Maximum Spacing Methods and Limit Theorems for Statistics Based on Spacings

Magnus Ekström

Department of Mathematical Statistics
Umeå University
1997
The maximum spacing (MSP) method, introduced by Cheng and Amin (1983) and independently by Ranneby (1984), is a general estimation method for continuous univariate distributions. The MSP method, which is closely related to the maximum likelihood (ML) method, can be derived from an approximation based on simple spacings of the Kullback-Leibler information. It is known to give consistent and asymptotically efficient estimates under general conditions and works also in situations where the ML method fails, e.g. for the three parameter Weibull model.

In this thesis it is proved under general conditions that MSP estimates of parameters in the Euclidian metric are strongly consistent. The ideas behind the MSP method are extended and a class of estimation methods is introduced. These methods, called generalized MSP methods, are derived from approximations based on sum-functions of $m$th order spacings of certain information measures, i.e. the $\phi$-divergences introduced by Csiszár (1963). It is shown under general conditions that generalized MSP methods give consistent estimates. In particular, it is proved that generalized MSP methods give $L^1$ consistent estimates in any family of distributions with unimodal densities, without any further conditions on the distributions. Other properties such as distributional robustness are also discussed. Several limit theorems for sum-functions of $m$th order spacings are given, for $m$ fixed as well as for the case when $m$ is allowed to increase to infinity with the sample size. These results provide a strongly consistent nonparametric estimator of entropy, as well as a characterization of the uniform distribution. Further, it is shown that Cressie's (1976) goodness of fit test is strongly consistent against all continuous alternatives.
Maximum Spacing Methods and Limit Theorems for Statistics Based on Spacings

Magnus Ekström

The thesis consists of a summary and the following papers:


Key words and phrases: Estimation, spacings, maximum spacing method, consistency, $\phi$-divergence, goodness of fit, unimodal density, entropy estimation, uniform distribution.
Maximum Spacing Methods and Limit Theorems for Statistics Based on Spacings

Magnus Ekström
# Contents

List of papers iv

Abstract v

Preface vii

1 Introduction 1

2 The maximum likelihood method 2
  2.1 Consistency ........................................... 2
  2.2 Asymptotic normality and efficiency ................ 3
  2.3 Miscellaneous remarks ............................... 4

3 Statistics based on spacings, with a view towards goodness of fit 5

4 The maximum spacing method 9
  4.1 Consistency ........................................... 12
  4.2 Asymptotic normality and efficiency ................ 14
  4.3 Goodness of fit and confidence regions .......... 15
  4.4 Miscellaneous remarks ............................... 15

5 Summary of the thesis 16
  5.1 Paper A ............................................. 16
  5.2 Paper B and Paper C ............................... 17
  5.3 Paper D ............................................. 20
  5.4 Paper E ............................................. 21
  5.5 Final remarks ...................................... 22

References 23

Papers A–E
List of papers

The present thesis is based on the following papers, referred to in the text by the letters A–E.


Abstract

The maximum spacing (MSP) method, introduced by Cheng and Amin (1983) and independently by Ranneby (1984), is a general estimation method for continuous univariate distributions. The MSP method, which is closely related to the maximum likelihood (ML) method, can be derived from an approximation based on simple spacings of the Kullback-Leibler information. It is known to give consistent and asymptotically efficient estimates under general conditions and works also in situations where the ML method fails, e.g. for the three parameter Weibull model.

In this thesis it is proved under general conditions that MSP estimates of parameters in the Euclidian metric are strongly consistent. The ideas behind the MSP method are extended and a class of estimation methods is introduced. These methods, called generalized MSP methods, are derived from approximations based on sum-functions of mth order spacings of certain information measures, i.e. the φ-divergences introduced by Csiszár (1963). It is shown under general conditions that generalized MSP methods give consistent estimates. In particular, it is proved that generalized MSP methods give \( L^1 \) consistent estimates in any family of distributions with unimodal densities, without any further conditions on the distributions. Other properties such as distributional robustness are also discussed. Several limit theorems for sum-functions of mth order spacings are given, for \( m \) fixed as well as for the case when \( m \) is allowed to increase to infinity with the sample size. These results provide a strongly consistent nonparametric estimator of entropy, as well as a characterization of the uniform distribution. Further, it is shown that Cressie's (1976) goodness of fit test is strongly consistent against all continuous alternatives.


Key words and phrases: Estimation, spacings, maximum spacing method, consistency, \( \phi \)-divergence, goodness of fit, unimodal density, entropy estimation, uniform distribution.
Preface

It is not easy to imagine jobs more exciting than doing research when things are going your way. However, when they are not, doing research can be very dull. For this reason, it is of great importance for a PhD student to have a good supervisor. I have been fortunate to have not only one, but two such advisors: my main supervisor Professor Bo Ranneby and my auxiliary supervisor Professor Dmitrii Silvestrov. I would like to thank them for their very helpful discussions and suggestions, as well as for their support.

Further, I want to thank all former and present colleagues working at the department of Mathematical Statistics at Umeå University, and at the department of Forest Resource Management and Geomatics at the Swedish University of Agricultural Sciences.

I especially wish to express my gratitude to Ingrid Westerberg Eriksson for her excellent and patient secretarial assistance. Many thanks also to Anitha Bergdahl, Yvonne Löwstedt and Anne-Maj Jonsson.

Further thanks to Dr. Jun Yu for valuable comments on the manuscript. I am also grateful to Professor Thomas E. Burk, Dr. Clarissa-Marie Claudel and Dr. Paul Haemig for improving the English language in the thesis.

I would like to express my appreciation to Dr. Yongzhao Shao and Professor Marjorie G. Hahn for supplying copies of their papers before publication.

To my parents and my two sisters, who have supported me emotionally as well in practical matters over the years, I express my deepest gratitude.

I conclude by saying thanks to all my friends who did not care about the thesis, who would not read it, and who had to put up with me while I wrote it.

Umeå, April 1997

Magnus Ekström
1 Introduction

A common problem in statistics is that of estimating the underlying distribution by a sample of independent random variables. Often there is an assumed candidate set, a model, of distribution functions. The distribution functions in the model may be indexed by a parameter \( \theta \), and in this case the problem becomes that of estimating \( \theta \).

This problem was first put into its modern form by James Bernoulli (in his posthumous work *Ars Conjectandi* (1713)), who considered estimation of a binomial parameter. Today the statistical literature on this topic is voluminous, and the most widely used "general" solution is that of maximum likelihood (ML). Although the ML method can be traced back to Lambert (1760) and Daniel Bernoulli (1777), it is generally agreed that Fisher (1912, 1922) introduced it as a general method of estimation.

Though widely used the ML method is not "best" under all circumstances. For instance, the ML method may behave badly when model parameters increase in number as the number of observations increases, see e.g. Neyman and Scott (1948), or when the likelihood functions are unbounded, see e.g. Kiefer and Wolfowitz (1956).

Many competitors of the ML method have been proposed over the years. The problem of unbounded likelihood functions inspired Cheng and Amin (1979, 1983) and independently Ranneby (1984) to develop a general estimation method called the maximum spacing (MSP) method (Cheng and Amin used the name maximum product of spacings method; because of the generalizations of the method introduced in this thesis, the name MSP method is more appropriate and will be used throughout the thesis). The MSP method in its original form is defined for continuous univariate distributions and can, just as the ML method, be derived from an approximation of the Kullback-Leibler information. There are many known situations in which the MSP method works better than the ML method and moreover, attractive properties such as consistency and asymptotic efficiency of the maximum spacing estimator (MSPE) closely parallel those of the maximum likelihood estimator (MLE) when it works well.

The content of the thesis is mainly concerned with the MSP method and some of its generalizations. It is focused on the problem of consistency of estimators. Since these methods are based on spacing statistics, an important part of the thesis is devoted to asymptotic properties of such statistics.

As the MSP method is closely related to the ML method, it is natural to compare these methods. In the next section we give a presentation of the ML method, including a survey of results regarding asymptotic properties of MLEs, such as consistency and asymptotic normality. In Section 3 we present a certain type of statistic based on spacings, and give a survey of its use in goodness of fit tests. The use of a particular spacing statistic, sums of logarithms of spacings, in
estimation problems, i.e. the MSP method, is described in Section 4. There we provide a detailed survey of results obtained since its introduction. In Section 5, the contents of the different papers included in the thesis are summarized.

2 The maximum likelihood method

Here we will only consider the ML method for independent, identically distributed (i.i.d.) random variables on the real line.

Let $F_\theta$ be a family of distribution functions on the real line, indexed by $\theta$ which belongs to some parameter space $\Theta \subseteq R^s$. Suppose that the distributions $F_\theta$ possess densities or mass functions $f_\theta(x)$, and assume that $\xi_1, ..., \xi_n$ are i.i.d. random variables from $F_{\theta_0}$. Define the likelihood function by

$$L_n(\theta) = f_\theta(\xi_1) \cdots f_\theta(\xi_n).$$

**Definition 1** Any value $\hat{\theta}_n \in \Theta$ maximizing $L_n(\theta)$ over $\Theta$ is called a maximum likelihood estimator of $\theta^0$.

As we already mentioned, Fisher (1912, 1922) introduced the ML method as a general method of estimation. It was also Fisher who named the method (1922) and set the stage for its general acceptance. Fisher gave proofs of asymptotic efficiency but no separate proof of consistency, although this could be regarded as implied by his proof of asymptotic efficiency.

In the following we will confine attention to the case where $f_\theta$ is a density.

2.1 Consistency

Hotelling (1930) and Doob (1934) gave early proofs of consistency. A more general proof was given by Cramér (1946). Cramér, who considered the one parameter case, proved under certain regularity conditions (e.g. that $f_\theta(\cdot)$ as a function of $\theta$ is three times differentiable) that with probability tending to one as $n \to \infty$, the likelihood equation $\frac{\partial \log L_n(\theta)}{\partial \theta} = 0$ has a solution, which converges to $\theta^0$ in probability.

Later Huzurbazar (1948) showed under general conditions, in the one parameter case, that even if the likelihood equation has many solutions it always has a unique consistent solution (for more on this result, see Perlman (1983)). For a corresponding result in the multiparameter case, see e.g. Foutz (1977).

It should be noted that the MLE $\hat{\theta}_n$ does not necessarily coincide with the consistent root guaranteed by Cramér’s (1946) theorem. Kraft and Le Cam (1956) gave an example in which Cramér’s conditions are satisfied, the MLE $\hat{\theta}_n$ exists, is unique, and satisfies the likelihood equations, yet is not consistent. Consequently, it is advantageous to establish the uniqueness of the likelihood equation roots
whenever possible. In Mäkeläinen et al. (1981) sufficient conditions for existence and uniqueness are given.

A consistency result more satisfactory than Cramér’s (1946) was given by Wald (1949). Wald gave his proof of strong consistency in a multiparameter context and considered approximate MLEs (AMLEs) \( \theta_n^* \) satisfying \( L_n(\theta_n^*) \geq c \sup_{\theta \in \Theta} L_n(\theta) \), for some \( 0 < c \leq 1 \), not merely some root of the likelihood equations. A cornerstone in Wald’s proof is the inequality \( \lim_{n \to \infty} n \log L_n(\theta) < \lim_{n \to \infty} n \log L_n(\theta^0) \), which holds almost surely if \( F_\theta \neq F_{\theta^0} \). If \( \Theta \) is finite, this inequality alone implies consistency of the MLE \( \tilde{\theta}_n \). In the general case, Wald assumed that \( \Theta \) is compact, and by familiar compactness arguments reduced the problem to the case in which \( \Theta \) contains a finite number of elements. Aside from the compactness assumption, Wald’s uniform integrability conditions imposed on \( \log f_\theta(\cdot) \) are often not satisfied. However, in comparison with Cramér (1946), Wald (1949) used no differentiability conditions on \( f_\theta(\cdot) \). Many improvements have been made in Wald’s (1949) approach toward MLE consistency; notably by Le Cam (1953), Kiefer and Wolfowitz (1956), Huber (1967) and Bahadur (1967, 1971). In these papers the conditions are imposed on the log likelihood ratio \( \log(f_\theta(\cdot)/f_{\theta^0}(\cdot)) \) rather than on \( \log f_\theta(\cdot) \). Also, these papers (except Le Cam (1953)) share the assumptions that

(a) there exists a “suitable compactification” of \( \Theta \) so that the log likelihood ratio may be extended without changing its supremum,

(b) the supremum of the log likelihood ratio is integrable (dominated).

Perlman (1972) gave conditions for strong consistency, based on dominance or semidominance by zero of the log likelihood ratio, which are weaker than the conditions in all these papers, except the paper by Le Cam (1953). Le Cam’s (1953) conditions are equivalent to those based on dominance. Further, Perlman (1972) discussed necessary and sufficient conditions for strong consistency of AMLEs. Wang (1985) found that the conditions used in Perlman (1972) usually fail in the nonparametric situation, and generalized some of Perlman’s results.

In Hoffman-Jørgensen (1992) all possible limit points of all (approximating) MLEs are characterized without imposing any conditions.

2.2 Asymptotic normality and efficiency

The first rigorous proof of asymptotic efficiency of MLEs, i.e. that

\[
\sqrt{n}(\tilde{\theta}_n - \theta) \overset{D}{\to} N(0, I(\theta)^{-1}),
\]

where \( I(\theta) \) is the Fisher information matrix with elements

\[
I_{j,k}(\theta) = E_\theta \left[ -\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f_\theta(\xi) \right], \quad j, k = 1, \ldots, s,
\]
is due to Cramér (1946). He considered the one parameter case and proved under certain regularity conditions that any consistent root of the ML equations is asymptotically efficient. His proof is based on a Taylor expansion of $\log L_n(\theta)$ and assumes a certain amount of smoothness in $f_\theta(\cdot)$ as a function of $\theta$, i.e. existence of derivatives up to order three. He also used conditions implying that $\int f_\theta(x)dx$ and $\int \partial \log f_\theta(x)/\partial \theta dx$ be differentiable under the integral sign and assumed that the random variable $\partial \log f_\theta(\xi_1)/\partial \theta$ has finite, positive variance.

Cramér's results were generalized to the multiparameter case by Chanda (1954) and Doss (1962, 1963). For a thorough analysis of Cramér's conditions see Kulldorff (1957), who also provided an alternative theorem not requiring the existence of third order derivatives of $f_\theta(\cdot)$. Daniels (1961) used only first order derivatives in his proof of asymptotic normality. However, according to Huber (1967), his proof is incorrect and an additional restrictive condition is required. Huber (1967) also provided a proof of asymptotic normality, and as in Daniels (1961) the conditions do not involve second or higher order derivatives of $f_\theta(\cdot)$. Further, in Huber (1967) it is not assumed that the underlying true distribution belongs to the family that defines the MLE, which is of importance in relation to questions of robustness.

For even weaker conditions assuring asymptotic normality of MLEs, we refer to Le Cam (1970), Ibragimov and Has'minskii (1981) and Hoffmann-Jørgensen (1992). In none of these is second order differentiability of $f_\theta(\cdot)$ with respect to $\theta$ required. Hoffmann-Jørgensen (1992) showed that a single, very weak kind of stochastic differentiability condition implies asymptotic normality of MLEs.

### 2.3 Miscellaneous remarks

Although the ML method has many appealing properties, it has deficiencies:

- It does not always provide consistent estimates. (See Examples 2-4 in Section 4.)

- There exist estimates with lower asymptotic variances, e.g. the classical example of super efficiency by Hodges (see Le Cam (1953)).

- Little can be said about small sample properties of MLEs. For instance, even if MLEs are asymptotically unbiased in regular cases, this is not generally true for finite samples. It is not generally clear if the removal of the bias from an MLE will "improve" the estimator. For arguments of bias correction of order $1/n$ in the context of second order efficiency of MLEs, see e.g. Rao (1961).
3 Statistics based on spacings, with a view towards goodness of fit

The uniform distribution on \([0,1]\) is of fundamental importance in statistics. For instance, if a random variable \(\xi\) is distributed according to a continuous distribution on \(R\), then \(F\) is its distribution if and only if \(F(\xi)\) is uniformly distributed. This important fact can be used to solve two important statistical problems:

**Goodness of fit:** To test whether an i.i.d. sample \(\xi_1, \ldots, \xi_n\) comes from a specified continuous distribution \(F\) on the real line or not, is equivalent to testing if \(F(\xi_1), \ldots, F(\xi_n)\) are generated from a uniform distribution on \([0,1]\) or not.

**Estimation:** For an i.i.d. sample \(\xi_1, \ldots, \xi_n\) with an unknown continuous distribution on the real line, the most likely candidate to have generated the sample is the distribution \(F\) which makes \(F(\xi_1), \ldots, F(\xi_n)\) "most uniform" according to some decision rule.

Denote the distribution that generated \(\xi_1, \ldots, \xi_n\) by \(F_\xi\), and denote the order statistics of \(\xi_1, \ldots, \xi_n\) by \(\xi_1(1) \leq \xi_1(2) \leq \ldots \leq \xi_1(n)\). Further, let \(\xi_1(0) = -\infty\) and \(\xi_1(n+1) = \infty\).

Suppose, for the moment, that the candidate set of possible distributions in the estimation problem consists of all continuous distributions on the real line. Also, if we for a moment let "most uniform" mean "most regular", then the solution of the estimation problem would be any distribution \(\hat{F}\) satisfying

\[
\hat{F}(\xi(i+1)) - \hat{F}(\xi(i)) = \frac{1}{n+1}, \quad i = 0, \ldots, n.
\]  

If we denote the empirical distribution function of the sample by \(F_n\), then we have

\[
\sup_{0 \leq x \leq 1} |\hat{F}(x) - F_n(x)| \leq \frac{1}{n+1},
\]

and so, by the Glivenko-Cantelli Theorem, \(\hat{F}\) converges weakly to the underlying true distribution \(F_\xi\). However, since \(F_\xi(\xi_1), \ldots, F_\xi(\xi_n)\) are random variables we can not expect that (1) holds for \(F_\xi\), and thus some other kind of criteria for "most uniform" should be used.

It should be noted that the empirical distribution function is useful also in the goodness of fit problem, for instance the Kolmogorov-Smirnov test and the Cramér-von Mises test.

A different way to solve both these problems is to use statistics based on spacings. The overlapping \(m\)th order spacings are defined by

\[
D_i^{(m)} = \xi(i+m) - \xi(i), \quad i = 0, \ldots, n - m + 1,
\]
and the non-overlapping $m$th order spacings are defined by

$$D_{i,m}^{(m)} = \xi_{(m(i+1))} - \xi_{(im)}, \quad i = 0, \ldots, \left[\frac{n-m+1}{m}\right],$$

where $[x]$ denotes the greatest integer less than or equal to $x$. When $m = 1$, $\{D_i\}$ or simply $\{D_i\}$ are called simple spacings. Hereafter, $\Psi$ will denote some "smooth", real valued function. The statistics based on spacings that we will take into account in this thesis are of the form

$$\sum_{i=0}^{n-m+1} \Psi \left( \frac{n+1}{m} D_i^{(m)} \right).$$

A version of this statistic on a circle with unit circumference is often considered. We will not always distinguish between these two cases here, since the goodness of fit tests based on these two versions are asymptotically equivalent if $m$ is not increasing rapidly with $n$.

For the use of these statistics for estimation we refer to the next two sections.

Now, consider the hypothesis-testing problem

$$H_0 : F_\xi = F, \quad H_A : F_\xi \neq F,$$

where $F$ is a specified continuous distribution function. Note that many other alternative hypotheses also are of interest, including parametric as well as non-parametric ones.

If $H_0$ is true, the probability integral transform $U_i = F(\xi_i), \ i = 1, \ldots, n$, takes the sample $\xi_1, ..., \xi_n$ into a sample of uniform random variables, so without loss of generality we assume throughout this section that $F$ in the hypothesis is the uniform distribution over $[0,1]$, and that $0 = \xi(0) < \xi(1) < \ldots < \xi(n+1) = 1$.

We now list some of the statistics of form (2) which have been used to provide goodness of fit tests. All of the statistics in this list were introduced for the case $m = 1$, and it should be noted that most of these tests have been introduced in connection with testing for exponentiality rather than as tests for uniformity. For instance, an exponential set of $n+1$ i.i.d. random variables $Y_1, ..., Y_{n+1}$ can be transformed into $n+1$ simple spacings, produced by $n$ uniform random variables, by the transformation $Y_i/\sum Y_i$ (in this case the null-hypothesis may be extended to the composite hypothesis of all exponential distributions). Moreover, the intervals between successive events of a Poisson process (which are exponentially distributed) conditioned on the number of events in a specified interval, are distributed like uniform simple spacings.

The examples of statistics of form (2) are as follows:

- $\Psi(x) = x^2$, suggested by Greenwood (1946).
\(\Psi(x) = x^r\), for \(r > 0\), suggested by Kimball (1950).

\(\Psi(x) = |x - 1|^2\), suggested by Irwin in the discussion of Greenwood (1946) and by Kimball (1947).

\(\Psi(x) = |x - 1|\), suggested by Kimball in the discussion of Greenwood (1946) and by Sherman (1950).

\(\Psi(x) = \log x\), suggested by Moran (1951) and by Darling (1953).

\(\Psi(x) = 1/x\), suggested by Darling (1953).

Let us take a closer look at the first test-statistic in the list for \(m = 1\), \(\sum((n + 1)D_i)^2\), the so called Greenwood statistic. Greenwood (1946) introduced it in connection with testing that the intervals between events were exponential, that is, that the times of occurrence of events constituted a Poisson process. Distributional properties of this statistic were investigated by Moran (1947, 1951). Large values of \(\sum((n + 1)D_i)^2\) indicate highly irregular spacings and small values indicate superuniform observations (i.e. the sample is too regular to be a uniform sample). Consequently, we can reject \(H_0\) for "large" and "small" values of Greenwood's statistic. For large \(n\), \(n^{-1/2}\sum((n + 1)D_i)^2\) is approximately \(N(2, 4)\), but its limiting distribution is attained very slowly. For upper and lower percentage points of the test for small samples, see Burrows (1979), Currie (1981) and Stephens (1981). Note that \(\sum((n + 1)D_i)^2 = \sum|(n + 1)D_i - 1|^2 + (n + 1)\), so Greenwood’s test is equivalent to the test suggested by Irwin and Kimball.

A disturbing feature of the tests based on \(\sum \Psi((n + 1)D_i)\) is that they are unable (asymptotically) to detect alternatives approaching the uniform at a faster rate than \(n^{-1/4}\) (see Chibisov (1961) and Sethuraman and Rao (1970)). Thus, in comparison with the Kolmogorov-Smirnov test, which can discriminate alternatives at a distance of order \(n^{-1/2}\) from the hypothesis, these tests have a poor asymptotic performance. However, it is known that within the class \(\sum \Psi((n + 1)D_i)\) of test statistics, Greenwood's test is asymptotically most powerful (see e.g. Sethuraman and Rao (1970)).

It is often of interest to test a composite null hypothesis like \(H_0 : F_{\xi} \in \{F_{\theta} : \theta \in \Theta\}\), for some family of absolutely continuous distributions \(F_{\theta}\) on \(R\), where \(\theta\) is a parameter vector belonging to some parameter set \(\Theta\). Cheng and Stephens (1989) considered this case by using the Moran-Darling statistic (with \(m = 1\)) and showed that this test statistic, under general conditions, has the same asymptotic distribution when the parameters must be estimated from the sample as when the parameters are actually known. In Wells et al. (1993) it is shown that this is true for a large class of statistics of form (2) based on simple spacings.

Cressie (1976) studied the statistic \(\sum \log((n + 1)D_i^{(m)}/m)\). He considered both the case where \(m\) is fixed as well as the case where \(m\) increases with \(n\), and gave asymptotic normality results for the test-statistic under the null hypothesis.
(see also Holst (1979)). The principal conclusion is that, at least asymptotically, increasing the value of $m$ increases the power of the test. Cressie showed that it is enough to use $\sum \log((n+1)D_i^{(m)}/m)$ with $m = 2$ to obtain a test which is asymptotically more powerful than Greenwood's test with $\sum((n+1)D_i)^2$. For results related to those of Cressie (1976), see Vasicek (1976) and Dudewicz and van der Meulen (1981).

In Cressie (1979), Kuo and Rao (1981) and Rao and Kuo (1984) it is shown that among the tests based on (2) with $m$ fixed, $\sum((n+1)D_i^{(m)}/m)^2$ is most powerful. The asymptotic theory here suggests that larger $m$ values are always better. But for fixed values of $m$, the spacing tests have no power against alternatives which are at a distance of $n^{-\delta}$ from $H_0$, where $\delta > 1/4$.

Hall (1986) showed that if $m \to \infty$ as $n \to \infty$ then the test statistic (2) can distinguish alternatives of distance $(mn)^{-1/4}$ from the uniform if $m$ does not diverge at a faster rate than $n^{1/2}$. The test actually becomes less powerful as $m$ increases beyond order $n^{1/2}$. This unpleasant behaviour can be eliminated by defining the test statistic on the circle rather than on the line. In this case, with $m$ properly choosen, i.e. $m/n \to \rho < 1$ where $\rho$ is irrational, the test detects alternatives of distance $n^{-1/2}$ from the uniform.

A test with non-overlapping spacings, i.e. with

$$\sum_{i=0}^{[(n-m+1)/m]} \Psi \left( \frac{n+1}{m} D_i^{(m)} \right),$$

(3)

was suggested by del Pino (1979). However it is clear that tests based on (2) will be at least as powerful as tests based on (3), see e.g. Cressie (1979), Rao and Kuo (1984) and Jammalamadaka et al. (1989). On the other hand, it follows that a non-overlapping spacing test of order $3m/2$ or larger is more efficient than a corresponding overlapping test of order $m$, and as pointed out in the latter paper, non-overlapping spacing tests are less complicated and easier to compute.

Asymptotic normality for statistics like (2) for uniform samples has been investigated by many authors, notably by Darling (1953) and Le Cam (1958) for $m = 1$, Holst (1979) and Beirlant et al. (1991) for finite $m \geq 1$, and Hall (1986) for $m \to \infty$.

The more difficult problem of deriving asymptotic results for statistics of type (2) for general distributions has been considered by Hall (1984) for finite $m$, and Khasimov (1989) and Van Es (1992) for $m \to \infty$. However, these results are obtained under quite restrictive assumptions, e.g. that the density of the underlying distribution is bounded away from zero on compact intervals.

For more extensive surveys on the distribution theory of statistics like (2) and (3), and discussions on applications in goodness of fit problems, see Pyke (1965, 1972) and D’Agostino and Stephens (1986).
We end this section by mentioning that in Holst and Rao (1981), Kuo and Rao (1981) and Wells et al. (1993) the more general class of statistics

\[ \sum_i \Psi_i \left( \frac{n+1}{m} D_i^{(m)} \right), \]

is considered. It has been found that tests based on statistics of this more general type can discriminate between alternatives converging to the uniform, even with finite values of \( m \), at a rate of \( n^{1/2} \) as in the Kolmogorov-Smirnov and Cramér-von Mises tests.

4 The maximum spacing method

The MSP method is a general method of estimating continuous, univariate distributions: an alternative to the ML method.

Let \( F_\theta \), where the unknown parameter vector \( \theta \) is contained in the parameter space \( \Theta \subseteq \mathbb{R}^s \), denote a family of continuous, univariate distribution functions. Let \( \xi_1, \ldots, \xi_n \) be i.i.d. random variables with distribution function \( F_\theta \), and denote the corresponding order statistics by \( \xi_{(1)} \leq \cdots \leq \xi_{(n)} \). Further, let \( \xi_{(0)} = -\infty \) and \( \xi_{(n+1)} \). Define

\[ S_n(\theta) = \frac{1}{n+1} \sum_{i=0}^{n} \log \left\{ (n+1) \left( F_\theta(\xi_{(i+1)}) - F_\theta(\xi_{(i)}) \right) \right\}. \]

The function \( S_n(\theta) \) can be seen as an analogue to the log likelihood function \( \log L_n(\theta) \). Note that \( S_n(\theta) \) is the Moran-Darling statistic used in goodness of fit tests.

**Definition 2** Any \( \hat{\theta}_n \in \Theta \) which maximizes \( S_n(\theta) \) over \( \Theta \) is called a maximum spacing estimator of \( \theta^0 \).

The MSP method was proposed by Cheng and Amin (1979, 1983) and independently by Ranneby (1984). The method has been derived from several different viewpoints. The argument in Cheng and Amin (1983) was that the maximum of \( (n+1)^{-1} \sum \log\{(n+1)D_i\} \) (the \( D_i \)'s representing the spacings \( F_\theta(\xi_{(i+1)}) - F_\theta(\xi_{(i)}) \)), under the constraint \( \sum D_i = 1 \), is obtained if and only if all the \( D_i \)'s are equal. This, in a rough sense, corresponds to our attempt to set \( \theta = \theta^0 \), when the \( D_i \)'s become identically distributed, e.g. the uniform spacings \( F_{\theta^0}(\xi_{(i+1)}) - F_{\theta^0}(\xi_{(i)}) \) should be "more nearly equal" than others. Ranneby (1984) derived the MSP method from an approximation of the Kullback-Leibler information (note that the ML method also can be derived from an approximation of the Kullback-Leibler information). In Titterington (1985) it was observed that the MSP method can be regarded as an ML approach based on grouped data. Shao and Hahn (1994)
proposed the MSP method upon reexamining Fisher's (1912) intuitive arguments behind the MLE.

In Shao and Hahn (1996b) the MSP method is extended so it can be applied for any family of univariate distributions, continuous or not. For families of purely discrete distributions with finitely many atoms, the extended version of Shao and Hahn (1996b) coincides with the ML method. Ranneby (1990) extended the MSP method to multivariate distributions, using Dirichlet cells (or to be more precise, inner circles of Dirichlet cells) as the multivariate counterpart of spacings.

Next, we give some examples to illustrate MSP and ML estimates.

Example 1. Let \( \xi_1, \ldots, \xi_n \) be i.i.d. \( U(0, \theta^0) \) (uniform on \((0, \theta^0)\)). Then the MLE is \( \hat{\theta}_n = \max(\xi_1, \ldots, \xi_n) \) and the MSPE is \( \hat{\theta}_n = (n + 1)\hat{\theta}_n/n. \) Both are consistent, and their large sample behaviours are described by

\[
\frac{n(\hat{\theta}_n - \theta^0)}{\theta^0 \cdot Y} \xrightarrow{D} \theta^0 \cdot Y \quad \text{and} \quad \frac{n(\hat{\theta}_n - \theta^0)}{\theta^0(Y - 1)} \xrightarrow{D} \theta^0(Y - 1) \quad \text{as} \quad n \to \infty,
\]

where \( Y \) is a standard exponential random variable. As is known, \( \hat{\theta}_n \) is the uniformly minimum variance unbiased (UMVU) estimator of \( \theta^0 \). Furthermore,

\[
E[n(\hat{\theta}_n - \theta^0)]^2/E[n(\hat{\theta}_n - \theta^0)]^2 \to 2, \quad n \to \infty.
\]

Thus, the MLE is not asymptotically optimal.

Example 2. Let

\[
F_\theta(x) = \frac{1}{2} \Phi(x) + \frac{1}{2} \Phi\left(\frac{x - \mu}{\sigma}\right), \quad \theta = (\mu, \sigma) \in R \times R^+,
\]

where \( \Phi(x) \) is the standard normal distribution function, and let \( \xi_1, \ldots, \xi_n \) be i.i.d. from \( F_\theta(x) \), \( \theta^0 = (\mu_0, \sigma_0) \in R \times R^+ \). Then the MLE of \( \theta^0 \) does not exist, since the likelihood function approaches infinity as, for example, \( \mu = \xi_1 \) and \( \sigma \downarrow 0 \). However, any \( \theta_n^* \in \Theta \) defined by

\[
S_n(\theta_n^*) \geq -c_n + \sup_{\theta \in \Theta} S_n(\theta)
\]

(4)

where \( 0 < c_n \) and \( c_n \to 0 \) as \( n \to \infty \), is a consistent estimator of \( \theta^0 \). Note that if an MSPE exists it satisfies (4).

Example 3. Let \( f_\theta(x) \) be the density of a three parameter Weibull distribution, i.e.

\[
f_\theta(x) = \beta \gamma^{-\beta} (x - \alpha)^{\beta-1} \exp\left\{-\left(\frac{x - \alpha}{\gamma}\right)^\beta\right\}, \quad x > \alpha,
\]

where \( \theta = (\alpha, \beta, \gamma) \), and let \( \xi_1, \ldots, \xi_n \) be i.i.d. from \( f_\theta(x) \), \( \theta^0 = (\alpha_0, \beta_0, \gamma_0) \). Consider the ML equations \( \partial \log L_n(\theta)/\partial \theta_j = 0 \) and the MSP equations \( \partial S_n(\theta)/\partial \theta_j = 0, j = 1, 2, 3. \)
(i) If $\beta > 2$ there are with a probability tending to one solutions $\hat{\theta}_n$ and $\check{\theta}_n$ of the ML and MSP equations, respectively, that are asymptotically normal with

$$\sqrt{n}(\hat{\theta}_n - \theta^0) \quad \text{and} \quad \sqrt{n}(\check{\theta}_n - \theta_0) \xrightarrow{D} N(0, I(\theta^0)^{-1}),$$

where $I(\theta)$ is the Fisher information matrix.

(ii) If $0 < \beta < 2$ there is with a probability tending to one a solution $\theta_n = (\alpha_n, \beta_n, \gamma_n)$ of the MSP equations with $\alpha_n - \alpha_0 = O_p(n^{-1/\beta})$, and $(\hat{\alpha}_n, \hat{\beta}_n)$ have the same asymptotic properties as the corresponding unique solution of the ML equations with a known $\alpha$.

For a solution $\check{\theta}_n = (\hat{\alpha}_n, \check{\beta}_n, \check{\gamma}_n)$ of the ML equations, however, property (ii) holds only for $1 < \beta < 2$. For $\beta < 1$ there is no consistent solution of the ML equations.

See Cheng and Amin (1983) for more details about this example, and for other examples of MSP estimation of distributions with a shifted origin.

The reason the ML method fails in Examples 2 and 3 is that the likelihood is unbounded. The function $S_n(\theta)$, however, is always bounded from above (by 0) and thus allows consistent estimates to be obtained by the MSP method. Further, the MLE may be inconsistent even when the likelihood function is bounded for any fixed sample size, as in the following example from Le Cam (1990) (a discrete version of this example was first given in Bahadur (1958)).

**Example 4.** Let $h(\cdot)$ be a continuous, strictly decreasing function defined on $(0, 1]$, with $h(x) \geq 1$ for all $0 < x \leq 1$ and satisfying

$$\int_0^1 h(x)dx = \infty. \quad (5)$$

Given a constant $0 < c < 1$, let $a_k, k = 0, 1, \ldots$ be a sequence of constants defined inductively as follows: $a_0 = 1$; given $a_0, \ldots, a_{k-1}$, the constant $a_k$ is defined by

$$\int_{a_k}^{a_{k-1}} (h(x) - c) dx = 1 - c. \quad (6)$$

It follows from (5) and (6) that this can be continued indefinitely and that $a_k \to 0$ as $k \to \infty$. Consider the sequence of densities

$$f_k(x) = \begin{cases} 
  c & \text{if } x \leq a_k \text{ or } x > a_{k-1} \\
  h(x) & \text{if } a_k < x \leq a_{k-1}
\end{cases}$$

and the problem of estimating the parameter $k^0$ on the basis of independent observations $\xi_1, \xi_2, \ldots, \xi_n$ from $f_k$.

Now, provided $h(\cdot)$ satisfies $h(x) \geq e^{x-2}$ for all sufficiently small $x$, the MLE exists and is unique with probability one but tends to infinity in probability, regardless of the true value $k^0$. Therefore the MLE is not consistent, although the likelihood functions are bounded for any fixed $n$. However, the MSP method works well here (see Shao and Hahn (1994) and Ekström (1994)).
4.1 Consistency

General consistency theorems are given in Ranneby (1984), Ekström (1994), Shao and Hahn (1996a) and Ghosh and Jammalamadaka (1996). In the following we call any \( \theta_n^* \) satisfying (4) an approximate MSPE (AMSPE).

In Ranneby's (1984) consistency proof of AMSPEs it is assumed that the densities \( f_\theta(x) \) have common support and are continuous functions of \( x \). He also used an identifiability condition in the strong sense and a continuity condition imposed on \( F_\theta \) as a function of \( \theta \). Furthermore, an additional "technical" condition was used. Ranneby (1984) also discussed some examples with inconsistent MLEs, e.g. Example 2. In Ekström (1994) a proof of consistency is given under weaker conditions than in Ranneby, i.e. the assumption that all densities \( f_\theta(x) \) should be continuous functions of \( x \) and have common support was not used. Also, a slightly weaker identification condition than in Ranneby (1984) was used and it was further shown that the "technical" condition used in Ranneby (1984) is not necessary. For instance, Examples 1, 3 (for AMSPEs rather than solutions of the MSP equation) and 4 are covered by the conditions in Ekström (1994), but not by those in Ranneby (1984). Moreover, in Ekström (1994) the more general case where the underlying true distribution does not belong to the family that defines the MSPE was considered.

In Shao and Hahn (1996a) results were obtained that apply to both parametric and nonparametric models. Let \( \mathcal{P} \) denote a given family of probability measures on \( \mathbb{R} \) dominated by the Lebesgue measure \( \mu \) (or any other dominating \( \sigma \)-finite measure with no atoms). Previous proofs assumed there is a given parametrization of \( \mathcal{P} \), say \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \). However, in Shao and Hahn (1996a) the probability measures \( P \) in \( \mathcal{P} \) are the unknown "parameters", and \( \mathcal{P} \) is the "parameter" space. Thus, for each \( P \in \mathcal{P} \), the corresponding density and distribution functions are denoted \( f_P \) and \( F_P \), respectively.

The following result, which supports the intuitive appeal of the MSPE, plays a significant role in Shao and Hahn's (1996a) proof of consistency.

Let \( \xi_1, ..., \xi_n \) be i.i.d. from \( P_0 \in \mathcal{P} \). Then, if \( P \neq P_0 \), \( P \in \mathcal{P} \),

\[
\lim_{n \to \infty} (S_n(P) - S_n(P_0)) \leq \int_{-\infty}^{\infty} \log \frac{f_P(x)}{f_{P_0}(x)} dF_{P_0}(x) < 0 \quad \text{a.s..} \tag{7}
\]

Moreover, if \( \mathcal{P} \) is finite, then the MSPE \( \hat{P}_n \) maximizing \( S_n(P) \) is consistent, i.e. \( \hat{P}_n \) converges weakly to \( P_0 \).

For the proof of consistency of the AMSPE \( P_n^* \), a compactification \( \overline{\mathcal{P}} \) of \( \mathcal{P} \) in the topology of vague convergence of subprobability measures, i.e. measures with total mass \( \leq 1 \), is used, like Bahadur (1971) did for the MLE. The conditions for consistency include a continuity condition and a weak identifiability condition (these conditions were also used, together with a condition of type (b) in Section
2.1, in Bahadur’s (1971) proof of consistency of AMLEs). Shao and Hahn (1996a) also assumed that for each \( P \), in the compactified version of \( \mathcal{P} \), \( \sup f_P(x) \), where the supremum is taken over small neighbourhoods of \( P \), is bounded on “large” sets; a condition of type (b) in Section 2.1 is not needed here. If in addition \( \overline{\mathcal{P}} \setminus \mathcal{P} \) is a closed set, they showed that an MSPE \( \hat{P}_n \) maximizing \( S_n(P) \) exists for all large \( n \). In particular, they showed that any AMSPE is \( L^1 \) consistent for any family \( \mathcal{P} \) of probability measures with unimodal densities. Note that many counterexamples of the MLE involve families of unimodal densities, e.g. Example 3.

Comparisons between the results of Ekström (1994) and Shao and Hahn (1996a) are not easily done since Shao and Hahn do not consider consistency of parameters in the Euclidian metric. To deduce consistency in the usual parametric situation from Shao and Hahn’s results, additional assumptions have to be made. Instead of doing this we present the following example from Bahadur (1971), in which the conditions for consistency in Shao and Hahn (1996a) fail, while the conditions in Ekström (1994) hold.

**Example 5.** For any positive integer \( k \), call the intervals \( (i/2^k, (i + 1)/2^k] \) for \( i = 0, 1, \ldots, 2^k - 1 \) dyadic intervals of rank \( k \). Let \( \mathcal{P}_k \) consist of all probability measures \( P \) with density functions \( f_P \) satisfying the following condition: for the positive integer \( k \), and \( k \) dyadic subintervals of \((0, 1]\) of rank \( k \), \( f_P \) is equal to one on these subintervals and on the interval \((k, k + 1 - k2^{-k}] \) and \( f_P \) is equal to zero elsewhere on \( R \). Let \( \mathcal{P}_0 \) be the uniform distribution on \((0, 1]\) and let \( \mathcal{P} = \bigcup_{k=0}^{\infty} \mathcal{P}_k \). Assume that we have a sample of i.i.d. observations from the distribution \( P^* \). Since \( \mathcal{P} \) is a countable set the model is easily parametrized. In Ekström (1994) it was shown that the given conditions for consistency of AMSPEs are fulfilled.

In Shao and Hahn (1996a) on the other hand, the probability measures \( P \) are treated as “parameters”, i.e. \( \Theta(P) \equiv \mathcal{P} \) and \( \Theta \equiv \mathcal{P} \), and \( \mathcal{P} \) is compactified in the topology of vague convergence of subprobability measures. To each measure \( Q \) in the compactification \( \overline{\mathcal{P}} \) of \( \mathcal{P} \) they define a “density” \( \gamma_Q \) (which may not integrate to one) by taking the limit of \( \sup f_P(x) \), where the supremum is taken over small neighbourhoods of \( Q \), as the radius of the neighbourhood tends to zero. The identifiability condition in Shao and Hahn (1996a) states that each such “density” \( \gamma_Q \) is not equal to the true underlying density a.e.. This condition is violated in the example above.

Ghosh and Jammalamadaka (1996) showed under general conditions, for the one parameter case, that with a probability tending to one the MSP equation \( \partial S_n(\theta)/\partial \theta = 0 \) has a solution, which converges to \( \theta^0 \) in probability. In contrast with the results discussed earlier in this subsection, Ghosh and Jammalamadaka assume that \( f_\theta(\cdot) \) is differentiable with respect to \( \theta \) (in an open interval \( I \subset \Theta \) containing \( \theta^0 \)). They also assume that the distributions have common support.
and that \( \int f_\theta(x)dx \) is twice differentiable under the integral sign. Further, they impose one additional assumption on the underlying distribution. They mention that their results generalize easily to the multiparameter case under similar conditions. However these kind of assumptions are not satisfied in Examples 1, 3, 4 and 5. On the other hand, Ghosh and Jammalamadaka's results are given for a class of estimation methods of which the MSP method is a special case (see Section 5.2 for further comments).

In Shao and Hahn (1996b), where the MSP method is extended so it can be applied for any univariate family of distributions, it is shown that for any family of distributions with a decreasing failure rate (i.e. \( \log(1 - F_\theta(x)) \) is convex on its support \([a, \infty)\), where \( a > -\infty \)), any (generalized) AMSPE is consistent.

It should be pointed out that a consistent MSPE does not always exist, as was shown in Shao, Wang and Xu (1996) where so called starshaped distributions were considered (a distribution function \( F \) on \([0, 1]\) is called starshaped if \( F(x)/x \) is nondecreasing on \((0, 1)\)). However, the MLE for a starshaped distribution function is also inconsistent (see Barlow et al. (1972)).

### 4.2 Asymptotic normality and efficiency

Asymptotic normality theorems for the MSPE \( \hat{\theta}_n \), i.e. that

\[
\sqrt{n}(\hat{\theta}_n - \theta^0) \xrightarrow{D} N(0, I(\theta^0)^{-1}),
\]

where \( I(\theta) \) is the Fisher information matrix, have been given by Ranneby (1985), Cheng and Stephens (1989), Shao and Hahn (1994) and Ghosh and Jammalamadaka (1996). The conditions used in Cheng and Stephens (1989) and Shao and Hahn (1994) are similar to those given in Cramér (1946) for the ML method, e.g. it is assumed that \( f_\theta(\cdot) \) as a function of \( \theta \) is three times differentiable. In Ranneby (1985) on the other hand, the existence of third order derivatives of \( f_\theta \) with respect to \( \theta \) is not required. However, the conditions of Ranneby (1985) are comparatively difficult to check.

A more satisfactory result is given in Ghosh and Jammalamadaka (1996). They show that any consistent root of the MSP equation \( \partial S_n(\theta)/\partial \theta = 0 \) is asymptotically normal under general assumptions. It is assumed that the distributions have common support, that \( S_n(\theta) \) is differentiable in an open neighbourhood \( I_0 \subset \Theta \) of \( \theta^0 \) and that \( \int f_\theta(x)dx \) is twice differentiable under the integral sign. They also impose an assumption on \( \partial f_\theta(F_\theta^{-1}(x))/\partial \theta|_{\theta=\theta^0} \). As mentioned in the previous subsection, Ghosh and Jammalamadaka (1996) give results for a class of estimation methods of which the MSP method is a special case (see Section 5.2 for further comments).

Because of the form of the asymptotic covariance matrix \( I(\theta^0)^{-1} \), the estimator \( \hat{\theta}_n \) is generally regarded as an asymptotically efficient estimator of \( \theta^0 \). The MSPE, just as the MLE, may also be "hyper"-efficient in the sense of having variance less than the usual order \( n^{-1} \), e.g. Examples 1 and 3(ii).
4.3 Goodness of fit and confidence regions

The set \( \{ F_{\theta^0}(\xi_{(i+1)}) - F_{\theta^0}(\xi_{(i)}) \} \) has the same distribution as a set of uniform spacings. Therefore \( \sqrt{n} (S_n(\theta^0) + \gamma)/(\pi^2/6 - 1) \), where \( \gamma \) is Euler's constant, is asymptotically normally distributed with mean 0 and variance 1 (see Darling (1953)). Further, in Cheng and Stephens (1989) it is shown under mild assumptions that \( S_n(\theta^0) \) and \( S_n(\hat{\theta}_n) \) have the same asymptotic distribution. Thus, as stated in Ranneby (1984) and Cheng and Stephens (1989) among others, the estimation problem of \( \theta^0 \) can be solved at the same time as a goodness of fit test using the function \( S_n \) (with estimated parameters).

Also, since \( S_n(\theta^0) \) has a distribution independent of the model, a \((1 - \alpha)\)100% confidence region for \( \theta^0 \) can be defined as all points \( \theta \) for which \( S_n(\theta) > s_\alpha \), where \( s_\alpha \) is the lower \( \alpha \) quantile point of the distribution of \( S_n(\theta^0) \) (see Roeder (1990, 1992) and Cheng and Traylor (1995)).

4.4 Miscellaneous remarks

Cheng and Amin (1983) gave a brief discussion on sufficiency of the MSPE and showed that in some situations an MSPE can be a function of sufficient statistics, while an MLE is not. However, in general, MSPEs will not necessarily be functions of a minimal sufficient statistic since, by the Neyman-Fisher Factorization Theorem, sufficient statistics are related to likelihood functions rather than distribution functions.

In Lind (1994) the connection to information theory is discussed. Cheng and Iles (1987) discussed "corrected" ML estimation and MSP estimation in nonregular problems, e.g. Example 3. Handling censored data is described in Cheng and Traylor (1995). They further discussed some weaknesses of the MSP method: the numerical efforts required in calculating MSPEs and the problem of tied observations. Roeder (1990) recommended using second order spacings instead of simple spacings, since they are more robust to near ties. In Roeder (1990, 1992) the MSP method is successfully used in semiparametric estimation of normal mixture densities.

For both MLEs and MSPEs, little can be said about small sample properties. However, some simulation studies have been performed comparing these methods. A study by Shah and Gokhale (1993) shows that the MSP method is superior to the ML method for many parametric configurations of the Burr XII family of distributions, described by

\[
F(x) = 1 - \left\{ 1 + \left( \frac{x - \mu}{\sigma} \right)^c \right\}^{-k}, \quad c, k, \sigma > 0, \ x > \mu.
\]
5 Summary of the thesis

5.1 Paper A

In the first thesis paper, Paper A, the AMSPE is shown to be strongly consistent, i.e. almost surely convergent, under the conditions of Ekström (1994). For comments on the conditions given in Ekström (1994), see Section 4.1.

The proof of strong consistency is approximately as follows: By introducing the random variables

\[ \eta_i(n) = (n + 1) \cdot "\text{the distance from } \xi_i \text{ to the nearest observation to the right of } \xi_i" \quad (\text{this distance is defined as } +\infty \text{ if } \xi_i = \max_{1 \leq j \leq n} \xi_j) \]

and

\[ z_i(n, \theta) = (n + 1) \left( F_\theta \left( \xi_i + \frac{\eta_i(n)}{n + 1} \right) - F_\theta (\xi_i) \right), \]

we can write

\[ S_n(\theta) = \frac{1}{n + 1} \log \left( (n + 1) F_\theta (\min_{1 \leq j \leq n} \xi_j) \right) + \frac{1}{n + 1} \sum_{i=1}^{n} \log z_i(n, \theta). \]

Note that \( T_n(\theta) = (n + 1)^{-1} \sum \log z_i(n, \theta) \) is a sum of identically distributed random variables. To avoid problems with small values, a truncated version \( T_n(M, \theta) \) of \( T_n(\theta) \), obtained by truncating each term \( \log z_i(n, \theta) \) from below by \(-M\), is introduced. Then it is shown that almost surely,

\[ \lim_{n \to \infty} S_n(\theta) \leq \lim_{M \to \infty} \lim_{n \to \infty} T_n(M, \theta) \leq \lim_{n \to \infty} S_n(\theta^0), \]

with equalities if and only if \( f_\theta(x) = f_{\theta^0}(x) \) a.e.. Further, by using Ranneby's (1984) continuity condition it follows that the convergence of \( T_n(M, \theta) \) is uniform in \( \theta \) as \( n \to \infty \). Finally, the strong convergence of the AMSPE is deduced by incorporating an identifiability condition.

The cumbersome step in the proof is the almost sure convergence of \( T_n(M, \theta) \) as \( n \to \infty \). But just as the strong law of large numbers for i.i.d. random variables plays a significant role in Wald's (1949) proof of strong consistency of AMLEs, the following result obtained in Paper A is a cornerstone to the proof of strong consistency of AMSPEs (i.e. it implies the almost sure convergence of \( T_n(M, \theta) \)).

Let \( h_n(\cdot, \cdot) \) be a real valued measurable function such that for some constant \( C \),

\[ \sup_{n,(x,y)\in\mathbb{R} \times \mathbb{R}^+} \left| h_n(x, y) \right| \leq C. \]

Then almost surely,

\[ \frac{1}{n} \sum_{i=1}^{n} \left( h_n(\xi_i, \eta_i(n)) - E[h_n(\xi_1, \eta_1(n))] \right) \to 0 \quad \text{as} \quad n \to \infty. \quad (8) \]
Note that for the special case \( h_n(\xi, \eta(n)) \equiv h(\eta(n)) \), this result becomes a strong limit theorem for spacing statistics of the form

\[
\frac{1}{n} \sum_{i=1}^{n+1} h\left((n+1) \left(\xi(i) - \xi(i-1)\right)\right).
\]

The proof of (8) is based on an investigation of the fourth order moments of \( n^{-1} \sum h_n(\xi, \eta(n)) \), together with an application of the Borel-Cantelli Lemma.

### 5.2 Paper B and Paper C

Ranneby (1984) asked the question whether it is possible to obtain better methods by approximating information measures other than the Kullback-Leibler information, such as the Hellinger distance. In Papers B and C a new class of estimation methods, called generalized maximum spacing (GMSP) methods, is derived from approximations based on spacings of so-called \( \phi \)-divergences. If \( \phi \) denotes an arbitrary convex function on the positive half real line, then the quantity

\[
I_\phi(P_\theta, P_{\theta^0}) = \int_{-\infty}^{\infty} f_{\theta^0}(x) \phi \left( \frac{f_\theta(x)}{f_{\theta^0}(x)} \right) \, dx
\]

is called a \( \phi \)-divergence of the distributions \( F_\theta \) and \( F_{\theta^0} \) (introduced by Csiszár (1963) as an information-type measure). Note that information measures such as the Kullback-Leibler information, the Jeffreys divergence and the Hellinger distance, are \( \phi \)-divergences or functions of a \( \phi \)-divergence.

The general idea behind the derivation of estimation methods from approximations of information measures is as follows: given a measure (a metric), e.g. the Kullback-Leibler information, of the distance between the distributions in our statistical model and the true underlying distribution a good inference method ought to make this distance as small as possible. Since the true distribution is not completely known we have to use our prior knowledge and observations to approximate the distance. Then we obtain our method for statistical inference by making the distance, in the approximation, as small as possible. Approximating different "metrics" we get different methods for statistical analysis.

It should be noted that these ideas are not new, e.g. Csiszar (1977) described how the distribution of a discrete random variable can be estimated by using approximations of \( \phi \)-divergences. Moreover, in Beran (1977) a general estimation method for absolutely continuous univariate distributions is based on an approximation of the Hellinger distance. In Beran's paper the Hellinger distance is approximated by using a kernel estimator of the underlying density function.

In Papers B and C another approach is provided. It is obtained by approximating \( I_\phi(P_\theta, P_{\theta^0}) \) by the spacing statistic

\[
\frac{1}{n} \sum_{j=0}^{n-m+1} \phi \left( \frac{n+1}{m} \left( F_\theta(\xi_{i+m}) - F_\theta(\xi_{i}) \right) \right).
\]
Although, at first sight, this approximation is not of the "plug in" type, its heuristic justification lies in the fact that $\frac{n+1}{m}(F_\theta(\xi(j+m)) - F_\theta(\xi(j)))$ (assume $m = 2k-1$ where $k$ is a positive integer) is a nonparametric estimator of $f_\theta(x)/f_\theta(x)$, $x \in [\xi(j+k-1), \xi(j+k))$. This estimator, a nearest neighbour density estimator, was introduced by Yu (1986).

If we define $\Psi(x) = -\phi(x)$, then the minimizing of the sum (9) is equivalent to maximizing

$$S_{\Psi,n}^{(m)}(\theta) = \frac{1}{n} \sum_{j=0}^{n-m+1} \Psi \left( \frac{n+1}{m}(F_\theta(\xi(j+m)) - F_\theta(\xi(j))) \right).$$

**Definition 3** Any $\hat{\theta}_{\Psi,n}^{(m)} \in \Theta$ which maximizes $S_{\Psi,n}^{(m)}(\theta)$ over $\Theta$ is called a generalized maximum spacing estimator (GMSPE) of $\theta^0$.

In both Paper B (for $m = 1$) and Paper C (for $m \geq 1$), consistency theorems for GMSPEs (or to be more specific, approximate GMSPEs) are given under general conditions for a large class of $\Psi$-functions. The conditions are closely related to those given in Ekström (1994) and Paper A. The ideas behind the proofs are similar to those behind the proof of strong consistency for AMSPEs, but with

$$\eta(n, m) = (n+1) \cdot \text{"the distance from } \xi_i \text{ to the } m \text{th nearest observation to the right of } \xi_i," \text{ (this distance is defined as } +\infty \text{ if } \xi_i \geq \xi_{(n-m+1)} \text{)}$$

instead of $\eta(n)$, where $m > 1$, and with $\Psi(x)$ instead of $\log x$. Also, since the theorems in Papers B and C are given in terms of convergence in probability rather than almost sure convergence, it suffices to consider the second order moments (of a truncated version) of $S_{\Psi,n}^{(m)}(\theta)$.

In Paper B we also discuss some (unpublished) results from Nordahl (1994) concerning asymptotic normality of GMSPEs based on simple spacings. It was found that the lower bound in the Cramér-Rao inequality is reached only for the GMSPEs, i.e. when $\Psi(x) = C_1 + C_2 x + C_3 \log x$ for some constants $C_1, C_2$ and $C_3 > 0$. Note that the estimator $\hat{\theta}_{\Psi,n}^{(1)}$ does not depend on the values chosen for $C_1, C_2$ and $C_3 > 0$, so we may choose $C_1 = C_2 = 0$ and $C_3 = 1$. Consequently, for $m \equiv 1$ we entail a loss of asymptotic efficiency (in regular problems) when we base the methods on information measures other than the Kullback-Leibler information. The case when $m > 1$ is an open question at this point. However, as pointed out in Paper C, it appears that for many choices of $\Psi$ one should allow $m$ to increase with $n$ (at some suitable rate).

Statistical inferences are based only in part upon observations. An equally important base is formed by prior assumptions about the underlying situation. There are assumptions about randomness and independence, about distribution models and so on. These kind of assumptions are not expected to hold exactly;
they are mathematically convenient rationalizations. Therefore it is desired that any statistical procedure possesses the following features:

- It should have reasonably good (optimal or nearly optimal) efficiency under the assigned model.
- Small deviations from the model assumptions should impair the performance only slightly.
- Somewhat larger deviations from the model assumptions should not cause a "catastrophe".

Procedures, e.g. estimation methods, satisfying these features are called robust. Extensive studies of robust procedures started with Tukey (1960) and others. For a general qualitative definition of robustness, see Hampel (1971).

In Papers B and C we discuss distributional robustness of GMSPEs, i.e. the behaviour of the estimators when the shape of the true underlying distribution deviates slightly from the assumed model. As in Nordahl (1992), who conducted simulation studies for some GMSPEs based on simple spacings for moderate sample sizes, we took a closer look at the case where the model is the normal distribution with unknown location parameter, but where the data are generated from an ɛ-contaminated normal distribution. Nordahl (1992) found that the GMSPE based on the Hellinger distance with \[ \Psi(x) = -(1 - \sqrt{x})^2 \] (or equivalently \[ \Psi(x) = \sqrt{x} \]) behaves "better" than the MLE and the MSPE, i.e. it is less influenced by the contaminating distribution. However, under the true model this GMSPE has a (asymptotic) variance which is approximately 9% larger than that of the MSPE. In Paper B we found that the choice \[ \Psi(x) = x^{0.1} \] gives an estimator which has robustness properties similar to the GMSPE based on the Hellinger distance, but with an asymptotic variance only approximately 0.6% larger than that of the MSPE under the true model. Further simulations (see Paper C) have shown that Nordahl's (1992) results on the GMSPE based on the Hellinger distance can be improved by using high order spacings. This is also true for the choice \[ \Psi(x) = x^{0.1}. \]

Note that with \[ \Psi(x) = x^r, \] where \( r > 0, \) \( S_{\Psi,n}^{(1)} \) is the statistic that Kimball (1950) suggested for goodness of fit tests, see Section 3.

GMSP methods based on simple spacings were introduced independently in Ghosh and Jammalamadaka (1996). As in Papers A and B they motivate the introduction of the methods by relating them to information measures like the Kullback-Leibler information and the Hellinger distance. In a simulation study of the three parameter Weibull model, i.e. Example 3, it is shown that the GMSP method based on \[ \Psi(x) = -|x - 1| \] can perform better than the MSP method in terms of mean squared error.
Under particular regularity conditions (described earlier in Sections 4.1 (consistency) and 4.2 (asymptotic normality)) Ghosh and Jammalamadaka show that GMSP estimates are consistent and asymptotically normal for a class of $\Psi$-functions. As noted in Section 4.1, their regularity conditions do not cover Examples 1, 3, 4 and 5. However, the conditions stated for consistency in Papers B and C hold in these Examples for many different $\Psi$-functions.

5.3 Paper D

In Paper D several strong limit results are given for sums of logarithms of high order spacings, i.e. for statistics of the form

$$L_n^{(m)}(\xi_1, \ldots, \xi_n) = \frac{1}{n} \sum_{j=0}^{n-m+1} \log \left( \frac{n+1}{m} \left( \xi_{(j+m)} - \xi_{(j)} \right) \right).$$

For all results, the order of the spacings is allowed to increase to infinity with the sample size. However, it should be pointed out that we do not require that the order must increase with $n$.

Note that $L_n^{(m)}$ is Cressie's (1976) statistic for goodness of fit tests (see Section 3). In Paper D it is proved that this goodness of fit test is strongly consistent against all continuous alternatives, for $m$ fixed as well as for $m$ increasing to infinity with $n$. It should also be noted that for $m$ even, $L_n^{(m)}$ is related to Vasicek's (1976) entropy estimator. Vasicek (1976), however, used $L_n^{(m)}$ for testing normality.

In the first part of Paper D, spacings of uniform random variables on the interval $[0,1]$ are considered. For example, we provide the following characterization of the uniform distribution.

If $\xi_1, \ldots, \xi_n$ are i.i.d. random variables on $[0,1]$ and if $m_n = o(n/\log n)$, then the sample is uniformly distributed if and only if

$$\lim_{n \to \infty} \left( L_n^{(m_n)}(\xi_1, \ldots, \xi_n) - \psi(m_n) + \log m_n \right) = 0 \quad \text{a.s.,}$$

where $\psi(x) = \frac{d}{dx} \log \Gamma(x)$ is the digamma function and $\Gamma$ the gamma function. This result generalizes a result of Shao and Hahn (1995) who considered the special case $m = 1$.

Further, for general i.i.d. random variables $\xi_1, \ldots, \xi_n$ with density function $f$ defined on a finite interval $[a, b]$, it is shown that if $m_n = o(n/\log n)$, then almost surely

$$\lim_{n \to \infty} \left( L_n^{(m_n)}(\xi_1, \ldots, \xi_n) - \psi(m_n) + \log m_n \right) \leq - \int_a^b f(x) \log f(x) \, dx.$$
The integral \( \int f(x) \log f(x) \, dx \) is known as the entropy of an absolutely continuous distribution \( F \) with density \( f \) (see Shannon (1948)). If, in addition, the density \( f \) above is bounded away from infinity, then \( L_n^{(m)}(\xi_1, \ldots, \xi_n) - \psi(m_n) + \log m_n \) actually becomes a nonparametric, strongly consistent estimator of entropy.

Van Es (1992) gave strong limit results for a statistic that has \( L_n^{(m)}(\xi_1, \ldots, \xi_n) \) as a special case, both for the case where \( m \) is fixed and where \( m/n \to \infty \), \( m/n \to 0 \), under the assumptions that \( f \) is bounded away from zero and infinity on its support. Thus, in comparison with Van Es (1992), we impose less restrictive assumptions on the density \( f \). Moreover, in Van Es (1992) the approach of Vasicek (1976), using Stieltjes sums, was used. In Paper D we adopt a different approach, related to that of Shao and Hahn (1995) who gave results for \( L_n^{(1)} \).

As mentioned earlier, \( L_n^{(m)}(\xi_1, \ldots, \xi_n) \) for \( m \) even is related to Vasicek’s (1976) entropy estimator. Beirlant and Van Zuijlen (1985), by using their Glivenko-Cantelli strong limit theorem for the distribution of uniform \( m \)th order spacings, gave a strong limit theorem for the case where \( m \to \infty \) and \( m = O(n^{1-\varepsilon}) \), for some \( 0 < \varepsilon < 1 \). However, they did not consider the case when \( m \not\to \infty \).

In the following we use the notation introduced in Section 4.1, but with

\[
S_n^{(m)}(P) = \frac{1}{n} \sum_{j=0}^{n-m+1} \log \left( \frac{n+1}{m} \left( F_P(\xi_{j+m}) - F_P(\xi_j) \right) \right)
\]

rather than \( S_n(P) \). In Paper D the following generalization of Shao and Hahn’s inequality (7) was given:

Let \( \xi_1, \ldots, \xi_n \) be i.i.d. from \( P_0 \in \mathcal{P} \). Then, if \( P \neq P_0 \), \( P \in \mathcal{P} \) and \( m_n = o(n/\log n) \), then almost surely

\[
\lim_{n \to \infty} \left( S_n^{(m_n)}(P) - S_n^{(m_n)}(P_0) \right) \leq \int_{-\infty}^{\infty} \log \frac{f_P(x)}{f_{P_0}(x)} \, dF_{P_0}(x) < 0. \tag{10}
\]

### 5.4 Paper E

The result (10) is a cornerstone to a proof given in Paper E of the approximate GMSPE (AGMSPE) \( P_{n,m_n}^* \), \( P_{n,m_n}^* \in \mathcal{P} \) for all \( n \), defined by

\[
S_n^{(m_n)}(P_{n,m_n}^*) \geq -c_n + \sup_{P \in \mathcal{P}} S_n^{(m_n)}(P),
\]

where \( c_n > 0 \) and \( c_n \to 0 \) as \( n \to \infty \). The AGMSPE \( P_{n,m_n}^* \), with \( m_n = o(n/\log n) \), is shown to be consistent under the conditions of Shao and Hahn (1996a). By using the inequality (10) the proof follows by a slight generalization of the corresponding proof of Shao and Hahn (1996a). No advantage to letting the order of the spacings tend to infinity when \( \Psi(x) = \log x \) is presently known, but
second-order spacings was recommended by Roeder (1990). She found through simulations that the second-order spacings are more robust to near ties, as expected, and that the estimators based on second-order spacings performed as well as those based on simple spacings.

Remark. The proof was actually given for estimates \( Q^*_{n,m_n} \in \mathcal{P} \) of \( P_0 \), defined by

\[
\lim_{n \to \infty} S_n^{(m_n)}(Q^*_{n,m_n}) \geq \lim_{n \to \infty} S_n^{(m_n)}(P_0) \quad \text{a.s.}
\]

\[
= \lim_{n \to \infty} (\psi(m_n) - \log m_n) \quad \text{a.s.,}
\]

(11)

where \( \psi(x) = \frac{d}{dx} \log \Gamma(x) \) is the digamma function and where \( \{m_n\} \) is a nondecreasing sequence of positive integers. But any AGMSPE defined above satisfies the inequality in (11). The equality in (11) follows from application of a result in Paper D and holds if \( m_n = o(n) \).

Paper E also presents consistency theorems for GMSPEs and AGMSPEs based on information measures other than the Kullback-Leibler information, but only for those based on simple spacings. A smaller class of \( \Psi \)-functions is considered than in Papers B and C, i.e. only \( \Psi \)-functions bounded from below were studied, with \( \Psi(x) = \log x \) as an exception. The proofs are related to those of Shao and Hahn (1996a), but they are based on the inequality

\[
\lim_{n \to \infty} S_{\Psi,n}^{(1)}(P) \leq \int_{-\infty}^{\infty} \int_{0}^{\infty} \Psi(yf_P(x)) f_P^2(x) e^{-yf_P(x)} dy dx \quad \text{a.s.,}
\]

(12)

rather than on Shao and Hahn's inequality (7). The inequality (12) is derived from the strong limit result (8) given in Paper A.

Shao and Hahn (1996a) proved that for any family \( \mathcal{P} \) of probability measures with unimodal densities, AMSPEs are \( L^1 \) consistent for the underlying, true unimodal density \( f_{P_0} \), without any further conditions on the distributions. That is,

\[
\lim_{n \to \infty} \int_{R} \left| f_{\Psi,n}^* - f_{P_0}(x) \right| dx = 0 \quad \text{a.s.}
\]

In Paper E we show that this result also holds for any AGMSPE based on a strictly concave \( \Psi \)-function bounded from below such that \( \Psi(x)/x \to 0 \) as \( x \to \infty \), e.g. it holds for the estimator based on the Hellinger distance (\( \Psi(x) = \sqrt{x} \)).

5.5 Final remarks

The thesis has mainly been concerned with the problem of estimating probability distributions. The methods of solving this problem, the MSP method and its generalizations, have all been based on sum-functions of spacings. The analysis
has been focused on the problem of consistency, but we have also discussed other desired properties of estimates like asymptotic efficiency and robustness. It has been shown that these estimation methods possess good properties and that they work in situations where the ML method does not. It should be pointed out that the conditions required for consistency of MSPEs and GMSPEs cover many of these types of situations.

We want to stress that we do not claim that the MSP method (or any of its generalizations) is a better estimation method than the ML method in general. We rather see the MSP method as an alternative to the ML method. Also, the question as to when the (generalized) MSPE works better than the MLE and vice versa is not easily answered.

As with all general methods the (generalized) MSP method has its weaknesses. For instance, the MSP method can be sensitive to round off errors and should be used with caution. An alternative is to use high order spacings that are less sensitive to near ties. Moreover, if we want to perform a goodness of fit test we gain power if we use high order spacings.

Another weakness of the (generalized) MSPE, in comparison with the MLE, is that it may not, in general, be a function of a sufficient statistic, since by the Neyman-Fisher Factorization Theorem sufficient statistics are related to densities (likelihoods) rather than distribution functions (on which $S_{\psi,n}^{(m)}$ is defined).

Lindsay (in the discussion of the paper by Cheng and Traylor (1995)) reported that the lack of a simple and natural generalization of spacings for multivariate data is a weakness of the MSP method. This problem was solved partially by Ranneby (1990) who proposed that Dirichlet cells should be used as the multivariate counterpart of spacings. Clearly, there are many unsolved problems in this rather difficult area.

Another unsolved problem is to show asymptotic normality for GMSPEs based on $m$th order spacings. Of particular interest is the case where the order of the spacings is allowed to increase to infinity with $n$ as well as the problem of finding an optimal rate of $m$.

At this point the (generalized) MSP method has almost exclusively been directed toward i.i.d. data (the report by Cheng and Traylor (1995) where censored data were discussed is an exception). We therefore see the necessity of giving general results for more realistic statistical data, e.g. for dependent data and for the case when the sample size is random. In order to obtain these more general results, we also see a need for very general limit theorems for statistics based on (high order) spacings (Dirichlet cells).

References


