Tatra Mt. Math. Publ. **17** (1999), 11–25

TATRA MOUNTAINS
Mathematical Publications

NONLINEAR ESTIMATION PROBLEMS

HENNING LÄUTER

ABSTRACT. The maximum likelihood or least squares estimators are solutions of (in general) nonlinear equations and those are very often ill-posed. Hence we consider in the first part the regularization of ill-posed problems and find that in linear cases an optimal regularization is to be found. In nonlinear problems an optimal regularization is unknown. In the second part we consider a nonparametric goodness of fit and find that the maximum likelihood estimator has an asymptotic expansion which is used for the computation of the power of the test.

In the last twenty or thirty years, many results on nonlinear estimation problems were published. I refer, for example, to the books of D. M. Bates & D. G. Watts [2], G. A. F. Seber & C. J. Wild [14] or G. J. S. Ross [13]. Here one finds many results on the asymptotic behavior of maximum likelihood estimations (m.l.e.) or least squares estimations (l.s.e.), and computational methods are included there. A. Pázman [11] discussed properties of nonlinear estimations in 'flat' models, he approximated the distribution of the l.s.e. and characterized the estimates in nonlinear regression by differential geometry methods.

Our aim is to describe some statistical methods where nonlinear estimation problems arise. At first we consider from a general point of view inverse problems, because m.l.e. or l.s.e. are solutions of such equations. Very often these equations are ill-posed and hence a regularization is necessary. In the second part we consider a test on densities and formulate conditions under which the m.l.e. can be substituted in the test statistic.

1. Inverse problems and their regularization

We start with some remarks on deterministic inverse problems. The exposition in this section is intended to describe the regularization of ill-posed problems. In linear models an optimal linear regularization is determined, but for

2000 Mathematics Subject Classification: Primary 62F10, 62G10; Secondary 62G20.
Keywords: estimation of parameters, ill-posed problems, goodness of fit tests, asymptotic power.

nonlinear regularizations the power depends on the whole distribution of the errors and an optimal regularization is not known.

There are many analogies between the regularization of ill-posed problems and descriptive statistical methods. In both fields, methods of solution and description that remain insensitive to random distractions are sought. Of the statistical literature especially the papers of Mallows [10], Wahba [15], Läuter–Pincus [7] and Härdle [5] are to be mentioned. Here, linear methods of smoothing are examined for their optimality and goodness. Some of these results are extended in this paper.

Notations. For an operator A, we denote its domain by $\mathscr{D}(A)$, the range by $\mathscr{R}(A) = \{A(\varphi) \colon \varphi \in \mathcal{X}\}$ and the null space by $\mathscr{N}(A)$. Let $\overline{\mathscr{R}(A)}$ be the topological closure of $\mathscr{R}(A)$.

If A is a linear operator, we write A^* for the adjoint and A^+ for the Moore-Penrose inverse operator. $P_{\mathcal{L}}$ stands for the orthogonal projection onto the closed space $\mathcal{L}$. We always require $\mathcal{X}$ and $\mathcal{Y}$ to be separable Hilbert spaces with inner products $(\cdot,\cdot)$ and norms $\|\cdot\|$, respectively.

Let A be a mapping: $\mathcal{X} \to \mathcal{Y}$. We consider the model

$$A(\gamma) = g \tag{1}$$

for data $g \in \mathcal{Y}$, parameters $\gamma \in \mathcal{X}$. Here, finite dimensional spaces or functional spaces are important special cases. Hadamard called the model (1) well-posed if it has a unique solution γ_g for any $g \in \mathcal{Y}$ and if γ_g is continuos in g. All other models are called ill-posed. The discontinuity is a limit case of instability.

A model (1) is called instable if $\gamma_1, \gamma_2 \in \mathcal{X}$ exist with

$$\|A(\gamma_1) - A(\gamma_2)\| \le 1/C \|\gamma_1 - \gamma_2\|$$

for $1 \ll C$, e.g., $C \ge 10^4$. There are close connections to the definition of stable systems given by G. Ross. Several nonlinear models are instable in the given sense.

Example. Growth curves.

Let $\gamma = (\beta_1, \beta_2, \beta_3)$ and

$$a(t,\gamma) = \frac{\beta_1}{1 + \beta_2 e^{\beta_3 t}}, \qquad 0 \le t \le 2.5\,.$$

Setting $\gamma_1 = (7.68, 3.68, 1.095)$, $\gamma_2 = (231271, 142711.7, 0.941)$, it results for $0 \le t \le 2.5$:

$$\frac{|\gamma_1 - \gamma_2|}{|a(t,\gamma_1) - a(t,\gamma_2)|} \ge 5.4 \cdot 10^6\,.$$

In this case $C \ge 5.4 \cdot 10^6$, so that these growth curves must be considered numerically instable.

We recognize numerical instability if $\frac{\partial}{\partial\gamma}a(t,\gamma) \to 0$ for $\|\gamma\| \to \infty$. With $A^{(n)}(\gamma) = \big(a(t_1,\gamma),\ldots,a(t_n,\gamma)\big)^*$ and if $G^{(n)}(\gamma) := \frac{\partial}{\partial\gamma}A^{(n)}(\gamma)$ exists, the ill-conditionedness of $G^{(n)}(\gamma)^* G^{(n)}(\gamma)$ is the reason for the instability and hence for the ill-posedness in finite dimensional models. If in (1), g is an element from a functional space $\mathcal{Y}$, then the model is equivalent to an infinite dimensional equation. Here $\mathcal{X}$ can be a finite or infinite dimensional space. In infinite dimensional equations, the ill-posedness can have many reasons.

If one wants to transform an ill-posed relation (1) by approximation of the equation into a well-posed problem, then the choice of a map $F\colon \mathcal{Y} \to \mathscr{R}(A)$ and the solution of the equation

$$A(\gamma) = F(g)\,, \qquad \gamma \in \mathcal{X}\,,\ g \in \mathcal{Y}, \tag{2}$$

must be considered.

Definition 1.1. Let $F\colon \mathcal{Y} \to \mathscr{R}(A)$ be a continuous operator such that

$$A(\gamma) = F(g)\,, \qquad \gamma \in \mathcal{X}\,,\ g \in \mathcal{Y}, \tag{3}$$

is well-posed. Then (3) is called a *regularization by data correction.*

Such regularizations are often used according to suitable interpretations. So one thinks of smoothing or projection when changing from g to $F(g)$. For characterization of the admitted transformations, let $\mathcal{S}_R$ and $\mathcal{S}_E$ be subsets of $\mathcal{Y}$ such that for positive constants c_1, c_2, with $c_2 < c_1$, it holds

$$F(\mathcal{S}_R) \subseteq \mathcal{S}_R \quad \text{and} \quad \big\|F(g)\big\| \geq c_1\|g\| \quad \text{for all} \quad g \in \mathcal{S}_R\,, \tag{4}$$

$$\big\|F(g)\big\| \leq c_2\|g\| \quad \text{for all} \quad g \in \mathcal{S}_E\,. \tag{5}$$

Here $\mathcal{S}_R$ describes a subset of $\mathcal{Y}$ that is mapped into itself by F, or in other words, that is reproduced. The case $F(g) = g$ for all $g \in \mathcal{S}_R$ has a special meaning here. The elements of $\mathcal{S}_E$ are functions that are almost eliminated by F. For example, one can think of $\mathcal{S}_R$ as a set of low-frequency or monotone functions and of $\mathcal{S}_E$ as high-frequency signals.

The usual least squares method can be described by (3) with an implicitly given F. Namely, if we approximate given observations (t_i, g_i), $i = 1,\ldots,n$, by a function $x(t,\gamma)$ at the points $t_1,\ldots,\, t_n$, then the least squares method

$$\sum_{i=1}^{n}\big(x(t_i,\gamma) - g_i\big)^2 \to \min_{\gamma} \tag{6}$$

leads to (3) for some F. Sometimes one uses in (6) some approximated data, e.g., by smoothing the data or detecting outliers. Denoting the approximated data by $k_n(t)$, one computes a solution of

$$\sum_{i=1}^{n}\big(x(t_i,\gamma) - k_n(t_i)\big)^2 \to \min_{\gamma}\,. \tag{7}$$

In all these cases we transform the data

$$(g_1, \dots, g_n) \xrightarrow{S} (k_n(t_1), \dots, k_n(t_n)) ,$$

and have to solve the equation

$$A(\gamma) = S(g) .$$

1.1 Optimal regularization.

Two types of models for (1) are considered now, that are not completely disjoint from each other but include different features. We admit deviations in the right hand side, which we describe by a probability law. We assume that g is disturbed additively by the errors, so that instead of $A(\gamma) = g$, the equation

$$A(\gamma) = g + \varepsilon \tag{8}$$

is to be solved. The solutions $\gamma_{g+\varepsilon}$ then should lie as close as possible to γ_g.

Model I. In (1) let $g \in \mathcal{Y}$ and A be an operator mapping into the Hilbert space $\mathcal{Y}$. Let g be disturbed additively by the random variable ε, with $\mathbf{E}|\varepsilon|^2 < \infty$. Then we have

$$\mathbf{E}(\varepsilon) = 0 \quad \text{and} \quad \mathbf{Var}\, \varepsilon = \sigma^2 W$$

for a known positive definite linear operator W. The admitted deviations of the right hand side are described by ε.

Model II. Equation (1) can be written equivalently as

$$a(t_i, \gamma) = g_i , \qquad i = 1, 2, \dots , \tag{9}$$

for design points $t_1, t_2, \dots$. The given admitted deviations in the data are described by additive disturbances ε_i. We assume that the ε_i are random variables with $\mathbf{E}(\varepsilon_i) = 0$ and equation (8) now corresponds to

$$a(t_i, \gamma) = g_i + \varepsilon_i , \qquad i = 1, 2, \dots \tag{10}$$

We write the first n equations in (10) using the n-dimensional vector notation as

$$A^{(n)}(\gamma) = Y^{(n)} := g^{(n)} + \varepsilon^{(n)}.$$

The models I and II show many analogies. But while in model II, the ε_i can be independent and identically distributed, for example, and so have a covariance matrix $\sigma^2 I_n$, in model I, the covariance of ε always differs from $\sigma^2 I_n$, since we assumed the realizations of ε to lie in the Hilbert space $\mathcal{Y}$. It is essential in both models that the admitted deviations in the data, with respect to which the solutions should be stable, are described by ε and ε_i, respectively.

1.1.1 Optimal regularizations in model I.

Regularizations by data correction of equation (8) have the form

$$A(\gamma) = F(Y) \tag{11}$$

for a continuous operator F on $\{g+\varepsilon(\omega)\colon g \in \mathcal{Y},\, \omega \in \Omega\}$. Since Y is a random, the solution of (11) becomes a random variable denoted by $\hat{\gamma}_F$. The aim now is to determine an operator F such that $\hat{\gamma}_F$ comes as close as possible to γ_g in the mean. The mean squared deviation of $\hat{\gamma}_F$ from γ_g is the linear operator $\mathrm{MSE}\hat{\gamma}_F$ satisfying

$$\big(\gamma, (\mathrm{MSE}\hat{\gamma}_F)\tilde{\gamma}\big) = \mathbf{E}\{(\gamma, \hat{\gamma}_F - \gamma_g)(\tilde{\gamma}, \hat{\gamma}_F - \gamma_g)\} \tag{12}$$

for any $\gamma, \tilde{\gamma} \in \mathcal{X}$. Let Q be a nonnegative functional defined on the set of operators $\mathrm{MSE}\hat{\gamma}_F$, $F \in \mathcal{F}$, for a set of admitted regularizations $\mathcal{F}$.

Definition 1.2. a) $\tilde{F} \in \mathcal{F}$ is called an *optimal regularization* if

$$Q\big(\mathrm{MSE}\hat{\gamma}_{\tilde{F}}\big) = \inf_{F \in \mathcal{F}} Q\big(\mathrm{MSE}\hat{\gamma}_F\big)\,.$$

b) The *relative efficiency* of F with respect to $F_1 \in \mathcal{F}$ is

$$\mathbf{eff}(F, F_1) = \frac{Q\big(\mathrm{MSE}\hat{\gamma}_{F_1}\big)}{Q\big(\mathrm{MSE}\hat{\gamma}_F\big)}\,.$$

It will turn out that an optimal regularization is determined by the operator A, the distribution of Y, the operator F, and Q. In general, the whole distribution of the distraction ε is important here. In linear models — that is with a linear operator A and a linear regularization F — only the first two moments from the distribution of Y determine the efficiency of F.

1.1.2 Linear models and linear regularizations.

We consider equation (11) for linear bounded operators A and F, that is

$$A\gamma = FY, \tag{13}$$

and assume that F is chosen such that (13) is well-posed. Let $\mathbf{Var}\, Y = \sigma^2 W$ be the covariance operator of Y with the known Hilbert-Schmidt operator W and an unknown factor σ^2. Assuming $A^+ F W^{1/2}$ to be bounded it follows that $\hat{\gamma}_F = (A^*A)^{-1}A^*FY$ is a solution of (13). If $A^+ F A \gamma_g \in \mathcal{X}$ then

$$\begin{aligned}\mathbf{E}(\hat{\gamma}_F) &= (A^*A)^{-1}A^*FA\gamma_g\,,\\ \mathbf{Var}\,\hat{\gamma}_F &= \sigma^2(A^*A)^{-1}A^*FWF^*A(A^*A)^{-1},\end{aligned}$$

and thus

$$\mathrm{MSE}\,\hat{\gamma}_F = (A^*A)^{-1}A^*\left[\sigma^2 FWF^* + (F-I)A\gamma_g\gamma_g^*A^*(F^*-I)\right]A(A^*A)^{-1}. \quad (14)$$

This representation makes clear that an optimal regularization depends on the distribution of Y only via the first two moments.

Let B be a self-adjoint linear operator on $\mathcal{X}$, ρ_* a given constant and μ a measure on $\mathcal{X}$ with

$$\mu\{\gamma\colon \gamma^*B\gamma \leq \sigma^2\rho_*^2\} = 1 \qquad \text{and} \qquad \int\limits_{\gamma^*B\gamma\leq\sigma^2\rho_*^2} \gamma\gamma^*\mu(\mathrm{d}\gamma) = \sigma^2\rho^2 C \qquad (15)$$

for a fixed linear self-adjoint operator C.

With (15) we calculate for the averaged risk

$$\int\limits_{\gamma^*B\gamma\leq\sigma^2\rho_*^2} \mathrm{MSE}\,\hat{\gamma}_F\mu(\mathrm{d}\gamma_g) =$$
$$= (A^*A)^{-1}A^*\left[\sigma^2 FWF^* + \sigma^2\rho^2(F-I)ACA^*(F^*-I)\right]A(A^*A)^{-1}.$$

Defining $Q\big(\mathrm{MSE}\,\hat{\gamma}_F\big) = \int\limits_{\gamma_g^*B\gamma_g\leq\sigma^2\rho_*^2} \mathrm{MSE}\,\hat{\gamma}_F\mu(\mathrm{d}\gamma_g)$, one obtains

$$Q\big(\mathrm{MSE}\,\hat{\gamma}_F\big) = \sigma^2\left[(A^*A)^{-1}A^*\left[FWF^* + \rho^2(F-I)ACA^*(F^*-I)\right]A(A^*A)^{-1}\right]. \quad (16)$$

Let $\mathcal{S}_R$ and $\mathcal{S}_E$ be given linear spaces in $\mathcal{R}(W^{1/2})\oplus\mathcal{N}(W)$. We denote the set of all linear operators by F, with

$$A^+FW^{1/2} \text{ is bounded}, \qquad A^+Fg\in\mathcal{X} \quad \text{for } g\in\mathscr{R}(W^{1/2})\oplus\mathcal{N}(W),$$
$$Fg = g \quad \text{for all } g\in\mathcal{S}_R, \qquad \text{and} \qquad Fg = 0, \quad \text{for all } g\in\mathcal{S}_E$$

by $\mathcal{F}$. Further we write P_R and P_E for the orthogonal projections onto $\mathcal{S}_R$ and $\mathcal{S}_E$, respectively.

For describing an optimal regularization, we denote $D = W + \rho^2 ACA^*$, $\mathcal{L} = (\mathcal{S}_R\oplus\mathcal{S}_E)^\perp$ and $\mathcal{K} = \mathscr{R}\big(D^{\frac{1}{2}}P_{\mathcal{L}}\big)$.

The next theorem shows how an optimal regularization can be constructed from a given regularization $F\in\mathcal{F}$.

Theorem 1.1. *The operator*

$$\tilde{F} = F - P_{\overline{\mathscr{R}(A)}}\big(FD - \rho^2 ACA^*\big)D^{-\frac{1}{2}}P_{\mathcal{K}}D^{-\frac{1}{2}} \qquad (17)$$

for any $F\in\mathcal{F}$ is an optimal regularization within $\mathcal{F}$.

Proof. With $F \in \mathcal{F}$ we have $\tilde{F} \in \mathcal{F}$ for $\tilde{F}$ in (17), because $P_{\mathcal{K}} D^{-\frac{1}{2}} P_R = 0$, $P_{\mathcal{K}} D^{-\frac{1}{2}} P_E = 0$. We get for any H with $\tilde{F} + H \in \mathcal{F}$

$$\begin{aligned} Q(\mathrm{MSE}_{\tilde{F}+H}) = Q\big(\mathrm{MSE}_{\tilde{F}}\big) &+ \sigma^2 A^+ H D H^* A^{+*} \\ &+ \sigma^2 A^+ H \big[D\tilde{F}^* - \rho^2 ACA^*\big] A^{+*} \\ &+ \sigma^2 A^+ \big[\tilde{F} D - \rho^2 ACA^*\big] H^* A^{+*} . \end{aligned} \tag{18}$$

Using $D^{-\frac{1}{2}} P_{\mathcal{K}} D^{-\frac{1}{2}} D P_{\mathcal{L}} = P_{\mathcal{L}}$, we compute

$$P_{\overline{\mathscr{R}(A)}}\big(\tilde{F} D - \rho^2 ACA^*\big) P_{\mathcal{L}} = 0 . \tag{19}$$

From $\tilde{F}, \tilde{F} + H \in \mathcal{F}$ follow $HP_R = 0$ and $HP_E = 0$, which means that $H^* = P_{\mathcal{L}} H^*$ holds. So from (19) we get $A^+ \big[\tilde{F} D - \rho^2 ACA^*\big] H^* = 0$ and from (18) we obtain

$$Q\big(\mathrm{MSE}_{\tilde{F}+H}\big) = Q\big(\mathrm{MSE}_{\tilde{F}}\big) + \sigma^2 A^+ H D H^* A^{+*}$$

for any H with $\tilde{F} + H \in \mathcal{F}$. Any $G \in \mathcal{F}$ can be represented in the form $G = \tilde{F} + H$, hence the assertion is proved. □

Conclusion 1.2. *If, especially,* $\mathcal{S}_E = \{0\}$ *and* $I \in \mathcal{F}$, *then*

$$\tilde{F} = I - P_{\overline{\mathscr{R}(A)}} D^{-\frac{1}{2}} P_{\mathcal{K}} D^{-\frac{1}{2}} \tag{20}$$

is an optimal regularization.

Proof. This follows from Theorem 1.1 with the possible choice $F = I \in \mathcal{F}$. □

1.2 Nonlinear regularizations.

We already mentioned that linear regularizations are not well suited for the elimination of rough errors in the data. But it turns out that for nonlinear regularizations no explicit expressions for the optimal regularizations are obtainable. In the following example we consider median smoothing and show how the efficiency of regularizations in simple models of the estimation of the mean is calculated.

Example. In the model $(\mathcal{X} = \mathbb{R}^1,\ \mathcal{Y} = \mathbb{R}^n)$, $Y = (Y_1, \dots, Y_n)'$ and

$$Y_i = \gamma + \varepsilon_i , \qquad i = 1, \dots, n , \tag{21}$$

γ is to be estimated. Here $A = \underline{1} = (1, \dots, 1)'$ is an n-vector. $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed. Equation (11) has the form

$$\gamma \cdot \underline{1} = F(Y), \tag{22}$$

the solution $\hat{\gamma}_F$ is one-dimensional, and

$$\mathrm{MSE}\,\hat{\gamma}_F = \mathbf{E}(\hat{\gamma}_F - \gamma)^2 .$$

Let $Q(a) = a$ for all $a \geq 0$.

a) Let $\varepsilon \sim \mathbb{N}(0, \sigma^2)$. Then $\hat{\gamma} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the maximum likelihood estimation which is asymptotically optimal and a best unbiased estimation for γ. Also $\hat{\gamma}$ is a solution of (22) for $F(Y) = F_1(Y) = \bar{Y} \cdot \underline{1}$ and thus the regularization belonging to $\hat{\gamma}$ is the arithmetic average.

b) Let $Y_i + \varepsilon_i$ be distributed by the double exponential law with the density

$$h(y_i) = \frac{1}{2} e^{-|y_i - \gamma|} .$$

Then $\tilde{\gamma} = \mathbf{med}(Y_1, \ldots, Y_n)$ is the asymptotically optimal maximum likelihood estimation. Also $\tilde{\gamma}$ is a solution of (22) for $F(Y) = F_2(Y) = \mathbf{med}(Y_1, \ldots, Y_n) \cdot \underline{1}$ and thus the regularization belonging to $\tilde{\gamma}$ is the median.

In the next example we show which loss or gain in efficiency the median causes as a method of regularization. Here we use results from Läuter – Pincus [7].

Example. We consider the model (21) and as a regularization we take the median smoothing:

$$F_3(Y) = \frac{1}{n}\sum_{i=1}^{n} U_i \cdot \underline{1} \qquad \text{for} \quad U_i = \mathbf{med}(Y_{i-1}, Y_i, Y_{i+1}), \quad i = 2, \ldots, n-1,$$

$$U_1 = U_n = \frac{1}{n-2}\sum_{i=2}^{n-1} U_i .$$

If G is the distribution of the Y_i, then the $U_2, \ldots, U_{n-1}$ are distributed identically and have the distribution $H(u) = 3G^2(u-\gamma) - 2G^3(u-\gamma)$. Whereas $Y_1, \ldots, Y_n$ are independent from each other, this is not valid anymore for $U_2, \ldots, U_{n-1}$. The setting of the initial and final values U_1 and U_n assures $\frac{1}{n}\sum_{i=1}^{n} U_i = \frac{1}{n-2}\sum_{i=2}^{n-1} U_i$ to hold, so that by the regularization a reduction to a $(n-2)$-dimensional model has happened. We obtain $\hat{\gamma}_{F_3} = \frac{1}{n-2}\sum_{i=2}^{n-1} U_i$. If $Y_i \sim \mathbb{N}(\gamma, \sigma^2)$, we get

$$\mathbf{E}\hat{\gamma}_{F_3} = \gamma , \quad \mathbf{Var}\;\hat{\gamma}_{F_3} = \frac{1.181n - 3.33}{(n-2)^2}\sigma^2 .$$

If the Y_i are distributed double exponentially, we have

$$\mathbf{E}\hat{\gamma}_{F_3} = \gamma , \quad \mathbf{Var}\;\hat{\gamma}_{F_3} = \frac{0.775n - 2.16}{(n-2)^2}\sigma^2 .$$

Since for the least squares estimation $\hat{\gamma}$ in (21) it holds $\mathbf{E}\hat{\gamma} = \gamma$ and $\mathbf{Var}\ \hat{\gamma} = \frac{1}{n}\sigma^2$ in both cases, we obtain for the relative efficiency of F_3 with respect to F_1 the following values:

$$\mathbf{eff}^N(n) := \frac{(n-2)^2}{n(1.181n - 3.33)} \quad \text{(normal distribution)},$$

$$\mathbf{eff}^{DE}(n) := \frac{(n-2)^2}{n(0.775n - 2.16)} \quad \text{(double exponential distribution)}.$$

TABLE 1. Efficiency of median smoothing for normal and double exponential distributions

n	5	10	20	50	100	∞
$\mathbf{eff}^N$	0.699	0.755	0.798	0.827	0.837	0.847
$\mathbf{eff}^{DE}$	1.05	1.14	1.21	1.26	1.27	1.29

Consequently, under normal distribution $\hat{\gamma}$ performs always better than $\hat{\gamma}_{F_3}$, and vice versa under double exponential distribution.

This example stands as a pattern for the questions that arise about regularizations. For being able to chose a proper regularization one has to know the model and, especially, the distribution of the variables.

1.3 Optimal regularizations in model II.

Starting from model (10)

$$a(t_i, \gamma) = g_i + \varepsilon_i\,, \qquad i = 1, 2, \dots\,,$$

or, equivalently, in n-dimensional vector notation as

$$A^{(n)}(\gamma) = Y^{(n)}\,, \tag{23}$$

the regularizations by data correction have the form

$$A^{(n)}(\gamma) = F^{(n)}\big(Y^{(n)}\big)\,. \tag{24}$$

Here we consider $\mathcal{X} = \mathbb{R}^k$ only. For fixed n we can use the concepts of section 1.1.1 without change. Hence here we are interested mainly in statements on the solutions for $n \to \infty$. One of the properties of interest is the consistency of regularizations.

DEFINITION 1.3.

A sequence of regularizations $\{F^{(n)}\}$ is called *consistent* if the solutions of (24) converge weakly to the solution γ_g of (9).

It comes out that the consistency of regularizations is determined by the model and the design $t_1, t_2, \dots$ (compare Auert – Läuter [1]).

THEOREM 1.3. *Let for all n the linear equation*

$$A^{(n)}\gamma = g^{(n)}$$

be solvable with the solution $\gamma^{(n)}$ and assume $\left(A^{(n)*}A^{(n)}\right)^{-1} \to 0_{k\times k}$. Assuming $\mathsf{E}Y^{(n)} = g^{(n)}$, **Var** *$Y^{(n)} = \sigma^2 I_n$. If the solutions of $A^{(n)}\gamma = F^{(n)}Y^{(n)}$ are asymptotically unbiased, and if $F^{(n)}F^{(n)*} \le m \cdot I_n$ for a constant m, then $\{F^{(n)}\}$ is consistent.*

Proof. The solution $\hat\gamma_{F^{(n)}}$ of $A^{(n)}\gamma = F^{(n)}Y^{(n)}$ has the form

$$\hat\gamma_{F^{(n)}} = \left(A^{(n)*}A^{(n)}\right)^{-1}A^{(n)*}F^{(n)}Y^{(n)}$$

and thus

$$\begin{aligned}\mathsf{E}\hat\gamma_{F^{(n)}} &= \left(A^{(n)*}A^{(n)}\right)^{-1}A^{(n)*}F^{(n)}A^{(n)}\gamma_g^{(n)},\\ \mathbf{Var}\ \hat\gamma_{F^{(n)}} &= \sigma^2\left(A^{(n)*}A^{(n)}\right)^{-1}A^{(n)*}F^{(n)}F^{(n)*}A^{(n)}\left(A^{(n)*}A^{(n)}\right)^{-1}.\end{aligned}$$

Because of $F^{(n)}F^{(n)*} \le m \cdot I_n$, one obtains

$$\mathbf{Var}\ \hat\gamma_{F^{(n)}} \le m\sigma^2\left(A^{(n)*}A^{(n)}\right)^{-1}.$$

Together with $\mathsf{E}\hat\gamma_{F^{(n)}} - \gamma^{(n)} \xrightarrow{\mathrm{P}} 0$, the statement follows. □

A necessary condition for $F^{(n)}F^{(n)*} \le m \cdot I_n$ is given in Auert – Läuter [1].

THEOREM 1.4. *If $F^{(n)} = \left(f_{ij}^{(n)}\right)$ and $\sum_{j=1}^{n}\left|f_{ij}^{(n)}\right| \le \rho$ for all i and n, and if for $|i-j| > m_1 > 0$ the inequality*

$$\left|f_{ij}^{(n)}\right| \le \frac{c}{|i-j|^{1+\alpha}}, \qquad \alpha > 0,\ c \ge 0,$$

is fulfilled, then there exists a constant m with

$$F^{(n)}F^{(n)*} \le m \cdot I_n.$$

In Auert – Läuter [1] general conditions for the asymptotic unbiasedness of the solutions of $A^{(n)}\gamma = F^{(n)}Y^{(n)}$ were formulated. These conditions affect the model, the design and the operator of regularization $F^{(n)}$. It turns out that, for example, polynomial smoothers and spline smoothers on tightening design points satisfy the conditions posed there.

Our message from this part is that in linear problems an optimal regularization is given and computable. Under nonlinearities, the efficiency of regularizations depends on the whole distribution. Therefore the estimation of the underlying distribution is necessary.

2. Nonparametric test of densities

Let $Y_1, \ldots, Y_n$ be a sample of i.i.d. random variables with density f. We wish to test whether f lies in some parametric family of density functions, i.e., we consider the problem of testing the hypothesis

$$H\colon f \in \mathcal{F} = \Big\{ f(\cdot, \vartheta)\colon\ \vartheta \in \Theta \subseteq \mathbb{R}^k \Big\} \tag{25}$$

against

$$K\colon f \notin \mathcal{F}\,.$$

As a test statistic, we propose the deviation of the well-known kernel estimate

$$\hat{f}_n(t) = (nh_n)^{-1} \sum_{i=1}^{n} K\left(\frac{t - Y_i}{h_n}\right) \tag{26}$$

from the expectation

$$\mathbf{E}\hat{f}_n(t) = e_{h_n}(t;\, f_\vartheta) = \int K(x) f(t - xh_n, \vartheta)\,\mathrm{d}x \tag{27}$$

of $\hat{f}_n(t)$ with respect to the density $f(\cdot, \vartheta) = f_\vartheta(\cdot)$ measured in the L_2-norm. In (26), K is a kernel function and h_n is the sequence of bandwidths. Since the parameter ϑ in (27) is unknown, we replace it by an estimator $\hat{\vartheta}_n$ which will be discussed later on.

The derivation of the limit distributions of our test statistic is based on the results of Bickel and Rosenblatt [3] and Hall [6] on the asymptotic behavior of global deviation measures of kernel density estimates, which we modify in the aspect of estimated parameters. Furthermore, to investigate the power of our test we consider local alternatives, and to handle the limit distributions under these alternatives we embed our test problem into a more general one: Let $Y_{n1},\ Y_{n2}, \ldots Y_{nn}$ be a sequence of i.i.d. r.v.'s with density f_n. The null hypothesis is $H\colon f_n \in \mathcal{F}$. Our local alternatives are of the form

$$K_n\colon f_n = f(\cdot, \vartheta_n) + w_n(\cdot) \qquad \text{with} \quad \vartheta_n = \vartheta + n^{-\beta}\psi\,, \tag{28}$$

where ψ is a given vector, $\beta > 0$ and w_n is a function tending to zero which will be specified later. Setting $w_n = \lambda_n w$ for some sequence λ_n tending to zero and some function w, we get Pitman alternatives, and for $w_n(\cdot) = \lambda_n w\big((\cdot - c)/b_n\big)$, with $b_n \to 0$, we obtain sharp peaks alternatives similar to those considered in Rosenblatt [12] and in Ghosh – Huang [4]. Rosenblatt and Ghosh – Huang proposed sharp peak alternatives to demonstrate that under these local alternatives, the limit of the power of the test based on the quadratic functional, which we will call L_2-test, can be greater than α even if the alternative converges to the hypothesis faster than $n^{-1/2}$. In Liero *et al.* [9] a comparison of the L_2-test with a test based on the maximal deviation on the grid, i.e., the L_∞-test, is given.

2.1 Asymptotic α-test.

Here we present two theorems on the limit distribution of the weighted integrated squared error of a kernel density estimate. These theorems are a slight generalization of results of Bickel and Rosenblatt [3] to the case that the underlying density depends on n and tends to a density f. It allows to derive the limit distribution of our test statistics under the null hypothesis and under local alternatives. First let us introduce some definitions and notations. For a function η, set

$$e_h(t,\eta) = \int K(x)\eta(t - xh_n)\,\mathrm{d}x\,, \qquad \text{and} \qquad v_h^2(t,\eta) = \int K^2(x)\eta(t - xh_n)\,\mathrm{d}x\,,$$

and define for $\mathcal{I} = [0,1]$

$$Q_n(\eta) = nh_n \int\limits_{\mathcal{I}} \big(\hat{f}_n(t) - e_h(t;\eta)\big)^2 a(t)\,\mathrm{d}t\,,$$

$$\mu(\eta) = \kappa \int\limits_{\mathcal{I}} \eta(t)a(t)\,\mathrm{d}t\,, \qquad \sigma^2(\eta) = 2\kappa^* \int\limits_{\mathcal{I}} \eta^2(t)a^2(t)\,\mathrm{d}t\,,$$

where a is a given weight function, $\kappa = \int K^2(x)\,\mathrm{d}x$ and $\kappa^* = \int (K*K)^2(x)\,\mathrm{d}x$, and

$$\hat{f}_n(t) = (nh_n)^{-1} \sum_{i=1}^{n} K\left(\frac{t - Y_{ni}}{h_n}\right)$$

is the kernel estimate based on a sample $Y_{n1}, Y_{n2}, \dots Y_{nn}$ of i.i.d.r.v.'s with density f_n. We write $\int$ for $\int\limits_{-\infty}^{+\infty}$. In our further investigations we make use of the following assumptions:

A1. The density f_n is bounded on $\mathbb{R}$ and continuous on a neighbourhood of the interval $\mathcal{I}$.

A2. The kernel function is bounded, of bounded variation, has a bounded support, and $\int K(x)\,\mathrm{d}x = 1$.

A3. Any $f \in \mathcal{F}$ is bounded on $\mathbb{R}$, continuous on a neighbourhood of the interval $\mathcal{I}$, and partially differentiable w.r.t. ϑ; $\nabla_\vartheta f(x,\vartheta)$ is bounded and uniformly continuous in both arguments.

A4. The weight function a is nonnegative, piecewise continuous, and bounded on $\mathbb{R}$.

A5. $h_n \to 0$, $nh_n \to \infty$.

As a generalization of Theorem 4.1 of Bickel and Rosenblatt [3] and Theorem 2 of Hall [6] we obtain

THEOREM 2.1. *Under A1, A2, A4, and A5 we have for $n \to \infty$ and for all x*

$$\mathrm{P}\left(\frac{h_n^{-1/2}(Q_n(f_n) - \mu(f_n))}{\sigma(f_n)} \le x\right) \longrightarrow \Phi(x). \tag{29}$$

To apply this limit statement to test whether our density lies in a parametric class of densities, one has to estimate the parameter ϑ. Here we suppose that there exists a sequence of parameter estimators $\hat{\vartheta}_n$ which tends to ϑ fast enough under the hypothesis, i.e., we assume that under H for each $\vartheta \in \Theta \subseteq \mathbb{R}^k$:

$$\sqrt{n}\,|\hat{\vartheta}_n - \vartheta| = O(1), \tag{30}$$

where $|\vartheta|$ is the Euclidean norm of ϑ. It can be shown that assumption (30) is satisfied under regularity conditions by maximum likelihood estimators in $\mathcal{F}$.

Theorem 2.1 and (30) motivate that $Q_n(f(\cdot, \hat{\vartheta}_n))$ is a suitable statistic for testing H against K. More exactly, this is expressed by the next theorem. We write $f(\cdot, \vartheta) = f_\vartheta(\cdot)$.

THEOREM 2.2. *Suppose that assumptions* **A2**, **A3**, **A4**, **A5** *and* (30) *hold and that $h_n(\log n)^\zeta \to 0$ for some $\zeta > k/2$. Then for $n \to \infty$ and all x*

$$\mathrm{P}\left(\frac{h_n^{-1/2}(Q_n(f_{\hat{\vartheta}_n}) - \mu(f_{\hat{\vartheta}_n}))}{\sigma(f_{\hat{\vartheta}_n})} \le x \,\middle|\, H\right) \longrightarrow \Phi(x). \tag{31}$$

Consequently an asymptotic α-test of H against K is provided by:

$$\text{Reject } H \qquad \text{iff} \qquad Q_n(f_{\hat{\vartheta}_n}) \ge \mu(f_{\hat{\vartheta}_n}) + u_\alpha h_n^{1/2} \sigma(f_{\hat{\vartheta}_n}), \tag{32}$$

where u_α is defined by $\Phi(u_\alpha) = 1 - \alpha$.

Notice, that this test is an analogue of the Bickel – Rosenblatt test considered in Ghosh – Huang [4], but here for the case of estimated parameters. We shall refer to (32) as the L_2-test. Under general conditions this test is consistent. For a more exact description of the power of this test we need local alternatives. Now we consider the asymptotic behavior of the power of both tests under Pitman and sharp peak alternatives, i.e., under local alternatives of the form

$$K_n : f_n(\cdot, \vartheta_n) = f(\cdot, \vartheta_n) + \lambda_n w\big((\cdot - c)b_n^{-1}\big) \tag{33}$$

with $\vartheta_n = \vartheta + n^{-\beta}\,\psi$ where ψ is a given vector, $\beta > 0$, and ϑ is fixed. λ_n is a sequence of positive numbers tending to zero, and $c \in \mathcal{I}$. For b_n tending to zero K_n is the sharp peak alternative, and K_n represents the Pitman alternative if $b_n \equiv b > 0$. For the function w we assume

A6 The function w has bounded support, without loss of generality $[-1, 1]$, and $\int w(x)\,\mathrm{d}x = 0$.

The alternatives (33) are generalizations of those proposed in Rosenblatt [12] and Ghosh – Huang [4].

Power of the L_2-test.

The power of the L_2-test under K_n is

$$\Pi(f_n) = \mathrm{P}\Big(Q_n(f_{\hat\vartheta_n}) \geq \mu(f_{\hat\vartheta_n}) + u_\alpha h_n^{1/2}\sigma(f_{\hat\vartheta_n})\,\big|\,K_n\Big)$$

and we have in addition to Theorem 2.1 the following

THEOREM 2.3. *Suppose that* **A1**, **A2**, **A3**, **A4**, **A5** *and* **A6** *are satisfied and that* $h_n b_n^{-1} \to 0$. *Under* K_n, *the m.l.-estimator* $\hat\vartheta_n$ *has the asymptotic expansion*

$$\vartheta_n - \vartheta = n^{-1/2} Z_n + \tau_n S + o_{\mathrm{P}}(\tau_n)\,, \tag{34}$$

where the vector Z_n *converges in distribution to a normally distributed vector with expectation zero,* S *is a nonrandom vector, and* τ_n *is a sequence of positive numbers. Put*

$$c_{1n} = n h_n^{1/2} \lambda_n^2\,, \qquad c_{2n} = h_n^{-1/2}\lambda_n\,.$$

a) *If* $b_n = 1$ *then*

$$\lim_{n\to\infty} : \big(\Pi(f_n) - \Phi(U_{n,\alpha})\big) = 0\,, \tag{35}$$

where

$$U_{n,\alpha} = \sigma^{-1}(f_\vartheta)\bigg(c_{1n}\int \big(w(t) - S^t\nabla_\vartheta f_\vartheta(t)\big)^2 a(t)\,\mathrm{d}t + \kappa c_{2n}\int \big(w(t) - S^t\nabla_\vartheta f_\vartheta(t)\big)a(t)\,\mathrm{d}t\bigg) - u_\alpha\,.$$

b) *Suppose* $b_n \to 0$. *If*

$$\rho = \int S^t\nabla_\vartheta f_\vartheta(t)a(t)\,\mathrm{d}t\,, \quad T_n = \begin{cases} -\kappa c_{2n} b_n \rho\,, & \text{if } \rho \neq 0\,, \\ -\kappa c_{2n} b_n^2 \int w(t)t\,\mathrm{d}t a'(c)\,, & \text{otherwise} \end{cases} \tag{36}$$

then (35) *with* $U_{n,\alpha} = \sigma^{-1}(f_\vartheta)\Big(c_{1n}b_n \int w(t)^2 a(t)\,\mathrm{d}t + T_n\Big) - u_\alpha\,.$

Acknowledgement. I would like to thank the referees for helpful comments and so for getting an improved paper.

REFERENCES

[1] AUERT, F.—LÄUTER, H.: *Some theoretical results on approximation methods*, Statistics **13** (1982), 265–293.

[2] BATES, D. M.—WATTS, D. G.: *Nonlinear Regression and its Applications*, J. Wiley, New York, 1988.

[3] BICKEL, P.—ROSENBLATT, M.: *On some global measures of the deviations of density function estimates*, Ann. Statist. **1** (1973), 1071–1095.

[4] GHOSH, B.—HUANG, W. M.: *The power and optimal kernel of the Bickel-Rosenblatt test for goodness of fit*, Ann. Statist. **19**, 999–1009.

[5] HÄRDLE, W.: *Applied nonparametric regression*, Econometric Society Monographs, Vol. 19, Cambridge, 1990.

[6] HALL, P.: *Central limit theorem for integrated square error of multivariate nonparametric density estimators*, J. Multivariate Anal. **14** (1984), 1–16.

[7] LÄUTER, H.—PINCUS, R.: *Mathematisch-Statistische Datenanalyse*, Oldenbourg Verlag, 1989.

[8] LÄUTER, H.—LIERO, H.: *Ill-posed inverse problems and their optimal regularization*, Preprint 57, Humboldt-Universität, Berlin, 1997.

[9] LIERO, H.—LÄUTER, H.—KONAKOV, V.: *Nonparametric vs. parametric goodness of fit*, Statistics **31** (1998), 115–149.

[10] MALLOWS, C. L.: *Some theory of nonlinear smoothers*, Ann. Math. Statist. **8** (1980), 695–715.

[11] PÁZMAN, A.: *Nonlinear Statistical Models*, Kluwer, Dordrecht, 1993.

[12] ROSENBLATT, M.: *A quadratic measure of deviation of two-dimensional density estimates and a test of independence*, Ann. Statist. **3** (1975), 1–14.

[13] ROSS, G. J. S.: *Nonlinear Regression*, Springer-Verlag, Berlin, 1990.

[14] SEBER, G. A. F.—WILD, C. J.: *Nonlinear Regression*, J. Wiley, New York, 1989.

[15] WAHBA, G.: *A survey of some smoothing problems and the method of generalized cross validation for solving them*, in: Applications of Statistics (P. R. Krishnaiah, ed.), North Holland Publ. Comp., Amsterdam, 1977.

Received May 21, 1998

Institute of Mathematics
University of Potsdam
Am Neuen Palais 10
D-14415 Potsdam
GERMANY

E-mail: laeuter@rz.uni-potsdam.de