Thesis for the Degree of Master of Science

The Effect of the Tail Part of the Density on the Choice of Priors and Kernels in Nonparametric Bayesian Goodness-of-Fit-Test

by

Tae Ung Uhm

Department of Statistics

The Graduate School

Pukyong National University

August 2014

The Effect of the Tail Part of the Density on the Choice of Priors and Kernels in Nonparametric Bayesian Goodness-of-Fit-Test

(비모수적 적합도 검정에서 밀도함수의 꼬리 부분이 사전분포와 커널 선택에 미치는 영향)

Advisor: Prof. Seongbaek Yi

by

Tae Ung Uhm

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in Department of Statistics, The Graduate School,
Pukyong National University

August 2014

The Effect of the Tail Part of the Density on the Choice of Priors and Kernels in Nonparametric Bayesian Goodness-of-Fit-Test

A dissertation

by

Tae Ung Uhm

Approved by:

_____

(Inho Park)

_____

(Maeng Seok Noh)

_____

(Seong Baek Yi)

August 22, 2014

Contents

List of Tables

List of Figures

비모수적 적합도 검정에서 밀도함수의
꼬리부분이 사전분포와 커널 선택에 미치는 영향

엄 태 웅

부 경 대 학 교 대 학 원 통 계 학 과

요약

본 논문에서는 밀도함수의 적합도검정절차로 베이즈인수를 비모수적인 방법으로 정의한 후 고전적인 방법을 사용하여 검정하는 절차를 제안한다. 여기서 베이즈인수는 두 개의 한계우도함수들의 비로 정의되는데 이 중 한 한계우도함수는 밀도함수의 커널추정량과 사용되는 평활모수의 사전분포의 곱을 적분하여 구해진다.

일반적으로 밀도함수의 커널 추정량은 평활모수의 선택에 가장 큰 영향을 받는 것으로 알려져 있다. 그러나 참 밀도함수의 꼬리부분이 두터운데 사용된 커널의 꼬리부분이 얇을 경우 추정된 우도는 좋은 성질을 갖지 못한 것으로 알려져 있다. 이에 본 연구에서는 참 밀도함수의 꼬리부분과 평활모수의 사전분포의 꼬리부분의 두터운 정도들이 어떤 관계가 있는지 살펴본다.

이 때 검정절차는 참 밀도함수의 귀무가설에서 계산된 백분위수들과 계산된 베이즈인수 값들을 비교하여 실행된다. 즉, 고전적인 검정절차를 따르게 된다. 또한 새롭게 제시된 검정통계량을 이용한 적합도검정방법을 기존의 적합도 검정방법과 비교하기 위해 몬테칼로 모의실험방법을 수행한다.

1. INTRODUCTION

The Bayesian approach to hypothesis testing was developed by Jeffreys in a 1935 paper and in a 1961 book Theory of Probability, where he elaborated a procedure of quantifying the evidence favoring a scientific claim. The key point was a measure, now termed the Bayes factor, which is the posterior odds of the null hypothesis when the prior distribution on the null is one-half.

In goodness of fit testing for a parametric null against a nonparametric alternative, Bayesian approaches have been investigated in the literature particularly in the context of Bayesian nonparametric testing of goodness of fit problem. However, because of the specification of an alternative model, and the computation of the Bayes factor in addition to theoretical issues such as Bayes factor consistency, Bayesian nonparametric goodness of fit testing is often regarded as a challenging inferential problem. In this regards, there have been several advances in Bayesian goodness of fit testing form nonparametric Bayesian point of view.

On the other hand, Bayes factors may also be used in frequentist fashion to test the fit of one model against another, an idea which appears to be due to Good (1957). The idea referred toas Bayes/non-Bayes synthesis has been proposed using the distribution of a Bayes factor and p-value in corresponding to the Bayes factor as a significance criterion (see e.g. Good (1992) for an extensive review). Much more recently Hart (2009) proposed using approximations to posterior probabilities to test lack of fit in the context of regression, and Albert (2010) revisited Good' s approach (Good, 1967) for Bayesian categorical data analysis.

Accordingly, in this research we also attempt to provide a kind of Bayes/non-Bayes compromise for hypothesis testing and propose a frequentist-Bayesian goodness of fit testing procedure of parametric null model against the nonparametric alternative using the Bayes factor in a nonparametric fashion. Specifically, a nonparametric Bayes factor is computed in which a kernel density estimate for the marginal likelihood under the nonparametric alternative is compared with a fitted parametric marginal likelihood under the null model from the Bayesian perspective while the testing procedure with the Bayes factor is performed in the

frequentist framework. That is, a pseudo Bayes, nonparametric procedure for testing the fit of a parametric model for a density is considered, and a Bayes factor in which a kernel density estimate is compared with a fitted parametric model is formulated with a suitable specification of prior distribution of the bandwidth in the kernel density estimate.

Although bandwidth selection is not a primary concern in this field, it is of some interest to know how the prior distributions of the bandwidth affect the posterior mode of the bandwidth. To this end we will consider a small simulation comparing the posterior modes corresponding to different choices of prior of the bandwidth.

In general, the quality of the kernel density estimate is known as to be primarily determined by the choice of bandwidth and only in a minor way by the choice of kernel (see e.g., Silverman (1986), Scott (1992), and Wand and Jones (1995)). However, when the true density is heavy-tailed and the kernel too light-tailed, the estimated likelihood of Bayes factor is known to behave poorly (Hall, 1987). As discussed in Hall (1987), the kernel needs to be chosen so that tail effects become negligible.

Therefore we need to choose the kernel in an appropriate way by the assumed null distribution. Hence, in this research, we examine these two issues on choice of the prior as well as the choice of the kernel for the goodness of fit testing purpose.

In order to complete the topic, in section 2 we review types of priors and kernels. In section 3 the nonparametric Bayes factor is defined and computed for some pairs of selected priors and kernels. We deal with the goodness of fit testing using the proposed nonparametric Bayes factor in frequentist fashion in section 4. Finally in section 5 we summarize the simulation results for the goodness of fit test and investigate further extension with the proposed Bayes factor.

2. PRIOR DISTRIBUTIONS

Central in Bayesian statistics is Bayes' theorem, which can be written as follows:

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta).$$

Given the likelihood function $f(x|\theta)$ and the prior $\pi(\theta)$, it is easy to calculate the posterior distribution of θ, $\pi(\theta|x)$, which is used for doing inference. An important problem in Bayesian analysis is how to define the prior distribution.

The prior distribution represents the information about an uncertain parameter that is combined with the probability distribution of new data to yield the posterior distribution, which in turn is used for future inferences and decisions involving the parameter. Thus we have two types of priors: informative and noninformative.

2.1 Noninformative Priors

If prior information about the parameter θ is available, it should be incorporated in the prior distribution.
If we have no prior information, we want a prior with minimal influence on the inference. We call such a prior a noninformative prior.

An important question is, how do we construct a noninformative prior? The Bayes/Laplace postulate, stated about 200 years ago says the following: When nothing is known about θ in advance, let the prior $\pi(\theta)$ be a uniform distribution, that is, let all possible outcomes of θ have the same probability. This is also known as the principle of insufficient reason.

Fisher did not support the Bayes/Laplace postulate. He argued that Not knowing the chance of mutually exclusive events and knowing the chance to be equal are two quite different states of knowledge. He accepted Bayes' theorem only for informative priors.

The fundamental problem by using the uniform distribution as our noninformative prior, is that the uniform distribution is not invariant under reparametrization. If we have no information about θ, we also have no information about for example $1/\theta$, but a uniform prior on θ does not correspond to a uniform prior for $1/\theta$. By the transformation formula, the

corresponding distribution for a one-to-one function $g(\theta)$ is given below:

$$\pi(\theta)=1,\ \phi=g(\theta)\ \Rightarrow\ \pi(\phi)=\left|\frac{d}{d\phi}g^{-1}(\phi)\right|.$$

Another problem with the uniform prior is that if the parameter space is infinite, the uniform prior is improper, which means, it does not integrate to one. This is however not always a serous problem, since improper priors often lead to proper posteriors. Now we look at the interpretation of noninformative priors.

2.2 Interpretation of Noninformative Priors

Kass and Wasserman (1996) stated two different interpretations of noninformative priors: 1) Noninformative priors are formal representations of ignorance. 2) There is no objective, unique prior that represents ignorance, instead noninformative priors are chosen by public agreement much like units of length and weight. In the second interpretation, noninformative priors are the default to use when there is insufficient information to otherwise define the prior. Today, no one use the first interpretation to claim that one particular prior is truly noninformative. The focus is on comparing different priors to see if any is preferable in some sense.

Box and Tiao (1973) define a noninformative prior as a prior which provides little information relative to the experiment. Bernardo and Smith (1994) use a similar definition, they say that noninformative priors have minimal effect relative to the data, on the final inference. They regard the noninformative prior as a mathematical tool, it is not a uniquely noninformative or object prior. These definitions are related to the second interpretation of Kass and Wasserman (1996).

Pericchi and Walley (1991) have a quite different view. They say that no single probability distribution can model ignorance satisfactory, therefore large classes of distributions are needed. They use the first interpretation of Kass and Wasserman (1996), but they realize that a single distribution is not enough. Therefore they introduce classes of prior distributions.

2.3 Invariant Noninformative Priors

In the previous introduction, we saw that the fundamental problem by using the uniform distribution as noninformative prior, is that it is not invariant under reparametrization. Now we will see how we can construct invariant noninformative priors.

One approach is to look for an invariance structure in the problem and let the prior have the same invariance structure. Mathematically, this means that the model and the prior should be invariant under action of the same group and we should use the right Haar measure as prior. The right Haar measure is the prior that is invariant to right multiplication with the group. For reasons not to be discussed here, we prefer the right invariant Haar measure instead of the left, as our noninformative prior. See for example Berger (1980) or Robert (1994) for a more throughly discussion of group invariance and Haar measures.

We illustrate the method by two simple examples.

Example 1. Location parameters:

Let X be distributed as $f(x-\theta)$, which is a location density, and θ is called a location parameter. A location invariant density is invariant to linear transformation. This means that $Y=X+a$ is distributed as $f(y-\phi)$ with $\phi=\theta+a$, that is X and Y have the same distribution, but with different location parameters. Since the model is location invariant, the prior distribution should be location invariant. Therefore:

$$\pi(\theta)=\pi(\theta-a) \quad \forall a \quad \Rightarrow \quad \pi(\theta)=1.$$

An invariant noninformative prior for a location parameter is the uniform distribution.

Another argument leading to the same result, is that since θ and ϕ are location parameters in the same model, they should have the same prior.

Example 2. Scale parameters:

Let X be distributed as $\frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$, which is a scale invariant density with scale parameter σ. That the distribution is scale invariant, means that $Y=cX$ has the same distribution as X, but with a different scale parameter. Since the density is scale invariant, the prior distribution should be scale invariant:

$$\pi(a) = \pi\left(\frac{a}{c}\right)\frac{1}{c} \;\forall a \in (0,\infty) \text{ and } c > 0.$$

This leads to

$$\pi(\sigma) = \pi\left(\frac{\sigma}{c}\right)\frac{1}{c} \;\; c > 0 \;\Rightarrow\; \pi(\sigma) = \frac{1}{\sigma}.$$

so the invariant noninformative prior for a scale parameter is $\pi(\sigma) = 1/\sigma$, which is an improper distribution.

We see that in both location and scale cases, the invariant noninformative prior is improper.

A difficulty with this method is that all problems do not have an invariance structure and the right Haar measure does not exist. In the following we present methods for finding invariant noninformative priors which do not take the structure of the problem into account.

2.3.1 Diffuse prior One of the most common priors is the flat, uninformative, or diffuse prior where the prior is simply a constant,

$$p(\theta) = \frac{1}{b-a}, \;\; \text{for } a \le \theta \le b.$$

This conveys that we have no a priori reason to favor any particular parameter value over another. With a flat prior, the posterior just a constant times the likelihood,

$$p(\theta|X) = C\, l(\theta|X).$$

and we typically write that $p(\theta|X) \propto l(\theta|X)$. In many cases, classical expressions from frequentist statistics are obtained by Bayesian analysis assuming a flat prior.

If the variable of interest ranges over $(0,\infty)$ or $(-\infty,+\infty)$, then strictly speaking a flat prior does not exist, as if the constant takes on any non-zero value, the integral does not exist. In such cases a flat prior (assuming $p(\theta|X) \propto l(\theta|X)$) is referred to as an improper prior.

2.3.2 Jeffreys' prior This method was described by Jeffreys (1946), and it is base on the Fisher information given by

$$I(\theta|X) = E_X\left(\frac{\partial \log f(\theta|X)}{\partial \theta}\right)^2 = -E_X\left(\frac{\partial^2 \log f(\theta|X)}{\partial \theta^2}\right).$$

Jeffreys' prior is defined as

$$\pi(\theta) \propto \sqrt{I(\theta|X)}.$$

A full discussion, with derivation, can be found in Lee (1997, Section 3.3). Jeffreys justified his method by the fact that it satisfies the invariant reparametrization requirement, shown by the following two equations:

$$I(\theta) = I(h(\theta))\,(h'(\theta))^2$$

$$\pi(\theta) \propto I(h(\theta))^{1/2}\,|h'(\theta)| = \pi(h(\theta))\,|h'(\theta)|.$$

In the last equation we recognize the transformation formula.

A motivation for Jeffreys' method is that the Fisher information $I(\theta)$ is an indicator of the amount of information brought by the model (observation) about θ. To favor the values for θ of which $I(\theta)$ is large is equivalent to minimizing the influence of the prior.

When the parameter θ is one-dimensional, the Jeffreys prior coincides with the right Haar measure when it exists.

Jeffreys' prior can be generalized to multidimensional parameters Θ by letting the prior be proportional to the square root of the determinant of the Fisher information matrix:

$$\pi(\Theta) \propto \sqrt{\det[I(\Theta|X)]}.$$

where $I(\Theta)$ is the Fisher Information matrix, the matrix of the expected second partials,

$$I(\Theta|X)_{ij} = -E_X\left(\frac{\partial^2 \log f(\Theta|X)}{\partial\theta_i \partial\theta_j}\right).$$

However, there are problems with this generalized Jeffreys' prior, as the following example, taken from Bernardo and Smith (1994) will show.

Example 3.

We let $X^n = \{X_1, \ldots, X_n\}$ be iid $N(\mu, \sigma^2)$. First we assume that the mean is known, and equal to zero. Then we have a scale density, and Jeffreys' noninformative prior for σ is given by $\pi(\sigma) = \sigma^{-1}$. With this choice of prior, the posterior of σ is such that

$$\sum_{i=1}^{n} X_i^2/\sigma^2 \sim \chi_n^2.$$

Then we assume that the mean is unknown. The two dimensional Jeffreys' prior for $\Theta = (\mu, \sigma)$ is now

$$\pi(\Theta) = \pi(\mu, \sigma) = \sigma^{-2}.$$

With this choice of prior, the posterior of σ is such that

$$\sum_{i=1}^{n}(X_i - \overline{X})^2/\sigma^2 \sim \chi^2_{n-1}.$$

This is however un acceptable, since we would expect to loose one degree of freedom when we estimate μ.

Jeffreys' advice in this case, and other location-scale families was to assume that μ and σ are independent a prior and use the one-dimensional Jeffreys prior for each of the parameters. Then the prior for $\Theta = (\mu, \sigma)$ becomes

$$\pi(\Theta) = \pi(\mu, \sigma) = \pi(\mu)\pi(\sigma) = 1 \cdot \frac{1}{\sigma} = \frac{1}{\sigma}.$$

which is also the right invariant Haar measure, and gives us the correct degrees of freedom.

2.3.3 Reference priors Another well-known class of noninformative priors, is the reference prior, first described by Bernardo (1979) and further developed by Berger and Bernardo (1989). The method for deriving the reference prior is also referred to as the Berger-Bernardo method.

The method leads to Jeffrey's prior in the one-dimensional case, but as we see later, it is advantageous to Jeffreys' method in the multidimensional case. The definition of a reference prior is the prior that maximizes the missing information in the experiment. The reference prior is derived as follows. Let $X^n = \{X_1, \cdots, X_n\}$ be iid random variables. Define the Kullback-Leibler distance between the posterior and the prior distribution as,

$$K_n(\pi(\theta \mid X^n), \pi(\theta)) = \int \pi(\theta \mid X^n) \log\left(\frac{\pi(\theta \mid X^n)}{\pi(\theta)}\right) d\theta.$$

Let K_n^{π} be the expected Kullback-Leibler distance with respect to X^n:

$$K_n^{\pi} = E_{X_n}\left[K_n(\pi(\theta \mid X_n), \pi(\theta))\right].$$

The missing information is now given as the limit of K_n^{π} as the number of observations, n goes to infinity. So we find the prior that maximizes

$$K_{\infty}^{\pi} = \lim_{n\to\infty} K_n^{\pi}.$$

Unfortunately, this limit is usually infinite. To overcome this difficulty, we find the prior π_n maximizing K_n^{π} and find the limit of the corresponding sequence of posteriors . Then the reference prior is given as the prior that produces the limiting posterior.

The Berger-Bernardo method can be extended to handle nuisance parameters. Then the parameter is given by $\theta=(\psi,\lambda)$ where ψ is the parameter of interest and is the nuisance parameter. We can write the prior for θ as

$$\pi(\psi,\lambda)=\pi(\lambda|\psi)\,\pi(\psi).$$

The idea is now to first define the conditional prior $\pi(\lambda|\psi)$ to be the reference prior for λ with ψ fixed. Then we find the marginal model

$$p(x|\theta)=\int p(x|\psi,\lambda)\,\pi(\lambda|\psi)\,d\lambda. \qquad (2.3.1)$$

and take the prior for ψ, $\pi(\psi)$ to be the reference prior based on the marginal model $p(x|\psi)$. There are some technical problems here, because the prior $\pi(\lambda|\psi)$ is often improper, and the integral (2.3.1) diverges. To accomplish this, we restrict the integral to a sequence of compact sets.

The method is invariant in choice of nuisance parameter. This seems reasonable, since the parameter of interest is independent of the nuisance parameter.

The method can also be generalized to multidimensional parameter spaces. Then we let the parameter vector $\theta=(\theta_1,\cdots,\theta_n)$ be ordered according to importance, with θ_1 being the most important parameter. We write the prior for θ as

$$\pi(\theta)=\pi(\theta_m|\theta_1,\cdots,\theta_{m-1})\cdots\pi(\theta_2|\theta_1)\,\pi(\theta_1).$$

and use the procedure above recursively. It should be noted that the ordering of the parameters is very important. Different orderings may lead to different reference priors. In some cases it might be difficult to chose the correct ordering. However, this method avoids the problem we saw with Jeffreys' multidimensional method.

2.3.4 Other methods Box and Tiao (1973) described a method based on something they called data-translated likelihoods. The method leads to

Jeffreys' prior. A likelihood function is data-translated if it can be written as $L_y(\phi) = f(\phi - t(y))$. They suggested to use a uniform prior when this is satisfied. An approximate data-translated likelihood was introduced to motivate for Jeffreys' general rule.

Jaynes (1968) suggested to select the prior that maximizes the entropy. This method is only good for discrete, finite parameter space. If no further constraints are imposed on the problem, this method gives the uniform prior. The method has been used successfully in many problems.

Welch and Peers (1963) developed a method called probability matching. They seek a prior $\pi(\theta)$ so that the posterior confidence interval for θ has coverage error $O(n^{-1})$ in the frequentist sense. This means that the difference between the posterior and frrequentist confidence interval should be small. Their method is equivalent to Jeffreys' prior when θ is one-dimensional. Tibshirani (1989) extended the method to be able to handle nuisance parameters.

2.4 Informative Priors

Informative prior distributions summarise the evidence about the parameters concerned from many sources and often have a considerable impact on the results.

Using informative priors distributions allows the incorporation of information available to researcher form the literature and in light of their experience. However, using informative priors may lead to problems because of the subjective beliefs. Unfortunately, even inf we wanted to use noninformative priors, the best method for choosing such priors is still an issue of considerable debate.

Here are two techniques used most frequently to develop informative prior distributions.

2.4.1 Pooling by Expert Opinions In principle, one of the most powerful methods for developing informative priors is to synthesise the information from a group of experts. Although the development of priors by consensus risks all the problems related to the impact of the subjective biases of the various parties in the assessment process (arguably priors developed using expert opinion are examples of "dreamt up" priors, to use an expression we used in the previous section), this approach can be

successful.

A potentially major problem with the development of priors by consensus is that different "experts" will suggest different priors. It is far from a trivial exercise (theoretically) to pool such priors to form a "consensus prior" (and it is impossible to include more than one prior for each parameter in a Bayesian assessment). We recommend that the various priors be multiplied together and then normalized because at least this procedure has the desirable property that the assessment results are independent of whether the priors are pooled and then the assessment conducted or whether assessments are conducted using each alternative prior in turn and the results then pooled. One very undesirable feature of this approach to pooling, however, is that if one expert believes that some parameter value/model has zero probability, the posterior is forced to be consistent with this opinion. Therefore, if this approach is to be used, our earlier advice that no plausible value for a parameter should be assigned zero probability should be followed.

2.4.2 Data summaries If the parameters of the model are chosen to be independent of the parameter that scales the population, data for other similar models can be used to construct priors for the model for which an assessment is needed. This approach to conducting priors is known as meta-analysis. Methods for constructing priors using data for other relevant models range from simply tabulating the estimates to hierarchical meta-analysis (Gelman et al., 1995). Simple tabulation methods can be extended by fitting a smooth functional form to the data and by weighting each estimate by a measure of its uncertainty and comparability to the model for which an assessment is required. Hierarchical meta-analysis (Gelman et al., 1995) is a more formal method for developing a prior for a parameter from values for that parameter for other similar models under the assumption that the models differ in that parameter.

"Selection bias" is a potential problem when developing a prior using meta analysis.

2.5 Conjugate Priors

In Bayesian probability theory, if the posterior distributions $\pi(\theta|x)$ are in the same family as the prior probability distribution $\pi(\theta)$, the prior and

posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function. For example, the Gaussian family is conjugate to itself (or self-conjugate) with respect to a Gaussian likelihood function: if the likelihood function is Gaussian, choosing a Gaussian prior over the mean will ensure that the posterior distribution is also Gaussian. This means that the Gaussian distribution is a conjugate prior for the likelihood which is also Gaussian. The concept, as well as the term conjugate prior, were introduced by Howard Raiffa and Robert Schlaifer (1961).

Consider the general problem of inferring a distribution for a parameter θ given some data $X^n=(X_1,...,X_n)$. From Bayes' theorem, the posterior distribution is equal to the product of the likelihood function $f(x|\theta)$ and prior $\pi(\theta)$, normalized (divided) by the probability of the data $f(x)$:

$$f(\theta|x)=\frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta')\pi(\theta')d\theta'}.$$

Let the likelihood function be considered fixed; the likelihood function is usually well-determined from a statement of the data-generating process. It is clear that different choices of the prior distribution $\pi(\theta)$ may make the integral more or less difficult to calculate, and the product $f(x|\theta)\pi(\theta)$ may take one algebraic form or another. For certain choices of the prior, the posterior has the same algebraic form as the prior (generally with different parameter values). Such a choice is a conjugate prior.

A conjugate prior is an algebraic convenience, giving a closed-form expression for the posterior: otherwise a difficult numerical integration may be necessary. Further, conjugate priors may give intuition, by more transparently showing how a likelihood function updates a prior distribution.

All members of the exponential family have conjugate priors. See Gelman et al. (2003) for a catalog. Table 2.1 gives the conjugate priors for several common likelihood functions.

Table 2.1 Conjugate Priors for Common Likelihood Functions.

Likelihood Function	Conjugate Prior
Binomial	Beta
Multinomial	Dirichlet
Poisson	Gamma
Normal	
μ unknown, σ^2 known	Normal
μ known, σ^2 unknown	Inverse Chi-Square
Multivariate Normal	
μ unknown, Σ known	Multivariate Normal
μ known, Σ unknown	Inverse Wishart

3. KERNELS IN DENSITY ESTIMATION

3.1 Weighting Function

Let $X_1,...,X_n$ be an independent and identically distributed sample drawn from some distribution with an unknown density f. From the definition of the probability density function, $f(x)$, of a random variable, X, one has that

$$P(x-h<X<x+h)=\int_{x-h}^{x+h} f(t)dt \approx 2hf(x).$$

and hence,

$$f(x)\approx \frac{1}{2h}P(x-h<X<x+h).$$

The above probability can be estimated by a relative frequency in the sample, hence

$$\hat{f}(x)=\frac{1}{2h}\frac{\text{number of observations} \in (x-h,x+h).}{n}$$

An alternative way to represent $\hat{f}(x)$ is,

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n} w(x-X_i, h). \qquad (3.1.1)$$

where $X_1,...,X_n$ are sample and,

$$w(t,h)=\begin{cases}\frac{1}{2h} & \text{for } |t|<h \\ 0 & otherwise\end{cases}.$$

The $\hat{f}(x)$ defined in (3.1.1) has the properties of a pdf, that is $\hat{f}(x)$ is non-negative for all x, and the area between $\hat{f}(x)$ and the x-axis is equal to one.

One way to think about (3.1.1) is to imagine that a rectangle (height $1/2h$ and width $2h$) is placed over each observed point on the x-axis. The estimate of the pdf at a given point is $1/n$ times the sum of the heights of all the rectangles that cover the point. By increasing h one increases the width of each rectangle and thereby increases the degree of smoothing.

Instead of using rectangles in (3.1.1) one could use other weighting functions, for example triangles,

$$w(t,h)=\begin{cases}\frac{1}{h}\left(1-\frac{|t|}{h}\right) & \text{for } |t|<h \\ 0 & \text{otherwise}\end{cases}.$$

The resulting $\hat{f}(x)$ is indeed a pdf. Note that here too larger values of h lead to smoother estimates $\hat{f}(x)$. Another alternative weighting function is the Gaussian,

$$w(t,h)=\frac{1}{\sqrt{2\pi}\,h}e^{-t^2/2h^2},\ -\infty<t<\infty.$$

3.2 Kernels

The above weighting functions, $w(t,h)$, are all of the form,

$$w(t,h)=\frac{1}{h}K\left(\frac{t}{h}\right),$$

where K is a function of a single variable called the kernel.

A kernel is a standardized weighting function, namely the weighting function with $h=1$. The kernel determines the shape of the weighting function. The parameter h is called the bandwidth or smoothing parameter. It determines the amount of smoothing applied in estimating $f(x)$. Six examples of kernels are given in Table 3.1.

Table 3.1 Six Kernels and their Efficiencies

Kernel	$K(t)$
Epanechnikov	$K(t)=\begin{cases}\frac{3}{4}\left(1-\frac{1}{5}t^2\right)/\sqrt{5} & \text{for } \|t\|<\sqrt{5} \\ 0 & \text{otherwise}\end{cases}$
Biweight	$K(t)=\begin{cases}\frac{15}{16}(1-t^2)^2 & \text{for } \|t\|<1 \\ 0 & \text{otherwise}\end{cases}$
Triangular	$K(t)=\begin{cases}1-\|t\| & \text{for } \|t\|<1 \\ 0 & \text{otherwise}\end{cases}$
Gaussian	$K(t)=\frac{1}{\sqrt{2\pi}}e^{-t^2/2},\ -\infty<t<\infty$
Rectangular	$K(t)=\begin{cases}1/2 & \text{for } \|t\|<1 \\ 0 & \text{otherwise}\end{cases}$

In general any function having the following properties can be used as a kernel,

$$\text{a)} \int K(z)dz = 1, \qquad \text{b)} \int zK(z)dz = 0, \qquad \text{c)} \int z^2K(z)dz = \sigma_k^2.$$

It follows that any symmetric pdf is a kernel. However, non-pdf kernels can also be used, e.g. kernels for which $K(z) < 0$ for some values of z. The latter type of kernels have the disadvantage that $\hat{f}(x)$ may be negative for some values of x.

Kernel estimation of pdfs is characterized by the kernel, K, which determines the shape of the weighting function, and the bandwidth, h, which determines the "width" of the weighting function and hence the amount of smoothing. The two components determine the properties of $\hat{f}(x)$. Considerable research has been carried out (and continues to be carried out) on the question of how one should select K and h in order to optimize the properties of $\hat{f}(x)$.

We are interested in estimating the shape of this function f. Its kernel density estimator is,

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

where $K(\bullet)$ is the kernel a symmetric but not necessarily positive function that integrates to one and $h > 0$ is a smoothing parameter called the bandwidth. A kernel with subscript h is called the scaled kernel and defined as $K_h(x) = 1/h\, K(x/h)$. Intuitively one wants to choose h as small as the data allow, however there is always a trade-off between the bias of the estimator and its variance; more on the choice of bandwidth below.

3.3 Properties of Kernel Estimators

3.3.1 Quantifying the accuracy of kernel estimators There are various ways to quantify the accuracy of a density estimator. We will focus here on the mean squared error (MSE) and its two components, namely bias and standard error (or variance). We note that the MSE of $\hat{f}(x)$ is a function of the argument x:

$$\begin{aligned} MSE(\hat{f}(x)) &= E(\hat{f}(x)-f(x))^2 \\ &= (E(\hat{f}(x)-f(x))^2 + E(\hat{f}(x)-E(\hat{f}(x)))^2 \\ &= Bias^2(\hat{f}(x)) + Var(\hat{f}(x)). \end{aligned}$$

A measure of the global accuracy of $\hat{f}(x)$ is the mean integrated squared error (MISE)

$$\begin{aligned} MISE(\hat{f}(x)) &= E\int_{-\infty}^{\infty}(\hat{f}(x)-f(x))^2 dx \\ &= \int_{-\infty}^{\infty} MSE(\hat{f}(x))dx \\ &= \int_{-\infty}^{\infty} Bias^2(\hat{f}(x))dx + \int_{-\infty}^{\infty} Var(\hat{f}(x))dx. \end{aligned}$$

We consider each of these components in term.

3.3.2 The Bias, Variance and Mean Squared Error of $\hat{f}(x)$

$$\begin{aligned} E(\hat{f}(x)) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}E\left[K\left(\frac{x-x_i}{h}\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\int_{-\infty}^{\infty}K\left(\frac{x-t}{h}\right)f(t)dt \\ &= \frac{1}{h}\int_{-\infty}^{\infty}K\left(\frac{x-t}{h}\right)f(t)dt. \end{aligned}$$

The transformation $z=\frac{x-t}{h}$, i.e., $t=hz+x$, $\left|\frac{dz}{dt}\right|=\frac{1}{h}$ yields

$$E(\hat{f}(x)) = \int_{-\infty}^{\infty}K(z)f(x-hz)dz.$$

Expanding $f(x-hz)$ in a Taylor series yields

$$f(x-hz)=f(x)-hzf'(x)+\frac{1}{2}(hz)^2f''(x)+o(h^2).$$

where $o(h^2)$ represents terms that converge to zero faster than h^2 as h approaches zero. Thus

$$E(\hat{f}(x)) = \int_{-\infty}^{\infty} K(z) f(x) dz - \int_{-\infty}^{\infty} K(z) hz f'(x) dz$$
$$+ \int_{-\infty}^{\infty} K(z) \frac{(hz)^2}{2} f''(z)\, dz + o(h^2)$$
$$= f(x) \int_{-\infty}^{\infty} K(z)\, dz - h f'(x) \int_{-\infty}^{\infty} z K(z) dz$$
$$+ \frac{h^2}{2} f''(x) \int_{-\infty}^{\infty} z^2 K(z) dz + o(h^2)$$
$$= f(x) + \frac{h^2}{2} k_2 f''(x) + o(h^2),$$

$$Bias(\hat{f}(x)) \approx \frac{h^2}{2} k_2 f''(x).$$

This depends on $\begin{cases} h & \text{Bias}(\hat{f}(x)) \to 0 \text{ as } h \to 0, \\ k_2 & \text{variance of the kernel,} \\ f & \text{curvature of the density at the point } x \end{cases}$

The variance of $\hat{f}(x)$ is given by

$$Var(\hat{f}(x)) = Var\left(\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h} \right) \right)$$
$$= \frac{1}{n^2 h^2} \sum_{i=1}^{n} Var\left[K\left(\frac{x - X_i}{h} \right) \right].$$

because the X_i, $i = 1, 2, \cdots, n$, are independently distributed. Now

$$Var\left(K\left(\frac{x - X_i}{h} \right) \right) = E\left(K\left(\frac{x - X_i}{h} \right)^2 \right) - \left(EK\left(\frac{x - X_i}{h} \right) \right)^2$$
$$= \int K\left(\frac{x - t}{h} \right)^2 f(t) dt - \left(\int K\left(\frac{x - t}{h} \right) f(t) dt \right)^2.$$

$$Var(\hat{f}(x)) = \frac{1}{n} \int \frac{1}{h^2} K\left(\frac{x - t}{h} \right)^2 f(t) dt - \frac{1}{n} \left(\frac{1}{h} \int K\left(\frac{x - t}{h} \right) f(t) dt \right)^2$$
$$= \frac{1}{n} \int \frac{1}{h^2} K\left(\frac{x - t}{h} \right)^2 f(t) dt - \frac{1}{n} (f(x) + Bias(\hat{f}(x)))^2.$$

Substituting $z = \frac{x - t}{h}$, one obtains

$$Var(\hat{f}(x)) = \frac{1}{nh} \int K(z)^2 f(x - hz) dz - \frac{1}{n} (f(x) + o(h^2))^2.$$

Applying a Taylor approximation yields

$$Var(\hat{f}(x)) = \frac{1}{nh}\int K(z)^2 f(x) - hzf'(x) + o(h)dz - \frac{1}{n}(f(x) + o(h^2))^2$$

Note that if n becomes large and h becomes small then the above expression becomes approximately:

$$Var(\hat{f}(x)) \approx \frac{1}{nh} f(x)\int K^2(z)dz.$$

We note that the variance decreases as h increases.

The above approximations for the bias and variance of $\hat{f}(x)$ lead to

$$\begin{aligned} MSE(\hat{f}(x)) &= Bias^2(\hat{f}(x)) + Var(\hat{f}(x)) \\ &= \frac{1}{4}h^4 k_2^2 f''(x)^2 + \frac{1}{nh} f(x) R(k). \end{aligned} \tag{3.3.1}$$

where $k_2 := \int z^2 K(z)dz$ and $R(k) := \int K(z)^2 dz$.

Integrating (3.3.1) with respect to x yields

$$MISE(\hat{f}) \approx \frac{1}{4}h^4 k_2^2 R(f'') + \frac{1}{nh} R(k). \tag{3.3.2}$$

Of central importance is the way in which $MISE(\hat{f})$ changes as a function of the bandwidth h. For very small values of h the second term in (3.3.2) becomes large but as h gets larger so the first term in (3.3.2) increases. There is an optimal value of h which minimizes $MISE(\hat{f})$.

3.3.3 Optimal Bandwidth Expression (3.3.2) is the measure that we use to quantify the performance of the estimator. We can find the optimal bandwidth by minimizing (3.3.2) with respect to h. The first derivative is given by

$$\frac{dMISE(\hat{f})}{dh} = h^3 k_2^2 R(f'') - \frac{1}{nh^2} j_2.$$

Setting this equal to zero yields the optimal bandwidth, h_{opt}, for the given pdf and kernel:

$$h_{opt} = \left(\frac{1}{n}\frac{\gamma(K)}{R(f'')}\right)^{1/5}. \tag{3.3.3}$$

where $\gamma(K) := j_2 k_2^{-2}$. Substituting (3.3.3) for h in (3.3.2) gives the minimal

MISE for the given pdf and kernel. After some manipulation this can be shown to be

$$MISE_{opt}(\hat{f}) = \frac{5}{4}\left(\frac{\beta(f)\, j_2^4\, k_2^2}{n^4}\right)^{1/5}. \qquad (3.3.4)$$

We note that h_{opt} depends on the sample size, n, and the kernel, K. However, it also depends on the unknown pdf, f, through the functional $\beta(f)$. Thus as it stands expression (3.3.3) is not applicable in practice. However, the "plug-in" estimator of h_{opt}, to be discussed later, is simply expression (3.3.3) with $\beta(f)$ replaced by an estimator.

3.3.4 Optimal Kernels The $MISE(\hat{f})$ can also be minimized with respect to the kernel used. It can be shown (see, e.g., Wand and Jones, 1995) that Epanechnikov kernel is optimal in this respect.

$$K(z) = \begin{cases} \frac{3}{4}(1-\frac{1}{5}t^2)/\sqrt{5} & \text{for } |t| < \sqrt{5} \\ 0 & otherwise \end{cases}.$$

This result together with (3.3.4) enables one to examine the impact of kernel choice on $MISE_{opt}(\hat{f})$. The efficiency of a kernel K, relative to the optimal Epanechnikov kernel K_{EP} , is defined as

$$Eff(K) = \left(\frac{MISE_{opt}(\hat{f})\, using\, K_{EP}}{MISE_{opt}(f)\, using\, K}\right)^{5/4} = \left(\frac{k_2^2\, j_2^4\, using\, K_{EP}}{k_2^2\, j_2^4\, using\, K}\right)^{5/4}.$$

The efficiencies for a number of well-known kernels are given in Table 3.1. It is clear that the selection of kernel has rather limited impact on the efficiency.

The rectangular kernel, for example, has an efficiency of approximately 93%. This can be interpreted as follows:
The $MISE_{opt}(\hat{f})$ obtained using an Epanechnikov kernel with $n = 93$ is approximately equal to the $MISE_{opt}(\hat{f})$ obtained using a rectangular kernel with $n = 100$.

3.4 Selection of the Bandwidth

Selection of the bandwidth of kernel estimator is a subject of

considerable research. We will outline four popular methods.

3.4.1 Selection with Reference to some given Distribution Here one selects the bandwidth that would be optimal for a particular pdf. Convenient here is the normal. We note that one is not assuming that $f(x)$ is normal; rather one is selecting h which would be optimal if the pdf were normal. In this case it can be shown that

$$R(f) = \int f''(x)^2 dx = \frac{3}{8\sqrt{\pi}}\sigma^{-5}.$$

and using a Gaussian kernel leads to

$$h_{opt} = \left(\frac{4}{3n}\right)^{1/5}\sigma \approx \frac{1.06\sigma}{n^5}. \tag{3.4.1}$$

The normal distribution is not a "wiggly" distribution; it is unimodal and bell-shaped. It is therefore to be expected that (3.4.1) will be too large for multimodal distributions. Secondly to apply (3.4.1) one has to estimate σ. The usual estimator, the sample variance, is not robust; it overestimates σ if some outliers (extreme observations) are present and thereby increases $\hat{h}_{opt}$ even more. To overcome these problems Silverman (1986) proposed the following estimator

$$\widehat{h_{opt}} = 0.9\hat{\sigma}/n^5. \tag{3.4.2}$$

where $\hat{\sigma} = \min\left(S, \frac{IQR}{1.34}\right)$, where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ and IQR is the interquartile range of the data. The constant 1.34 is derived from the fact that for a $N(\mu, \sigma^2)$ random variable X, one has $P\{|X-\mu| < 1.34\,\sigma\} = 0.5$. The expression (3.4.2) is used as the default option in the R function "density". It is also used as a starting value in some more sophisticated iterative estimators for the optimal bandwidth.

3.4.2 Cross-Validation The technique of cross-validation will be discussed in more detail in the chapter on model selection. At this point we will only outline its application to the problem of estimating optimal bandwidths. By definition

$$MISE(\hat{f}) = \int (\hat{f}(x) - f(x))^2 dx$$
$$= \int \hat{f}(x)^2 dx - 2\int \hat{f}(x) f(x) dx + \int f(x)^2 dx.$$

The third term does not depend on the sample or on the bandwidth. An approximately unbiased estimator of the first two terms is given by

$$\widehat{MCV}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\int \hat{f}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{f_{-1}}(x_i).$$

where $\hat{f}_{-i}(x)$ is the estimated density at the argument x using the original sample apart from observation x_i. One computes $\widehat{MCV}(\hat{f})$ for different values of h and estimates the optimal value, h_{opt}, using the h which minimizes $\widehat{MCV}(\hat{f})$.

3.4.3 Plug-in estimator The idea developed by Sheather and Jones (1991) is to estimate h from (3.3.2) by applying a separate smoothing technique to estimate $f''(x)$ and hence $\beta(f'')$. For details see, e.g. Wand and Jones (1995), section 3.6. An *R* function to carry out the computations is available in the *R* library "sm" of Bowman and Azzalini (1997).

4. BAYES FACTOR

4.1 Bayes Factor

In the classical hypothesis testing framework, we have two alternatives. The null hypothesis H_0 that the unknown parameter θ belongs to some set or interval $\Theta_0(\theta \in \Theta_0)$, versus the alternative hypothesis H_1 that θ belongs to the alternative set $\Theta_1(\theta \in \Theta_1)$. Θ_0 and Θ_1 contain no common elements $(\Theta_0 \cap \Theta_1 = \oslash)$ and the union of Θ_0 and Θ_1 contains the entire space of values for θ(i.e., $\Theta_0 \cup \Theta_1 = \Theta$).

In the classical statistical framework of the frequentists, one uses the observed data to test the significant of a particular hypothesis, and (if possible) compute a $p-$value (the probability p of observing the given value of the test statistic if the null hypothesis is indeed correct). Hence, at first blush one would think that the idea of a hypothesis test is trivial in a Bayesian framework, as using the posterior distribution,

$$\Pr(\theta > \theta_0) = \int_{\theta_0} p(\theta|x)d\theta \text{ and } \Pr(\theta_0 < \theta < \theta_1) = \int_{\theta_0} p(\theta|x)d\theta.$$

The kicker with a Bayesian analysis is that we also have prior information and Bayesian hypothesis testing addresses whether, given the data, we are more or less inclined towards the hypothesis than we initially were. For example, suppose for the prior distribution of θ is such that $\Pr(\theta > \theta_0) = 0.10$, while for the posterior distribution $\Pr(\theta > \theta_0) = 0.05$. The later is significant at the 5 percent level in a classical hypothesis testing framework, but the data only doubles our confidence in the alternative hypothesis relative to our belief based on prior information. If $\Pr(\theta > \theta_0) = 0.50$ for the prior, then a 5% posterior probability would greatly increase our confidence in the alternative hypothesis. Hence, the prior probabilities certainly influence hypothesis testing.

To formalize this idea, let,

$$p_0 = \Pr(\theta \in \Theta_0|x),\ p_1 = \Pr(\theta \in \Theta_1|x).$$

denote the probability, given the observed data x, that θ is in the null (p_0) and alternative (p_1) hypothesis sets. Note that these are posterior

probabilities. Since $\Theta_0 \cap \Theta_1 = \oslash$ and $\Theta_0 \cup \Theta_1 = \Theta$, it follows that $p_0 + p_1 = 1$. Likewise, for the prior probabilities we have,

$$\pi_0 = \Pr(\theta \in \Theta_0),\ \pi_1 = \Pr(\theta \in \Theta_1).$$

Thus the prior odds of H_0 versus H_1 are π_0/π_1, while the posterior odds are p_0/p_1.

The Bayes factor B_0 in favor of H_0 versus H_1 is given by the ratio of the posterior odds divided by the prior odds,

$$B_0 = \frac{p_0/p_1}{\pi_0/\pi_1} = \frac{p_0 \pi_1}{p_1 \pi_0}.$$

The Bayes factor is loosely interpreted as the odds in favor of H_0 versus H_1 that are given by the data. Since $\pi_1 = 1 - \pi_0$ and $p_1 = 1 - p_1$, we can also express this as,

$$B_0 = \frac{p_0(1-\pi_0)}{\pi_0(1-p_0)}.$$

Likewise, by symmetry note that the Bayes factor B_1 in favor of H_1 versus H_0 is just,

$$B_1 = 1/B_0.$$

When the hypotheses are simple, say $\Theta_0 = \theta_0$ and $\Theta_1 = \theta_1$, then for $i = 0,1$,

$$p_i \propto p(\theta_i) p(x|\theta_i) = \pi_i p(x|\theta_i).$$

Thus,

$$\frac{p_0}{p_1} = \frac{\pi_0 p(x|\theta_0)}{\pi_1 p(x|\theta_1)}.$$

and the Bayes factor (in favor of the null) reduces the,

$$B_0 = \frac{p(x|\theta_0)}{p(x|\theta_1)}.$$

which is simply a likelihood ratio.

When the hypotheses are composite (containing multiple members), things are slightly more complicated. First note that the prior distribution of θ conditioned on H_0 vs. H_1 is,

$$p_i(\theta) = p(\theta)/\pi_i \quad \text{for } i = 0,1. \tag{4.1.1}$$

as the total probability $\theta \in \Theta_i = \pi_i$, so that dividing by π_i normalizes the

distribution to integrate to one. Thus,

$$\begin{aligned} p_i &= \Pr(\theta \in \Theta_i | x) = \int_{\theta \in \Theta_i} p(\theta | x) d\theta \\ &\propto \int_{\theta \in \Theta_i} p(\theta) p(x | \theta) d\theta \\ &= \pi_i \int_{\theta \in \Theta_i} p(x | \theta) p_i(\theta) d\theta. \end{aligned}$$

where the second step follows from Bayes' theorem and the final step follows from Equation (4.1.1), as $\pi_i p_i(\theta) = p(\theta)$. The Bayes factor in favor the null hypothesis thus becomes,

$$B_0 = \left(\frac{p_0}{\pi_0}\right)\left(\frac{\pi_1}{p_1}\right) = \frac{\int_{\theta \in \Theta_0} p(x | \theta) p_0(\theta) d\theta}{\int_{\theta \in \Theta_1} p(x | \theta) p_1(\theta) d\theta}.$$

which is a ratio of the weighted likelihoods of Θ_0 and Θ_1.

A compromise between Bayesian and classical hypothesis testing was suggested by Lindley (1965). If the goal is to conduct a hypothesis test of the form $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ and we assume a diffuse prior, then a significance test of level α follows by obtaining a $100(1-\alpha)\%$ HDR for the posterior and rejecting the null hypothesis if and only if θ is outside of the HDR. See Lee (1997) are further discussions on hypothesis testing (or lack thereof) in a Bayesian framework.

4.2 Nonparametric Bayes Factor

Let $X_n = (X_1, \cdots, X_n)$ be a random sample of size n from an unknown probability density f, and we wish to test whether the density f, against the non-parametric alternative,

$$H_0 : f \in \Gamma_0 = \{f_0(\bullet | \theta) | \theta \in \Theta\}, \ \textit{against} \ \ H_1 : f \in \Gamma_1 - \Gamma_0.$$

Suppose that X_n is from an unknown probability density $f \notin \Gamma_1 - \Gamma_0$, and define the following cross-validated, nonparametric likelihood:

$$L(h | X_n) = \prod_{i=1}^{n} \hat{f}_i(X_i | h), \hat{f}_i(x_i | h) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{x_i - X_j}{h}\right), \tag{4.2.1}$$

Let π_h be a prior for h and π_0 a prior for θ, both of which are assumed to be proper. Define the following nonparametric Bayes factor B_n of two

marginal likelihoods, the marginal nonparametric likelihood of (4.2.1) for the numerator (cf. Vexler et al. (2013) for the Bayes factor in terms of the ratio of empirical likelihoods) and the marginal parametric likelihood for the denominator,

$$B_n = \frac{\int_0^\infty L(h|X_n)\,\pi_h(h)dh}{\int_\Theta \prod_{i=1}^n f_0(X_i|\theta)\,\pi_0(\theta)d\theta}.$$

4.2.1 The Role of Prior and Kernel Estimate Prior to selecting the smoothing parameter for the nature of the noninformative general prior satisfaction as much as possible but there is no improper the weak point of noninformative prior, proper prior that was used. To this end, Richardson and Green (1997) is propsed Gaussian of the kernel estimator is called mixture model of the normal distribution for the prior h^2 paying attention to what is proposed inverse-gamma. Thus, prior for h is considered prior derived from the inverse-gamma prior π_{1h} of h^2. The prior for h induced by assuming that h^2 is inverse-gamma is,

$$\pi_{1h}(h|\alpha,\beta) = \frac{2\beta^\alpha}{\Gamma(\alpha)}\frac{1}{h^{2\alpha+1}}\exp\left(-\frac{\beta}{h^2}\right)I_{(0,\infty)}(h). \qquad (4.2.2)$$

This prior is noninformative when α and β are smaller. However, the appropriate choice can be as informative. Alternative to the prior of (4.2.2), the inverse-gamma prior for h using a factor $\exp(-h^{-1})$ instead of h^{-2},

$$\pi_{2h}(h|\alpha,\beta) = \frac{2\beta^\alpha}{\Gamma(\alpha)}\frac{1}{h^{\alpha+1}}\exp\left(-\frac{\beta}{h}\right)I_{(0,\infty)}(h).$$

We may choose α and β small as in the case of π_{1h}, say $\alpha = 1$ and $\beta = 1$, but a sample-size dependent inverse-gamma prior π_{2nh} can also be considered with $\beta = B_n$,

$$\pi_{2nh}(h|\alpha,\beta_n) = \frac{2\beta_n^\alpha}{\Gamma(\alpha)}\frac{1}{h^{\alpha+1}}\exp\left(-\frac{\beta_n}{h}\right)I_{(0,\infty)}(h).$$

We consider simulation comparing the posterior modes corresponding to different choices of prior of the bandwidth h, π_{1h} with $\alpha = 1/2$ and $\beta = 1$, π_{2h} with $\alpha = 1$ and $\beta = 1$, and π_{2nh},

$$\pi_{1h}(h)= \frac{2}{\sqrt{\pi}}\frac{1}{h^2}\exp\left(-\frac{1}{h^2}\right)I_{(0,\infty)}(h),$$

$$\pi_{2h}(h)= \frac{1}{h^2}\exp\left(-\frac{1}{h}\right)I_{(0,\infty)}(h),$$

$$\pi_{2nh}(h)= \frac{\log 2}{n^{1/5}}\frac{1}{h^2}\exp\left(-\frac{\log 2}{n^{1/5}h}\right)I_{(0,\infty)}(h),$$

Typically, properties of the kernel estimators is known that smoothing parameter h is depended rather than kernel. But case of the cross-validated likelihood is not exactly certainly.

According to Hall(1987), true density is heavy-tailed and in contrast, if the kernel is light-tailed, properties of estimated nonparametric likelihood is not good. Thus Hall(1987) is proposed that the effect of the tail part could not infect and as a general rule, if the tail part of the density function of is the thicker, the kernel must be the thicker thus the Hall(1987) is proposed that use to kernel K_0.

The following practical choice was suggested by Hall (1987):

$$K_0(z)= \{(8\pi e)^{1/2}\Phi(1)\}^{-1}\exp\left[-\frac{1}{2}\{\log(1+|z|)\}^2\right], -\infty < z < \infty,$$

where Φ denotes the standard normal distribution function and $\{\sqrt{8\pi e}\,\Phi(1)\}^{-1} \doteq 0.1438$. Note that tails of $K_0(z)$ decrease more slowly than $\exp(-|z|^k)$ for any $k>0$ (Hall, 1987). Tail part of K_0 induced slowly than tail part of the normal distributions because case of heavy-tailed density function is appropriate kernel than gaussian kernel.

The purpose of this study is that assumed density function under H_0 in the form of listed above two types of prior and of a kernel selection look a goodness of fit test of the effect.

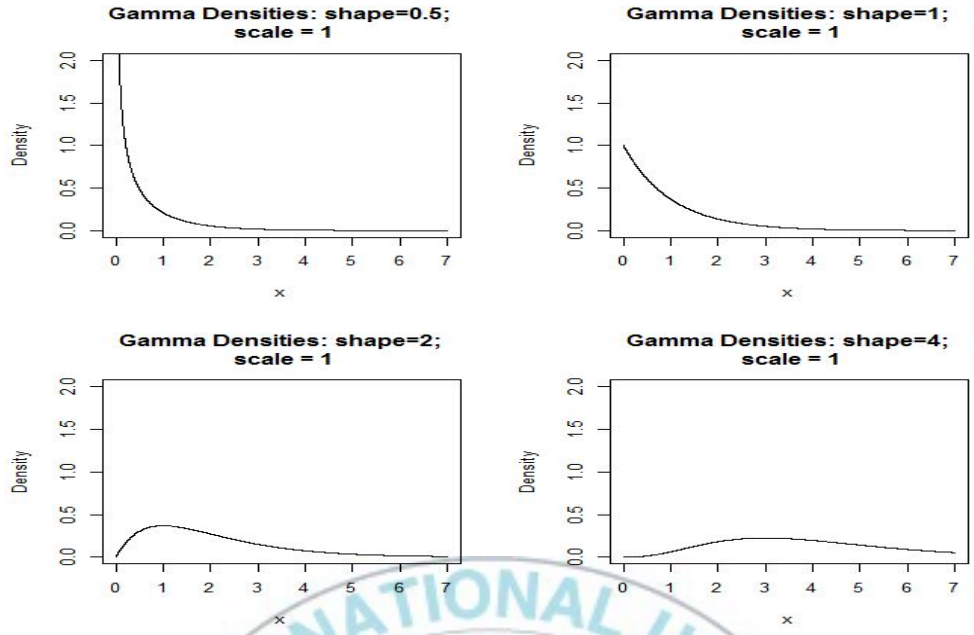

Figure 4.1. Gamma Density for Various Parameters.

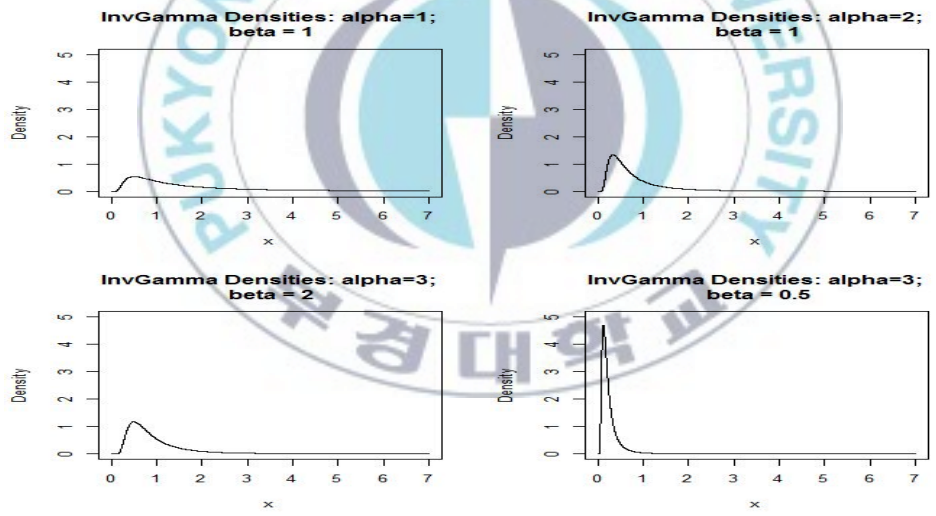

Figure 4.2. Inverse Gamma Density Various Parameters.

5. SIMULATIONS

Firstly, we consider random sample of size n = 100 from the standard normal distribution. And then we compute B_{n1} and repeat the perform 100,000 times. Thus, based on K = 105 samples of $B_{nk}'s$, $k=1,\cdots,10^5$, we approximate the null distribution of B_n and use them as a empirical null distribution. Indeed, six empirical distributions of $\log B_n$ were obtained using two different kernels and three different prior distributions, and four of them based on two by two combinations of the kernel (normal kernel and K_0) with the prior of $h(\pi_{1h}$ and $\pi_{2nh})$ are illustrated in Table 5.2. As shown the summary statistics of B_n's in Table 5.1, the distributions of $\log B_n$ are the summary statistics of four empirical distributions are negative in the general. Note that the purpose of evaluating the empirical null distribution is to perform the GOF testing for the normal data using the nonparametric Bayes factor, B_n. Bayes factors which support the null hypothesis in other words it means that the consistency of Bayes factors as the property.

Table 5.1 Summary Statistics of $\log B_n$, $n=100$

Kernel/ Prior	Summary Statistics					
	Min	Q1	Median	Mean	Q3	Max
Normal/π_{1h}	−4.888	−2.395	−1.890	−1.656	−1.168	10.530
Normal/π_{2nh}	−5.466	−0.410	0.543	0.812	1.749	14.158
K_0/π_{1h}	−31.083	−28.916	−27.757	−27.386	−26.257	−7.279
K_0/π_{2nh}	−5.798	−2.927	−1.886	−1.627	−0.617	17.414

Table 5.2 illustrated empirical type I error probabilities based on the proposed GOF using two different kernel functions $Normal(\phi)$, K_0 and two different prior distributions π_{1h}, π_{2nh} in comparison with Shapiro−Wilk(SW) and Anderson−Darling(AD) tests as well as a GOF test based on different sample sizes $n=25, 50, 100$ and 150 with the $\alpha=0.05$.

Table 5.2 Empirical Type I Error Probabilities ($\alpha = 0.05$)

sample size	Proposed Tests with B_n (Kernel/Prior)				GOF Tests	
	(ϕ/π_{1h})	(ϕ/π_{2nh})	(K_0/π_{1h})	(K_0/π_{2nh})	SW	AD
$n=25$	0.0565	0.0585	0.0051	0.0550	0.0625	0.0450
$n=50$	0.0560	0.0530	0.0515	0.0510	0.0500	0.0460
$n=100$	0.0540	0.0605	0.0485	0.0520	0.0395	0.0480
$n=150$	0.0505	0.0590	0.0400	0.0565	0.0525	0.0620

Table 5.3 illustrated the empirical powers of the GOF tests among several non-normal alternatives when the sample size is $n=100$. Note that Mix distribution in Table 5.2 denotes the half-and-half normal mixtures with $\frac{1}{2}N(-2,1)+\frac{1}{2}N(2,1)$, and DE(0,1) in Table 5.3 denotes the double exponential distribution with a location parameter 0 and a scale parameter 1. The proposed GOF tests are more powerful than two popular GOF tests, SW and AD and much more powerful than the nonparametric likelihood ratio test. Among the proposed four GOF tests with B_n, the test with K_0 kernel and prior distribution $\pi_{1h}(h)$ is the most powerful in all cases as shown with bold faced decimals in Table 5.3.

Table 5.3 Empirical Power Analysis ($n=100$ $\alpha=0.05$)

Alter. Dist.	Proposed Tests with B_n (Kernel/Prior)				GOF Tests		
	(ϕ/π_{1h})	(ϕ/π_{2nh})	(K_0/π_{1h})	(K_0/π_{2nh})	SW	AD	NL_n
$t(20)$	0.1585	0.0665	**0.2365**	0.1170	0.1055	0.0825	0.0645
$t(10)$	0.4005	0.1955	**0.5525**	0.3230	0.2300	0.1755	0.1340
$t(8)$	0.5415	0.3050	**0.6865**	0.4575	0.3100	0.2305	0.1780
$t(5)$	0.8500	0.6680	**0.9270**	0.7900	0.6220	0.4845	0.3710
$t(3)$	0.9910	0.9695	**0.9970**	0.9850	0.8820	0.8510	0.7590
$t(2)$	1.0000	0.9995	**1.0000**	1.0000	0.9860	0.9830	0.9635
MIX	1.0000	1.0000	**1.0000**	1.0000	0.9980	1.0000	0.9985
DE(0,1)	0.9785	0.9400	**0.9915**	0.9705	0.7775	0.8230	0.6410

6. CONCLUSION

In this study, a nonparametric frequentist-Bayes procedure was proposed for testing the goodness of fit of a parametric model. The test statistic is based on a nonparametric Bayes factor B_n. which compares the two marginal likelihoods corresponding to a kernel estimate and the parametric model. The marginal likelihood for the kernel estimate is obtained by proposing a prior for the bandwidth of kernel estimate, and then integrating the product of this prior and a leave-one-out kernel likelihood.

We considered the numerical comparison of the effect of different choices of the prior distribution in terms of the bandwidth selection purpose by Monte Carlo simulations with different sample sizes.

While Bayesian principles are used to construct the statistic, the test is done in frequentist way by comparing the Bayes factor with its null percentiles. Monte Carlo was used to compare the power of the new test with that of the existing goodness-of-fit testing procedures in the important case of testing for normality.

In terms of the large-sample behavior of B_n, we could establish the consistency of Bayes factor B_n that ensures the proposed GOF test with B_n favors the true model from which the data is generated.

The nonparametric GOF we proposed in the research could be generalized into several directions that include the multivariate data.

References

Albert, J.H. (2010), "Good smoothing". In: Chen, M.-H., Dey, D. K., Muller, P., Sun, D., Ye, K.(Eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, New York: Springer.

Berger, J. (1980), *Statistical Decision Theory,* New York: Springer-Verlag.

Berger, J. and Bernardo, J. (1989), "Estimating a Product of Means: Bayesian Analysis with Reference Priors," *Journal of the American Statistical Association,* 84, 200-207.

Bernardo, J. (1979), "Reference Posterior Distributions for Bayesian Inference," *Journal of the Royal Statistical Society,* B41, 113-147.

Bernardo, J.M. and Smith, A.F.M. (1994), *Bayesian Theory,* New York: Wiley.

Box, G. and Tiao G. (1973), *Bayesian Inference in Statistical Analysis,* New York: John Wiley and Sons.

Gelman, A., Carlin, B.P., Stern, H.S. and Rubin, D.B. (1995), B*ayesian Data Analysis,* London: Chapman and Hall.

Good, I.J. (1957), "Saddle-Point Methods for the Multinomial Distribution," *The Annals of Mathematical Statistics,* 28, 861-881.

Good, I.J. (1967), "A Bayesian Significance Test for Multinomial Distributions(With discussion)," *Journal of the Royal Statistical Society,* B29, 399-431.

Good, I.J. (1992), "The Bayes/Non-Bayes Compromise: a Brief Review," *Journal of the American Statistical Association,* 87, 597-606.

Hall, P. (1987), "On Kullback-Leibler Loss and Density

Estimation," *The Annals of Statistics,* 15, 1491-1519.

Hart, J.D. (2009), "Frequentist-Bayes Lack-of-Fit Tests Based on Laplace Approximations," *Journal of Statistical Theory and Practice,* 3, 681-704.

Raiffa, H. and Schlaifer, R. (1968), "Applied Statistical Decision Theory," Cambridge: M.I.T. Press.

Jaynes, E. (1968), "Prior Probabilities," *IEEE Transactions on Systems Science and Cybernetics,* 4, 227-241.

Jereys, H. (1946), "An Invariant form for the Prior Probability in Estimation Problems," *Proceedings of the Royal Society,* A186, 453-461.

Kass, R. and Wasserman, L. (1996), "The Selection of Prior Distributions by Formal Rules," *Journal of the American Statistical Association,* 91, 1343-1370.

Lindley, D.V. (1965), *Introduction to Probability and Statistics from a Bayesian Viewpoint* (2 Volumes), Cambridge: University Press.

Lee, P.M. (1997), *Bayesian Statistics: An Introduction*, 2nd ed. London: Arnold.

Pericchi, L. and Walley, P. (1991), "Robust Bayesian Credible Intervals and Prior Ignorance," *International Statistical Review,* 59, 1-23.

Richardson, S., Green, P.J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society,* B59, 731-792.

Robert, C. (1994), *The Bayesian Choice,* New York: Springer.

Scott, D.W. (1992), *Multivariate Density Estimation. Wiley Series in*

Probability and Mathematical Statistics: Applied Probability and Statistics, New York: John Wiley and Sons.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis,* London: Chapman and Hall.

Tibshirani, R. (1989), "Noninformative Priors for One Parameter of Many," *Biometrika,* 76, 604-608.

Vexler, A., Deng, W. and Wilding, G.E. (2013), "Nonparametric Bayes Factors Based on Empirical Likelihood Ratios," *Journal of Statistical Planning and Inference,* 143, 611-620.

Wand, M.P., Jones, M.C. (1995), *Kernel Smoothing,* New York: Chapman and Hall.

Welch, B. and Peers, H. (1963), "On Formulae for Confidence Points Based on Integrals of Weighted Likelihoods," *Journal of the Royal Statistical Society,* B25, 318-329.