NONPARAMETRIC METHODS IN SEQUENTIAL ANALYSIS

Pranab Kumar Sen

Departments of Biostatistics and Statistics
University of North Carolina
Chapel Hill, NC 27599 USA

ABSTRACT

Most of the other chapters in this volume are devoted to various parametric models arising in the classical sequential analysis pertaining to statistical decision theory including sequential testing as well as (point and interval) estimation problems. Though of relatively recent origins, various nonparametric models have also been worked out for these traditional problems in sequential analysis. A systematic account of these robust nonparametric methods is presented in the current chapter.

Keywords and phrases: Aligned rank statistics; asymptotic efficiency; asymptotically minimum risk estimation; ASN; bounded-width confidence interval; cost function; influence function; jackknifing; L-estimator; loss function; martinglae; M-estimator; OC; optimal stopping; pseudovariables; R-estimator; robustness; SLRT; SPRT; shrinkage estimator; statistical functionals; stopping number; weak convergence.
1. INTRODUCTION

The past twenty five years have witnessed a phenomenal growth of the literature on nonparametric methods in sequential analysis (to be referred to, in the sequel, as sequential nonparametrics). Flexibility of the model (with respect to the form of the underlying probability distributions), robustness (against departures from model based assumptions, error contaminations and/or outliers), and (asymptotic) efficiency considerations dominate the scenario in sequential nonparametrics. As such, in this review, due attention is paid to each of these three important aspects.

By now, sequential tests based on robust nonparametric statistics are well regarded as good competitors of the conventional sequential probability ratio tests (SPRT) or sequential likelihood ratio tests (SLRT) based on specific parametric (distributional) models. A departure from such an assumed model may have serious damaging effects on the optimality properties of the SPRT/SLRT, that is, a parametric SPRT/SLRT may not retain its prescribed strength and/or good average sample number (ASN) property when the actual distribution is different from the assumed one. Even a small or moderate deviation may have noticeable effect on the operating characteristics (OC) of such parametric sequential tests, so that they are not generally robust. On the other hand, in a nonparametric setup, generally, the underlying distribution is of unspecified form, and it is only assumed that it belongs to a certain family of distributions. This feature endows room for model flexibility and enhances the scope for robustness. A similar picture relates to the sequential estimation problems. In a sequential confidence interval problem, the objective is to
obtain an interval which would contain the true parameter with a prescribed coverage probability and further the length of this interval be bounded from above by a prescribed (small) number. Naturally, the determination of a sample size and suitable statistics (leading to the confidence limits) for which both these requirements are met may depend on the underlying distribution. As such, parametric solutions may lack robustness. On the other hand, nonparametric confidence intervals are based on "inversions" of suitable (robust) nonparametric test statistics, and hence, may retain their robustness property to a greater extent. In minimum risk estimation problems, it is quite natural to incorporate a cost function (depending on the sample size) along with a conventional loss function (depicting the sampling variability of the estimator) in the formulation of a risk function, and minimisation of this risk (with respect to the choice of the sample size) leads to the minimum risk point estimator. The intricacy of such a risk leads to an optimal sample size which generally depends on the underlying distribution, and hence, a non-sequential optimal solution may not exist. In an asymptotic setup (allowing the cost per unit sample to be small), multi-stage or sequential procedures work out neatly. However, in this context too, the usual parametric procedures may suffer from the usual lack of robustness property, and nonparametric procedures have better robustness properties. In either of the (testing/estimation) problems, there are certain models where parametric formulations may not appear to be very appropriate while their nonparametric counterparts work out very conveniently [viz., statistical functionals]. Judged from these perspectives, sequential nonparametrs seem to have a much wider appeal than their parametric counterparts.

The three basic sequential problems in a nonparametric setup are
considered in this chapter. There are, however, certain other basic problems in sequential nonparametrics which deserve mention. Nonparametric repeated significance tests are most noteworthy, and these are discussed in detail in Chapter 7 (along with their parametric counterparts). In the context of progressively censored schemes (PCS) in clinical trials, life testing procedures and in some other follow up studies, time-sequential inference procedures occupy a central place. Time-sequential estimation procedures have been considered in Chapter 28 [by Gardiner and Susarla]. Also, in the context of clinical trials, some other specific time-sequential (viz., group-sequential) tests have been considered; see for example, Chapter 27 [by Whitehead], for a nice review of the applied aspects. A more comprehensive account of the methodology of some of these nonparametric time-sequential testing procedures is given in Chapter 8. Asymptotically efficient (sequential) adaptive estimation procedures have been discussed in Chapter 21 [by Husková], and hence, we shall not repeat these here. Finally, in Chapter 24 [by Zacks], a nice account of the change-point models is given. In the concluding section of this chapter, we shall provide some complementary results on nonparametric tests for some change-point models. Although asymptotically distribution-freeness and robustness aspects underline the scope of applicability of sequential nonparametrics in a very broad setup, there are various special sectors where exact distribution-freeness property holds and adds additional strength to the relative advantages of the sequential nonparametrics over their parametric counterparts.
2. NONPARAMETRIC SEQUENTIAL TESTS

A sequential probability ratio test (SPRT) for testing a simple null hypothesis \( H_0 \) against a simple alternative \( H_1 \) has a prescribed strength \((\alpha, \beta)\) [i.e., the probabilities of type I and II errors are bounded by some preassigned numbers \( \alpha (0 < \alpha < 1) \) and \( \beta (0 < \beta < 1) \) respectively], and, in addition, among all tests which have the same strength \((\alpha, \beta)\), it has the optimality property that it has the minimum average sample size (ASN). However, if we look at the picture a bit more carefully, we may observe that the attainment of the prescribed strength by a SPRT is very much dependent on the basic fact that the assumed form and the actual distribution of the random variables are the same. Or in other words, the SPRT may not be usually robust against plausible departures from the model based assumptions. The ASN of a SPRT is even more vulnerable to such departures, so that the Wald-Wolfowitz optimality of a SPRT may hold only under the assumed model. In the fixed sample size case, for a general class of statistical hypotheses (of invariance), genuinely distribution-free (nonparametric) tests exist (whose type I error probability does not depend on the underlying distribution), and within this class, there exist suitable tests which are (at least, asymptotically) optimal (i.e., most powerful) against specific sub-classes of distributions. A further remark: Adaptive nonparametric tests exist which are asymptotically efficient for all distributions belonging to a suitable class; we may refer to Chapter 21 (by Husková) for some good accounts of these developments.

Although exact distribution-free (EDF) nonparametric tests have significance level independent of underlying distributions (belonging to a class), their power functions (for specific types of parametric alternatives, like location/scale/regression ones) are generally dependent
on the underlying distribution. Even asymptotically, this dependence is evident from the presence of a scale factor (in the asymptotic power function) which is typically a unknown functional of the underlying distribution. An exception to this pattern is the so called "Lehmann alternatives" or more generally the "Cox proportional hazard models" where the power function depends on the hazard ratio only (and the null hypothesis relates to a constant hazard ratio (equal to one)). Thus, simultaneous attainment of the strength \((\alpha, \beta)\) by a nonparametric sequential test (for an unknown form of the underlying distribution) may not be feasible, excepting for such 'proportional hazard models'. For this reason, some of the earlier nonparametric sequential tests were considered for such "Lehmann alternatives"; we may refer to Savage and Sethuraman (1966) for some nice account of these developments. Even in the asymptotic case, besides the termination of such sequential tests with probability one, only some bounds on the ASN etc. have been derived. From a mathematical point of view, there are some challenging problems related to the general asymptotics of such sequential tests. However, from a statistical stand point, these developments are mostly of academic interest only. Hence, in this review, we shall not enter into any discussion on them.

At the second phase, some nonparametric sequential tests were also considered for the case where the null hypothesis relates to a nonparametric one (viz., equality of two distribution functions) but the alternative hypothesis to a specific parametric one (viz., two normal distributions differing only in locations). In such a case, a log-likelihood ratio type test statistic can be constructed (in the sequential case) incorporating the joint distribution of the ranks of the observations under the competing alternatives. Such procedures also have not met the light of full success
or popularity mainly due to the arbitrariness in this formulation and their
mathematical complications. We may refer to Savage and Sethuraman (1972)
for some of these details.

There are other nonparametric sequential tests which are structurally
quite similar to the SPRT, share the same properties (in an asymptotic
setup), and, at the same time, are robust. We shall consider these tests in
detail. As in Sen (1981), we shall find it convenient to present the
methodology separately for two types of tests which are structurally
somewhat different. These two types are respectively nonparametric
estimators based sequential tests and tests based on aligned nonparametric
statistics. To motivate these procedures, let us consider the simplest
parametric model: \{X_i, i \geq 1\} i.i.d.r.v.'s with N(\theta, \sigma^2) distribution. H_0 :
\theta = \theta_0, H_1 : \theta = \theta_1 > \theta_0 \text{ and } \theta_0, \theta_1 \text{ and } \sigma^2 \text{ are all specified. Let } A = \theta_1 - \theta_0 (>0).

The classical SPRT [due to Wald (1947)] is based on two
constants \(b = \log B \) and \( a = \log A \), where \( 0 < B < 1 < A < \infty \) and \( B \geq \beta/(1-\alpha) \),
\( A \leq (1-\beta)/\alpha \), for the preassigned strength \((\alpha, \beta)\). For each \(n \geq 1\), the
log-likelihood ratio is given by
\[
\log \lambda_n = A \sigma^{-2} \{X_1^n \left( X_i - \frac{1}{2}(\theta_0 + \theta_1) \right) \}. 
\]
(2.1)
The sequential procedure consists in drawing observations so long as
\[
b < \log \lambda_n < a, \ n \geq 1; \quad (2.2)
\]
if, for the first time, for \(n = N\), \(\log \lambda_N\) is \(\geq a\) (or \(< b\)), then sampling is
stopped at that stage, and \(H_1\) (or \(H_0\)) is accepted. Thus, the stopping
number \(N\) is defined by
\[
N = \inf \{n \geq 1 : \log \lambda_n \in (b, a) \}, \quad (2.3)
\]
and, for every (fixed) \(\theta\), it is known that
\[
P_\theta(N > n) \rightarrow 0 \ \text{as} \ n \rightarrow \infty, \quad (2.4)
\]
i.e., the process terminates with probability one. In this setup, it is
known that the probabilities of Type I and II errors are bounded by \( \alpha \) and \( \beta \) respectively (ensured by the assumed bounds for A and B and the normality of the underlying d.f.). Under the usual (Wald-) supposition that the excesses over the boundaries are negligible (justified when A is small), the Type I and II error probabilities are close to \( \alpha \) and \( \beta \) respectively. Before we examine the robustness aspects of this SPRT, we consider the case where in the same model \( \sigma^2 \) is unspecified. There are various modifications of the SPRT in this case, discussed in Wald (1947), Ghosh (1970) and other texts. We consider the natural contender: The Sequential likelihood ratio test (SLRT) considered by Bartlett (1946), Cox (1963) and others. For every \( n \geq 2 \), we let

\[
s_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2; \quad \bar{x}_n = n^{-1} \sum_{i=1}^{n} x_i.
\]

Parallel to (2.1), we consider the modified log-likelihood ratio statistics:

\[
\log \lambda_n^* = \Delta s_n^{-2} \sum_{i=1}^{n} (x_i - \frac{1}{2}(\theta_0 + \theta_1)) = \sigma^2 s_n^{-2} \log \lambda_n, \quad n \geq 2.
\]

Note that \( s_n^2 \) is a strongly consistent estimator of \( \sigma^2 \), and hence, for \( \Delta \) small enough, we may use the same stopping rule as in (2.2)–(2.3) with the \( \log \lambda_n \) replaced by \( \log \lambda_n^* \). In fact, for \( \Delta \to 0 \), the SPRT and SLRT share the common properties. For \( \Delta \) not so small, one may use some generalized SPRT (SLRT) procedures based on the sequence \( \{\log \lambda_n^*; \quad n \geq 2\} \).

Let us now examine the robustness properties of the SPRT/SLRT for the normal mean problem treated above. Sequential tests are usually judged by their OC (operating characteristic) and ASN functions. Recall that \( L(\theta) \), the OC, is the probability of accepting \( H_0 : \theta = \theta_0 \) when the true parameter value is \( \theta \). Thus, in order that a (sequential) test has strength \( (\alpha, \beta) \), we must have

\[
L(\theta_0) \geq 1 - \alpha \quad \text{and} \quad L(\theta_1) \leq \beta.
\]

Also, the ASN is \( E_{\theta} N \), where the stopping number \( N \) is defined as in (2.3).
Both of these measures depend on the distributional properties of \( \{ \log \lambda_n \) (or \( \log \lambda_n^* \)), \( n \geq 1 \) (or 2), and these, in turn, depend on \( (\theta, \sigma^2) \). When the \( X_i \) are i.i.d. r.v.'s with a normal \( (\theta, \sigma^2) \) distribution, \( \{ \sigma^{-1} \sum_{i=1}^{n} X_i - n\theta ; n \geq 1 \} \) has the discrete time parameter Brownian motion on \((0, \infty)\), so that for small values of \( \Delta(>0) \), (2.1)-(2.2) may as well be approximated by a continuous time parameter Brownian motion with a linear drift \( (n\Delta/\sigma)(\theta - \frac{1}{2}(\theta_0 + \theta_1)/\sigma, n \geq 1) \). As such, the general theory developed by Dvoretzky, Kiefer and Wolfowitz (1953) can be used to provide an adequate approximation to the OC function. If we write \( \theta = \theta_0 + \varphi\Delta \), where \( \varphi \) lies in a compact interval containing both 0 and 1 as innerpoints, then letting \( L^0(\varphi) = \lim_{\Delta \to 0} L(\theta_0 + \varphi\Delta) \), it follows that

\[
L^0(\varphi) = \begin{cases} 
\frac{(A^{1-2\varphi} - 1)}{(A^{1-2\varphi} - B^{1-2\varphi})}, & \varphi \neq 1/2 \\
\log A/(\log A - \log B), & \varphi = 1/2
\end{cases}
\]  

(2.8)

where \( A, B \) are defined as in (2.2). Similarly, we have on letting

\[
\zeta(\varphi, \sigma^2) = \begin{cases} 
[\log B \cdot L^0(\varphi) + \log A(1-L^0(\varphi))]\sigma^2/(\varphi\frac{1}{2}), & \varphi \neq 1/2 \\
(\log B)(\log A)\sigma^2, & \varphi = 1/2
\end{cases}
\]

(2.9)

\[
\lim_{\Delta \to 0} [A^2 \cdot \mathbb{E}(N | \theta = \theta_0 + \varphi\Delta)] = \zeta(\varphi, \sigma^2).
\]

(2.10)

for all \( \varphi \) belonging to a compact interval.

Let us now examine the robustness of (2.8) and (2.10) when the underlying distribution may not be strictly normal. In this context, we may note that (through the Skorokhod-Strassen embedding of a Wiener process) the distributional approximation of \( \{ \sigma^{-1} \sum_{i=1}^{n} X_i - n\theta ; n \geq 1 \} \) by a Wiener process holds for a broad class of distributions having finite second moment. Thus, (2.8) holds for a broad class of distributions, and the assumption of normality is not that crucial. However, there are certain points worth pondering in this context. First, for a normal distribution,
(2.8) holds even for moderately small values of $\Lambda$, while the rate of convergence in the non-normal case may depend very much on the tail behavior of the underlying distribution. This lack of uniformity of the convergence in (2.8) is a genuine concern in practice where the form of the underlying distribution may rarely be known. Secondly, for the SLRT, one needs to use the sequence $\{s_n^2, n \geq 2\}$ of estimators of $\sigma^2$, and it is well known that $s_n^2$ is not a very robust estimator. Thus, a small departure from the assumed normal distribution may have a comparatively more serious effect on the SLRT than the SPRT. This lack of robustness aspect is more apparent with the ASN function in (2.9)-(2.10). It is clear from the definition in (2.10) that not only it involves the distribution of the stopping number $N$ but also it rests on the moment convergence of $(\Lambda^2 N)$. The distribution of $N$ [viz., (2.3)] is quite sensitive to any departure from normality of the underlying distribution and the convergence in first moment of $\Lambda^2 N$ is even less robust.

The weak convergence of the partial sum process (or its embedding) holds for a general class of nonparametric (and robust) estimators and test statistics. This general methodology has been systematically studied in Sen (1981). Some general sequential tests are based on this methodology. We present here a general account of these sequential procedures.

2.1. Sequential tests based on nonparametric and robust estimators.  
Instead of the normal mean ($\theta$) problem, consider now a general parameter $\theta$ (which may even be treated as a functional of the underlying (unknown) distribution function $F$), and assume that there exists a sequence $\{T_n\}$ of suitable estimators of $\theta$ (having some nice properties). At the minimal, we assume that there exists a parameter $\sigma^2_{\theta}$ (which may depend on $\theta$ or $F$), such that as $n \to \infty$,

$$n^{1/2}(T_n - \theta) \overset{\mathcal{D}}{\to} N(0, \sigma^2_{\theta}) \quad \text{when } \theta \text{ holds.} \quad (2.11)$$
Generally, $\sigma_0^2$ is not known, but there exists a sequence $\{s_n^2\}$ of (strongly) consistent estimators of $\sigma_0^2$. Then, for testing the null hypothesis $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1 = \theta_0 + \Delta$, $\theta_0$ and $\Delta(>0)$ specified, in (2.1), we replace the role of $\sum_{i=1}^{n} X_i$ by $nT_n$ and conceive of the following stopping number: For some $n_0 \geq 2$,

$$
N = N_\Delta = \min(n \geq n_0 : n \Delta s_n^2(T_n - \frac{1}{2}(\theta_0 + \theta_1)) \in (b,a)). \quad (2.12)
$$

where $a,b$ are defined as in (2.2); we accept $H_0$ or $H_1$ according as $N_\Delta(\Delta(s_{H_\Delta}^2(T_{n_\Delta}^n - \frac{1}{2}(\theta_0 + \theta_1)))$ is $\leq b$ or $\geq a$. Let us look at (2.12) and examine the battery of regularity conditions which may be needed to establish the various properties of such a sequential test.

(a) **Termination of the test.** Note that for any fixed $\theta$,

$$
P_\theta(N_\Delta > n) = P_\theta(\text{m.d.}(T_m^n - \frac{1}{2}(\theta_0 + \theta_1))/s_m^2 \in (b,a), m \in [n_0,n])
$$

$$
\leq P_\theta(\text{m.d.}(T_n^m - \frac{1}{2}(\theta_0 + \theta_1))/s_m^2 \in (b,a))
$$

$$
= P_\theta(n\text{-d.}(\theta_n^m - \theta^2) + n\text{-d.}(\theta_n - \frac{1}{2}(\theta_0 + \theta_1))) \in \frac{1}{\sqrt{n}} s_n^2(b,a). \quad (2.13)
$$

Thus, if (2.11) holds and $s_n^2 \rightarrow \sigma_0^2$, in probability a $n \rightarrow \infty$, then $\frac{2}{\sqrt{n}} s_n^2/\sigma_0 \rightarrow 0$, in probability, as $n \rightarrow \infty$, so that (2.13) converges to 0, as $n \rightarrow \infty$. Thus, the asymptotic normality in (2.11) and the (weak) consistency of $s_n^2$ ensure that the test based on $N_\Delta$ in (2.12) terminates with probability one (for any fixed $\Delta$).

(b) **OC function.** For this study, we take recourse to an asymptotic situation where we let $\Delta \rightarrow 0$. Since $\Delta$ relates to the distance between the null hypothesis and alternative specifications of $\theta$, in such an asymptotic case, the OC function will be nondegenerate only in a small neighborhood of $(\theta_0, \theta_1)$, where the effective diameter of this neighborhood is also proportional to $\Delta$. Thus, we specifically let
\[ \theta = \theta_0 + \varphi \Delta, \text{ where } \varphi \in \text{compact } J = [-K, K], \] (2.14)

for some \( K > 1 \). In this setup, we shall synonymously denote \( P_{\theta} \) by \( P_{\varphi} \).

Then, note that as in (2.8), the OC \( L(\theta) \) is given by

\[
L(\theta) = P_{\theta} \left( m \Delta \left[ T_m - \frac{1}{2}(\theta_0 + \Theta_1) \right] / s_m^2, \ m \geq n_0, \right. \\
\text{first goes below } \log b \\
\left. \text{before crossing } \log a \right) \\
= P_{\varphi} \left( m \Delta / \sigma_\theta^2 + m \Delta^2 / 2 \sigma_\theta^2, \ m \geq n_0, \right. \\
\text{first goes below} \\
(s_m^2 / \sigma_\theta^2) \log b \text{ before going above} \\
(s_m^2 / \sigma_\theta^2) \log a). \] (2.15)

Thus, if (i) \( \Delta^2 n_0 \to 0 \) as \( \Delta \to 0 \), (ii) \( s_n^2 / \sigma_\theta^2 \to 1 \), a.s., as \( n \to \infty \), and (iii) \( m(\Delta / \sigma_\theta)(T_m - \theta) / \sigma_\theta, \ m \geq n_0 \), weakly converges to a Wiener process with zero mean (function), then again we can appeal to the Dvoretzky, Kiefer and Wolfowitz (1953) result, and conclude that as \( \Delta \to 0 \), \( L(\theta_0 + \varphi \Delta) \to L^0(\varphi) \), where \( L^0(\varphi) \) is defined by (2.8). This shows that the OC function is asymptotically (as \( \Delta \to 0 \)) insensitive to the form of the underlying distribution \( F \), provided we have the conditions (i), (ii) and (iii) on \( \Delta \), \( n_0, T_m, m \geq n_0 \), and \( s_m^2 \) satisfied. These conditions hold for a general class of nonparametric and robust estimators, and hence, (2.8) extends to this general class as well. Note that if the excess over the boundaries are negligible, then for the stopped process, \( (\Delta / s_{N_\Delta}^2) N_\Delta (T_n - \frac{1}{2}(\theta_0 + \Theta_1)) \) can only assume the two values \( b(\log B) \) and \( a(\log A) \), and the probabilities for these two realizations are given by \( L(\theta) \) and \( 1 - L(\theta) \), respectively. Actually, for nonparametric or other robust statistics, \( \{T_n\} \) is much more well behaved (in the tail) than \( \{X_n\} \), and hence, this basic assumption of negligibility of the excess over the boundaries generally holds when \( \Delta \to 0 \). On the other hand, the presence of \( s_{N_\Delta}^2 \) (instead of \( \sigma_\theta^2 \)) is rather disturbing, and the strong convergence of \( s_n^2 \) to \( \sigma_\theta^2 \) may not be enough in this context. A
slightly more stringent condition, viz., \( \forall \eta > 0, \exists \alpha > 0, \) such that
\[
P_\theta (|s_n^2 \sigma_\theta^2 - 1| > \eta) \leq K n^{-1-\delta}, \forall n \geq n_0, \tag{2.16}
\]
(uniformly in \( \theta \) in a neighborhood of \( \theta_0 \)), eliminates the problem.

Similarly, in the parametric case [see (2.6)], \( \sum_{i=1}^{n} (X_i - \frac{1}{2}(\theta_0 + \theta_1)) \) is a sum of independent (i.d) r.v.'s, so that the process has independent increments.

On the other hand, for nonparametric or robust statistics, \( n(T_n - \frac{1}{2}(\theta_0 + \theta_1)) \) may not have, in general, independent (and homogeneous) increments. This drawback can largely be avoided by providing suitable martingale approximations for \( n(T_n - \theta) \) (when \( \theta \) holds), and using then the standard martingale theorems as in the parametric case. As such, we may need a similar rate of convergence results (as in (2.16)) for \( n(T_n - \theta) \) and the martingale part \( T_{n, \theta}^\infty \). Further, on \( T_{n, \theta}^\infty \), we may need the usual condition that
\[
n^{-1} E_\theta (T_{n, \theta}^\infty 2) \to \sigma_\theta^2, \text{ as } n \to \infty. \tag{2.17}
\]
\[
n^{-1} E_\theta (\max_{k \leq n} |T_{k, \theta}^\infty - T_{k-1, \theta}^\infty|^2) \to 0, \text{ as } n \to \infty. \tag{2.18}
\]

and, for some \( r > 2 \),
\[
\lim \{n^{-r/2} E_\theta |T_{n, \theta}^\infty|^r\} < \infty. \tag{2.19}
\]

all uniformly in \( \theta \) in some neighborhood of \( \theta_0 \). Under these additional regularity conditions, (2.9)-(2.10) extend to a broad class of nonparametric and robust statistics [viz., Sen (1981, 1985)]. The asymptotic theory works out quite well for small to moderate values of \( \Delta \).

Let us now illustrate this methodology with some special type of nonparametric or robust tests. First, we consider the case of Hoeffding's (1948) U-statistics in a slightly general form treated in Sen (1977). Let \( \theta = \theta(F) \) be an estimable parameter [i.e., \( \theta(F) = \int \cdots \int \phi(x_1, \ldots, x_m) dF(x_1) \cdots dF(\ell_m) \), for some kernel \( \phi(\cdots) \) of degree \( m \)], and set \( \xi \)
- 13 -

where the form of \( g(\cdot) \) is known. Consider then the null hypothesis 

\[ H_0 : \xi = \xi_0 \text{ against an alternative } H_1 : \xi = \xi_1 = \xi_0 + \Delta, \text{ for some specified } \xi_0 \text{ and } \Delta(>0). \]

Note that

\[ \theta(F) = E_F \phi(X_1, \ldots, X_m), \quad \forall F \in \mathcal{F}, \tag{2.20} \]

and a symmetric, unbiased and optimal estimator of \( \theta(F) \) is

\[ U_n = \binom{n}{m}^{-1} \sum_{i=1}^{n} \phi(X_{i_1}, \ldots, X_{i_m}), \quad n \geq m. \tag{2.21} \]

Thus, for \( \xi \), a natural estimator is \( g(U_n) = T_n \), say. Let \( T^{(1)}_{n-1} \) be the estimator \( (g(U^{(1)}_{n-1})) \) based on a sample of size \( n-1 \) (deleting the ith observation \( X_i \) from the set \( X_1, \ldots, X_n \)), and let \( T_{n,i} = nT_n - (n-1)T^{(1)}_{n-1}, \quad i=1, \ldots, n \). Let then

\[ s_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (T_{n,i} - T_n)^2, \quad n \geq m+1, \tag{2.22} \]

where \( T_n = n^{-1} \sum_{i=1}^{n} T_{n,i} \) is the jackknifed version of \( T_n \). Thus, \( s_n^2 \) is the usual jackknifed variance estimator. As in Sen (1977), \( s_n^2 \) can be represented in terms of a set of U-statistics of degrees \( 2m-1, \ldots, 1 \) and some coefficients depending on \( g(\cdot) \). As such, it follows as in Sen (1977) that whenever \( g'(\theta) < \infty \) and \( E_F \phi^{2} < \infty \), as \( n \) increases,

\[ |s_n^2 - n \text{ Var}(T_n)| \rightarrow 0 \text{ almost surely}, \tag{2.23} \]

where \( n \text{ Var}(T_n) \rightarrow (g'(\theta))^2 n^{-2} \xi_1 \) and \( \xi_1 \) is an estimable parameter of degree \( 2m \). Further, the weak invariance principle for the usual U-statistics, established by Miller and Sen (1972), hold for the \( T_n \) as well [Sen (1977)].

This will ensure that for any fixed \( \Delta(>0) \), the sequential procedure based on \( (T_n, s_n^2) \) terminates with probability one, and the OC function (for \( \Delta \rightarrow 0 \)) converges to the limit in (2.8). To verify (2.16), we may need that \( E_F |\phi|^r \)

\[ < \infty, \text{ for some } r > 4. \]

Finally, we may use the well-known Hoeffding (1961) decomposition of \( U_n \) (in terms of a linear combination of martingales), and hence, a Taylor's expansion on \( g(U_n) \) (around \( g(\theta) \)) will lead to (2.17)
(2.19). As such, the SLRT type procedure extends directly to a broader class of (functions of) Hoeffding's U-statistics whenever \( A \) is small. Because of the close affinity of von Mises' functions \( V_n = n^{-m} \sum_{i=1}^{n} \sum_{i=1}^{m} \phi(X_{i_1}, \ldots, X_{i_m}) \) to \( U_n \) [viz., Ch. 3 of Sen (1981)], one may use the \( T_n^{**} = g(V_n) \) instead of the \( T_n \) (and the corresponding jackknifed variance estimator \( s_n^{**2} \) instead of \( s_n^2 \)), and obtain a parallel test procedure sharing the same asymptotic properties.

Let us next consider a more general class of functions which may not be as much structured as \( U_n \) or \( V_n \). Consider the case of a functional \( \theta(F) \) of the underlying d.f. \( F \). Since the sample d.f. \( F_n \) is a consistent, unbiased and efficient estimator of \( F \), it may be quite natural to base a sequential test for \( H_0 : \theta(F) = \theta_0 \) vs. \( H_1 : \theta(F) = \theta_1 = \theta_0 + \Delta \) on the sequence \( T_n = \theta(F_n), n \geq n_0 \). Under certain "smoothness" condition on \( \theta(\cdot) \), an expansion of \( \theta(F_n) \) around \( \theta(F) \) yields a representation of the form

\[
T_n = \theta(F_n) = \theta(F) + n^{-1} \sum_{i=1}^{n} \theta'_i(F; X_i) + \text{remainder term} \tag{2.24}
\]

where \( \theta'_i(F; X_i) \) is the so called influence function. If we let \( \|F_n - F\| = \sup\{|F_n(x) - F(x)| : x \in \mathbb{R}\} \), then under appropriate regularly conditions, it is known that in (2.24), (remainder term)/\( n\|F_n - F\| \to 0 \), as \( n \to \infty \), so that

\[
n(T_n - \theta(F)) = \sum_{i=1}^{n} \theta'_i(F; X_i) + o(n\|F_n - F\|), \tag{2.25}
\]

where by the well known results on the classical Kolmogorov-Smirnov statistics \( n\|F_n - F\| = O(1) \), in probability/rth mean, for every \( r > 0 \). There are diverse regularity conditions under which (2.25) holds. Among these, the compact or Hadamard-differentiability of \( \theta(\cdot) \) at \( F \) deserves special mention. Such a condition can be verified and it ensures the validity of (2.25). Note further that
\(|k[F_k(F^{-1}(u))-u], \ 0 \leq u \leq 1, \ k \geq 1\)

(2.26)

Kiefer process,

\[
\begin{aligned}
&
\end{aligned}
\]

and this provides suitable invariance principles for the \(n||F_n-F||_n \geq n_0\).

Similarly, if we assume that \(\theta_1(F;X_1)\) is square integrals (it has mean 0),
then the usual Wiener process embedding holds for the first term on the
right hand side of (2.25). All those can be invoked to establish suitable
martingale approximation for the left hand side of (2.25). Next to estimate
\(\sigma^2 = E[\{\theta_1(F;X_1)\}^2]\), we may as well use the classical jackknifed variance
estimator [viz., Sen (1988)]. Defining this estimator as in (2.21), we have
under no extra regularity conditions that \(s_n^{*2} \to \sigma^2\) a.s., as \(n \to \infty\). For
(2.16) to hold, we may similarly assume that for some \(r > 4, E[|\theta_1(F;X_1)|^r] < \infty\).
Having obtained these results, there is no problem in verifying
(2.16)-(2.19). As such, the asymptotic theory of sequential tests
formulated before extends to such Hadamard differentiable statistical
functionals under very comparable regularity conditions. As special cases
of such functionals, we may refer to the following:

(a) **Trimmed mean.** Let \(X_{n:1} \leq \ldots \leq X_{n:n}\) be the ordered r.v.'s
    corresponding to \(X_1, \ldots, X_n\). For some \(\alpha : 0 < \alpha < \frac{1}{2}\), on letting
    \(k = k_n = \lfloor n\alpha \rfloor\), an \(\alpha\)-trimmed mean is defined as

\[
\bar{X}_n(\alpha) = (n-2k)^{-1} \sum_{j=k+1}^{n-k} X_{n:j}
\]

(2.27)

If the distribution of \(X\) is symmetric about \(\theta\), then \(\bar{X}_n(\alpha)\) is an unbiased,
robust estimator of \(\theta\). It is also a "smooth" functional in the sense
described before.

(b) **Winsorized Mean.** This is defined by

\[
\bar{X}_n(\alpha) = n^{-1}\{k(X_{n:k} + X_{n:n-k+1}) + \sum_{j=k+1}^{n-k} X_{n:j}\}
\]

(2.28)
It shares the robustness properties along with the trimmed mean.

(c) **BLUE: L-functionals/L-estimators.** Among the class of linear functions of the order statistics, one may choose a particular one having the minimum variance — in the literature this is known as the BLUE. Basically, we have an estimator

\[ L_n = \sum_{j=1}^{n} c_{nj} X_{n:j} \tag{2.29} \]

where the \( c_{nj} \) depend on \((n,j)\) and some other conditions pertaining to the population density function. If we let \( c_{nj} = n^{-1} \phi_n(j/n) \), \( 1 \leq j \leq n \), where \( \phi_n(u) \), \( 0 < u < 1 \) is a suitable score function, then we may rewrite \( L_n \) as \( \int x \phi_n(F_n(x)) dF_n(x) \), and hence, it is termed a linear (L-) functional or L-estimator. Inducing good amount of robustness, asymptotically optimal choice of \( \phi_n(\cdot) \) can be made, and, generally, such functionals are Hadamard-differentiable under quite general regularity conditions.

(d) **M- and R-estimators of location/regression.** For a symmetric d.f. \( F \) (around \( \theta \)) and a skew-symmetric score function \( \psi(x) \), \( x \in \mathbb{R} \), the solution to the equation

\[ \int \psi(x-t) dF(x) = 0 \tag{2.30} \]

is given by \( t = \theta \). As such, choosing a suitable \( \psi \), one may estimate \( \theta \) by equating

\[ n^{-1} \sum_{i=1}^{n} \psi(X_{i} - t) = 0; \tag{2.31} \]

the solution \( \hat{\theta}_n \) [of (2.31)] is termed an M-estimator of \( \theta \). Under quite general regularity conditions, for \( \hat{\theta}_n \), an asymptotic representation as in (2.25) works out well. Similarly, using a signed-rank statistic on the \( X_{i} - t \) and equating that to 0, one obtains an R-estimator of location. These are both robust estimators. For R-estimators, (2.25) may need a bounded score function, and this can be avoided in the alternative method to be considered in Section 2.2.
2.2. Sequential tests based on aligned rank and robust statistics. In (2.12), we have actually mimicked (2.1)-(2.2) with the role of $\sum_{i=1}^{n} X_i$ being replaced by $n T_n$, $n \geq n_0$. An alternative look at (2.1) reveals that on defining the aligned observations $X_i - \frac{1}{2}(\theta_0 + \theta_1)$ as $X_i^*, i \geq 1$, log $\lambda_n = (\Lambda/\sigma)(\sigma^{-1} (X^*_1 + \ldots + X^*_n))$, $n \geq n_0$. This immediately suggests that instead of $\sigma^{-1} (X^*_1 + \ldots + X^*_n)$, some robust (viz., rank) statistic can be used and a sequential testing procedure can be developed along the same line as in Section 2.1. This was the motivation of sequential rank tests for location considered by Sen and Chosh (1974); see also Chapter 9 of Sen (1981) for some general developments. We again illustrate such aligned sequential tests with the help of rank tests and M-tests, which are known to be robust.

Let us consider the model: $X_i, i \geq 1$, i.i.d.r.v.'s with the d.f. $F(x-\theta)$, where $\theta$ is real and $F$ is symmetric about 0. As in before (2.1) considered $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1 = \theta_0 + \Lambda$, $\Lambda > 0$, where $\theta_0$ and $\Lambda$ (or $\theta_1$) are both specified (but not the form of $F$). Define

$$X^*_1 = X_1 - \frac{1}{2}(\theta_0 + \theta_1), i \geq 1,$$

$$S^*_n = \sum_{i=1}^{n} \text{sign} X^*_1 a_n(R^*_n), n \geq 1,$$

where $a_n(1) \leq \ldots \leq a_n(n)$ are suitable scores and

$$R^*_n = \sum_{j=1}^{n} I(|X^*_j| \leq |X^*_1|), i=1,\ldots,n$$

are the ranks of the absolute aligned observations. If we choose $a_n(i) = i/(n+1), 1 \leq i \leq n$ (Wilcoxon scores), $S^*_n$ in (2.33) reduces to the (aligned) Wilcoxon signed rank statistic. Similarly, if $a_n(1) = \text{expected value of the}$ $i$th order statistic in a sample of size $n$ from the chi d.f. with 1 degree of freedom (1 $\leq i \leq n$), then $S^*_n$ is the normal scores signed rank statistic. For $a_n(k) = 1, V_k$, (2.33) reduces to the sign statistic. Under (2.14), as $\Lambda \to 0$, $[(S^*_n - \gamma (\varphi - \frac{1}{2})n\Lambda)/\sigma^*], n \geq n_0$ (where $\sigma^* = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} a^2_n(1)$) converges in law to a standard Brownian motion [on $R^+$], where
\[ \gamma = \gamma(F) = \text{a functional of } F \text{ depending on the} \]

score generating function. \hfill (2.35)

For example, for the Wilcoxon case, \( \gamma = \int_{-\infty}^{\infty} f^2(x) \, dx \), where \( f \) is the density function corresponding to the d.f. \( F \). Further, \( D_n \), an estimator of \( \gamma \) (based on \( X_1, \ldots, X_n \)) can also be obtained by considering a confidence interval for \( \theta \) based on \( S_n^w \). [This will be considered in the next section.] Finally, note that \( \sigma^2 \) is a known quantity. With these notations, let us define

\[ Z_n = \Lambda \frac{D_n}{S_n^w / \sigma^2} \quad n \geq n_0. \] \hfill (2.36)

Then parallel to (2.12), we define here the stopping number \( N \) as

\[ N_\Lambda = \min\{n \geq n_0 : Z_n \in (b, a)\}. \] \hfill (2.37)

and we accept \( H_0 \) or \( H_1 \) according as \( Z_n \) is \( \leq b \) or \( \geq a \). It may be noted that under fairly general regularity conditions, \( D_n \to \gamma \), in probability (as \( n \to \infty \)), and also, when \( \theta = \theta_0 + \frac{1}{2} \Lambda \), \( n^{-\Lambda} S_n^w \) converges in law to a normal variable with zero mean and variance \( \sigma^2 \), while \( n^{-\Lambda} b \Lambda \to 0 \) and \( n^{-\Lambda} a \Lambda \to 0 \) as \( n \to \infty \) (for any given \( \Lambda > 0 \)), so that for \( \theta = \theta_0 + \frac{1}{2} \Lambda \), process terminates with probability 1. For \( \theta \neq \theta_0 + \frac{1}{2} \Lambda \), we may use the fact that \( n^{-1} S_n^w \) converges in probability to a nonzero quantity (depending on \( \theta \)), while \( n^{-1} a \) and \( n^{-1} b \) both converge to 0, as \( n \to \infty \). Hence again the process terminates with probability one. Further, by virtue of the asymptotic linearity results studied by Jurecková (1971), Sen and Ghosh (1971, 1974) and others, it follows that as \( \Lambda \to 0 \), over the range of \( n : n \Lambda^2 = O(1) \), for \( \theta = \theta_0 + \varphi \Lambda \),

\[ \Delta[S_n^w - S_n^0 + n(\varphi - \frac{1}{2}) \Lambda \gamma] \Rightarrow 0, \] \hfill (2.38)

where \( S_n^0 \) is the signed-rank statistic based on the \( X_i^0 = X_i - \frac{1}{2}(\theta_0 + \theta_1) \), \( i = 1, \ldots, n \), and where for \( S_n^0 \), the weak convergence to a (zero-mean) Brownian process holds [viz., Sen (1981)]. As such, the limiting (as \( \Lambda \to 0 \)) OC function is again given by (2.8), and under an additional uniform integrability condition, for the ASN, (2.9)-(2.10) hold with \( \sigma^2 \) replaced by
\( \sigma^2 \gamma^2 \). Thus, the usual (aligned) signed-rank statistics may be incorporated to construct suitable sequential tests, and these will be more robust than their parametric competitors. To stress this robustness aspect, let us side by side, consider sequential tests based on M-estimators and aligned M-statistics. Instead of the aligned signed-rank statistic in (2.33), we consider an aligned M-statistic
\[
M_n^* = \sum_{i=1}^{n} \psi(x_i - \frac{1}{2}(\theta_0 + \theta_1)), \quad n \geq 1.
\]
where \( \psi \) is a skew-symmetric function (i.e., \( \psi(x) + \psi(-x) = 0 \), \( \forall x \)): the specific Huber-score function is given by
\[
\psi(x) = \begin{cases} 
 x \text{ if } |x| \leq K \\
 k \text{ sign } x, \text{ if } |x| > k,
\end{cases}
\]
where \( K \) is a prefixed positive number. Usually \( \psi \) is chosen as a bounded function (to curtail the influence of outliers). Also, usually \( \psi \) is chosen as monotone, although there are notable examples where this monotonicity is not the case. With the M-estimator \( \hat{\theta} \) (or \( \theta \)) defined in (2.31), we also let
\[
s_n^* = n^{-1} \sum_{i=1}^{n} \psi^2(x_i - \hat{\theta}_n);
\]
\[
D_n^* = n^{-1} \sum_{i=1}^{n} [\psi(x_i - \hat{\theta}_n + a_n) - \psi(x_i - \hat{\theta}_n - a_n)]/2a_n,
\]
where \( a_n \) is an arbitrary positive number such that \( a_n = o(n^{-\frac{1}{2}}) \), as \( n \to \infty \). We may take \( a_n = n^{-\frac{1}{2}} s_n^* \). Then, we take
\[
Z_n^* = \Delta_n^* D_n^* M_n^*/s_n^* \gamma^2, \quad n \geq n_0
\]
and proceed as in the case of sequential signed-rank test. The termination probability, limiting OC function and ASN are the same (with \( \sigma^2 \) being replaced by the asymptotic variance of \( n^\frac{3}{2}(\hat{\theta}_n - \theta) \)). In the case of aligned signed-rank statistics, we need to estimate only one unknown parameter (\( \gamma \) by \( D_n \)), while in the case of aligned M-statistics, we need two estimators \( D_n^* \).
and $s_{2n}^2$. However, from robustness considerations, often, suitable (aligned) sequential $M$-test may perform better than the (aligned) sequential signed-rank tests. The problem of testing regression parameters (including the two sample location model as a special case) based on suitable aligned rank or $M$-statistics has also been treated in the literature; we may refer to Sen (1981) for some account of these developments.

We conclude this section with the remark that (2.9)-(2.10), as extended to a wider class in Sections 2.1 and 2.2, justify the adoption of the classical Pitman efficiency measure in the sequential case. However, (2.10) for such nonparametric or robust methods may require extra regularity conditions for the limit to exist. These may be avoided by comparing the limiting distributions of $(A^2 N_A)$ for two competing methods. This leads to the same definition of the (asymptotic) Pitman efficiency in the sequential case. For details, we may refer to Sen (1981, Ch. 9).
3. MINIMUM RISK SEQUENTIAL NONPARAMETRIC ESTIMATION

Let \( \{X_i, \ i \geq 1\} \) be a sequence of i.i.d.r.v.'s with a distribution function (d.f.) \( F \), defined on \( \mathbb{E}^p \), for some \( p \geq 1 \). In a parametric model, the functional form of \( F \) is assumed to be known, and the unknown algebraic constants associated with this form are regarded as parameters. In a nonparametric setup, \( F \) is of unknown form and it is assumed only that \( F \) belongs to a suitable family (\( \mathcal{F} \)) of d.f.'s. For example, \( \mathcal{F} \) may be the class of all continuous \( F \) and \( \mathbb{E}^p \), or all \( F \) (diagonally) symmetric about an origin, which is taken as the location parameter of \( F \). In general, in a nonparametric formulation, we take

\[
\theta = \theta(F) = \text{a functional of the d.f. } F; \quad (3.1)
\]

as in Section 2, the mean, variance, moments etc. are all particular cases of \( \theta(F) \) in (3.1).

Based on a sample \( \{X_1, \ldots, X_n\} \), of size \( n \), let \( T_n = T(X_1, \ldots, X_n) \) be a nonparametric estimator of \( \theta(F) \), i.e., \( T_n \) is not based on any specific form of \( F \). If \( F_n \) be the sample (empirical) d.f. (based on \( X_1, \ldots, X_n \)), then \( F_n \) is a natural (nonparametric) estimator of \( F \), so that one may choose \( T_n = \theta(F_n) \) as a natural estimator of \( \theta \); this is usually termed a von Mises functional. There are other estimators [such as the U-statistics], which we have briefly presented in Section 2. It is, however, not necessary to restrict ourselves only to \( \theta(F_n) \) or some U-statistics. The mean square error (MSE) of \( T_n \), whenever exists, is defined by

\[
v_n(F) = E_F[(T_n - \theta(F))^2], \ n \geq n_0, \ F \in \mathcal{F}, \quad (3.2)
\]

where \( n_0 \) is the smallest sample size for which \( v_n(F) \) exists. Such an \( n_0 \) may generally depend on \( F \). Instead of the MSE, we might have considered the mean absolute deviation (MAD) of \( T_n \) (i.e., \( E_F|T_n - \theta(F)| \)) or some other
measure of average variation of $T_n$ from $\theta(F)$. For simplicity, we consider only the case of MSE; the others can be treated similarly.

In general, $\nu_n(F)$ depends on the d.f. $F$ as well as on the sample size $n$. In reality, drawing of sample observations entails cost (of sampling), and hence, in a proper formulation of the loss function, the cost of sampling should be emphasized. For this reason, we take the loss of estimating $\theta(F)$ by $T_n$ as

$$\mathcal{G}(T_n, \theta(F)) = (E_F L(T_n, \theta(F)))$$

$$= \nu_n(F) + cn, \ n \geq n_0. \quad (3.4)$$

From what has been discussed after (3.2), it follows that, in general, (3.3) is a function of $c, n$ and the unknown d.f. $F$. For every $c > 0$, let us define

$$n_0^c = \min\{m \geq n_0 : \mathcal{G}(T_m, \theta(F)) = \inf \mathcal{G}(T_n, \theta(F))\}. \quad (3.5)$$

so that $n_0^c$ is an optimal sample size in the sense that $T_{n_0}^0$ has the minimum risk (MR) property in the sense of (3.5). We term $T_{n_0}^0$ a MRE (estimator) of $\theta(F)$.

Since the $\mathcal{G}(T_n, \theta(F))$ depend, in general, on the unknown $F(\in \mathcal{F})$, $n_0^c (= n_0^c(F))$ is also dependent on the unknown $F$. As such, based on any specific sample size, say $n^\ast$, $T_{n^\ast}$ may not have the MRE property, simultaneously for all $F \in \mathcal{F}$. Also, the minimum risk $\mathcal{G}(T_{n_0}^0, \theta(F))$ depends on $c$ as well on the unknown $F$, and hence, a fixed sample size estimation procedure may not work out. For this reason, a sequential procedure may be sought to reach for a solution (in, at least, some well defined asymptotic sense). Even in the parametric case (viz., the mean of a normal distribution with unknown variance), a fixed sample size procedure may not work out, and hence, in a nonparametric setup, the prospect for a fixed
sample size procedure seems to be even less bright.

As is the case with Hoeffding's (1948) U-statistics or von Mises' differentiable statistical functionals, for a general nonparametric estimator \( T_n \), under quite general regularity conditions, we have

\[
v_n(F) = E_F(T_n - \theta(F))^2 = n^{-1} v(F) + \text{terms of the order } o(n^{-1}),
\]

(3.6)

where \( v(F) \) is itself an (unknown) functional of the d.f. \( F \). Thus, in (3.4), we have

\[
\mathcal{J}(T_n, \theta(F)) = n^{-1} v(F) + cn + o(n^{-1}).
\]

(3.7)

so that when \( c \), the cost per unit sampling, is small, we obtain that

\[
n_c^0 \sim \{c^{-1} v(F)\}^{1/2} \quad \text{and} \quad \mathcal{J}(T_{n_c^0}, \theta(F)) \sim 2\{c^{-1} v(F)\}^{1/2},
\]

(3.8)

where \( a \sim b \) means that the ratio \( a/b \) converges to 1. Thus, to estimate \( \theta(F) \) in an optimal way (when \( c \) is small), one possibility is to incorporate suitable nonparametric estimators of \( v(F) \) in an usual framework of a stopping rule on which a proper sequential estimation procedure would rest. We shall discuss this approach here.

Suppose that based on \( X_1, \ldots, X_n \), \( V_n \) is a suitable nonparametric estimator of \( v(F) \), \( n \geq n_0 \), and suppose that

\[
V_n \to v(F) \quad \text{a.s., as } n \to \infty \quad (\forall \, F \in \mathcal{J}).
\]

(3.9)

Then, motivated by (3.8) and (3.9), for every \( c > 0 \), we may formulate, a stopping number \( N_c \) as

\[
N_c = \min\{n \geq n_0 : n^2 \geq c^{-1}(V_n + n^{-h})\},
\]

(3.10)

where \( h \) is a positive number, and the incorporation of \( n^{-h} \) is primarily to prevent a very early stopping should \( V_n \) be very small. Then, for every \( c > 0 \), we have the sequential estimator \( T_{N_c} \) whose risk is given by
\( \mathcal{F}^*(c, F) = \mathcal{F}(T_{N_c}^*, \theta(F)) = E\{ (T_{N_c}^* - \theta(F))^2 \} + c E N_c, \quad c > 0. \) \hspace{1cm} (3.11)

where \( E \) stands for \( E_F \), expectation under \( F \). We say that the nonparametric sequential estimator \( T_{N_c}^* \) is asymptotically (as \( c \downarrow 0 \)) minimum risk estimator (AMRE) (of first order) if
\[
\lim_{c \downarrow 0} \left( \frac{\mathcal{F}^*(c, F)}{\mathcal{F}(T_{N_c}^*, \theta(F)))} \right) = 1, \quad \forall F \in \mathcal{F}. \hspace{1cm} (3.12)
\]

Indeed, for a general class of parameters \((\theta(F))\), (3.12) holds under fairly general regularity conditions. A general review of this first order AMRE in a nonparametric setup is given in Sen (1981, Chp. 10) [see also Sen (1985, Ch. 4)]. The past few years have witnessed some further developments in this arena encompassing general multiparameter estimation problems and incorporating the Stein-rule estimation theory. We shall therefore review the AMRE procedures in this general framework.

We consider a multi-parameter setup, where \( \bar{\theta}(F) = (\theta_1(F), \ldots, \theta_p(F))' \) and \( T_n = (T_{n1}, \ldots, T_{np})' \), for some \( p \geq 1 \), where \( T_n \) is based on the sample \((X_1, \ldots, X_n)\) of size \( n \), and the \( X_i \) are \( q \)-dimensional \( r \)-vectors for some of \((\geq 1)\). Then, the squared error loss in (3.3) extends in a natural manner to a quadratic error loss
\[
L(T_n, \bar{\theta}(F)) = \| T_n - \bar{\theta}(F) \|_Q^2 + cn
\]
\[
= (T_n - \bar{\theta}(F))' Q(T_n - \bar{\theta}(F)) + cn. \hspace{1cm} (3.13)
\]

where \( Q \) is a given positive definite (p.d.) matrix. Thus, the risk in estimating \( \bar{\theta}(F) \) by \( T_n \) is given by
\[
\mathcal{F}_c(T_n, \bar{\theta}(F)) = cn + E_F[(T_n - \bar{\theta}(F))' Q(T_n - \bar{\theta}(F))]
\]
\[
= cn + \text{Trace}(Q E_F[(T_n - \bar{\theta}(F))(T_n - \bar{\theta}(F))'])
\]
= cn + \text{Trace}(Q \, \nu_n(F)), \text{ say.} \quad (3.14)

where \( \nu_n(F) \) is the dispersion matrix of \( \mathcal{T}_n \). This definition allows us to extend (3.5) to the multi-parameter case as well. As in (3.6), we may write, under fairly general regularity conditions,

\[
u_n(F) = n^{-1} \, \nu(F) + o(n^{-1}), \quad n \geq n_0, \quad (3.15)\]

where \( \nu(F) \) is a \( p \times p \) matrix of unknown functionals of \( F \). Then, the right hand side of (3.14) reduces to \( cn + n^{-1} \text{Trace}(Q \, \nu(F)) + o(n^{-1}) \), so that parallel to (3.8), we have here (as \( c \downarrow 0 \)),

\[
\frac{c^0}{c} \sim \{n^{-1} \text{trace}(Q \, \nu(F))\}^{1/2} \quad \text{and} \quad \mathcal{G}(\mathbb{T}_n^0, \mathbb{G}(F)) \sim 2(c \text{trace}(Q \, \nu(F)))^{1/2} \quad (3.16)
\]

As in (3.9), we may construct suitable nonparametric estimators of \( \nu(F) \) (denote by \( V_n \)), such that (3.5) holds (coordinatewise). Then, we may introduce the stopping number

\[
N_c = \min\{n \geq n_0 : n^2 \geq c^{-1} \{\text{trace}(Q \, \nu_n) + n^{-h}\}, \quad c > 0. \quad (3.17)
\]

For the sequential estimator \( \mathcal{T}_{N_c} \), as in (3.11), we define

\[
\mathcal{G}(c, F) = c \, E \, N_c + E \{\|T_{N_c} \mathbb{G}(F)\|^2\}, \quad c > 0. \quad (3.18)
\]

With these notations, the definition of the first order AMRE property in (3.12) extends to the multiparameter case directly. Thus, for this multi-parameter AMRE problem in a nonparametric setup, the basic problem is to construct suitable estimators \( \{\nu_n\} \) of \( \nu(F) \), for which (3.9) and (3.12) both hold. We illustrate this by reference to some important class of estimators. We shall mainly emphasize on the classical jackknife method of estimating \( V_n \), although some other methods (including bootstrapping) may as well be used in some of these problems.
From the base sample $(X_1, \ldots, X_n)$ (of size $n$), we drop the $i$th observation $(X_i)$, and denote the estimator based on this subsample of size $n-1$ by $\hat{T}_{n-1}^{(i)}$ for $i=1, \ldots, n$. We then define the vector of pseudovariables by

$$T_{n, i} = n \hat{T}_n - (n-1) \hat{T}_{n-1}^{(i)}, \quad i=1, \ldots, n. \quad (3.19)$$

Let

$$\bar{T}_n^n = n^{-1} \sum_{i=1}^n T_{n, i}; \quad (3.20)$$

$$V_n^n = (n-1)^{-1} \sum_{i=1}^n (T_{n, i} - \bar{T}_n^n)(T_{n, i} - \bar{T}_n^n)' \quad (3.21)$$

Then, $\bar{T}_n^n$ is the classical jackknifed version of $\hat{T}_n$ and is known to have a smaller bias. $V_n^n$ is the jackknifed covariance (matrix estimator of $\gamma$). In (3.17), under quite general regularity conditions, $V_n^n$ can be replaced by $V_n^n$ and then the first order AMRE property in (3.12) holds.

For U-statistics and von Mises' functionals, Sen and Ghosh (1981) showed that for the jackknifed variance estimator $V_n^n$ (in the single parameter case) (3.9) holds, and their treatment readily extends to the multiparameter case as well. Actually, $V_n^n$ can itself be expressed as a linear combination of several U-statistics, each of which (by virtue of the reverse martingale property) possesses the a.s. convergence property. Further, if the underlying kernel has finite moments up to the order $r$, for some $r > 4$, then (3.12) holds. The a.s. convergence result for $V_n^n$ has also been established for a general class of (Hadamard or compact-) differentiable statistical functionals, and this class contains $M$-estimators of location (with bounded score functions) as well as some $L$-estimators of location [viz., Sen (1988)]. It also includes $R$-estimators of location with bounded score functions. But, for unbounded score functions, generally, $M$- and $R$-estimators do not belong to this class, so that a different approach is needed. For $R$-estimator of location, in the uniparameter case, $v(F)$ in
(3.6) is defined by
\begin{equation}
v(F) = (\int_0^1 \phi^2(u) \, du) / (\int_{-\infty}^\infty f(x) \, d\Phi^*(F(x)))^2, \tag{3.22}
\end{equation}
where \( \phi = \{\phi(u) = \phi^* \left[ \frac{1+u}{2} \right], \quad 0 \leq u < 1 \} \) is the score generating function (known) and the density function \( f \) (as well as \( F \)) are unspecified. Thus, to estimate \( v(F) \) consistently, we need to estimate \( \gamma = \int_{-\infty}^\infty f(x) \, d\Phi^*(F(x)) \, dx \) consistently. This does not entail any serious restriction on \( F \) (or \( \phi \)) – the asymptotic linearity of signed-rank statistics in shift parameter provides a convenient means of estimating \( \gamma \). This has been exploited in Chapter 10 of Sen (1981). In the case of multivariate location vector, \( y(F) \) can be written as
\begin{equation}
y(F) = \frac{1}{D} \Gamma(F) \frac{1}{D}, \tag{3.23}
\end{equation}
where \( D = \text{Diag}(d_1, \ldots, d_p) \), the \( d_j \) are \( \gamma \)-functionals for the marginal densities \( f_j \) and score functions \( \phi_j \), where \( \Gamma(F) \) is the generalized grade covariance matrix – it can be very conveniently estimated by the permutational covariance matrix of the signed-rank statistics. Also, the coordinate-wise estimates of the \( d_j \) can be readily incorporated in the estimation of \( D \). The details have been worked out in Sen (1984).

For \( M \)-estimator of location, in the uniparameter case, \( v(F) \) in (3.6) is defined by
\begin{equation}
v(F) = \int_{-\infty}^\infty \psi^2(x) \, dF(x) / (\int_{-\infty}^\infty f(x) \, d\psi(x))^2, \tag{3.24}
\end{equation}
where \( \psi = \{\psi(x), \quad -\infty < x < \infty \} \) is the score function. Again, the numerator is consistently estimated by \( n^{-1} \sum_{i=1}^n \psi^2(x - \hat{\theta}_n) \), while the denominator can be consistently estimated by using the asymptotic linearity of \( M \)-statistics in shift parameter. [See (2.41) and (2.42) in this context]. In the multivariate location model, as in the case of \( R \)-estimators, for the marginal densities \( f_j \), the functionals \( \int f_j \, d\psi_j, \quad 1 \leq j \leq p \), can be estimated consistently coordinate-wise, while the covariance matrix \( E[\psi \psi'] \) can also be
estimated from the residual vectors. Hence, the AMRE property holds under parallel regularity conditions.

There is an interesting feature of this multiparameter estimation problem. To motivate the problem, let us consider the simple parametric model where a p-vector \( \mathbf{X} \) has a multinormal distribution with unknown mean vector \( \mathbf{\theta} \) and a known dispersion matrix, which we may take without any loss of generality as \( \mathbf{I}_p \). Then \( \mathbf{X} \) is the usual maximum likelihood estimator (MLE) of \( \mathbf{\theta} \). Suppose now that we consider a quadratic risk function \( \mathbb{E}_\mathbf{\theta}(Q(\mathbf{T}\mathbf{-}\mathbf{\theta})) \)

where \( \mathbf{T} \) is an estimator of \( \mathbf{\theta} \) and \( Q \) is a given positive definite matrix. Then Stein (1966) has shown that for \( p > 3 \), \( \mathbf{X} \) does not have the smallest risk, and there exists some other estimators which dominate \( \mathbf{X} \) (for all \( \mathbf{\theta} \)); such estimators are known as the Stein-rule or shrinkage estimators. This Stein phenomenon holds even when the dispersion matrix of \( \mathbf{X} \) is of the form \( \sigma^2 \mathbf{Y} \) where \( \mathbf{Y} \) is known and \( \sigma^2 \) is unknown, or the more general case of an arbitrary covariance matrix \( \mathbf{\Sigma} \); we may refer to Chapter 20 [by M. Ghosh] for some of these details. Incorporating the cost of sampling, for a sample of size \( n \), we may take the risk function as in (3.13) and formulate a stopping rule as in (3.17). A natural question arises in this context: If we consider a shrinkage version of the sequential estimator \( \mathbf{T}_{N_C} \), then can \( \mathbf{T}_{N_C} \) be dominated by its shrinkage version? An affirmative answer to this query has been provided by Ghosh, Nickerson and Sen (1987), for the multivariate normal mean problem. This (exact) dominance result extends to a much wider class of estimation problem in the same asymptotic setup treated before.

Corresponding to an estimator \( \mathbf{T}_n \) of a parameter-vector \( \mathbf{\theta}(\mathbf{F}) \), we define the jackknifed dispersion matrix estimator by \( \mathbf{Y}_{\mathbf{n}}^* \) [See (3.21)]. Further, for the given (p.d.) matrix \( Q \), we define

\[
d_n = \text{smallest characteristic root of } Q \mathbf{Y}_{\mathbf{n}}^*. \tag{3.25}
\]
Also, let
\[ \mathcal{L}_n = n(\mathcal{I}_n)'(\mathcal{I}_n^T)^{-1}(\mathcal{I}_n). \]  \hfill (3.26)

Then, we may consider the following type of shrinkage estimators.
\[ \mathcal{T}_n^S = (1 - k \frac{\mathcal{L}_n}{n}) \mathcal{L}_n^{-1} \mathcal{Q}^{-1} \mathcal{Y}^{-1} \mathcal{Y}^{-1} \mathcal{L}_n. \]  \hfill (3.27)

where \( k : 0 < k < 2(p-2) \), \( p \geq 3 \), is a positive constant (the shrinkage factor). [In (3.27), we have conceived of the null pivot (\( \mathcal{Q} \)); if we choose any other pivot (\( \theta_0 \), known), then in (3.26) and (3.27), we replace \( \mathcal{T}_n^S \theta_0 \) and replace \( \mathcal{T}_n^S \theta_0 \) by \( \mathcal{T}_n^S \theta_0 \) too.]

The presence of \( \mathcal{L}_n^{-1} \) in (3.27) may create some difficulties in the computation of the asymptotic risk (i.e., \( E_\theta[(\mathcal{T}_n^S \theta_0)^T \mathcal{Q}(\mathcal{T}_n^S \theta_0)] + cn \)). For example, if \( \mathcal{L}_n \) in (3.26) has a non-zero probability of being equal to \( 0 \) (however small it may be), the asymptotic risk of the shrinkage estimator is \( +\infty \). Even if \( P_\theta(\mathcal{L}_n = 0) = 0 \), the rate of convergence of \( P_\theta(\mathcal{L}_n \leq \ell) \) (as \( \ell \downarrow 0 \)) determines whether the asymptotic risk is finite or not; the situation is quite comparable to the estimation of the reciprocal of a binomial parameter, where the reciprocal of the sample proportion (still the MLE) does not have any finite moment of any positive order. In nonparametric shrinkage estimation theory this difficulty has been avoided by incorporating the asymptotic distribution of \( n^T(\mathcal{T}_n^S \theta_0) \) in the computation of the asymptotic risk - this is termed the asymptotic distributional risk (ADR). Thus, if \( G_\theta^\mathcal{Y}(x) \) stands for the asymptotic distribution of \( n^T(\mathcal{T}_n^S \theta_0) \) when \( \theta \) holds, then the ADR is given by \( \frac{1}{n} \text{trace}(Q" \mathcal{Y}^\mathcal{Y}) + cn \), where \( \mathcal{Y}^\mathcal{Y} = \int \int x x' dG_\theta^\mathcal{Y}(x) \) is the dispersion matrix of the asymptotic d.f. \( \mathcal{G}_\theta^\mathcal{Y} \). In the light of this ADR criterion \( \mathcal{T}_n^S \) dominates \( \mathcal{T}_n \) for a broad class of nonparametric and robust estimators. In this context, it may also be mentioned for that large \( n \), the effective domain of \( \theta \) for which there is
some improvement due to shrinkage estimation is given by $A_n = \{ \theta : n^\xi \| \theta - \theta_0 \| = O(1) \}$, where $\theta_0$ is the adopted pivot. Beyond this Pitman-neighborhood of $\theta_0$, the ADR of $T_n^S$ and $T_n$ are asymptotically the same.

The ADR criterion and the Pitman-neighborhood of the pivot both play a fundamental role in sequential shrinkage estimation. Recall that $n_c^0$, the optimal sample size for the ANRE problem, is $O(c^{-\xi})$. Keeping this in mind, we conceive of a sequence $\{A(c)\}$ of nested neighborhoods:

$$A(c) = \{ \theta : \| \theta - \theta_0 \| \leq c \}, \quad c > 0, \quad (0 < A < \infty). \tag{3.28}$$

Further, for every $c > 0$, we consider a stopping variable $N_c$, defined as in (3.17), and corresponding to the shrinkage estimator $T_n^S$ in (3.27), we adopt the sequential shrinkage version $T_{N_c}^S$. Then, for every $\theta \in A(c)$, we consider the asymptotic distribution of $c^{-\xi} [T_{N_c}^S - \theta]$ (which exists under very general regularity conditions), and incorporate the same in the computation of ADR. We repeat the same procedure for $T_{N_c}$. Then, in the light of the ADR, $T_{N_c}^S$ dominates $T_{N_c}$. A general treatment of this asymptotic dominance results for sequential shrinkage nonparametric estimators is given in Sen (1989). That study mainly deals with differentiable statistical functions for which jackknifed dispersion matrix estimators (i.e., $V_n^M$) work out well. For $R$-estimators and $M$-estimators of location (multivariate case), this approach may call for bounded score functions. Hence, alternative methods of estimating this dispersion matrix, such as the ones discussed in (2.41)-(2.42) and after (3.18), can be used to extend the scope of applicability to unbounded score functions as well.

There is one important question regarding the stopping number $N_c$ in (3.17) which we shall elaborate a bit further. If we look at (3.17), we may observe that $N_c$ has been mainly motivated by the asymptotic optimality of $n_c^0$. But $n_c^0$ does not retain its asymptotic optimality in the multivariate
case (i.e., when \( p \geq 3 \)), as it can be readily verified by working with the asymptotic risk of the corresponding Stein-rule version (which is smaller).

Thus, while considering sequential shrinkage (multi-parameter) estimators, it may be quite natural to obtain the form of the asymptotically optimal sample size (say, \( n_c^{OS} \)) and then to define the stopping number (say, \( N_c^S \)) in an analogous manner. In this context, we may note that \( n_c^0 \) does not depend on \( \hat{\theta} \), but only on the triplet \((c, Q, \psi(F))\), so that the solution is \( \hat{\theta} \)-invariant. On the other hand, the (asymptotic) dispersion matrix of a shrinkage estimator of \( \theta \) depends on \( \hat{\theta} \), even when \( \theta \) is confined to a set \( \Lambda(c) \), as in (3.28). Thus, \( n_c^{OS} \) will not only depend on \((c, Q, \psi(F))\) but also on \( \hat{\theta} (\in \Lambda(c)) \). This dependence is generally of a highly complex form, and hence, a formulation of \( N_c^S \), by analogy with \( n_c^{OS} \), will be generally quite complicated. The lack of \( \hat{\theta} \)-equivariance of \( n_c^{OS} \) (and \( N_c^S \)) may cause further difficulties too. However, the following argument provides a good justification for adopting \( N_c^S \) in sequential shrinkage estimation. Recall that from what we have discussed before,

\[
T_{N_c}^S \text{ dominates } T_{N_c}^N \text{ (in ADR), as } c \downarrow 0.
\]  

(3.29)

On the other hand, whenever \( \{N_c^S\} \) is properly defined and satisfies the basic condition that

\[
N_c^S / n_c^{OS} \to 1, \text{ in probability, as } c \downarrow 0.
\]  

(3.30)

it can be shown by similar manipulations that

\[
T_{N_c}^S \text{ dominates } T_{N_c}^N \text{ (in ADR), as } c \downarrow 0.
\]  

(3.31)

Thus, combining (3.29) and (3.31), we have

\[
T_{N_c}^S \text{ dominates } T_{N_c}^N \text{ is ADR, as } c \downarrow 0.
\]  

(3.32)

Thus, as regards the asymptotic dominance of a sequential shrinkage estimator over its classical counterpart is concerned, (3.29) provides the
key to (3.32). However, in general, the study of the extent of improvement due to using $N^S_C$ (instead of $N_C$) may stumble into extra complications because of the usually complicated form of $N^S_C$. From robustness considerations (of $s_n$ and $T_n$), the AMRE problem treated here provides good clues for choices of suitable class of estimators, and the solutions parallel the classical nonsequential case.
4. NONPARAMETRIC SEQUENTIAL CONFIDENCE SETS

As in (3.1), we conceive of a parameter \( \theta(F) \), a functional of the underlying (unknown) distribution \( F \). Our prime objective is to locate an interval, say, \( I_n \), based on the sample observations \( X_1, \ldots, X_n \) (of a sample of size \( n \)), such that

(i) \[ P_F\{ \theta(F) \in I_n \} \geq 1 - \alpha, \quad (4.1) \]

(ii) width of \( I_n \leq 2d, \quad d > 0, \quad (4.2) \]

where \( \alpha (0 < \alpha < 1) \) and \( d \) are preassigned numbers. \( 1 - \alpha \) is termed the coverage probability (or confidence coefficient) and \( 2d \) is the width of the interval \( I_n \). We refer to Chapter 11 (by J. Jurecková) for a general motivation and a good account of both parametric and nonparametric sequential confidence sets. Based on robustness considerations, we shall mainly confine ourselves to the nonparametric case.

As in Section 3, let \( T_n = T(F_n) \) be an estimator of \( \theta(F) \) based on the empirical d.f. \( F_n \) [of \( (X_1, \ldots, X_n) \)]. We assume that as \( n \to \infty \),

\[ n^{1/2}(T_n - \theta(F)) \sim \mathcal{N}(0, v^2(F)), \quad (4.3) \]

where \( v(F) : 0 < v(F) < \infty \) is itself a functional of \( F \). We also assume that there exists a sequence \( \{ V_n^X \} \) of estimators of \( v^2(F) \). [For example, we have already stressed on the jackknifed variance estimator in Section 3.] Note that (4.3) ensures that for large \( n \),

\[ P\{ n^{1/2}|T_n - \theta(F)| \leq \tau_{\alpha/2} v(F) \} \approx 1 - \alpha \quad (4.6) \]

where \( \tau_{\alpha/2} \) is the upper 50 \( \alpha\% \) point of the standard normal distribution. So that, if in (4.2), \( d(>0) \) is chosen sufficiently small, we may set

\[ n_d = \min(n : \tau_{\alpha/2}^2 v^2(F)d^{-2} \leq n), \quad (4.5) \]

and obtain that as \( d \downarrow 0 \),

\[ P_F\{ T_n_d - d \leq \theta(F) \leq T_n_d + d \} \to 1 - \alpha, \quad (4.6) \]

so that both (4.1) and (4.2) hold for the interval \( I_{n_d} = [T_{n_d} - d, T_{n_d} + d] \).
But, the very definition of \( n_d \) in (4.5) reveals that \( n_d \) depends on the unknown \( F \) through \( u(F) \). Hence, no \( n_d \) can not satisfy (4.6), simultaneously for all \( F \) belonging to a class \( \mathcal{F} \). This calls for plausible adoption of sequential procedures. In view of the assumed consistency of \( \{ V^*_n \} \) (as an estimator of \( \nu^2(F) \)) and (4.5), we may consider a stopping number

\[
N_d = \inf \{ n \geq n_0 : n d^2 \geq \tau^2 \alpha / 2 \nu^*_n \}, \quad d > 0. \tag{4.7}
\]

and define \( T_{N_d} \) by \( T_n \) for \( n = N_d \). Let then

\[
I_{N_d} = [T_{N_d} - d, T_{N_d} + d], \quad d > 0. \tag{4.8}
\]

Note that for \( I_{N_d} \), by definition, (4.2) holds. Thus, the basic problem is to show that as \( d \downarrow 0 \),

\[
P_F \{ \theta(F) \in I_{N_d} \} \rightarrow 1 - \alpha, \quad \forall F \in \mathcal{F}, \tag{4.9}
\]

for a suitable class \( \mathcal{F} \). In the literature, this is known as the asymptotic (as \( d \downarrow 0 \)) consistency property. It may also be shown that under suitable regularity conditions,

\[
N_d / n_d \rightarrow 1 \quad \text{as} \quad d \downarrow 0, \quad \text{in a suitable norm}. \tag{4.10}
\]

[for example, \( N_d / n_d \rightarrow 1 \) a.s., as \( d \downarrow 0 \) or \( E N_d / n_d \rightarrow 1 \) as \( d \rightarrow 0 \).] This property is termed the asymptotic efficiency of the sequential procedure. This shows that for \( d \) sufficiently small, \( N_d \) is 'close to' the optimal \( n_d \), and hence, the two procedures share the common efficiency. Looking at (4.4), we may observe that the choice of \( T_n \) may again be guided by some robustness considerations, and we shall see later on that such robustness properties percolate through the sequentialization. At this stage, we may refer to Chapter 10 of Sen (1981), and omitting details we may present the following.

Suppose that \( \{ V^*_n \} \) in (4.7) satisfies the condition that

\[
V^*_n \rightarrow \nu^2(F) \quad \text{a.s., as} \quad n \rightarrow \infty. \tag{4.11}
\]
and the sequence \( \{T_n\} \) satisfies the Anscombe (1952) condition
\[
\max_{m: |m-n| \leq \delta n} \left\{ n^{\delta} |T_m - T_n| \right\} \leq 0 \quad \text{as} \quad n \to \infty, \quad (4.12)
\]
Then, the asymptotic consistency in (4.9) holds. So also (4.10) holds (in the a.s. mode of convergence).

If we want to establish that
\[
\lim_{d \to 0} \{E N_d)/n_d \} = 1, \quad \forall \ F \in \mathfrak{F}, \quad (4.13)
\]
then we may need some additional conditions on \( \{V_n^x\} \).

(I) Suppose that there exists a sequence \( \{Z_i\} \) of i.i.d.r.v. such that (i) the \( Z_i \) are nonnegative, (ii) \( E Z_1 \) exists, and (iii) \( V_n^x \leq (n-m)^{-1} \sum_{i \leq n} Z_i, \quad \forall \ n \geq n_0 > m \). In this setup, we take \( E Z_1 = v^2(F) \). Then (4.13) holds.

(II) Suppose that \( V_n^x \) is expressible as a linear combination of reversed (sub-) martingales, so that
\[
E(\sup_{n \geq n_0} V_n^x) < \infty, \quad \text{for some} \ n_0 \geq 2, \quad (4.14)
\]
then (4.13) holds.

(III) Suppose that for some \( r > 1 \) (not necessarily an integer),
\[
E\left\{ n^{\delta} |V_n^x - v^2(F)|^{2r} \right\} \leq c_r < \infty, \quad \forall \ n \geq n_0. \quad (4.15)
\]
Then, (4.13) holds. If it is also possible to replace (4.15) by a probability inequality: \( P\{|V_n^x - v^2(F)| > \varepsilon\} \leq c_\varepsilon n^{-r}, \quad \text{for every} \ n \geq n_0, \quad \text{where} \ r > 1 \ (\forall \ \varepsilon > 0, \ c_\varepsilon < \infty) \). In any case, (III) is more restrictive than (I) or (II), and in a majority of the cases, it may be possible to incorporate (I) or (II) and to avoid the extra moment condition in (III). In (4.7), often \( \tau_{a/2}^2 \) is replaced by a sequence \( \{a_n^2\} \) where \( a_n \to \tau_{a/2} \) as \( n \to \infty \), and the conclusions made above remain intact. Also, \( n \) may be replaced by a monotone function \( \psi(n) \), and parallel results hold.

The procedure described above works out well for a general class of statistical functionals including Hoeffding's (1948) U-statistics, von Mises (1947) differentiable statistical functions and Hadamard differentiable
statistical functionals; L-estimators and M-estimators (with bounded score functions) are all members of this class. However, in general, R- and M-estimators are implicitly defined statistical functions, and for unbounded score functions, verification of (4.11) or (I), (II) or (III) may require more complicated analysis. Moreover, the very alignment procedure yielding the point estimators also yield suitable confidence intervals. Such sequential confidence intervals were considered by Geertsema (1970), Sen and Ghosh (1971) and Ghosh and Sen (1971, 1972), among others. We refer to the Chapter (11) by Jurecková for some technical details. Essentially, by aligning the observations so that the aligned rank (or M-) statistic equals to the upper 50 $\alpha$ and lower 50 $\alpha$ points of its null distribution, one obtains a lower ($\hat{\theta}_{L,n}$) and upper ($\hat{\theta}_{U,n}$) confidence limit, from the given sample of size $n$; the corresponding confidence coefficient is $1-\alpha_n$, where $\alpha_n \rightarrow \alpha$, the preassigned value, as $n \rightarrow \infty$. Let then

$$N_d = \min\{n \geq n_0 : \hat{\theta}_{U,n} - \hat{\theta}_{L,n} \leq 2d\}, \ d > 0. \quad (4.16)$$

Note that, by definition, $I_{N_d} = \{\hat{\theta}_{L,n}, \hat{\theta}_{U,n}\} \, N_d$ satisfies (4.2). So the crux of the problem is to verify (4.1) and (4.13). The verification of these results depend on

(i) the a.s. convergence of $\sqrt{n}(\hat{\theta}_{U,n} - \hat{\theta}_{L,n})$ to a non-negative functional $\gamma(F)$ of the d.f. $F$,

(ii) asymptotic normality of $n^{1/2}(\hat{\theta}_{U,n} - \theta)$, and

(iii) Anscombe (1952) condition on these statistics.

Under quite general regularity conditions on the score function and the d.f. $F$, these conditions have all been verified [viz., Chapter 10 of Sen (1981)]. In particular, linear rank statistics based on normal scores and signed-rank statistics based on absolute normal scores all belong to this class.

Let us sketch briefly the multi-parameter extensions. Suppose that $\hat{\theta} = \hat{\theta}_1 + \cdots + \hat{\theta}_p$...
is a parametric vector \((\theta_1, \ldots, \theta_p)\)' is a parametric vector \((\theta_1, \ldots, \theta_p)\)' where each \(\theta_j\) is a suitable (estimable) parameter, and let \(T_n = (T_n^1, \ldots, T_n^p)\)' be an estimator of \(\theta\) based on a sample of size \(n > n_0\). Suppose that as \(n\) increases,

\[
\frac{n^\frac{1}{2}(T_n - \theta)}{\Sigma} \overset{\mathcal{D}}{\rightarrow} \mathcal{N}_p(0, \Sigma)
\]  

(4.17)

where \(\Sigma\) is a p.d. matrix of unknown parameters. Let \(\chi^2_{p, \alpha}\) be the upper 100\(\alpha\)% point of the central chi square d.f. with \(p\) degrees of freedom. Then, by (4.17), we claim that

\[
\lim_{n \to \infty} P_\theta \{ n(T_n - \theta)' \Sigma^{-1}(T_n - \theta) \leq \chi^2_{p, \alpha} \} = 1 - \alpha,
\]

(4.18)

where the confidence coefficient \(1 - \alpha\) is preassigned \((0 < \alpha < 1)\). Then, note that for any \(\ell \in \mathbb{R}^p\),

\[
\sup \{ (\ell' (T_n - \theta))^2 : \ell' \ell = 1 \}
\]

\[
= \sup \{ ((\ell' (T_n - \theta))^2 / (\ell' \Sigma \ell) ((\ell' \Sigma \ell) / (\ell' \ell)) : \ell' \ell = 1 \}
\]

\[
\leq \sup \left\{ \frac{(\ell' (T_n - \theta))^2}{\ell' \Sigma \ell} : \ell \neq 0 \right\} \cdot \sup \left\{ \frac{\ell' \Sigma \ell}{\ell' \ell} : \ell \neq 0 \right\}
\]

\[
= \chi_1^2(\Sigma^{-1}(T_n - \theta)(T_n - \theta)') \cdot \chi_1(\Sigma)
\]

\[
= (T_n - \theta)' \Sigma^{-1}(T_n - \theta) \cdot \chi_1(\Sigma)
\]

where \(\chi_1(A)\) stands for the largest characteristic root of \(A\) and where the last step of (4.19) follows from the fact that \((T_n - \theta)(T_n - \theta)')\) is of rank 1, so that \(\chi_1(\Sigma^{-1}(T_n - \theta)(T_n - \theta)') = \text{Trace}(\Sigma^{-1}(T_n - \theta)(T_n - \theta)') = (T_n - \theta)' \Sigma^{-1}(T_n - \theta)\). Therefore, from (4.18) and (4.19), we obtain that

\[
\lim_{n \to \infty} P \{ (\ell' (T_n - \theta))^2 \leq \frac{1}{n} (\ell' \ell) \chi_1(\Sigma) \chi^2_{p, \alpha} \forall \ell \neq 0 \}
\]

\[
= \lim_{n \to \infty} P \{ n(T_n - \theta)' \Sigma^{-1}(T_n - \theta) \leq \chi^2_{p, \alpha} \}
\]

\[
= 1 - \alpha, \text{ the preassigned coverage probability. (4.20)}
\]

Thus, if we define for every \(d > 0\),
\[ n_d = \min\{n \geq n_0 : n \geq \text{ch}_1(\bar{\xi}) \chi_{p,\alpha}^2 / d^2\}, \]

then from (4.20) and (4.21), we have

\[ \lim_{d \downarrow 0} P\{\|z'(T_n - \theta)\| \leq d, \forall \tilde{z} : \|\tilde{z}\| = 1\} = 1 - \alpha, \]  

(4.22)

so that, we obtain a simultaneous confidence region for \( \theta \) for which the maximum width is \( 2d (\geq 0) \) and the limiting (as \( d \downarrow 0 \)) coverage probability is \( 1 - \alpha \). In this setup, \( \bar{\xi} \) is unknown, and hence, \( n_d \), defined by (4.21), depends on \( d(>0) \), \( \chi_{p,\alpha}^2 \) as well as the unknown \( \bar{\xi} \). Let us now assume that there exists a sequence \( \{V_n\} \) of estimators of \( \bar{\xi} \), such that

\[ V_n \to \bar{\xi} \text{ almost surely, as } n \to \infty. \]  

(4.23)

Motivated by (4.21) and (4.23), we may consider the following stopping number:

\[ N_d = \min\{n \geq n_0 : n \geq \text{ch}_1(V_n) \chi_{p,\alpha}^2 / d^2\}, \]  

(4.24)

Consider then the following confidence set

\[ I_{N_d} = \{\theta : \tilde{z}'T_{N_d} - d \leq \tilde{z}'\theta \leq \tilde{z}'T_{N_d} + d, \|\tilde{z}\| = 1\} \]  

(4.25)

It is clear from the definition in (4.25) that the diameter of \( I_{N_d} \) is equal to \( 2d \), so that we need to show that

\[ \lim_{d \downarrow 0} P\{\theta \in I_{N_d}\} = 1 - \alpha \]  

(4.26)

(i.e., the asymptotic consistency holds), and that

\[ N_d / n_d \to 1 \text{ a.s., as } d \downarrow 0 \]  

(4.27)

(or in the first mean), so that the asymptotic efficiency holds.

A closer look at (4.23) ensures that \( \text{ch}_1(V_n) \to \text{ch}_1(\bar{\xi}) \text{ a.s., as } n \to \infty \), and hence, by (4.21) and (4.24), we have that (4.27) holds. Suppose further that the Anscombe condition in (4.12) holds for the vector case i.e., \( |T_m - T_n| \) being replaced by \( \|T_m - T_n\| \). Then (4.17) extends to the case of the
stopping number $N_d$, i.e., for $T_{N_d}$, and this along with (4.23) would imply (4.26). To extend (4.13) in this vector case, we may virtually follow the same steps as in (I), (II) or (III) [after (4.13)]. For (I), no modification is needed, excepting that the $Z_i$ are themselves p.d. matrices, whose expectation is $\mathcal{X}$. In (II), we need to work with the $\text{ch}_1(V_n)$, so that the reversed sub-martingale property would remain in tact [as $\text{ch}_1(V)$ is a convex function of $V$]. In (III), in (4.15), we similarly need to replace $V_n^{-1} V(F)$ by $\text{ch}_1(V_n) \text{ch}_1(\mathcal{X})$. Thus, the discussions made after (4.15) all pertain to this multiparameter case. In particular, if $V_n$ is a version of the jack-knifed dispersion matrix of $\mathcal{Y}_n$, then under the usual first order Hadamard-differentiability condition, (4.23) holds, while under appropriate moment convergence properties of the vector of influence functions (for the components of $\mathcal{Y}_n$), for the $\mathcal{Y}_n$. For multivariate $R$-estimators of location, $\mathcal{X}$ is typically of the form (3.23), and under appropriate regularity conditions on the score functions, an estimator $\hat{\Gamma}_n$ of $\Gamma(F)$ [typically, the permutation dispersion matrix] satisfies the condition that $\hat{\Gamma}_n \rightarrow \Gamma(F)$ a.s. as $n \rightarrow \infty$. For the diagonal matrix $\mathcal{D}$, the vector of coordinatewise estimators can be obtained as in the univariate case. These results ensure the a.s. convergence of $V_n$ to $\mathcal{X}$, and (4.14) or (4.15) can similarly be obtained for $\text{ch}_1(V_n)$. For some details, we may refer to Sen (1984). For $M$-estimators of location, a very similar methodology works out: For $V_n$ we may take $\hat{\Delta}_n^{-1} \hat{\Gamma}_n$ where

$$\hat{\Gamma}_n = n^{-1} \sum_{i=1}^n \psi(X_i - \hat{\theta}_n) [\psi(X_i - \hat{\theta}_n)]'$$

(4.28)

and $\hat{\Delta}_n = \text{Diag}(\hat{d}_{n1}, \ldots, \hat{d}_{np})$ with $\hat{d}_{nj}$ defined as in (2.42).

For parameters other than locations, the same technique applies. Verification of (I), (II) and (III) [after (4.13)] or their multiparameter
extensions can most conveniently be made either by employing the moment convergence or suitable reversed martingale characterizations. For some related results, we refer to Chapter 10 of Sen (1981).
5. NONPARAMETRICS IN CHANGE-POINT MODELS

Detection and change-point problems in sequential setups have been treated in Chapter 24 [by S. Zacks]. He has classified the sequential procedures into three categories: Bayes/Sequential stopping rules, CUSUM procedures and tracking methods. Bayes theory dominates the scenario and this in turn calls for suitable parametric families of distributions which are readily amenable to Bayesian analysis. The past few years have witnessed some fruitful developments in suitable nonparametric setups, and a brief review of these developments will be made in this section.

In a conventional setup, a sample of \( n \) observations \( X_1, \ldots, X_n \) relates to identically distributed and independent random variables. In a variety of situations, these observations are gathered at time-points \( t_1, \ldots, t_n \) respectively, where \( t_1 < \ldots < t_n \). (The quality control setup is a typical representation of this type.) In this context, due to some extraneous causes, a change in the distribution may take place at some point of time. Thus, we may formulate a model that \( X_i \) has the d.f. \( F_i \), \( i=1, \ldots, n \). Ideally, we would have (no change):

\[
F_1 = \ldots = F_n = F \quad \text{(possibly unknown)}, \tag{5.1}
\]

and we may conceive of a time point \( \tau \) (change point), such that for some (unknown) \( m(1 \leq m < n) \),

\[
t_m < \tau < t_{m+1} \quad \text{and} \quad F_1 = \ldots = F_m \neq F_{m+1} = \ldots = F_n. \tag{5.2}
\]

so that a test for this change-point model amounts to testing the homogeneity of \( F_1, \ldots, F_n \) against the composite alternative in (5.2). Such a test can be made in a purely nonsequential setup, or in a quasi-sequential setup (where the maximum possible sample number is prefixed as \( n \)). Also, in this context, one may consider a parametric model where the \( F_i \) have all a common specified form (possibly differing in the associated parameters or
algebraic constants), or a more general nonparametric model which allows
more flexibility for these unknown d.f.'s. Such quasi-sequential procedures
are quite similar to repeated significance tests, and hence, we shall
consider them in Chapter 7 (dealing with a general class of repeated
significance tests).

A somewhat more general problem (having a stronger sequential flavor)
is the so called sequential detection problem, treated by Page (1964) and
Shiryaev (1963, 1978), among others. The model is as follows: Assume that
for some positive integer K (possibly, \( + \infty \)), \( X_1, \ldots, X_K \) are independent and
identically distributed with a d.f. \( F \), and \( X_{K+j}, j \geq 1 \) are i.i.d.r.v. with a
d.f. \( G(\neq F) \). In this setup, if \( K < \infty \), then the problem is to raise an
alarm, while for \( K = + \infty \), there is no harm in continuing drawing
observations, and any alarm raised would be a false one. Thus, one would
like to introduce a stopping number \( N \), such that

(i) if actually \( K < \infty \), then \( (N-K)^+ = \max\{0,N-K\} \) should be as small as
possible (i.e., quickest detection), and

(ii) if \( K = + \infty \), the probability of a false alarm, i.e., \( P\{N < \infty | K = + \infty \} \) should be as small as possible.

Now \( N \) is a r.v. and \( K \) is unknown, and hence, a characterization of the
"smallness" of \( (N-K)^+ \) has to be made in a stochastic way. One possibility
is to take

\[
P\{N > K+r \mid K < \infty \} = P_K(r), \quad r > 0, \tag{5.3}
\]
as a suitable measure of the "smallness" of \( (N-K)^+ \); in this setup, the
choice of an appropriate \( r \) remains a vital issue. Another possibility is to
take

\[
E((N-K)^+ | K < \infty) \tag{5.4}
\]
as a suitable measure. Recall that by the usual formula for expectation, we
have

$$E\{(N-K)^+ \mid K < \infty\} = \sum_{r \geq 0} P\{N \geq K+r \mid K < \infty\} \sum_{r \geq 0} P_K(r),$$

(5.5)

so that (5.4) is an integrated version of (5.3). Note that whereas (5.3) is defined properly, in order that (5.5) exists, we need that $E[N|\bar{K} < \infty] < \infty$, and this may demand extra regularity conditions. Similarly, if $K = +\infty$, for every $n < \infty$, $P\{N < n|K = +\infty\}$ should be small, $E[N|K = +\infty] = \sum_{r\geq 0} P\{N \geq r \mid K = +\infty\}$ should be large (if not $+\infty$).

Whatever criteria we decide to choose in a proper formulation of the "quickest detection" problem (without too many false alarms), we may keep in mind that (i) for the stopping number $N$, the event $[N=n]$ depends on $\{X_i, i \leq n\}$, and (ii) the actual probability distribution of $N$ is dependent on $K$. In this setup, for any given $K$, the formulation (and distribution) of the stopping number $N$ may rest on parametric or nonparametric considerations, while, one way of handling the change point $(t_K)$ is to take recourse to some Bayes solution wherein some prior distribution $\pi(\cdot)$ is attached to $K$. Such Bayes formulations are discussed in Chapter 24 [by Zacks]. In this consideration, the likelihood ratio plays a basic role, and simple forms for the prior distribution $\pi(\cdot)$ (of $K$) lead to suitable optimal detection procedures. Shiryaev (1978, pp. 200) considered a diffusion process approximation for the change point problem for the drift of a drifted Wiener process $W = \{W(t), t \in R^+\}$, where

$$W(t) = \begin{cases} \sigma W_0(t), & t \leq K (=\tau), \\ \theta(t-\tau) + \sigma W_0(t), & t > \tau; \end{cases}$$

(5.6)

$W_0(t), t \in R^+$ is a standard Wiener process, $\sigma(>0)$ is a scale factor, and $\theta$ is the drift parameter; $\tau=K$ may be identified as the change point. Fortunately, the prescribed solution [see Section 1 of Chapter 24] remains applicable in a broad class of nonparametric procedures. Basically, in the
parametric case, the log-likelihood ratio process is known to have the weak (as well as almost sure) representation of a drifted Wiener process, so that this diffusion process approximation should be quite adequate for large values of $K$ (or expected $K$) and for a broad range of underlying distributions. We have seen in earlier sections that similar Wiener process approximations work out well for various nonparametric and robust statistics (and estimators), so that the diffusion process approximation presented in Chapter 24 are adaptable for such statistics too. For some details of these invariance principles, we may refer to Lombard (1983), Sen (1982, 1983, 1985), Telksnys (1986), and others. In practice, these approximations are generally adequate.
REFERENCES


