixed and known in advance) 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

$$T_k = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l(Y_j) \right\} L \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l(Y_j) \right\}^T. \tag{3.8}$$

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

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 $l(Y)$; in classical score tests theory one had an analogous situation. However, our definitions allow us 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 in some sense a mathematical formalization of the idea how P_0 should be connected to Y_i's.

Example 2 (continued). It is possible to define by the formula (3.8) a version of the partial likelihood statistic PL_m for the case when θ is multidimensional or even infinite dimensional. In [9] 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 3. If for SNT-statistic T_k defined by (3.1) additionally $E_0\, l(Y) = 0$, then T_k is obviously NT-statistic also. 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 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 3.1). 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 thesis.

Example 4. Statistical inverse problems. The most well known example here is the deconvolution problem. It 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. In Chapter 2 we constructed data-driven score tests for the problem.

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 $k \le d(n)$ 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

$$l(y) = \frac{\frac{\partial}{\partial\theta}\left(\int_{\mathbb{R}} f_\theta(s)\, h(y-s)\, ds\right)\Big|_{\theta=0}}{\int_{\mathbb{R}} f_0(s)\, h(y-s)\, ds} \tag{3.9}$$

and define the corresponding test statistic U_k by the formula (3.8). Under regularity conditions from Chapter 2 all conditions of Definition 9 are satisfied and U_k is an NT-statistic.

Example 5. 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 un-

known. It is the aim to test the null hypothesis of independence

$$H_0: \quad D(x,y) = F(x)G(y), \quad x,y \in \mathbb{R}, \tag{3.10}$$

against a wide class of alternatives. The following construction was proposed in [25].

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 [25] is

$$T_k = \sum_{j=1}^{k} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_j\left(\frac{R_i - 1/2}{n}\right) b_j\left(\frac{S_i - 1/2}{n}\right) \right\}^2, \tag{3.11}$$

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

The selection rule proposed in [25] to choose the number of components k in T_k was

$$S = \min\left\{k:\ 1 \le k \le d(n);\ T_k - k\log n \ge T_j - j\log n,\ j = 1,2,\ldots,d(n)\right\}. \tag{3.12}$$

This selection rule satisfies Definition 8, and so the data-driven statistic T_S from [25] is a data-driven NT-statistics. □

Yet even a more general definition can be useful. The following notion is a complete generalization of the notion of SNT-statistics.

Definition 10. Suppose we have n random observations $Y_1, \ldots, Y_n$. Let k be 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 7. Let L be some (fixed and known in advance) symmetric

$k \times k$ matrix. Put

$$CT_k = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l(Y_j) \right\} L \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l(Y_j) \right\}^T . \tag{3.13}$$

We call CT_k the *complicated statistic of Neyman's type* (or CNT-statistics).

3.5 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 3.2. 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. What is important in our theory is that the new (possibly obtained by a complicated transformation) sequence $Y_1, Y_2, \ldots$ obeys the conditions of consistency theorems of this Chapter. Then it is possible to build consistent tests of some 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 the 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

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

$$n^{-1/2} \sum_{j=1}^{n} (l(Y_j) - E_P\, l(Y)) \to_d \mathcal{N}(0, L^{-1}), \tag{3.14}$$

where L is defined by (3.7) and $\mathcal{N}(0, L^{-1})$ denotes the $k-$dimensional normal dis-

tribution 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. Because of that in consistency theorems of this Chapter we are not using the full generality of Definition 9. But 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 9.

We assume additionally (without loss of generality) that

$$\lim_{n\to\infty} d(n) = \infty\,. \tag{3.15}$$

Remark 3.2. Assumption (3.15) is 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 Chapter, 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 9. To avoid possible confusion with the statement of the next theorem, we have to modify our notations a little bit. We remind that in Definition 9 L is a $k \times k-$matrix. Below we will sometimes need to denote it by L_k in order to stress the model dimension. Accordingly, ordered eigenvalues of L_k will be denoted by $\lambda_1^{(k)} \geq \lambda_2^{(k)} \geq \ldots \geq \lambda_k^{(k)}$. We have the sequence of matrices $\{L_k\}_{k=1}^{\infty}$ and each matrix has its own eigenvalues. That is why below we will use a more precise notation. When it will be possible, we will use the simplified notation from Definition 9.

Theorem 3.3. *Let* $\langle C \rangle$ *and (3.15) holds and*

$$\lim_{n\to\infty} \sup_{k\leq d(n)} \frac{\pi(k,n)}{n\lambda_k^{(k)}} = 0\,. \tag{3.16}$$

Then

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

Remark 3.4. Condition (3.16) means that not only n tends to infinity, but that it is also possible for k to grow infinitely, but at the controlled rate.

Proof. (Theorem 3.3). By the law of large numbers, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} l_K(Y_i) \;\to_P\; C_P \neq 0. \tag{3.17}$$

By Lemma 2.7

$$\begin{aligned} T_K &= \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\overrightarrow{l}(Y_i)\right\} L_k \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\overrightarrow{l}(Y_i)\right\}^T \\ &\geq \lambda_K^{(k)}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\overrightarrow{l}(Y_i)\right\|^2 \\ &\geq \lambda_K^{(k)}\cdot\frac{1}{n}\left(\sum_{i=1}^{n} l_K(Y_i)\right)^2. \end{aligned} \tag{3.18}$$

By (3.17)

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

and because K, C_K are constants determined by fixed P, condition (3.16) yields

$$T_K - \pi(K,n) \;\to_P\; \infty \quad \text{as} \quad n\to\infty\,. \tag{3.19}$$

On the other hand, by (3.14)

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

where $\mathcal{N}$ is a $(K-1)-$dimensional multivariate normal distribution with 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 (3.19)

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

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. □

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

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

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 (see Appendix).

Lemma 3.5. *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 condition $\langle A \rangle$, but even a much slower rate is not a problem. The main theorem of this section is

Theorem 3.6. *Let* $\langle A\rangle$, $\langle C\rangle$, *(3.15) and (3.16) holds and*

$$d(n) = o(r_n) \quad as \quad n \to \infty\,. \tag{3.20}$$

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

Proof. (Theorem 3.6). Let $x > 0$. Since $T_j > T_K$ if $j > K$ and (3.15) holds, we get by Theorem 3.3 that

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

Now by Lemma 3.21 and (3.20) we get

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

□

3.6 The null hypothesis.

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

Definition 11. Let $\{T_k\}$ be a sequence of NT-statistics and S be a selection rule for it. Suppose that $\lambda_1 \ge \lambda_2 \ge \dots$ are ordered eigenvalues of L, where L is defined by (3.7). 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)
$$\lim_{n\to\infty} \sup_{k\le u_n} \frac{s(k,n)}{n\lambda_k^{(k)}} = 0\,,$$
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 $k \ge 2$
$\lim_{k\to\infty} t(k,n) = \infty$ for every fixed n.

2. $s(k,n) \le \pi(k,n) - \pi(1,n) \le t(k,n)$ 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 $l = (l_1, \ldots, l_k)$ from Definition 9

$$\bar{l}_j := \frac{1}{n} \sum_{i=1}^{n} l_j(Y_i)\,, \tag{3.21}$$

$$\bar{l} := (\bar{l}_1, \bar{l}_2, \ldots, \bar{l}_k) \tag{3.22}$$

and, using notation L from Definition 9, a quadratic form

$$Q_k(\bar{l}) = (\bar{l}_1, \bar{l}_2, \ldots, \bar{l}_k)\, L\, (\bar{l}_1, \bar{l}_2, \ldots, \bar{l}_k)^T\,. \tag{3.23}$$

The first reason for the new notation is that $T_k = Q_k(\bar{l})$, where T_k is the statistic from Definition 9. It is more convenient to formulate and prove Theorem 3.7 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 3.8, where Q_k is the central object.

Below we use the notation of Definitions 9 and 11.

Definition 12. Let S be 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 11.

2. (B)
$$P_0\left(n\, Q_k(\bar{l}) \geq y\right) \;\leq\; \varphi(k; y)$$
for all $k \geq 1$ and $y \in [s(k,n); t(k,n)]$, where P_0 is as in Definition 11.

We call φ a *proper majorant* for (large deviations of) the statistic T_k. Equivalently, we say that (large deviations of) the statistic T_k 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 the test statistic of interest. NT-statistics are no exception from this. In order to prove consistency of an NT-test, one has to choose some specific inequality to use in the proof. In the consistency theorem part of the regularity assumptions on the model and the value of $d(n)$ are determined by this choice. 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 for the problem. Sometimes it can be more desirable to have a better rate for $d(n)$ by the cost of more restrictive regularity assumptions determined by the use of a strong probabilistic inequality, and sometimes it is better to use simple inequality that puts less restrictions on the applicability of the test but gives worse rate for $d(n)$. The meaning of Definitions 11 and 12 and Theorem 3.9 below is that one can be sure in advance that whatever inequality he choose, he will succeed in proving 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 3.9.

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, inequalities of Chebyshev, Prohorov, or Dvoretzky-Kiefer-Wolfowitz for dependent data, or other large deviations inequalities.

Theorem 3.7. *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*

$$d(n) \;\le\; \min\{u_n\,, m_n\}\,. \tag{3.24}$$

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

Proof. (Theorem 3.7). If $S \ge K$, then $T_k - T_1 \ge \pi(k,n) - \pi(1,n)$ for some $K \le k \le d(n)$ and so, equivalently,

$$\begin{aligned}&\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l(Y_i)\right\} L \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l(Y_i)\right\}^T \\ &\quad - \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l_1(Y_i)\right\}^2 \{E_0[l_1(Y)]^T l_1(Y)\}^{-1} \;\ge\; \pi(k,n) - \pi(1,n)\end{aligned} \tag{3.25}$$

for some $K \le k \le d(n)$, where $l = (l_1, l_2, \ldots, l_k)$. We can rewrite (3.25) in terms of the notation (3.21)-(3.23) as follows:

$$\begin{aligned}&(\sqrt{n}\;\bar{l}_1, \ldots, \sqrt{n}\;\bar{l}_k)\, L\, (\sqrt{n}\;\bar{l}_1, \ldots, \sqrt{n}\;\bar{l}_k)^T \\ &\quad = n\,(\bar{l}_1, \ldots, \bar{l}_k)\, L\, (\bar{l}_1, \ldots, \bar{l}_k)^T \;\ge\; \frac{n\,\bar{l}_1{}^2}{E_0\, l_1{}^2} + \big(\pi(k,n) - \pi(1,n)\big),\end{aligned} \tag{3.26}$$

for some $K \le k \le d(n)$. Denote $\Delta(k,n) := \pi(k,n) - \pi(1,n)$; then with the help of (3.23) we rewrite (3.26) as

$$n\, Q_k(\bar{l}) \;\ge\; \Delta(k,n) + \frac{n\,\bar{l}_1{}^2}{E_0\, l_1{}^2}\,, \tag{3.27}$$

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

$$\begin{aligned} P_0(S \ge K) \;&\le\; P_0\big((3.25) \text{ holds for some } K \le k \le d(n)\big) \\ &=\; P_0\big((3.27) \text{ holds for some } K \le k \le d(n)\big)\end{aligned}$$

$$\leq \; P_0\big(n\,Q_k(\bar{l}) \;\geq\; \Delta(k,n) \text{ for some } K \leq k \leq d(n)\big)\,.$$

But now by condition (B) we have

$$\begin{aligned} P_0(S \geq K) \;&\leq\; P_0\big(n\,Q_k(\bar{l}) \;\geq\; \Delta(k,n) \text{ for some } K \leq k \leq d(n)\big) \\ &\leq\; \sum_{k=K}^{d(n)} P_0\Big(n\,Q_k(\bar{l}) \;\geq\; \Delta(k,n)\Big) \\ &\leq\; \sum_{k=K}^{d(n)} \varphi\big(k;\Delta(k,n)\big)\,, \end{aligned} \tag{3.28}$$

if only $d(n) \leq \min\{u_n, m_n\}$ (see Definition 11). 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 $t > 0$

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

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

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

therefore

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

and

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

This completes the proof. □

Remark 3.8. In Definition 12 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 11 and still prove a modified version of Theorem 3.7. But usually it happens that if the penalty does not satisfy all the conditions of Definitions 11 and 12, then T_S has the same distribution under both alternative and null hypotheses and the test is inconsistent. Then, formally, the conclusions of Theorem 3.7 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 3.9. *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), (3.15) and (3.16) 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).*

Proof. (Theorem 3.9). Follows from Theorems 3.3, 3.6 and 3.7 and our definition of consistency. □

As the first application we have the following result.

Theorem 3.10. *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*

$\|(l_1(Y_1),\ldots,l_k(Y_1))\|_k \le M(k)$ *a.e., where* $\|\cdot\|_k$ *is the norm of the* $k-$*dimensional Euclidean space. Assume* $\pi(k,n)-\pi(1,n)\ge 2k$ *for all* $k\ge 2$ *and*

$$\lim_{n\to\infty}\frac{M(d(n))\,\pi(d(n),n)}{\sqrt{n}}=0. \tag{3.29}$$

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

Proof. (Theorem 3.10) The SNT-statistic T_S is an NT-statistic with $L_k=E_{k\times k}$ and $\lambda_1^{(k)}=\ldots=\lambda_k^{(k)}=1$. Therefore Theorem 3.7 is applicable. Put (in Theorem 3.7) $s(k,n)=\sqrt{2k}$, $t(k,n)=\sqrt{n}M(k)^{-1}$. 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 (3.29) guarantees that the Prohorov inequality is applicable and, moreover, that (B) holds with

$$\varphi(k;y)=\frac{150210}{\Gamma(k/2)}\left(\frac{y^2}{2}\right)^{\frac{k-1}{2}}\exp\left\{-\frac{y^2}{2}\left(1-\frac{M(k)\,y}{\sqrt{n}}\right)\right\}. \tag{3.30}$$

Since φ is exponentially decreasing in y under (3.29), 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 (3.29) holds.

□

3.7 Applications.

Example 1 (continued). As a simple corollary, we derive the following theorem that slightly generalizes Theorem 3.2 from [22].

Theorem 3.11. *Let* T_S *be the Neyman's smooth data-driven test statistic for the case of simple hypothesis of uniformity. Assume that* $\pi(k,n)-\pi(1,n)\ge 2k$ *for all* $k\ge 2$ *and that for all* $k\le d(n)$

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

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 3.10. □

Theorem 3.10 can be also be applied to other statistical problems.

Remark 3.12. In my point of view, the 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 $d(n) \equiv 10$. I think that an applied statistician should not try too much to increase $d(n)$ as much as possible for each n.

Example 5 (continued). In [25] the following consistency result was established.

Theorem 3.13. *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*

$$E_{\mathbb{P}}\, b_j(F(X))b_j(G(Y)) \neq 0 \tag{3.31}$$

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* $\mathbb{P}$*).*

Let us take a look at this result in view of the theory of NT-statistics. Consistency condition $\langle C\rangle$ requires that there exists $K = K_{\mathbb{P}}$ such that $E_{\mathbb{P}}\, l_K \neq 0$, i.e.

$$E_{\mathbb{P}}\, b_j\left(\frac{R_i - 1/2}{n}\right) b_j\left(\frac{S_i - 1/2}{n}\right) \neq 0. \tag{3.32}$$

For continuous F and G (3.32) is asymptotically equivalent to (3.31) 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 3.6 is applicable to get a result similar to Theorem 3.13. We do not go into technical details here. □

3.8 Quadratic forms of P-type.

Now we introduce another abstract notion concerning quadratic forms.

Definition 13. Let $Z_1, Z_2, \ldots, Z_n$ 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 $y \in [s(k,n); t(k,n)]$

$$P_F\left(n\,Q\left(\frac{Z_1 + Z_2 + \ldots + Z_n}{n} - E_F Z_1\right) \geq y\right) \leq \varphi(k;y)\,, \tag{3.33}$$

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

$$\varphi(k;y) \;=\; C_1\,\varphi_1(k)\,\varphi_2(\lambda_1, \lambda_2, \ldots, \lambda_k)\,y^{k-1}\,\exp\left\{-\frac{y^2}{C_2}\right\}, \tag{3.34}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ 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 sometimes shortly say that $Q(x)$ is of $P-$type for Z_i's.

If $Z_1, \ldots, Z_n$ 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 13 is meant to incorporate all the cases when Prohorov's inequality or some of its variations holds. Thus, Definition 13 is just some specification of the general condition (B) from Theorem 3.7. It is useful in the sense that it shows which kind of majorating functions φ could (and typically would) occur in condition (B).

As an example we state the following theorem that is a direct consequence of Theorem 3.9 .

Theorem 3.14. *Suppose that for T_S condition $\langle A \rangle$ holds, L is of P-type for the distribution function of the vector $\{(l_1(Y_1), \ldots, l_k(Y_1))\}_{i=1}^{n}$ 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).*

In general, there is no simple sufficient conditions for L to be of P-type. 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 a much 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 [2] or [14]. 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.

On the side note, most of the conditions for the existence of such approximation of $Q(\bar{l}(Y))$ are rather technical and very specific on the structure of L (e.g., imposing sometimes assumptions on the 5 largest eigenvalues of L). See series of papers by Gotze, Bentkus, Tikhomirov and references therein.

3.9 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 9. Therefore the concept of NT-statistics needs to be modified to be applicable in case of composite hypotheses. The following definition can be helpful.

Definition 14. 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

$$L^{(0)} = \{E_0[l(Y)]^T l(Y)\}^{-1} . \tag{3.35}$$

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 matrix with finite elements such that for the sequence $\{L_k\}$ it holds that

$$\left\| L_k - L^{(0)} \right\| = o_{P_0}(1) . \tag{3.36}$$

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 $\varepsilon > 0$

$$P_0^n \left(\frac{1}{\sqrt{n}} \left\| \sum_{i=1}^{n} (l_j^* - l(Y_j)) \right\| \geq \varepsilon \right) \to 0 \quad \text{as} \quad n \to \infty , \tag{3.37}$$

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

$$GT_k = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l_j^* \right\} L_k \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l_j^* \right\}^T . \tag{3.38}$$

We call GT_k the *generalized statistic of Neyman's type* (or GNT-statistic). Let selection rule S satisfy Definition 9. We call GT_S the data-driven GNT-statistic.

Remark 3.15. Now it is not obligatory to know functions $l_1, \ldots, l_k$ explicitly (in Definition 9 we assumed that we knew those functions). It is only important that we should be able to choose reasonably good L and l_j^*'s. In classical theory of efficient testing we had efficient score test statistics with estimated scores. Definition 14 generalizes this idea.

Remark 3.16. Establishing (3.37) in parametric problems is usually not difficult and can be done if a $\sqrt{n}-$consistent estimate of the nuisance parameter is available (see examples in Chapter 2). In semiparametric models finding estimators of score function satisfying (3.37) is more difficult and not always possible, but there exist effective methods for constructing such estimates. Often sample splitting technic is helpful. See, e.g., [35], [36], [27] for general results related to the topic. Some authors (e.g., [19]) constructed efficient score tests with estimated scores for some semiparametric problems. See Example 9 below.

Example 7. If $Y_1, \ldots, Y_n$ are equally distributed and T_k is an NT-statistic, then T_k is also a GNT-statistic. Indeed, put in Definition 14 $L := L^{(0)}$ and $l_j^*(Y_1, \ldots, Y_n) := l_j(Y_1)$.

Example 8. Let $X_1, \ldots X_n$ 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 [17] the data-driven score test for testing H_0 was constructed using score test for composite hypotheses from [10]. Here we briefly describe the construction from [17]. Let F be the distribution function corresponding to f and set

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

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

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

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

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

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

$$W_k(\widehat{\beta}) = n\, Y_n^T(\widehat{\beta}) \{ I + R(\widehat{\beta}) \}\, Y_n(\widehat{\beta}). \tag{3.39}$$

As follows from the results of [10], Section 9.3, pp.323-324, in a regular enough situation $W_k(\widehat{\beta})$ satisfies Definition 14 and is a GNT-statistic. Practically useful sets of such regularity assumptions are given in [17].

Example 9. Consider the problem described in Example 4, 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 (see Chapter 2):

$$\dot{l}_{\theta_0,\eta_0}(y) = \big(\dot{l}_{\theta_0}(y),\, \dot{l}_{\eta_0}(y) \big)\,, \tag{3.40}$$

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.

$$\dot{l}_{\theta_0}(y) = \frac{\frac{\partial}{\partial\theta} \Big(\int_{\mathbb{R}} f_\theta(s)\, h_{\eta_0}(y-s)\, ds \Big)\Big|_{\theta=\theta_0}}{\int_{\mathbb{R}} f_{\theta_0}(s)\, h_{\eta_0}(y-s)\, ds}\, 1_{[y:\, g(y;(\theta_0,\eta_0))>0]}\,, \tag{3.41}$$

$$\dot{l}_{\eta_0}(y) = \frac{\frac{\partial}{\partial\eta} \Big(\int_{\mathbb{R}} f_{\theta_0}(s)\, h_{\eta}(y-s)\, ds \Big)\Big|_{\eta=\eta_0}}{\int_{\mathbb{R}} f_{\theta_0}(s)\, h_{\eta_0}(y-s)\, ds}\, 1_{[y:\, g(y;(\theta_0,\eta_0))>0]}\,. \tag{3.42}$$

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

$$I(\theta,\eta) \;=\; \int_{\mathbb{R}} \dot{l}^{T}_{\theta,\eta}(y)\, \dot{l}_{\theta,\eta}(y)\; dG_{\theta,\eta}(y)\,, \tag{3.43}$$

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:

$$I(\theta_0,\eta_0) \;=\; \begin{pmatrix} I_{11}(\theta_0,\eta_0) & I_{12}(\theta_0,\eta_0) \\ I_{21}(\theta_0,\eta_0) & I_{22}(\theta_0,\eta_0) \end{pmatrix}, \tag{3.44}$$

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 (see Chapter 2):

$$l^{*}_{\theta_0}(y) \;=\; \dot{l}_{\theta_0}(y) \;-\; I_{12}(\theta_0,\eta_0)\, I^{-1}_{22}(\theta_0,\eta_0)\, \dot{l}_{\eta_0}(y)\,, \tag{3.45}$$

and the efficient Fisher information matrix for θ is

$$I^{*}_{\theta_0} \;=\; E_{\theta_0,\eta_0} l^{*T}_{\theta_0}\, l^{*}_{\theta_0} \;=\; \int_{\mathbb{R}} l^{*}_{\theta_0}(y)^{T}\, l^{*}_{\theta_0}(y)\, dG_{\theta_0,\eta_0}(y)\,. \tag{3.46}$$

Then the efficient score test statistics for the composite deconvolution problem from Chapter 2 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 plugged estimators satisfy (3.36) and (3.37).

Example 10. The following semiparametric example belongs to [19]. 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, $v(x) = (v_1(x), \ldots, v_q(x))$ 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 $u_1(x), u_2(x), \ldots$. Set

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

$$+\frac{1}{\tau}[y - v(x)\beta^T][m_1 - m_2V^{-1}M],$$

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

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

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 (3.36) and (3.37) are satisfied, therefore W_k becomes a GNT-statistic and its asymptotic properties can be studied by the method of this thesis. □

A general consistency theorem for GNT-statistics is required. Sometimes in statistical literature authors do not prove consistency of their tests. They just study the

null distribution and show simulated examples were the test performs well. One of the reasons for this is, probably, that without a general consistency theorem one has to perform a proof of consistency anew for every particular problem. This becomes difficult in such cases where sample splitting, tricky estimators and huge formulae are involved. Therefore, in my opinion, for most of the semi- and nonparametric problems general consistency theorems are the most convenient tool for proving consistency of NT- and GNT-tests. If one has a general consistency theorem analogous to Theorem 3.9 for NT-statistics, then at least some consistency result will follow automatically.

Now we prove consistency theorems for GNT-statistics. First, note that Definitions 11 and 12 are meaningful for a sequence of GNT-statistics $\{GT_k\}$, if only instead of L we use in Definition 11 and in (3.23) the matrix $L^{(0)}$ from Definition 14.

Theorem 3.17. *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 (for R_k) 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 $GT_S = O_{P_0}(1)$.*

Proof. (Theorem 3.17). Consider the auxiliary random variable

$$R_k = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l(Y_j) \right\} L^{(0)} \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l(Y_j) \right\}^T . \tag{3.47}$$

This is not a test statistic, but formally this random variable satisfies Definition 9. Therefore Theorem 3.7 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 14, we get the statement of the theorem by the Slutsky lemma. □

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

$$\langle \mathbf{C1} \rangle \qquad \left\| L - L^{(0)} \right\| = o_P(1) \tag{3.48}$$

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 $\varepsilon > 0$

$$P^n\left(\frac{1}{\sqrt{n}}\left\|\sum_{i=1}^{n}(l_j^* - l(Y_j))\right\| \geq \varepsilon\right) \to 0 \quad \text{as} \quad n \to \infty\,. \tag{3.49}$$

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

Theorem 3.18. *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 (for R_k). Assume that conditions $\langle A\rangle$, (3.15) and (3.16) 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 (3.49).*

Proof. (Theorem 3.18). Consider the random variable R_k defined in the proof of Theorem 3.17. Theorems 3.3, 3.6 and 3.7 are valid for the random variable R_S. Under assumptions of the theorem $GT_S \to R_S$ in $P-$probability, and we get the statement of the theorem by the Slutsky lemma. □

Remark 3.19. Some relaxation of assumptions (3.48) and (3.49) 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 is defined by (3.47). This is 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 3.20. Sometimes in the literature on nonparametric testing 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 can be measured by this rate. See [20], [38], e.g.. We do not consider rates at this stage of the development of our theory, but it is possible to consider local alternatives in this general setup as well. This remains to be investigated.

3.10 Appendix

Proof. (Lemma 3.5) 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

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

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

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

Lemma 3.21.

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

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

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

$$
=\; P\bigg(-x_n - C_K \,\leq\, \frac{1}{n}\sum_{i=1}^{n}\big(l_K(Y_i) - E_P l_K(Y_i)\big) \,\leq x_n - C_K\,\bigg)\,.
$$

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

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

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

(for all $n \geq$ some n_K)

$$\leq \quad P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\big(l_K(Y_i)-E_Pl_K(Y_i)\big)\Big|\ \geq\ \frac{C_K}{2}\Big)\ =\ O\Big(\frac{1}{r_n}\Big)$$

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

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

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

and then we proceed analogously to the previous case. □

In the proof of Theorem 3.10 we use the following theorem from [33].

Theorem 3.22. *Let $Z_1, \ldots, Z_n$ 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 $\|Z_1\|_k \leq L$ a.e. Then, for $2k \leq y^2 \leq nL^{-2}$, we have*

$$Pr\Big(\|n^{-1/2}\sum_{i=1}^{n}Z_i\|_k \geq y\Big) \leq \frac{150210}{\Gamma(k/2)}\Big(\frac{y^2}{2}\Big)^{\frac{k-1}{2}}\exp\Big\{-\frac{y^2}{2}\Big(1-\eta_n\Big)\Big\},$$

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

Chapter 4

Appendix I. Score tests

In this Appendix we list some basic definitions and theorems related to efficient estimation and score tests. We start with basic definitions related to regular parametric models and scores. We mainly follow here the classical book [3]. The reader interested in more general treatment of the topic and stronger results should consult another classical book [16].

Let μ be a fixed $\sigma-$finite measure on $(\mathbf{X}, \mathcal{B})$, and let M_μ be all probability measures on $(\mathbf{X}, \mathcal{B})$ dominated by μ. Suppose that we have have the *parametrization* map $\theta \to P_\theta$ and that $\theta \in \Theta \subseteq \mathbb{R}^k$, where Θ is the *parameter* set. Let $\mathbf{P} = \{P_\theta, \theta \in \Theta\}$.

We introduce the following two important parametrization maps p and s as follows:

$$p \equiv \frac{dP}{d\mu}, \quad s \equiv \sqrt{p}, \quad p(\theta) \to s(\theta)\,.$$

The map p serves as an embedding of $\mathbf{P}$ into $L_1(\mu)$ and the map s is an embedding of $\mathbf{P}$ into $L_2(\mu)$.

Definition 15. θ_0 is a *regular point* of the parametrization $\theta \to P_\theta$ if θ_0 is an interior point of Θ, and

(1) The map $\theta \to s(\theta)$ from Θ to $L_2(\mu)$ is Frechet differentiable at θ_0

(2) The $k \times k$ matrix $\int \dot{s}(\theta_0)\dot{s}^T(\theta_0)\, d\mu$ is nonsingular.

Definition 16. A parametrization $\theta \to P_\theta$ is *regular* if

(1) Every point of Θ is regular

(2) The map $s \to \dot{s}_i(\theta)$ is continuous from Θ to $L_2(\mu)$ for $i = 1, \ldots, k$.

We call **P** a *regular parametric model* if it has a regular parametrization.

Note that (1) of the last definition implies that Θ is open.

Definition 17. Define the *score function* $\dot{l}$ of an observation by

$$\dot{l}(\theta) \;=\; 2\,\frac{\dot{s}(\theta)}{s(\theta)}\,1_{[s(\theta)>0]} \;=\; \frac{\dot{p}(\theta)}{p(\theta)}\,1_{[p(\theta)>0]}\,. \tag{4.1}$$

Define the *Fisher information matrix* of θ by

$$I(\theta) = 4\int \dot{s}(\theta)\dot{s}^T(\theta)\,d\mu \;=\; \int \dot{l}(\theta)\dot{l}^T(\theta)\,dP_\theta\,. \tag{4.2}$$

Proposition 4.1. *Suppose Θ is open and for all θ :*

(1) $p(x,\theta)$ is continuously differentiable in θ for $\mu-$almost all x with gradient $\dot{p}(\theta)$

(2) $|\dot{l}(\theta)| \in L_2(P_\theta)$

(3) $I(\theta)$ is nonsingular and continuous in θ.

Then, if we define

$$\dot{s}(\theta) = \frac{1}{2}p^{-1/2}(\theta)\dot{p}(\theta)\,1_{[p(\theta)>0]}\,, \tag{4.3}$$

the parametrization $\theta \to P_\theta$ is regular with $\dot{s}(\theta)$ from (5.32) as Frechet derivative of $s(\theta)$.

Regularity of θ is enough to guarantee a score function identity which is basic to the Cramer - Rao information bound calculation:

$$\int \dot{l}(\theta)\,dP_\theta \;=\; 0\,. \tag{4.4}$$

Definition 18. The log-likelihood of $(X_1, \ldots, X_n)$ is defined by

$$L_n(\theta) \;=\; \sum_{i=1}^{n} l(X_i, \theta) \tag{4.5}$$

and the score function of $(X_1, \ldots, X_n)$ by

$$S_n(\theta) \,=\, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}(X_i, \theta)\,. \tag{4.6}$$

Here one has some inconsistency in standard terminology, since one has two different definitions of score function: one for single observation X_i and another for $(X_1, \ldots, X_n)$.

Proposition 4.2. *Suppose that* $\mathbf{P} = \{P_\theta, \theta \in \Theta\}$ *is a regular parametric model. Then uniformly in* $\theta \in \mathcal{K}$ *for compact* $\mathcal{K} \subset \Theta$ *it holds*

$$L_\theta(S_n(\theta)) \,\to\, \mathcal{N}(0, I(\theta))\,, \tag{4.7}$$

where $\mathcal{N}$ *is the multivariate normal distribution and the sign* L_θ *means convergence in law for the case when* θ *is the true parameter value.*

Now let $\nu : \mathbf{P} \to \mathbb{R}^m$ be a Euclidean parameter, where $\mathbf{P}$ is a general (not necessarily parametric) model.

Definition 19. T is an *asymptotically linear estimate* of ν if there exists

$$\Psi : \mathbf{X} \times \mathbf{P} \to \mathbb{R}^m$$

such that for all $P \in \mathbf{P}$

$$|\Psi(\cdot, P)| \in L_2(P), \tag{4.8}$$

$$\int \Psi(x, P)\, dP = 0, \tag{4.9}$$

$$T_n = \nu(P) + n^{-1} \sum_{i=1}^{n} \Psi(X_i, P) + o_P(n^{-1/2}). \tag{4.10}$$

We call $\Psi(\cdot, P)$ the *influence function* of T.

We can identify parameter ν with the parametric function $q : \Theta \to \mathbb{R}^m$ defined by

$$q(\theta) = \nu(P_\theta).$$

Definition 20. Fix $P = P_\theta$ and suppose q has a total differential matrix $\dot{q}_{m\times k}$ at θ. Define $I^{-1}(P|\nu, \mathbf{P}) = \dot{q}(\theta)I^{-1}(\theta)\dot{q}^T(\theta)$ to be the *information bound* for ν and $\widetilde{l}(\cdot, P|\nu, \mathbf{P}) = \dot{q}(\theta)I^{-1}(\theta)\dot{l}(\theta)$ to be the *efficient influence function* for ν.

The notation of the previous definition is confusing, but it still remains to find a better one.

Information inequality. If T is uniformly Gaussian regular, then

$$\Sigma(P_\theta, T) \geq I^{-1}(P_\theta|\nu, \mathbf{P}) \tag{4.11}$$

in the order on nonnegative definite matrices. Equality holds if and only if T is uniformly efficient. □

Asymptotic optimality theorem. If T is uniformly regular and l is a bowl-shaped loss function, then

$$\liminf_{n\to\infty} E_\theta l(\sqrt{n}(T_n - q(\theta))) \geq El(Z_\theta), \tag{4.12}$$

where $Z_\theta \sim \mathcal{N}(0, I^{-1}(P_\theta|\nu, \mathbf{P}))$.

This theorem has a broad range of applications. It covers such important special cases as quadratic loss function and zero-one loss function, for example.

The next proposition (from [3], p.39) is important for proving efficiency of estimators.

Proposition 4.3. *Suppose that T_n is an asymptotically linear estimator at θ_0 of $\nu(P_\theta) = q(\theta)$ with influence function ψ where $q : \Theta \to \mathbb{R}^m$. Then*

A. T_n is (Gaussian) regular at θ_0 iff *$q(\theta)$ is differentiable at θ_0 with derivative $\dot{q}(\theta_0)$ and, with $\widetilde{l} \equiv \widetilde{l}(\cdot, P_{\theta_0}|\nu, \mathbf{P})$,*

$$(*) \qquad \psi - \widetilde{l} \quad \perp \quad \dot{\mathbf{P}} = [\dot{l}_1, \dot{l}_2],$$

where $(*)$ is equivalent to

$$E_0\psi\dot{l}^T = \dot{q}(\theta_0).$$

B. *If T_n is regular, then $\psi \in \dot{\mathbf{P}}^m$* iff

$$\psi = \widetilde{l} = \dot{q}(\theta_0) I^{-1}(\theta_0) \dot{l}(\theta_0).$$

Chapter 5

Appendix II. Neyman's smooth tests

This Appendix contains some basic definitions related to the data-driven Neyman's smooth tests of fit. We mainly cite below [18].

When testing H_0 against H_A w.l.o.g. attention can be restricted to i.i.d. $X_1, \ldots, X_n$ with values in $[0,1]$, with the null hypothesis being that the X_i are uniform on $[0,1]$. Suppose an alternative is from the standard exponential family, i.e. has the form

$$g_k(x;\theta) = \exp\left\{ \sum_{j=1}^{k} \theta_j \phi_j(x) - \psi_k(\theta) \right\}, \tag{5.1}$$

where ϕ_j's are orthonormal in $L_2[0,1]$ with $\phi_0 \equiv 1$ and $\psi_k(\theta)$ is a normalizing constant. Then Neyman's smooth test statistic with k components is then given by

$$N_k = \sum_{j=1}^{k} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_j(X_i) \right\}^2. \tag{5.2}$$

Note that this Neyman's statistic is a special case of the score statistic defined in Appendix I.

Inglot, Kallenberg and Ledwina in used in their papers selection rule S in order to automatically select "good" model dimension k. The correspondingly modified test N_S is called Neman's smooth data-driven goodness-of-fit test. In this Thesis we

sometimes use the abbreviation "DDN-test".

Consider now another testing problem when in the setup above the alternative hypothesis H_A is just a simple hypothesis, i.e. we are testing hypothesis $P = P_0$ against the only possible alternative $P = P_A$, where P_0 and P_A are some distribution functions. Suppose these distributions have densities, and denote them by p_0 and p_A correspondingly.

Definition 21. The statistic

$$NP_n = \left\{ nVar_{P_0} \log p_A(X) \right\}^{-1/2} \left\{ \sum_{j=1}^{k} \left[\log p_A(X_i) - E_{P_0} \log p_A(X) \right] \right\} \tag{5.3}$$

is a standardized version of the logarithm of the Neyman-Pearson test statistic for P_0 against P_A.

Chapter 6

Appendix III. Basic definitions related to Asymptotic Efficiency

6.1 Historical remarks.

Here we say only a couple of words concerning the history of the notion of efficiency. For more details see the classical book [32] by Nikitin.

It was pointed out by Kendall and Stuart [26] that the notion of asymptotic efficiency of tests is more complicated than the asymptotic efficiency of estimates. Various approaches to this notion were identified only in the late forties and early fifties, i.e. 20 years later than for the estimation theory.

6.2 Basic classical definitions.

The definitions of this section are formulated following the way Nikitin does in his book. However, I omitted some definitions from Nikitin's book, and so this section is not self-contained and can be used only as a brief reminder. I collected only basic classical definitions of AREs in this section.

Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables having the distribution P_θ from some parametric family determined by parameter θ taking on values in a parametric set Θ. Here Θ is not necessarily finite dimensional space, but it can also be a functional space or any other set. The situation is therefore essentially nonparametric. We assume in addition that Θ is a topological space.

Consider the problem of testing the hypothesis

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

against the alternative

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

on the basis of observations $X_1, X_2, \ldots, X_n$.

Suppose we use a sequence of statistics $\{T_n\}$, where $T_n = T_n(X_1, X_2, \ldots, X_n)$, and assume large values of T_n to be significant.

Define for any $\beta \in (0,1)$ and $\theta \in \Theta_1$ a real sequence $c_n := c_n(\beta, \theta)$ such that the following inequality holds

$$P_\theta(T_n > c_n) \leq \beta \leq P_\theta(T_n \geq c_n). \tag{6.1}$$

Then

$$\alpha_n(\beta, \theta) := \sup\{P_{\theta'}(T_n \geq c_n) : \theta' \in \Theta_0\}$$

is the minimal *size* of the test based on $\{T_n\}$ for which the power at the point θ is not less than β. Now we can define for any level of significance α, $0 < \alpha < \beta$, the positive integer

$$N_T(\alpha, \beta, \theta) := \min\{n : \ \alpha_n(\beta, \theta) \leq \alpha \text{ for all } m \geq n\}.$$

We see that $N_T(\alpha, \beta, \theta)$ is the minimal sample size necessary for the test at a level α, based on $\{T_n\}$, to have the power not less than β at the point θ.

Definition 22. Suppose that for testing H against A we have two sequences of test statistics $\{T_n\}$ and $\{V_n\}$. Define by $e_{V,T}(\alpha, \beta, \theta)$ the *relative efficiency* of the sequence $\{V_n\}$ with respect to $\{T_n\}$ in the following way:

$$e_{V,T}(\alpha, \beta, \theta) := N_T(\alpha, \beta, \theta)/N_V(\alpha, \beta, \theta). \tag{6.2}$$

A value $e_{V,T}(\alpha, \beta, \theta)$ larger than 1 means that for given α, β, θ the sequence $\{V_n\}$ is better to use than $\{T_n\}$ because the first sequence requires less observations for reaching the power β for the level α and the alternative value θ. Therefore relative efficiency is a meaningful statistical notion. However, $e_{V,T}(\alpha, \beta, \theta)$ depends on 3 parameters and two sequences of statistics, which makes it too difficult to calculate

relative efficiency in most cases. Lucky enough, in many cases different limiting values of $e_{V,T}(\alpha,\beta,\theta)$ still provide us useful statistical information. Therefore we introduce the following three fundamental definitions.

Definition 23. If for $\beta \in (0,1)$ and $\theta \in \Theta_1$ there exists the limit

$$e^{B}_{V,T}(\beta,\theta) := \lim_{\alpha \downarrow 0} e_{V,T}(\alpha,\beta,\theta), \tag{6.3}$$

it is called the *Bahadur ARE of the sequence* $\{V_n\}$ *with respect to* $\{T_n\}$.

Definition 24. If for $\alpha \in (0,1)$ and $\theta \in \Theta_1$ there exists the limit

$$e^{HL}_{V,T}(\beta,\theta) := \lim_{\beta \uparrow 1} e_{V,T}(\alpha,\beta,\theta), \tag{6.4}$$

it is called the *Hodges-Lehmann ARE of the sequence* $\{V_n\}$ *with respect to* $\{T_n\}$.

Definition 25. If for $0 < \alpha < \beta < 1$ and $\theta \to \theta_0 \in \partial\Theta_0$ (in a certain topology on Θ) there exists the limit

$$e^{P}_{V,T}(\beta,\theta) := \lim_{\theta \to \theta_0} e_{V,T}(\alpha,\beta,\theta), \tag{6.5}$$

it is called the *Pitman ARE of the sequence* $\{V_n\}$ *with respect to* $\{T_n\}$.

It is also difficult to calculate these three types of ARE, but it is still much easier than to deal with the definition 21. There also exist the intermediate approaches to measuring the ARE not coinciding with the above ones. For example, Chernoff ARE (see [8]), intermediate (or Kallenberg) ARE (see [21], and also Appendix IV). For other definitions see Rubin and Sethuraman [34], and Borovkov and Mogulskii [7]. It seems that it is much less proven about these intermediate AREs than about the fundamental notions.

Chapter 7

Appendix IV. Intermediate Efficiency and Optimality

7.1 Intermediate efficiency.

In this section we give for the ease of reference some basic definitions related to the intermediate (or Kallenberg) efficiency and formulate several important theorems about intermediate optimality of the DDN-test (see Appendix II).

Consider a probability space $(\mathbb{X}, \mathcal{B})$. Let $S = (X_1, X_2, \ldots)$ be a sequence of i.i.d. random variables having the distribution P_θ from some parametric family determined by parameter θ taking on values in a parametric set Θ.

Consider the problem of testing the hypothesis

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

against the alternative

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

on the basis of observations $X_1, X_2, \ldots, X_n$.

Definition 26. A family $\{\phi_{N;\alpha}; N \in \mathbb{N}, 0 < \alpha < 1\}$ is called a *family of (randomized) tests* of H_0, if for each $N \in \mathbb{N}$ and $0 < \alpha < 1$ the function $\phi_{N;\alpha}(s) = \phi_{N;\alpha}((x_1, x_2, \ldots))$ is a measurable function of $x_1, x_2, \ldots, x_N$ only, with values in

$[0,1]$, satisfying

$$\sup_{\theta_0\in\Theta_0} E_{\theta_0}\phi_{N;\alpha}(s) \le \alpha. \tag{7.1}$$

The next definition was introduced by Kallenberg (see [21]). Let α_n be a sequence of levels such that $\alpha_n > 0$ and for some $0 \le \tau < 1$

$$\lim_{n\to\infty} \alpha_n = \lim_{n\to\infty} n^{-\tau}\log\alpha_n = 0. \tag{7.2}$$

Let $\{\theta_n\}$ be a sequence of alternatives with

$$\lim_{n\to\infty} H(\theta_n,\Theta_0) = 0, \quad \lim_{n\to\infty} nH^2(\theta_n,\Theta_0) = \infty. \tag{7.3}$$

Here $H(\theta,\Theta_0) = \inf_{\theta_0\in\Theta_0} H(\theta,\theta_0)$ and $H(\theta,\theta_0)$ denotes the Hellinger distance between the probability measures P_θ and P_{θ_0}. Let $V_n^{(i)}$, $i = 1,2$ be some test statistics and let $\phi^{(i)}_{n;\alpha_n}$ be a sequence of test functions for H_0 against A, rejecting H_0 for large values of $V_n^{(i)}$ (see Definition 25). Assume additionally that

$$0 < \liminf_{n\to\infty} E_{\theta_n}\phi^{(2)}_{n;\alpha_n} \le \limsup_{n\to\infty} E_{\theta_n}\phi^{(2)}_{n;\alpha_n} < 1. \tag{7.4}$$

Define

$$N_{V^{(2)},V^{(1)}}(n,\theta_n) = \inf\{N : E_{\theta_n}\phi^{(1)}_{N+k;\alpha_n} \ge E_{\theta_n}\phi^{(2)}_{n;\alpha_n} \text{ for all } k \ge 0\} \tag{7.5}$$

and set

$$e_{V^{(2)},V^{(1)}} = \lim_{n\to\infty} \frac{N_{V^{(2)},V^{(1)}}(n,\theta_n)}{n}. \tag{7.6}$$

Definition 27. In the notations above, if the limit $e_{V^{(2)},V^{(1)}}$ exists and doesn't depend on $\{\theta_n\}$ and $\{\alpha_n\}$, we say that the *asymptotic τ intermediate efficiency* of $V^{(2)}$ with respect to $V^{(1)}$ equals $e_{V^{(2)},V^{(1)}}$.

7.2 Intermediate optimality.

The next definition applies only in the setup of Appendix II. We use the notation of that Appendix as well.

Definition 28. We shall say that a test statistic V_n is *asymptotically τ efficient* if its asymptotic τ intermediate efficiency with respect to the Neyman-Pearson test exists and is equal 1. If V_n is τ efficient for some τ, we shall say that V_n is *asymptotically (intermediate) optimal*.

It is usually said in statistical folklore that "the DDN-test is asymptotically optimal against essentially all alternatives" and that "intermediate efficiency fills the gap between Pitman's and Bahadur's efficiencies". Now we formulate the full theorem about intermediate optimality of the DDN-test. We cite this theorem from [18].

The following additional conditions on the model of Appendix II are imposed.

$$\int_0^1 \phi_j(x)\,dx = 0, \quad j = 1, 2, \ldots. \tag{7.7}$$

$$\sup_x |\phi_j(x)| < \infty. \tag{7.8}$$

$$\max_{1\le j\le k} \sup_x |\phi_j(x)| = O(k^\omega) \text{ for some } \omega \ge 0. \tag{7.9}$$

For some r such that $r > \omega + 3/2$, let $\{m_n\}$ be a sequence such that for $n \to \infty$

$$m_n \to \infty \quad \text{and} \quad m_n = O(n^{1/(2r+1)}). \tag{7.10}$$

Assume $\{P_n\}$ is a sequence of distributions on $[0,1]$ possessing densities p_n with respect to λ. By P_0 we denote throughout this section the uniform distribution on $[0,1]$.

Suppose that there exists M such that, for sufficiently large n,

$$e^{-M} \le p_n(x) \le e^M, \quad x \in [0,1].$$

Now set $\phi = (\phi_1, \ldots, \phi_{m_n})$ and let $\circ$ stand for the inner product in $\mathbb{R}^{m_n}$. We assume

that there exists $\theta = (\theta_1, \ldots, \theta_{m_n})$ such that

$$\gamma_{m_n} = \| \log p_n - \theta \circ \phi \|_\infty \text{ is bounded,} \tag{7.11}$$

$$\Delta_{m_n} = \| \log p_n - \theta \circ \phi \|_2 = O\left(\frac{1}{m_n^r}\right), \tag{7.12}$$

where $\| \cdot \|_\infty$ denotes the supremum norm and $\| \cdot \|_2$ is the $L_2[0,1]$ norm.

Additionally we suppose that, as $n \to \infty$,

$$H(p_n, p_0) \to 0 \quad \text{and} \quad m_n^r H(p_n, p_0) \to \infty, \tag{7.13}$$

where H denotes the Hellinger distance.

A set of $\{p_n\}'$s satisfying the three above assumptions shall be denoted by $\mathcal{P}_{m,r}$. For a given alternative p_n set

$$e_{0,n} = E_{P_0} \log p_n(X), \quad v_{0,n}^2 = \int_0^1 \log^2 p_n(x)\, dx.$$

Define

$$Y_{n,i} = v_{0,n}^{-1}\{\log p_n(X_i) - e_{0,n}\}.$$

Assume now that there exist positive constants B, C', C'' such that for all complex h, $|h| < B$ and for all n

$$C' \le |E_{P_0} \exp(hY_{n,1})| \le C''. \tag{7.14}$$

Define $\mathcal{P}_{m,r}^*$ to be the *subset* of $\mathcal{P}_{m,r}$ for which $\gamma_{m_n} \to 0$ as $n \to \infty$ and (7.14) is satisfied. We will be concerned with a smaller set of the form

$$\mathcal{D}(\mu, \nu) = \big\{ \{p_n\} \in \mathcal{P}_{m,r}^* : n^\mu H^2(p_n, p_0) \to 0 \text{ and } n^\nu H^2(p_n, p_0) \to \infty \big\}.$$

Now we are finally ready to finish the formulation of the theorem.

Theorem 7.1. *Assume that r is a fixed number satisfying the above conditions.*

Take $m_n = Cn^{1/\eta}$, *where* $\eta \geq 2r+1$ *and* C *is an arbitrary constant. Let* μ, ν *be any numbers satisfying* $(3+2\omega)/\eta < \mu < \nu < 2r/\eta$. *Then, for testing against all sequences of alternatives from* $\mathcal{D}(\mu,\nu)$ *and* $\tau = (\eta - 2\omega - 1)/\eta$, *the DDN-statistic is asymptotically optimal in the intermediate sense.*

Inglot, Kallenberg and Ledwina proved other theorems on the topic as well, but I think this one already gives an impression about this type of results. The set $\mathcal{P}^*_{m,r}$ consists of alternatives which obey a set of restrictions concerning growth of their Fourier coefficients. It can be argued if these restrictions are strong or not, but anyway there exist an infinite number of directions of alternatives which are not covered by this theorem. For example, one can construct an infinite-dimensional set of such "bad" alternatives even using only contamination alternatives of the form $p_n(x) = 1 + n^{-\xi}g(x)$, where g is bounded. It is enough to choose g's with Fourier coefficients growing faster than the optimality theorem allows. It is even easier to construct such an example using theorem 4.6 from [22], which is somewhat analogous to the above optimality theorem but deals specifically with contamination alternatives.

It will be an interesting task to give an estimate of how big is a subset of alternatives for which some intermediate optimality result holds, in comparison with all possible alternatives of interest. More specifically, if one takes only contamination alternatives mentioned above and the optimality theorem 4.6 from [22], how big is actually the set of "good" g's in comparison with the Sobolev space W_2^1 (even if we do not care about the speed parameter ξ)?

I think one shouldn't say "the DDN-test is asymptotically optimal against *essentially all* alternatives" and that "intermediate efficiency *fills the gap* between Pitman's and Bahadur's efficiencies" while the above question remains unanswered.

Bibliography

[1] O. Barndorff-Nielsen. *Information and exponential families in statistical theory*. John Wiley & Sons Ltd., Chichester, 1978. Wiley Series in Probability and Mathematical Statistics. 13

[2] V. Bentkus and F. Götze. Optimal rates of convergence in the CLT for quadratic forms. *Ann. Probab.*, 24(1):466–490, 1996. 57

[3] 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. 5, 11, 12, 21, 23, 26, 67, 70

[4] 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. 5, 9, 39

[5] P. J. Bickel and M. Rosenblatt. On some global measures of the deviations of density function estimates. *Ann. Statist.*, 1:1071–1095, 1973. 9

[6] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001. 40, 41

[7] A. A. Borovkov and A. A. Mogul′skiĭ. *Bolshie ukloneniya i proverka statisticheskikh gipotez*, volume 19 of *Trudy Instituta Matematiki [Proceedings of the Institute of Mathematics]*. "Nauka" Sibirsk. Otdel., Novosibirsk, 1992. 77

[8] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statistics*, 23:493–507, 1952. 77

[9] D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975. 38, 42

[10] D. R. Cox and D. V. Hinkley. *Theoretical statistics*. Chapman and Hall, London, 1974. 5, 59, 60

[11] 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. 9

[12] 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. 9, 39

[13] J. Fan. Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.*, 91(434):674–688, 1996. 9, 39

[14] F. Götze and A. N. Tikhomirov. Asymptotic distribution of quadratic forms. *Ann. Probab.*, 27(2):1072–1098, 1999. 39, 57

[15] H. Holzmann, N. Bissantz, and A. Munk. Density testing in a contaminated sample. *J. Multivariate Anal.*, 98(1):57–75, 2007. 9

[16] 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. 5, 26, 67

[17] T. Inglot, W. C. M. Kallenberg, and T. Ledwina. Data driven smooth tests for composite hypotheses. *Ann. Statist.*, 25(3):1222–1250, 1997. 29, 59, 60

[18] T. Inglot and T. Ledwina. Asymptotic optimality of data-driven Neyman's tests for uniformity. *Ann. Statist.*, 24(5):1982–2019, 1996. 5, 10, 73, 81

[19] 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. 14, 26, 59, 61

[20] 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. 64

[21] W. C. M. Kallenberg. Intermediate efficiency, theory and examples. *Ann. Statist.*, 11(1):170–182, 1983. 77, 80

[22] W. C. M. Kallenberg. The penalty in data driven Neyman's tests. *Math. Methods Statist.*, 11(3):323–340 (2003), 2002. 9, 14, 39, 55, 83

[23] 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. 5, 9, 16, 39

[24] 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. 5

[25] W. C. M. Kallenberg and T. Ledwina. Data-driven rank tests for independence. *J. Amer. Statist. Assoc.*, 94(445):285–301, 1999. 44, 56

[26] M. G. Kendall and A. Stuart. *The advanced theory of statistics. Vol. 2: Inference and relationship.* Second edition. Hafner Publishing Co., New York, 1967. 75

[27] C. A. J. Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.*, 15(4):1548–1562, 1987. 59

[28] 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

[29] T. Ledwina. Data-driven version of Neyman's smooth test of fit. *J. Amer. Statist. Assoc.*, 89(427):1000–1005, 1994. 5, 37

[30] J. Neyman. "smooth test" for goodness of fit. *Skand. Aktuarietidskr.*, 20:150–199, 1937. 5, 37

[31] 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. 5

[32] Y. Nikitin. *Asymptotic efficiency of nonparametric tests*. Cambridge University Press, Cambridge, 1995. 5, 75

[33] A. V. Prohorov. Sums of random vectors. *Teor. Verojatnost. i Primenen.*, 18:193–195, 1973. 66

[34] H. Rubin and J. Sethuraman. Bayes risk efficiency. *Sankhyā Ser. A*, 27:347–356, 1965. 77

[35] A. Schick. On asymptotically efficient estimation in semiparametric models. *Ann. Statist.*, 14(3):1139–1151, 1986. 26, 59

[36] A. Schick. A note on the construction of asymptotically linear estimators. *J. Statist. Plann. Inference*, 16(1):89–105, 1987. 59

[37] S. L. Sclove and J. Van Ryzin. Estimating the parameters of a convolution. *J. Roy. Statist. Soc. Ser. B*, 31:181–191, 1969. 30

[38] V. G. Spokoiny. Adaptive and spatially adaptive testing of a nonparametric hypothesis. *Math. Methods Statist.*, 7(3):245–273, 1998. 64

Curriculum Vitae

Name:	Mikhail Langovoy
Geburtsdatum:	26.12.1980
Geburtsort:	Sankt Petersburg, Russland
1986 - 1994	Grundschule 2 in Kirovsk, Russland
1994 - 1996	Akademisches Gymnasium an der Staatlichen Universität Sankt-Petersburg, Russland
1996 - 2001	Mathematikstudium, Studienrichtung Reine Mathematik, Abteilung Algebra und Zahlentheorie, Didaktik-Zusatzausbildung für die Lehre an Hochschulen Staatliche Universität Sankt-Petersburg
1997 - 1998	Fachkurse in den Fächern Mathematik und Informatik am Mathematischen Institut der Russischen Akademie der Wissenschaften, Sankt-Petersburg
Juni 2001	Abschluss: Diplom mit Summa cum laude
2001 - 2002	Doktorand an der Staatlichen Universität Sankt-Petersburg Abteilung Algebra und Zahlentheorie
2003 - 2004	Mathematiker und Business-Analyst in der Industrie, Sankt-Petersburg
seit September 2004	Doktorand am Institut für Mathematische Stochastik der Universität Göttingen bei Herrn Prof. Dr. Axel Munk mit dem Ziel Promotion
seit September 2004	Teilnahme am interdisziplinären Promotionsstudiengang "Identifikation in mathematischen Modellen"