COMPARISON OF SPLINE APPROXIMATION WITH THE MODIFIED LIKELIHOODS IN THE PRESENCE OF NUISANCE PARAMETER

AHMET SEZER
ANADOLU UNIVERSITY
Department of Statisticss
Yunus Emre Kampusu Eskisehir
TURKEY
a.sezer@anadolu.edu.tr

Abstract: The likelihood functions from independent studies can be easily combined, and the combined likelihood function serves as a meaningful indication of the support the observed data give to the various parameter values. This fact has led many scientists to suggest using the likelihood function as a summary of post-data uncertainty concerning the parameter. Indeed, likelihood functions have several desired properties . They are objective, in that they depend only on the agreed-upon model and the data. They are also flexible, allowing us to combine information about competing models across studies. However, a serious difficulty arises because likelihood functions may not be expressible in a compact, easily-understood mathematical form suitable for communication or publication. For example, likelihood functions in mixture models may only be computable for individual values of the parameters and otherwise cannot be given in ”closed form”. To overcome this difficulty, we propose to approximate log-likelihood functions by using piecewise polynomials governed by a minimal number of parameters

Key–Words: likelihood,Approximation LaTeX

1 Introduction

In parametric statistical inference, the goal is to make inference about the unknown value of a parameter θ. Suppose the observed data $y = (y_1, y_2, \ldots, y_n)$ can be modeled as independent observed values of a random variable Y distributed according to the density $f(y;\theta)$, $\theta \in \Omega$. The likelihood function of θ based on the data y is given by;

$$L(\theta) = \prod_{i=1}^{n} f(y_i;\theta). \tag{1}$$

If θ_1 and θ_2 are two possible values of θ and such that $L(\theta_1) > L(\theta_2)$, then the probability of observing the data y is greater when the true parameter value is θ_1 than when the true parameter value is θ_2. The maximum likelihood estimator(MLE) is the parameter value for which the data are most likely. Intuitively, the MLE is a reasonable choice for an estimator. Under certain regularity conditions MLE have several important properties when the sample size is large:

1. Consistency: As the sample size increases, the MLE converges in probability to the true parameter value, e.g., $\hat{\theta} \to \theta$. (Wald,1949).

2. Invariance: If g(θ) is any function of the unknown parameters of the distribution, then the MLE of g(θ) is $g\left(\hat{\theta}\right)$.

3. Asymptotic normality and efficiency: As the sample size increases, the sampling distribution of the MLE converges to a normal distribution with mean θ and variance equal to $\frac{1}{nI(\theta)}$ where $I(\theta)$ is Fisher's information for one observation. A proof of the asymptotic normality of the maximum likelihood estimator is given in Cramer (1946). The result that MLE's have the minimum possible asymptotic variance has been studied by Kalianpur and Rao (1955) and Bahadur (1964). The choice between using the observed and expected information for estimating Fisher's information of the maximum likelihood estimate has been considered in Efron and Hinkley (1978).

Because the MLE has many desirable properties as mentioned above, it is a reasonable choice for a point estimator of the parameter θ. Finding the MLE is relatively straightforward in one parameter models; one can use calculus or simply plot the likelihood function. In large samples, central limit and saddle-point approximation theory suggest that spline functions having just two or three knots may be used to approximate the log- likelihood function for a single parameter, with the approximating polynomial being quadratic inside and linear outside of the knots. Natural cubic splines, which are piecewise cubic inside of the extreme knots and linear outside of these knots, are frequently used in non-parametric regression and data mining. These two facts suggest the use of natural cubic splines to approximate the log-likelihood functions.

A non-trivial problem is how to extend this theory to approximate the log-likelihood for two or more parameters. Of course, if one of the two parameters is viewed as a “nuisance parameter” of little interest, we can reduce to the one- parameter model after the elimination of that nuisance parameter by a variety of methods proposed in the literature. If we include both parameters in the likelihood to be approximated, the approximation becomes more complicated. In order to draw inferences regarding the parameter of interest in such a two-parameter model, we must deal with the nuisance parameter. General likelihood-based methods have been proposed for the elimination of a nuisance parameter to focus on the structural parameter only. Some of these methods are: profile likelihood, marginal likelihood, conditional likelihood and integrated likelihood. Before discussing the relationships among these methods, we will introduce each method briefly.

1.1 Profile Likelihood

The simplest likelihood approach to eliminating nuisance parameters is to replace them with their maximum likelihood estimates, leading to profile likelihood. Suppose that the model can be identified by parameters (θ, λ), where θ is a parameter of interest and λ is a nuisance parameter. To eliminate λ from the likelihood, it is replaced by the maximum likelihood estimator of λ, keeping θ fixed. The resulting likelihood function; $L_p\left(\theta\right) = \sup_\lambda L\left(\theta, \lambda\right)$ is called the profile likelihood function and $\log L_p\left(\theta\right)$ is called the profile log-likelihood function.

$L_p\left(\theta\right)$ can be used as if it were an ordinary likelihood to produce asymptotically correct inference about θ. It is the simplest method, but it does not take into account the uncertainty due to lack of knowledge of the nuisance parameter and can be misleading in both precision and location (Severini,1998b). In large samples, replacing λ by its maximum likelihood estimate has relatively minor effect on inferences regarding θ. However in small samples, replacing λ by the maximum likelihood estimator may have a large effect on inference, particularly when there are several nuisance parameters in the model.

1.2 Integrated likelihood

The integrated likelihood is obtained for each fixed value of the parameter of interest by integrating out the nuisance parameter with respect to a weighting function. Let $\pi\left(\lambda\right)$ denote a weighting function defined on Λ, the space of possible values λ. Then the integrated likelihood function for θ is given by;

$$L\left(\theta;\pi\right) = \int_\Lambda L\left(\theta;\lambda\right)\pi\left(\lambda\right)d\lambda \qquad (2)$$

Inference about θ will be based on $L\left(\theta;\pi\right)$, or equivalently on the logarithm of the integrated log-likelihood function $logL\left(\theta;\pi\right)$. In this study, we will consider the use of a uniform weighting function $\pi\left(\lambda\right)$=1; the resulting integrated likelihood function will be denoted L$(\theta;U)$ to emphasize that the uniform weight function applies to λ. A comprehensive discussion of the use of integrated likelihood methods in the Bayesian approach is given by Berger, Liseo and Wolpert (1998). Integrated likelihood functions have the advantage that, unlike conditional or marginal likelihoods, they are generally available and, in principle, are relatively easy to determine, although sophisticated computational methods may be needed to evaluate the integrals that arise.

In the Bayesian literature, the noninformative prior plays an important role. When we apply the Bayesian approach, we may not have any prior information about the parameter. In this situation, the statistician tries to find a prior that provides as little information about the parameter of interest as possible. One of the most important noninformative priors is the Jeffreys’ prior which is equal to the square root of the Fisher information $I\left(\theta\right)$(Jeffreys 1939).

Sweeting (1995) has shown that one-parameter methods which operate solely on the likelihood L(θ) can also be used with an integrated likelihood. Examples of such methods are: (i) using the mode $\hat{\theta}$ of the likelihood as the estimate of θ, and (ii) using,

$$C = \left\{ \theta : -2log \left(L \left(\hat{\theta} \right) / L \left(\theta \right) \right) < X_p^2 \left(1 - \alpha \right) \right\} \tag{3}$$

as an approximate 100 $(1-\alpha)$ confidence set for θ, where $X_p^2(1-\alpha)$ is the $(1-\alpha)$th quantile of the chi-squared distribution with p degrees of freedom (for a scalar θ, p=1). By using these facts, we will be able to compare the accuracy of inference about θ based on the integrated likelihood with that based on other modified likelihoods (profile likelihood and conditional likelihood) by comparing the coverage probabilities of confidence intervals of the form (2.3) based on such modified likelihoods.

1.3 Marginal Likelihood

Another way to eliminate a nuisance parameter is to construct a likelihood function based on a statistic **T** having the property that the distribution of **T** depends only on θ. Then we may form a genuine likelihood function for θ based on the density function of **T**; such a likelihood function is called a marginal likelihood function, since it is based on the marginal distribution of **T**. Marginal likelihoods were considered in detail by Kalbfleisch and Sprott (1970). The main drawback of this approach is that we may not be using all of the available information about θ in the data.

Suppose that there exists a statistic **T**= **T(y)** such that the density of the data **Y** may be written

$$P(y;\theta,\lambda) = P(t;\theta)\, P(y|t;\theta,\lambda)\,. \tag{4}$$

In (2.4), the marginal likelihood function based on t is given by

$$L(\theta;t) = P(t;\theta)\,, \tag{5}$$

Whereas the joint likelihood function for (θ,λ) is given by P$(y;\theta,\lambda)$.

1.4 Conditional Likelihood

Another approach to eliminating nuisance parameters can be applied whenever there exists a statistic **S** = **S (y)** such that the conditional distribution of the data **y** given **S**=s depends only on θ. In this case, we may form a genuine likelihood function for θ based on the conditional density of **y** given **S**=s ; this is called a conditional likelihood function. Suppose that the data can be transformed to the vector (t, s) such that;

$$P(t,s:\theta,\lambda) = P(t|s;\theta)\, P(s;\theta,\lambda)\,. \tag{6}$$

The statistic **S** is a sufficient statistic for θ in the model with λ held fixed. A likelihood function for θ may be based on P$(t|s;\theta)$ which does not depend on λ; the resulting conditional likelihood function for θ is a genuine likelihood function. The use of conditional likelihood inference in models with many nuisance parameters was discussed by Anderson (1970), Kalbfleisch and Sprott (1970) and van der Vaart (1988).

2 Polynomial Splines

We may have models with more than one parameter where various components of the parameter vector have different levels of interest. Consider a model parameterized by a two-dimensional vector of parameters (θ,λ), where θ is the parameter of interest (called a structural parameter) and λ is a nuisance parameter. For models with only one parameter, θ, inference may be based on the likelihood function L(θ). However, when a nuisance parameter is present, it is not possible to use the likelihood function to directly compare different values of θ. Indeed, the greater the dimension of the nuisance parameter, the greater its potential effect on the conclusions regarding the parameter of interest (Berger, Liseo and Wolpert, 1998).
Polynomials have played an important role in approximation theory for many years. The main drawback of polynomials for approximation purposes is that the class is relatively inflexible. Polynomials work well on sufficiently small intervals, but when we go to larger intervals, severe oscillations often appear. In order to achieve a class of approximating functions with greater flexibility, we can divide up the interval of interest into smaller pieces. In this chapter we will justify use of the cubic spline approximation for one parameter models.

Polynomial splines are piecewise polynomials of some degree **d**. The breakpoints marking a transition from one polynomial to the next are referred to as “knots”. A piecewise polynomial function $f(y)$ is obtained by dividing the domain of Y into contiguous intervals and f can be separated by a polynomial in each interval. In the literature on approximation theory, the term ”linear spline” is applied to a continuous piecewise linear function.

Similarly, the term ”cubic spline” is reserved for piecewise cubic functions having two continuous derivatives, allowing jumps in the third derivative at the knots. It is common in statistics to require a simple approximation for a smooth relationship between response and predictor variable. Such relationship may be known but complicated or unknown. The cubic spline functions are very popular in data mining to serve for this job.

Given a maximum polynomial degree **d** and a knot vector **t**, the collection of polynomial splines having s continuous derivatives form a linear space. For example the collection of linear splines with knot sequence $(t_1, ..., t_k)$ is spanned by the functions

$$1, y, (y - t_1)_+, \ldots, (y - t_k)_+ . \qquad (7)$$

where $(.)_+$ = max(., 0). This set is called the truncated power basis of the space. Classical cubic spline have d=3 and s=2 so that the basis has elements

$$1, y, y^2, (y - t_1)^3_+, \ldots, (y - t_k)^3_+ \qquad (8)$$

However the truncated power functions (3.1) and (3.2) are known to have rather poor numerical properties. In linear regression problems, for example, the condition number of the design matrix deteriorates rapidly as the number of knots increases (Hansen and Kooperberg,2002).

Extended linear models (ELMs) were defined as a theoretical tool to understand the properties of spline-based procedures in a large class of estimation problems (Hansen,1994; Huang 2001). This class contains all of the standard generalized linear models as well as density and conditional density estimation, hazard regression, censored regression and spectral density estimation.

Friedman (1991) introduced multivariate adaptive regression splines (MARS), which is a polynomial spline methodology for estimating regression functions. Multivariate adaptive regression splines (MARS) is a method for flexible modeling of high dimensional data. The MARS method has become very popular in data mining because it does not assume or require any particular type of relationship between predictor variables and response variable. However, it has more power and flexibility to model relationships that are nearly additive. The multivariate adaptive regression splines (MARS) model can be written as

$$f(y; \beta) = \sum_{j=1}^{J} \beta_j B_j(y) \qquad (9)$$

for a given set of basis functions $B_1(y)$ ....$B_J(y)$. The unknown parameters $\beta_1 \cdots \beta_J$ in MARS are estimated using least squares.

In functional ANOVA, spline basis elements and their tensor products are used to construct the main effects and interactions. Stone (1994) gave the first theoretical treatment of convergence of spline estimation with functional ANOVA decompositions.

Most of the early applications of splines were focused mainly on curve estimation. These tools also have proved effective for multivariate problems. In the context of density estimation, the log-spline procedure of Kooperberg and Stone (1991) shows excellent spatial adaptation, capturing the full height of spikes without overfitting smoother regions. Note that approximation of densities by log-spline is similar to approximation of log-likelihoods in the sense that the argument y of the density plays the same role as the value of the chosen parameter values θ in log- likelihood estimation.

Kooperberg and Stone (1991,1992) modeled the log-density as a natural cubic spline. Like the log-spline in density estimation, log-likelihood can be modeled as a natural cubic spline. Indeed in large samples, central limit and saddle-point approximation theory suggest that cubic splines may be used to approximate the log-likelihood functions.

Natural cubic splines are twice continuously differentiable, piecewise polynomials defined relative to a knot sequence $t = (t_1, ...t_k)$. Within each interval $[t_1, t_2], ...[t_{K-1}, t_K]$, natural cubic splines are cubic polynomials, but on $(L, t_1]$ and $[t_K, U)$ (beyond the first and last knots) they are forced to be linear. We assume that log-likelihood can be written in the form:

$$logL(y; \beta) = \sum_{j=1}^{J} \beta_j B_j(y_j) \qquad (10)$$

where J is the number of basis functions. The basis of natural cubic spline with K knots is $B_1(y) = 1$, $B_2(y) = y$, $B_{k+2}(y) = d_k(y) - d_{K-1}(y)$, $k = 1, ..., K-2$, where

$$d_k(y) = \frac{(y - t_k)^3_+ - (y - t_K)^3_+}{t_K - t_k}. \qquad (11)$$

Another basis representation of the natural cubic splines is the B-spline basis
(De Boor,1978). B-splines are constructed from polynomial pieces joined at certain values of y. Once the knots are given, it is easy to compute the B-splines

recursively, for any desired degree of the polynomial.

An important question in spline modeling is to decide number of knots in the model. The choice of knots has been a subject of much research; too many knots lead to overfitting of the data, too few knots lead to underfitting. Some authors have proposed automatic schemes for optimizing the number and the position of knots (Friedman and Silverman, 1989 ; Kooperberg and Stone 1991). Kooperberg and Stone (1991) found that it matters less where the knots are placed than how many knots are chosen. Fortunately, equally spaced fixed knots ensure that there are enough data within each region to get sensible fits. This choice also guards against outliers overly influencing the fitted curve. In this study we will determine the appropriate numbers of knots by the AIC method and will assume that knots are equally spaced.

The knot selection methodology in log-spline density estimation involves initial knot placement, stepwise knot addition, stepwise knot deletion and final model selection based on the Akaike information criteria (AIC). Log-spline density estimation is discussed by Stone (1990), who used a basis of the form $1, B_1(y,), ...B_J(y,)$ of natural cubic spline functions, where $j = k-1$, $t = (t_1, ..., t_k)$. Let t be fixed and G denote the J-dimensional span of the functions $B_1, ...B_J$, so that any $g \in G$ is of the form;

$$g(y, \beta, t) = \beta_1 B_1(y,t) + ... + \beta_J B_J(y,t). \tag{12}$$

Then, density functions on (L, U) of the form

$$f(y : \beta, t) = \exp[g(y : \beta, t) - C(\beta, t)] \tag{13}$$

$$= \exp(\beta_1 B_1(y,t) + ... + \beta_j B_j(y,t) - C(\beta, t)) \tag{14}$$

(called a log-Spline Model) are used to approximate a member of exponential family density functions of interest. Given a random sample $Y_1, ...Y_n$ of size n from a distribution on (L, U) having an unknown density function in an exponential family, the log-likelihood function corresponding to the log-spline model is given by

$$L(\beta, t) = \sum_i \sum_j \beta_i B_j(Y_i, t) - nC(\beta, t) \tag{15}$$

Then, the maximum likelihood estimate of β is calculated by arg max $L(\beta, t)$. The MLE's of β and fixed choice of knots **t** can be found efficiently in reasonably-sized problems through simple Newton-Raphson iteration.

2.1 MODEL SELECTION CRITERION IN POLYNOMIAL SPLINE

To approximate the log-likelihood functions, first we will sample n values of the parameter vector uniformly from a chosen compact convex region in the parameter space. Next, the log-likelihood for each choice of parameter vector is calculated. Then the log-likelihood is taken to be the response and the elements of θ to be the predictors, and in a regression model we will fit by natural cubic splines. However before going through this process, we should determine the optimum number of knots that we include in the models. Clearly the number of knots will determine the complexity of the model.

In this chapter we describe and illustrate the methods that we used to select the number of knots in spline models. We assume that we have the target variable Y(log-likelihood), a vector of inputs x(parameter vector) and the prediction model $\hat{f}(x)$. The loss function for measuring errors between Y and $\hat{f}(x)$ is denoted by $L\left(Y, \hat{f}(x)\right)$ and is the squared error loss function.

$$L\left(Y, \hat{f}(x)\right) = \left(Y - \hat{f}(x)\right)^2 \tag{16}$$

The test error (Err) is the expected prediction error over an independent test sample.

$$Err = E[L\left(Y, \hat{f}(x)\right)] \tag{17}$$

We estimate the test error from the training error and training error as the average loss over the training sample:

$$\overline{err} = \frac{1}{N} \sum_{i=1}^{N} L\left(Y, \hat{f}(x_i)\right) \tag{18}$$

Training error decreases as model complexity increases, finally dropping to zero if we increase the model complexity enough. For this reason, training error is not a good estimate of the test error. Fortunately, there is an optimal model complexity that gives the minimum test error. Indeed Akaike and Bayesian information criterion trade between bias and variance to get optimum level of the test error.

2.2 Akaike Information Criterion

Akaike information criterion(AIC) is one of the most popular methods to determine the optimum number of variables in the ideal model. Assume that we have the linear model, $Y = f(x) + \epsilon$ where $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma_\epsilon^2$. We can derive an expression for the expected prediction error of a regression fit $\hat{f}(x)$ at an input point $X = x_0$ using squared error loss

$$Err(x_0) = E[(Y - \hat{f}(x_0))^2 | X = x_0] \tag{19}$$

$$= \sigma_\epsilon^2 + [E\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 \tag{20}$$

$$= \sigma_\epsilon^2 + Bias^2\left(\hat{f}(x_0)\right) + Var\left(\hat{f}(x_0)\right) \tag{21}$$

$$= IrreducibleError + Bias^2 + variance. \tag{22}$$

For a linear model fit $\hat{f}_p(x) = \hat{\beta}^T X$, where the parameter vector β with p components is fit by least squares,

$$Err(x_0) = \sigma_\epsilon^2 + [\hat{f}(x_0) - E\hat{f}_p(x_0)]^2 + \|h(x_0)\|^2 \sigma_\epsilon^2 \tag{23}$$

$h(x_0)$ is the N-vector of linear weights that produce the fit $\hat{f}_p(x_0) = X_o^T \left(X^T X\right)^{-1} X^T Y$, and hence

$$Var[\hat{f}_p(x_0)] = \|h(x_0)\|^2 \sigma_\epsilon^2. \tag{24}$$

This variance changes with x_0 and its average(over the sample values x_i) is $\frac{p}{N}\sigma_\epsilon^2$ and the in-sample error is

$$\frac{1}{N}\sum_{i=1}^{N} Err(x_i) = \sigma_\epsilon^2 + \frac{1}{N}\sum_{i=1}^{N}[f(x_i) - E\hat{f}(x_i)]^2 + \frac{p}{N}\sigma_\epsilon^2. \tag{25}$$

Here the model complexity is directly related to the number of parameters p. We define optimism as the expected difference between Err_{in} and the training error $e\bar{r}r$. Let Y^{new} indicates that we observe N new response values at each of the training points x_i, $i = 1, ..., N$

$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N} E_y E_{Y^{new}} L\left(Y_i^{new}, \hat{f}(x_i)\right). \tag{26}$$

$$optimism = Err_{in} - E_y\left(\overline{err}\right). \tag{27}$$

For a squared error loss function, we can show that,

$$optimism = \frac{2}{N}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) \tag{28}$$

where Cov indicates covariance. Then we have the important relation

$$Err_{in} = E_y\left(\overline{err}\right) + \frac{2}{N}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) \tag{29}$$

for the additive error model $y = f(x) + \epsilon$,

$$Err_{in} = E_y\left(\overline{err}\right) + 2\frac{d}{N}\sigma_\epsilon^2 \tag{30}$$

Simply, the AIC is an estimate of Err_{in} when a log-likelihood loss function is used.

$$AIC = -2loglik + 2\frac{d}{N}\sigma_\epsilon^2 \tag{31}$$

To use AIC for model selection in spline modeling, we choose the model giving smallest AIC over the set of models. Let $f_\alpha(x)$ be the set of models indexed by a tuning parameter α and denote by $\overline{err}(\alpha)$ and $d(\alpha)$ the number of parameters for each model, then we can define

$$AIC(\alpha) = \overline{err}(\alpha) + 2\frac{d(\alpha)}{N}\hat{\sigma}_\epsilon^2 \tag{32}$$

The function AIC(α) provides an estimate of the test error and we choose the tuning parameter $\hat{\alpha}$ that minimizes the test error. Typically, with the spline models the tuning parameter is the number of knots in the spline model. The number of knots controls the complexity of the spline model and we want to find the number of knots that minimizes the test error. In this study, before approximating the log-likelihoods by cubic splines, first we determine the appropriate number of knots by AIC and then we will equally space the knots to obtain the cubic spline approximation of log-likelihoods.

2.3 The Bayesian Information Criterion

The Bayesian information criterion (BIC) is applicable where the fitting is carried out by maximization of log-likelihood. The form of Bayesian information criterion (BIC) is

$$BIC = -2loglik + (logN)\,d \tag{33}$$

Under the Gaussian model, assuming the variance σ_ϵ^2 is known, we can write BIC as,

$$BIC = \frac{N}{\sigma_\epsilon^2}[\overline{err} + (logN)\,\frac{d}{N}] \tag{34}$$

BIC tends to penalize complex models more heavily than Akaike information criterion, giving preference to simpler models in selection. Our simulation results indicate that to approximate the log-likelihood of the exponential distribution both AIC and BIC agreed on 1, as the appropriate number of knots for the natural cubic spline model.

3 ASSESSING THE QUALITY OF THE APPROXIMATION

One of the most important tasks for us is to assess the quality of the approximation. We have determined three criteria to manage this job. In this chapter we are using one parameter distribution (exponential distribution) to show how to assess the approximation.

1) Coverage probability of the highest density region of the approximated
log-likelihood function.

2) Mean squared error (MSE) of the estimator obtained as the maximum over the parameter of the approximation of the log-likelihood

3) Average interval length of the highest density region.

Before presenting our results, first we will show how to find the region from which we select parameter values, and then we will introduce the algorithms that we used to find the coverage probability from the highest density region (HDR) and mean squared error (MSE) of the approximated log-likelihood functions.

To find the region for the parameter of interest we have the following algorithm. Recall that in chapter 2, we have the equation (2.3):

$$C = \left\{\theta : -2log\left(L\left(\hat{\theta}\right)/L\left(\theta\right)\right) < X_p^2\left(1-\alpha\right)\right\} \tag{35}$$

1) First find the maximum likelihood estimator of the parameter.

2) Plug in the maximum likelihood value in (2.3) for $\widehat{\theta}$

3) Solve (2.3) to find the lower and upper limit of the interval of parameter of interest.

4) Finally, parameter values are sampled uniformly from the interval between the lower and upper limits found in step(3).

3.1 Highest Density Region of the Approximated Log-likelihood Function

Statistical methods summarize a probability distribution by a region of the sample space covering a specified probability. One method of selecting such a region is to contain points of relatively high density. Hyndman (1996) proposed a simple method for computing a highest density region. If we have a distribution $f\left(y\right)$, we would like to find the region $R\left(y\right)$ that satisfies

$$(i)\int_{R(y)} f\left(y\right)dy = 1-\alpha \tag{36}$$

$$(ii) SizeR\left(y\right) \leq SizeR^{'}\left(y\right) \tag{37}$$

for any region $R'\left(y\right)$ which satisfies $\int_{R'(y)} f\left(y\right) \geq 1-\alpha$ such a region is called the highest density region (HDR) of the distribution $f\left(y\right)$.

It follows from the definition that the HDR has the smallest possible volume in the sample space of y. One of the most important advantages of using HDR is that the mode is contained in every HDR.

For normal distributions, the high density region (HDR) coincides with the usual probability region symmetric about the mean and this is also true for all unimodal and symmetric distributions. Another characteristic of HDR is that it can contain disjoint intervals when the underlying distribution is multimodal.

There have been several suggestions for constructing the HDR from a general bounded and continuous univariate density $f(y)$. Wright (1986) proposed an algorithm which includes numerical integration of $f(y)$. Hyndman (1996) developed a density quantile approach that computes the HDR. In this study we adapted Hyndman's idea to find the highest density region of the approximated log-likelihood function. We use the following algorithm to find the HDR of the approximated log-likelihoods;

1) Use (2.3) to sample the parameter values from the interval of parameter of interest.

2) Determine the appropriate number of knots for the log-likelihood function.

3) Approximate the log-likelihood function by a spline model by treating the likelihood values as response and parameter values as independent variable(s).

4) Find the fitted values of the approximated spline model for chosen θ values.

5) Treat those fitted values as the height and find the total area under the approximated function by summing those heights.

6) Divide each height by the total area to make the total area 1 under the spline function.

7) Next step is finding the biggest height(mode) and add the next biggest height and continue this until we reach the given confidence level.

8) Finally, claim the corresponding minimum and maximum parameter values as the lower and upper bound of the highest density region of approximated log-likelihood.

Assume that we want to approximate the log-likelihood of the exponential distribution:

$$f(y;\theta) = \frac{1}{\theta} exp \frac{-y}{\theta} \tag{38}$$

It is easy to show that the maximum likelihood estimator of θ is $1/\overline{y}$. We plug in $1/\overline{y}$ for $\widehat{\theta}$ in (2.3). Then we solve (2.3) to find the lower and upper limits of the interval. Finally, we sample the parameter values uniformly from this region.

Table 1 shows 90% coverage probabilities from the HDR of natural cubic spline and B-spline approximations of the log-likelihood of the exponential distribution. The first value is the coverage probability calculated from natural cubic spline approximation and number in bold is the coverage probability from B-spline approximation. Both methods produce very similar results and those results are reasonably close to the asserted confidence level(α=90) of the interval. Standard deviation of coverage probabilities is around .003. The result of this simulation study indicates that both the natural cubic spline and B-spline approximation provide excellent coverage probabilities except when the sample size is 10.

Table 1: 90% coverage probabilities from the Highest Density Region(HDR) of approximated log-likelihood of the exponential distribution with θ=2

SAMPLE SIZE	$\theta = 2$	$\theta = 5$	$\theta = 10$
n=10	0.8752 **0.8773**	0.8851 **0.8844**	0.8677 **0.8703**
n=20	0.8900 **0.8904**	0.9053 **0.9056**	0.9079 **0.9075**
n=30	0.9061 **0.9061**	0.8992 **0.8991**	0.9075 **0.9074**
n=50	0.8953 **0.8952**	0.9027 **0.9028**	0.9052 **0.9052**

3.2 The Mean Squared Error Of The Approximated Log-likelihood Function

In sufficiently large samples, the log-likelihood function is known to be approximately a quadratic form in a neighborhood of the MLE of θ. In the tails (outside of a region around the MLE of θ), the likelihood function is approximately linear in either θ or in the natural parameter of an exponential family saddlepoint approximation. Saddlepoint approximations have been used successfully to approximate the tails of distributions; discussion of many such applications are given by Reid (1988), Goutis and Casella (1999), and Huzurbazar (1999).

By approximating the log-likelihood function using regression methodology, we have the advantage that an estimate of the mean squared error(MSE) of the structural parameter can be reported along with the approximation. The mean squared error (MSE) of an estimator W of a parameter θ is the function of θ defined by $E_\theta (W - \theta)^2$. MSE measures the average squared difference between the estimator W and the parameter θ.

MSE incorporates two components, one measuring the variability of the estimator(precision) and the other measuring its bias(accuracy).

$$E_\theta\left(W-\theta\right)^2 = VAR_\theta W + \left(BIAS_\theta W\right)^2 \quad (39)$$

To find an estimator with good MSE properties, we need to find estimators that control both variance and bias. It is obvious that unbiased estimators do the best job of controlling bias and the MSE is equal to the variance when the estimator is unbiased. Since we approximate the log-likelihood function from the density distribution with known parameter, we can calculate the exact mean square error of the maximum likelihood estimator. We will compare the exact MSE, with the empirical MSE from the approximated log-likelihood. To find the estimated MSE from the approximated log-likelihood function, we have the following algorithm:

1) Use (2.3) to select parameter values from the interval of parameter of interest.

2) Calculate the log-likelihood values for chosen parameter values from step 1.

3) Treat the log-likelihood values as response and parameter values as independent variable and find the appropriate number of knots.

4) Approximate the log-likelihood by the cubic spline function with appropriate number of knot(s).

5) Find the fitted values corresponding to the chosen parameter values.

6) Find the biggest fitted value (mode) and claim the corresponding θ value as the ”MLE” estimator of the unknown true parameter.

7) Repeat 1-6 to find the empirical MSE by using;

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\theta-\hat{\theta}_i\right)^2 \quad (40)$$

where $\hat{\theta}_i$ is the estimation of the mle from the ith simulation, θ is known true parameter and N is the simulation size.

In our simulation, we generated 10,000 samples from an exponential distribution for each sample size and parameter value. Selecting such a large simulation size(10,000) will allow us to obtain small binomial standard deviations of the parameter estimates. Table 2 shows the empirical MSE of the maximum likelihood estimator (mle) from the cubic spline approximations. The first value is MSE of natural cubic spline approximation and the number in bold is MSE of B-spline approximation.

Table 3 presents the exact MSE of the mle from the exponential distribution. As it is expected, the MSE of the approximations results are getting closer to the expected MSE as the sample size increases.

Since we generate data from the exponential distribution with known parameter, we can calculate the exact MSE of the maximum likelihood estimator. It is easy to show that for the exponential distribution $1/\overline{y}$ is the mle of θ. Furthermore it can be shown that

$$E[1/\overline{y}] = \frac{n\theta}{n-1} \quad (41)$$

$$E[(1/\overline{y})^2] = \frac{n^2\theta^2}{(n-1)(n-2)} \quad (42)$$

From (3.14) we can write MSE of the mle;

$$MSE(1/\overline{y}) = E[(1/\overline{y})^2] - (E[1/\overline{y}])^2 + (E[1/\overline{y}]-\theta)^2 \quad (43)$$

$$MSE(1/\overline{y}) = \frac{n^2\theta^2}{(n-1)(n-2)} - \frac{n^2\theta^2}{(n-1)^2} + (\frac{n\theta}{(n-1)^2} - \theta)^2 \quad (44)$$

After some simplifications;

$$MSE(1/\overline{y}) = \frac{(n+2)\theta^2}{(n-1)(n-2)} \quad (45)$$

Now we can calculate the exact MSE of the maximum likelihood estimator of the exponential distribution. Table 3 gives the exact MSE’s of the maximum likelihood estimator of the exponential distributions for different sample sizes and parameter values.

The result of the simulation study reveal that both the natural cubic spline approximation and cubic B-spline approximation provide an accurate point estimates of the known true parameter. Estimated MSE results are consistent with the coverage probability results in the sense that for large sample sizes difference between exact MSE and estimated MSE's are getting smaller and smaller.

Table 2: Empirical MSE's of approximated log-likelihoods of the exponential distribution

SAMPLE SIZE	$\theta = 2$	$\theta = 5$	$\theta = 10$
n=10	0.572 **0.577**	4.371 **4.412**	14.79 **15.18**
n=20	0.234 **0.232**	1.63 **1.667**	6.98 **6.93**
n=30	0.169 **0.174**	1.007 **1.014**	4.29 **4.27**
n=50	0.093 **0.094**	0.537 **0.538**	2.152 **2.153**

Table 3: Exact MSE's of maximum likelihood estimator of the exponential distribution

SAMPLE SIZE	$\theta = 2$	$\theta = 5$	$\theta = 10$
n=10	0.667	4.17	16.67
n=20	0.257	1.608	6.43
n=30	0.157	0.985	3.94
n=50	0.088	0.552	2.21

4 Discussion and Comparison

The simplest approach to eliminating nuisance parameters is to replace them with their maximum likelihood estimates, leading to the profile likelihood. Many examples of misleading behavior of the profile likelihood have been given, leading to various corrections of the profile likelihood. Among the proposed corrections are modified profile likelihood (Barndroff-Nielsen, 1983) and the conditional profile likelihood (Cox and Reid, 1987).

The profile likelihood and integrated likelihood are not genuine likelihood functions. That is, an integrated likelihood function or a profile likelihood function do not, in general, correspond to a likelihood function arising from an observed statistic. However, integrated likelihood is closely related to the profile likelihood function in the sense that the first order approximations of procedures based on the uniform integrated likelihood function are the same as the first-order approximations of procedures based on the profile likelihood function.

Severini (1998b) has shown that the profile likelihood function can be viewed as an estimate of the genuine likelihood function. Indeed the profile likelihood can be used as if it were an ordinary likelihood to produce asymptotically $(n \to \infty)$ correct inferences about a structural parameter. Severini (1998b) has also shown that the modified profile likelihood can be derived as an approximation to either the conditional or marginal likelihood when either of the latter likelihoods exist.

In large samples, there are unlikely to be large differences between the results based on these modified likelihood methods. However for a given small sample size, the results may differ. Our goal is to compare the validity of each likelihood method in terms of the coverage probabilities of confidence regions derived from them, using a region of the form (2.3), when the sample size is small. In this study, we will apply profile, conditional and integrated likelihood methods to the two-parameter gamma distribution for the situation where the shape(θ) and scale(λ) parameters are regarded as the structural and nuisance parameters respectively. The probability density of the two-parameter Gamma distribution is given by;

$$f\left(y;\theta,\lambda\right) = \frac{1}{\Gamma(\theta)\lambda^{\theta}} y^{(\theta-1)} \exp^{(-y/\lambda)} y > 0, \quad (46)$$

for parameters $\theta, \lambda > 0$. We proceeded as follows: First, we eliminated the nuisance parameter (λ) from the two-parameter gamma distribution by the integrated, conditional and profile likelihood methods. Then we generated 10,000 replications of various sample sizes from the two-parameter gamma distribution with known parameter values θ=2 and λ=3. Generating the data from the density with known parameters allows us to determine coverage probabilities for confidence intervals of the form (2.3) obtained from each of the modified likelihood methods for different sample sizes. The results are shown in Table 1. Binomial standard deviations belong to each coverage probability are shown in bold numbers.

As it is expected for small sample sizes(n=7,n=15, n=20), profile likelihood is not as accurate as integrated likelihood and conditional likelihood. However the integrated likelihood produces coverage probabilities very close to the nominal level (0.95) for all sample sizes.

Table 4: 95% coverage probabilities from the modified likelihoods of the gamma density

SAMPLE SIZE	INT LIK.	COND LIK.	PROFLIK.
n=7	0.9476	0.9422	0.9223
n=15	0.9478	0.9442	0.9285
n=20	0.9495	0.9461	0.9326
n=30	0.9518	0.9481	0.9364

References:

[1] Anderson, E.B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Society B* **32**, 283-301.

[2] Bahadur, R.R. (1964). On Fisher’s bound for asymptotic variances.*Annal of Mathematical Statistics* **35**, 1545-52.

[3] Basu, D. (1978). On the elimination of nuisance parameters.*Journal of the American Statistical Association*, **72**, 355-66.

[4] Choon, O. Hoong L. and Huey, S (2008). A Functional Approximation Comparison between Neural Networks and polynomial regression. *Wseas Transactions on Mathematics*, **7**, 2008, 353-363

[5] Demiralp, M. (2009). Applications of High Dimensional Model Representations to Computer Vision. *Wseas Transactions on Mathematics*, **8**, 2009, 184-192

[6] Efron, B and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected fisher information (with discussion). *Biometrika* **65**, 457-487.

[7] Kalianpur, G and Rao, C. R. (1955). On Fisher’s lower bound to the asymptotic variance of a consistent estimate. *Sankhya* **15**, 331-42.

[8] Severini, T.A. (2000). *Likelihood methods in statistics*. New York: Oxford.

[9] Wald, A. (1949). Note on the consistency of the maximum likelihood estimator. *Annals of Mathematical Statistics* **20**, 595-601.

[10] Tanner, M.A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions*, 3rd edition. New York: Springer.

[11] Zainuddin, Z. and Pauline, O (2008) Function approximation Using Artificial Neural Network. *Wseas Transactions on Mathematics*, **7**, 2008, 333-338