Bayesian Analysis (2017) **12**, Number 3, pp. 653–677

Nonparametric Goodness of Fit via Cross-Validation Bayes Factors

Jeffrey D. Hart[*‡] and Taeryon Choi[†§]

Abstract. A nonparametric Bayes procedure is proposed for testing the fit of a parametric model for a distribution. Alternatives to the parametric model are kernel density estimates. Data splitting makes it possible to use kernel estimates for this purpose in a Bayesian setting. A kernel estimate indexed by bandwidth is computed from one part of the data, a training set, and then used as a model for the rest of the data, a validation set. A Bayes factor is calculated from the validation set by comparing the marginal for the kernel model with the marginal for the parametric model of interest. A simulation study is used to investigate how large the training set should be, and examples involving astronomy and wind data are provided. A proof of Bayes consistency of the proposed test is also provided.

MSC 2010 subject classifications: primary 62G10, 62F15; secondary 62G05.

Keywords: bandwidth selection, Bayes factor, consistency, cross validation, goodness-of-fit tests, kernel density estimates.

1 Introduction

Nonparametric testing of the fit of a parametric model for a distribution has a long and rich history in frequentist statistics; see, e.g., Rayner et al. (2009). However, the literature on Bayesian goodness-of-fit tests is much smaller. Müller and Quintana (2004) and Tokdar et al. (2010) review some Bayesian approaches to goodness of fit, an important one being that of Berger and Guglielmi (2001) based on Pólya trees. More recently Tokdar and Martin (2013) have proposed a Bayesian test of normality versus a Dirichlet process mixture alternative. The purpose of the current paper is to introduce a Bayesian approach to goodness of fit that has the virtues of (i) simplicity, and (ii) transparency to users unfamiliar with the somewhat daunting notions of Dirichlet processes and Pólya trees.

Given data $\boldsymbol{X} \equiv (X_1, \ldots, X_n)$ from a density f, a kernel estimator of $f(x)$ has the form

$$\hat{f}(x|\boldsymbol{X}, h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

*Department of Statistics, Texas A&M University, TX, USA, hart@stat.tamu.edu

†Department of Statistics, Korea University, Seoul, Korea. Corresponding author, trchoi@korea.ac.kr

‡Research of Jeffrey D. Hart was supported by NSF grant DMS-0604801.

§Research of Taeryon Choi was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2013R1A1A2074463).

 DOI: 10.1214/16-BA1018

where K is an appropriate kernel, typically a finite variance, unimodal density that is symmetric about 0, and h is a positive bandwidth. Kernel estimators have reached the point of being a familiar means of describing the distribution of a data set. They are no more intimidating than their primitive cousin the histogram, and have the appeal of being smooth, like one envisions the density from which the data are drawn.

Kernel estimators are attractive in the nonparametric goodness-of-fit problem since they are nonparametric estimators of the underlying density. Given a fitted parametric density, it seems natural to test the fit of this density by seeing how close it is to a kernel estimate. Such an approach in the frequentist realm dates at least to Bickel and Rosenblatt (1973). From a Bayesian point of view, however, it is not immediately clear how kernel estimates could be used in the goodness-of-fit problem. The Bayesian approach requires models for the underlying density that are well defined prior to data collection. By its nature, though, a kernel estimate only becomes a model after the data are observed. Kernel estimates do, nonetheless, have a connection with Bayesian methodology. Ferguson (1983) showed that when n is large and K is a Gaussian density, estimate (1) approximates a posterior predictive density in a model where the densities are mixtures of normals and the prior for the parameters of the mixtures is a Dirichlet process.

In this paper we use data splitting to sidestep the problem that kernel estimates are not a priori models. Given a random sample $X_1, \ldots, X_n$, suppose one randomly splits the data set into two parts, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Let $\hat{f}(\cdot\,|\boldsymbol{X}_1, h)$ be a kernel estimate computed from the data $\boldsymbol{X}_1$. The key idea of this paper is that *the collection of densities* $\{\hat{f}(\cdot\,|\boldsymbol{X}_1, h) : h > 0\}$ *comprises a parametric model (with parameter* h*) for the data* $\boldsymbol{X}_2$. To test the fit of a parametric model, such as normality, one may compute a Bayes factor based on marginal distributions of $\boldsymbol{X}_2$ corresponding to $\{\hat{f}(\cdot\,|\boldsymbol{X}_1, h) : h > 0\}$ and the parametric model of interest. The kernel model for $\boldsymbol{X}_2$ should be a reasonably good one, since $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ come from the same distribution. So, if the parametric model is wrong, the Bayes factor should favor the kernel model. If the parametric model is correct, then the parametric estimate is more efficient than the best kernel estimate, and the Bayes factor should favor the null model.

An appealing aspect of our approach is that it does not require specification of alternative parametric models. In a traditional Bayesian, nonparametric test of a parametric hypothesis, the alternative would be a rich collection of parametric models. The prior distributions for such rich collections do not always reflect sensible beliefs about the alternative, and hence the resulting Bayes factor does not necessarily have the desirable properties that Bayes factors have when testing one parametric model versus another. In contrast, under the alternative our approach only involves one parameter, the bandwidth h of the kernel estimate, and so choosing a prior is relatively straightforward. If the null hypothesis is rejected in our approach, one may consider the kernel density estimate that led to this conclusion to seek guidance as to an appropriate parametric model for the underlying density. Such an approach has long been promoted by advocates of kernel estimates.

Our method based on data splitting has an obvious connection to cross-validation. Indeed, we term the Bayes factor computed from the partitioned data as a *cross-validation Bayes factor*, or CVBF. Our idea may be regarded as a Bayesian analog

of the method proposed by van der Laan et al. (2004) for choosing the bandwidth of a kernel estimator. The most common version of cross-validation in density estimation is leave-one-out, in which an estimate using all but one data value is evaluated at the deleted observation. In contrast, and using the notation of the previous paragraph, van der Laan et al. (2004) advocate computing a likelihood by evaluating $\hat{f}(\cdot|\boldsymbol{X}_1, h)$ at the observations in $\boldsymbol{X}_2$, and then choosing the bandwidth that maximizes this likelihood. The use of cross-validation in Bayesian model selection has been previously considered by Alqallaf and Gustafson (2001).

The purpose of this paper is to explore the use of CVBFs for testing goodness of fit. The paper is by no means a comprehensive study of this idea. We do, however, establish conditions under which our test is Bayes consistent, and we present simulations and real data examples that indicate considerable promise for the method. The rest of the paper proceeds as follows. The method is described in detail in Section 2, and choosing a prior for the bandwidth of the kernel estimator is discussed in Section 3. The effect of the size of the training set $\boldsymbol{X}_1$ is investigated by simulation in Section 4, and real data examples are the subject of Section 5. A theorem on Bayes consistency of the proposed goodness-of-fit test is provided in Section 6, and concluding remarks given in Section 7.

2 The method

We describe the method in the context of checking the fit of a parametric model for a probability distribution. The method could also be applied to other testing problems, such as lack-of-fit in regression, although we do not pursue that possibility here. For $n > k$, suppose that $\boldsymbol{X}_1 = (X_{1,1}, \ldots, X_{1,k})$ and $\boldsymbol{X}_2 = (X_{2,1}, \ldots, X_{2,n-k})$ are independent random samples from density f, and define $\hat{f}(\cdot|\boldsymbol{X}_1, h)$ to be a kernel estimate as follows:

$$\hat{f}(x|\boldsymbol{X}_1, h) = \frac{1}{kh} \sum_{i=1}^{k} K\left(\frac{x - X_{1,i}}{h}\right),$$

where K is an appropriate kernel, typically a finite variance, unimodal density that is symmetric about 0, and h is a positive bandwidth. Of interest is testing the fit of M_0, a parametric model for f. Given $\boldsymbol{X}_1$, $M_1(\boldsymbol{X}_1) = \{\hat{f}(\cdot|\boldsymbol{X}_1, h) : h > 0\}$ is another parametric family of densities for the data $\boldsymbol{X}_2$. Assuming only smoothness of f, one can expect a member of $M_1(\boldsymbol{X}_1)$ to be relatively close to f, especially when k is large. So, we may test the fit of M_0 by computing a Bayes factor from the data set $\boldsymbol{X}_2$ that compares M_0 and $M_1(\boldsymbol{X}_1)$. The posterior of h given $\boldsymbol{X}_2$ is

$$\pi(h|\boldsymbol{X}_2) = \prod_{i=1}^{n-k} \hat{f}(X_{2,i}|\boldsymbol{X}_1, h)\pi(h)/m(\boldsymbol{X}_2|M_1(\boldsymbol{X}_1)),$$

where $\pi(\cdot)$ is the prior for h and

$$m(\boldsymbol{X}_2|M_1(\boldsymbol{X}_1)) = \int_0^\infty \prod_{i=1}^{n-k} \hat{f}(X_{2,i}|\boldsymbol{X}_1, h)\pi(h)\, dh.$$

If $m(\boldsymbol{X}_2|M_0)$ is the marginal of $\boldsymbol{X}_2$ assuming that M_0 is the correct model, then a Bayes factor would be $m(\boldsymbol{X}_2|M_1(\boldsymbol{X}_1))/m(\boldsymbol{X}_2|M_0)$.

Given a single data set $\boldsymbol{X} = (X_1, \ldots, X_n)$, the previous idea may be applied by randomly splitting the data set into two parts. Using the parlance of cross-validation, these two sets are called the *training* and *validation* sets. The kernel estimate is computed from the training set and the Bayes factor from the validation set. We shall refer to a Bayes factor computed in this way as a *cross-validation Bayes factor*, or CVBF. In principle one could obtain Bayes factors for all splits corresponding to a given k, and then compute either an arithmetic or geometric mean of all the Bayes factors. The idea of averaging Bayes factors has been used by Berger and Pericchi (1996) in defining intrinsic Bayes factors. Unless n is very small, it would be prohibitive to consider all $\binom{n}{k}$ splits for a given k. Instead, it seems to be sufficient to randomly choose a large number (say 1000 to 10,000) of random splits, compute a Bayes factor for each and average the results. We also believe it is worthwhile to consider the distribution of the Bayes factors.

What should one use for k? Ideally one would choose k so that, with high probability, the resulting Bayes factor is less than 1 under H_0 and considerably larger than 1 under alternatives. Calibration of CVBF may be performed by investigating how it behaves when data are generated from the null model. Doing so has somewhat the same flavor as methodology proposed by Xu et al. (2011), who suggest that data splitting be used to calibrate Bayes factors. Their approach uses a training sample to construct posteriors which are used as priors for models that are fitted to the validation data. The calibration is aimed at finding priors that contain a given amount of information. Calibration in our setting consists of determining a value of k such that CVBF has a very small probability of exceeding, say, 0.50 when the null hypothesis is true. Having determined a k that produces the desired CVBF behavior, one would then have to accept the resulting behavior in the event that H_0 is false. This is similar to what a frequentist does in choosing a test to have a small type I error probability, and then accepting the fact that test power will not be good unless the alternative is sufficiently discrepant from the null.

Another consideration in choosing k is Bayes consistency of the test. In Section 6 we establish conditions under which a test based on CVBF for a single random split is Bayes consistent. A condition required to prove our result is that k tend to ∞, but at a much slower rate than n.

3 Prior for the bandwidth

In the sequel, we will always use a Gaussian kernel for K. The density estimation literature (e.g., Silverman, 1986) tells us that the Gaussian kernel is nearly optimal in a mean squared error sense. We do not claim that a similar property of the Gaussian kernel is true in the current context. However, the point of this paper is to introduce ideas and to demonstrate the potential of CVBFs, and hence we defer an investigation of kernel effect to future work.

Our experience with CVBFs indicates that choosing a good prior for the bandwidth h can be quite important. A good prior will produce stable Bayes factors that tend to be much less than 1 when the null hypothesis is true and much larger than 1 when the null hypothesis is false. For this purpose we have found that a prior of the following form works quite well:

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

where $\beta > 0$. When $\pi(h|\beta)$ is used as the prior for the bandwidth of kernel estimate $\hat{f}(\cdot\,|\boldsymbol{X}_1, h)$, we take β to be $R/1.35$, where R is the interquartile range of the validation data $\boldsymbol{X}_2$. Using the validation data to tune the prior for h is consistent with the practice of using the same data set to compute parameters of reference priors and a Bayes factor.

Two aspects of prior (2) make it appealing. First of all, it is proper, which is a necessary condition for a Bayes factor to be well-defined. Secondly, the prior tends to 0 as h tends to 0. As a prior for the bandwidth of an *observed* kernel estimate, it seems that this should almost be a requirement, since one can plot the estimate for different values of h and identify an h below which the corresponding empirical kernel models are clearly unsatisfactory. On the other hand, if one uses prior (2) with the same β for all n, then for all n sufficiently large the prior is giving very low probability to values of h that are a priori the most likely under H_0. (This is because an asymptotically optimal choice of h tends to 0 as n tends to ∞.) Such a prior may seem odd, but is in the same spirit as the idea of a *non-local* prior espoused by Johnson and Rossell (2010). Suppose Θ_0 and Θ_1 are mutually exclusive and exhaustive subsets of a parameter space Θ. When testing $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$, a *local alternative prior*, in the terminology of Johnson and Rossell (2010), is one that assigns nonzero probability to Θ_0, and a non-local alternative prior assigns probability 0 to Θ_0. Johnson and Rossell (2010) advocate the use of non-local alternative priors that are quite small for values of θ that are in Θ_1 but close to Θ_0. They prove that certain priors of this type can induce an exponential rate of convergence of a Bayes factor to 0 when H_0 is true, while not disturbing the exponential convergence rate of the Bayes factor under alternatives. We will show in Section 6 that a Bayes factor using prior (2) also converges to 0 at an exponential rate under H_0.

Another motivation for (2) comes from an idea as in Berger and Pericchi (1996) for circumventing the problem of using noninformative priors in model selection. Berger and Pericchi (1996) suggest using a small part of the data to compute a posterior based on a noninformative prior, and to then use this posterior as a prior for the remainder of the data. One uses the smallest number of observations such that the posterior for those observations is proper. Arguably, a reasonable noninformative prior for h would be the improper prior $h^{-1}I_{(0,\infty)}(h)$, which results from an invariance-under-scale-transformations argument. Suppose one computes a kernel density estimate using a Gaussian kernel and a single observation X_i. A posterior can be formed by then evaluating that estimate at an independent data value X_j, and multiplying by h^{-1}. This posterior is proportional to

$$h^{-2} \exp\left(-\frac{(X_i - X_j)^2}{2h^2}\right),$$

which obviously has the same form as (2). Our data-driven choice of β^2 targets $E[(X_i - X_j)^2]/2$ when the observations are Gaussian.

Though our subsequent results will only use the prior (2), we have experimented extensively with other priors, for example gamma distributions with shape parameters less than 1. Our finding is that priors that do not tend to 0 as h tends to 0 produce CVBFs that are somewhat unstable. In particular, the CVBFs are sensitive to the choice of k. A choice of k that is acceptable under the null hypothesis is not necessarily a good choice under alternative hypotheses. In contrast, we show numerically in the next section that CVBFs based on (2) seem to perform well for any k under the null hypothesis, which means that one is free to choose k so that it is optimal for alternatives.

4 Choice of training set size

An important consideration in using CVBFs is the size k of the training set. Ideally one would consider this problem theoretically, but at this early stage of our investigation we study it using simulation. We consider testing the fit of two parametric models: normal and Laplace. For each data set generated, values of k equal to $0.20n, 0.30n, \ldots, 0.80n$ were considered, and a CVBF for each k was computed by averaging the log(CVBF) values from 100 random data splits. Whether testing for a normal or Laplace density, data were generated from three different distributions. These were normal, Laplace and skew-normal distributions when testing for normality, and Laplace, normal and skew-Laplace when testing the fit of a Laplace density.

Two sample sizes, $n = 200$ and 500 were considered, and five hundred independent data sets were generated for each of the twelve combinations of n and testing/distribution scenario. The skew-normal density considered was

$$f_{\mathrm{SN}}(x) = 2\phi(x)\Phi(5x),$$

where ϕ and Φ are the pdf and cdf of a standard normal distribution, respectively, and the skew-Laplace density was

$$f_{\mathrm{SL}}(x) = \left(\frac{\alpha}{\alpha+\beta}\right) f_{\mathrm{E}}(-x|\alpha) + \left(\frac{\beta}{\alpha+\beta}\right) f_{\mathrm{E}}(x|\beta),$$

where $f_{\mathrm{E}}(\cdot\,|\theta)$ is an exponential density with mean θ, and we take $\alpha = 1/4$ and $\beta = 1$.

Whether testing the fit of a normal or Laplace distribution, our prior for the location and scale parameters, (μ, σ), has the form

$$\pi_0(\mu,\sigma) = \frac{1}{\sigma} g_1\left(\frac{\mu - m}{\sigma}\right) \frac{1}{\gamma} g_2\left(\frac{\sigma}{\gamma}\right),$$

where g_1 and g_2 are densities with respective supports the real line and the positive reals, m is a location estimate and γ a scale estimate. When testing normality, we take g_1 to be standard normal,

$$g_2(s) = \frac{2}{\sqrt{\pi}} \cdot \frac{1}{s^2} e^{-1/s^2} I_{(0,\infty)}(s),$$

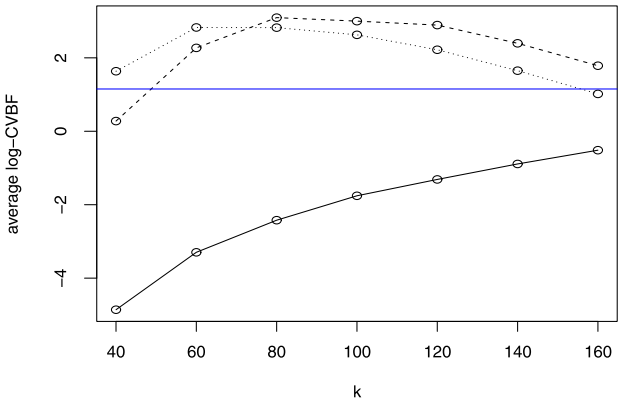

Figure 1: Averages of log-CVBF as a function of the training set size. The null hypothesis tested is normality, and the sample size is $n = 200$. The solid, dashed and dotted lines correspond to normal, Laplace and skew-normal data, respectively. The horizontal line indicates Jeffreys' cutoff between substantial and insubstantial evidence against the null hypothesis.

m to be the sample mean and $\gamma = \hat{\sigma}/\sqrt{2}$, where $\hat{\sigma}$ is the sample standard deviation. The resulting prior is a unit-information reference prior, in that it is centered at the observed data and contains an amount of information equivalent to that in a single observation.

When testing the fit of a Laplace density, we take g_1 to be a standard Laplace distribution, $g_1(x) = \exp(-|x|)/2$, and g_2 an inverse gamma distribution: $g_2(s) = s^{-2}e^{-1/s}I_{(0,\infty)}(s)$. For γ we use $R/\log 2$, and for m the sample median. As in the normal case, this prior is a unit-information reference prior. Finally, as prior for h we use (2) with $\beta = R/1.35$. The data used in computing the parameters of the priors for both null and kernel models were always the validation data $\boldsymbol{X}_2$, i.e., the data from which the Bayes factor is computed. It is important to note that defining the priors as we do *makes our Bayes factors invariant to the location and scale of the density from which the data were generated.*

The results of our simulations are summarized in Figures 1–4. We desire two things from a Bayes factor: that it be smaller than 1 when the null hypothesis is true, and quite a bit larger than 1 when the null is not true. Of course, on the log scale, 0 is the value representing indifference between the two models. With this in mind, Figures 1–4 show that the average log(CVBF) always increased with k when the null hypothesis was true, but never exceeded 0. Jeffreys (1961) considered Bayes factors larger than $\sqrt{10} = 3.16$ to be "substantial evidence" against H_0. When the null hypothesis was false, the only time the geometric mean of the CVBF was smaller than 3.16 was when testing normality at $n = 200$. This occurred with Laplace data at $k = 40$ and skew-normal data at $k = 160$, both of which are extreme choices for k. Under alternatives, the average log(CVBF) was either approximately flat or increasing between $k = 0.2n$ and $k = 0.4n$, and then decreasing from $k = 0.4n$ to $0.8n$. This suggests that $0.4n$ is a good choice for k in the 8 scenarios where H_0 was false. At $k = 0.4n$, the smallest geometric mean for CVBF

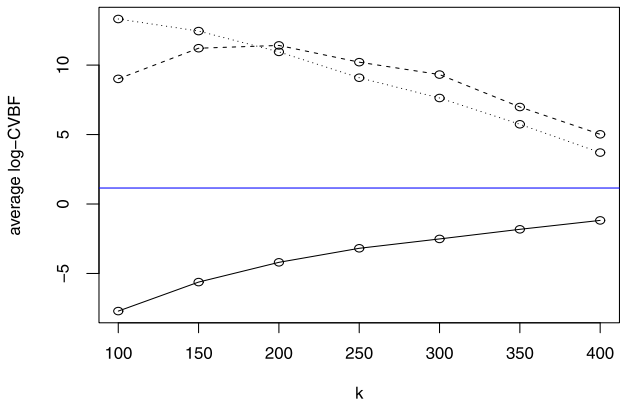

Figure 2: Averages of log-CVBF as a function of the training set size. The caption for this case is identical to that in Figure 1 except that $n = 500$.

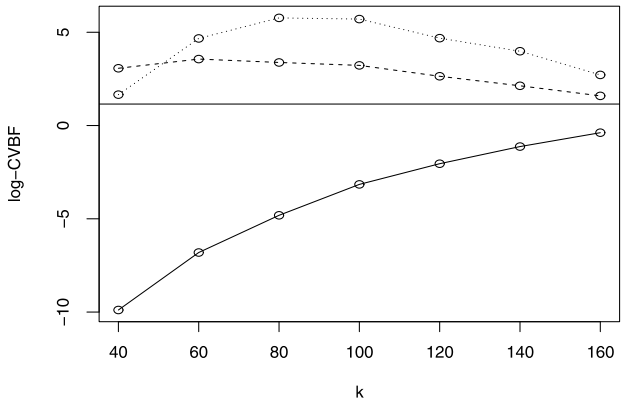

Figure 3: Averages of log-CVBF as a function of the training set size. The null hypothesis tested is that the data are Laplace, and the sample size is $n = 200$. The solid, dashed and dotted lines correspond to Laplace, normal and skew-Laplace data, respectively. The horizontal line indicates Jeffreys' cutoff between substantial and insubstantial evidence against the null hypothesis.

over these 8 scenarios was 16.81. On the other hand, at $k = 0.4n$ the geometric mean of CVBF was no larger than 0.089 for any of the four scenarios where H_0 was true.

On the basis of our simulation results, it seems reasonable to suggest using a value of k between $0.3n$ and $0.4n$, at least for sample sizes between 200 and 500. Of course, much more work needs to be done on the problem of choosing the training set size. In order for CVBF to be fully efficient, it may be necessary that k/n tend to 0 as n tends to ∞. Only then is CVBF computed on a data set whose size is asymptotic to n. Our simulation results suggest that choosing k very small is not a problem when the null

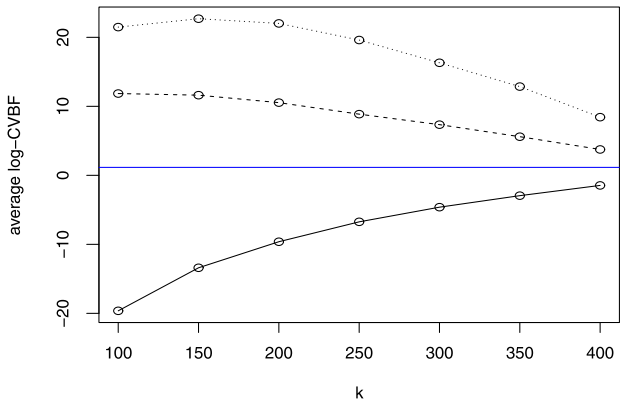

Figure 4: Averages of log-CVBF as a function of the training set size. The caption for this case is identical to that in Figure 3 except that $n = 500$.

hypothesis is true. However, if H_0 is false, it is conjectured that k should tend to ∞, albeit at a rate slower than n.

It would be interesting to develop a theory that would indicate an optimal rate at which k should grow with n. A possible definition of an optimal k is as follows. First, determine a set of k-values, call it $\mathcal{K}$, such that for each $k \in \mathcal{K}$ the expected value of log(CVBF) under H_0 is no larger than a given threshold (which would certainly be less than 0). For a specific alternative, define the optimal k to be a member of $\mathcal{K}$ that maximizes expected value of log(CVBF) when the data come from that alternative.

5 Examples

5.1 Analysis of planetary nebula luminosity data

Here we analyze observations of planetary nebula luminosity in the Messier 31, or *Andromeda*, galaxy. For background on the data analyzed, the reader is referred to Ciardullo et al. (1989) and Ciardullo et al. (2002). The frequency distribution of planetary nebula luminosities is referred to as a *planetary nebula luminosity function*, or PNLF. From a visual inspection of a histogram of Messier 31 data, Ciardullo et al. (1989) claimed that the PNLF of Messier 31 does not follow a power law. We wish to validate this claim using our CVBF.

The available data from Messier 31 are brightness readings from 238 planetary nebulae, which can be downloaded from `http://astrostatistics.psu.edu/datasets/plan_neb.html`. These readings are measured on a scale such that smaller values correspond to brighter objects. Ciardullo et al. (1989) note that dimmer readings are subject to more measurement error. Therefore, they used only observations smaller than 22 to fit the PNLF curve. In statistical parlance, their data are censored. Let $X_1, \ldots, X_n$ be the actual observations and $Y_1, \ldots, Y_n$ be the underlying data that are free of errors. When

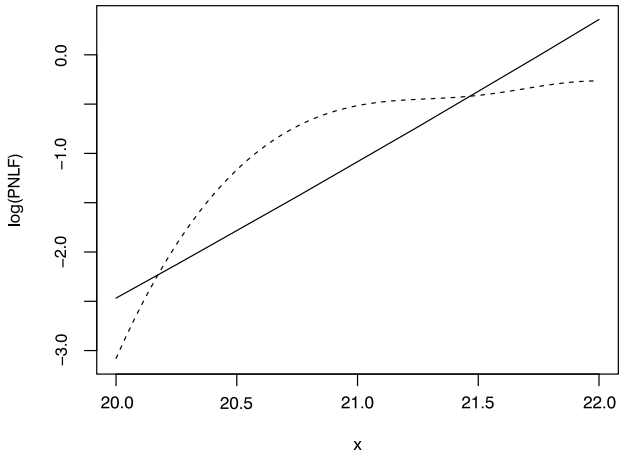

Figure 5: Estimates of log-density of luminosity. The solid line is a fitted power law curve that maximizes the posterior density. The dashed line is the log of a kernel density estimate. Each estimate is computed from 61 luminosity readings smaller than 22.

$X_i < 22$, it is assumed that $X_i = Y_i$, but if $X_i > 22$ then we only know that $Y_i > 22$. We wish to test whether the PNLF curve follows a power law at brightnesses less than 22. Plots of kernel and power law estimates of log-density are shown in Figure 5. Of the 238 observed luminosities, only 61 were smaller than 22.

The null density of Y_i is such that

$$f_0(y) = (1-p)f(y|\alpha), \quad y < 22,$$

where $p = P(Y_i > 22)$ and

$$f(y|\alpha) = \frac{\alpha\, 22^\alpha}{(44-y)^{\alpha+1}} I_{(-\infty,22)}(y). \tag{3}$$

The parameters p and α are unknown, with $\alpha > 0$. Let $\boldsymbol{X}_1$ be an arbitrary subset of k of the 238 observations, and let $\boldsymbol{X}_2$ be the other $238-k$ observations. Under the null hypothesis, the likelihood function for $\boldsymbol{X}_2$ is

$$L(\alpha, p) = p^{n_T}(1-p)^{238-k-n_T} \prod_{x_i \in S} f(x_i|\alpha),$$

where n_T is the number of observations among $\boldsymbol{X}_2$ that are at least 22 and S contains the observations in $\boldsymbol{X}_2$ that are smaller than 22. Now let $\hat{f}(\,\cdot\,|\boldsymbol{X}_1, h)$ be a kernel estimate computed from the observations in $\boldsymbol{X}_1$ that are smaller than 22. Treating such kernel estimates as a model for $\boldsymbol{X}_2$, the kernel likelihood has the form

$$L_K(h, p) = p^{n_T}(1-p)^{238-k-n_T} \prod_{x_i \in S} \hat{f}(x_i|\boldsymbol{X}_1, h).$$

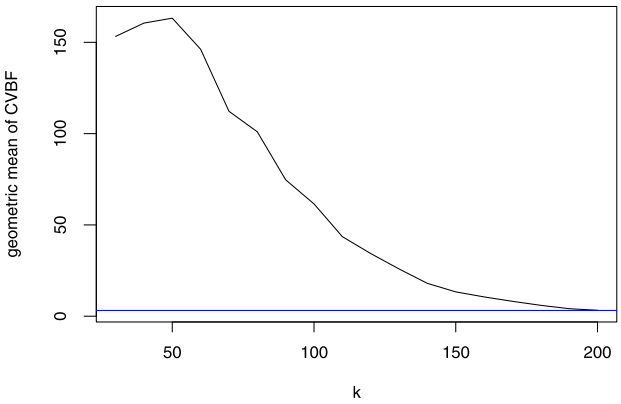

Figure 6: CVBF as a function of k, the size of the training set, for the luminosity data. Each value is the geometric mean of 1000 Bayes factors. The flat line indicates the smallest value that is considered substantial evidence against H_0.

If p is assumed a priori independent of h and α, then a Bayes factor is simply

$$B = \left[\int_0^\infty \pi(h) \prod_{x_i \in S} \widehat{f}(x_i|\boldsymbol{X}_1, h)\, dh\right] \left[\int_0^\infty \pi_0(\alpha) \prod_{x_i \in S} f(x_i|\alpha)\, d\alpha\right]^{-1}.$$

In other words, the Bayes factor has exactly the same form it would have if the uncensored data were a random sample of size $238 - k - n_T$ from density $(1-p)^{-1} f(y) I_{(-\infty,22)}(y)$, where f is the true density of Y_i. To compute the Bayes factor we assumed that the prior for h is $\pi(\cdot\,|\beta)$, as defined in Section 3. We used a gamma prior for α, which is a conjugate prior in this case, taking the shape parameter to be 2 and the rate parameter equal to $M = \sum_{x_i \in S} \log((44 - x_i)/22)/(238 - k - n_T)$. This prior has an amount of information equivalent to one observation from the power law density.

As is well-known, kernel estimates are subject to edge effects when the underlying density does not tend to 0 at the endpoints of its support. Such is the case here at $x = 22$ because of the data censoring. We thus used a reflection technique (Silverman, 1986, p. 30) to compute $\hat{f}(\cdot\,|\boldsymbol{X}_1, h)$. Suppose there are m observations in $\boldsymbol{X}_1$ that are less than 22. If x_i is any one of these m, then the reflection technique involves defining a new data value larger than 22 that is the same distance from 22 as is x_i. One then computes a standard kernel estimate based on the new data set of $2m$ observations, and multiplies it by 2, to ensure that it integrates to 1 over $(-\infty, 22)$.

Initially we considered the effect of using different choices for k, the size of the training set. At each of 18 choices for k ranging from 30 to 200, we considered 1000 random data splits, computed CVBF for each split, and then found the geometric mean of the 1000 Bayes factors. The results are summarized in Figure 6. The largest evidence against H_0 occurred when k was 50, which is 21% of $n = 238$. The flat line is at $\sqrt{10}$,

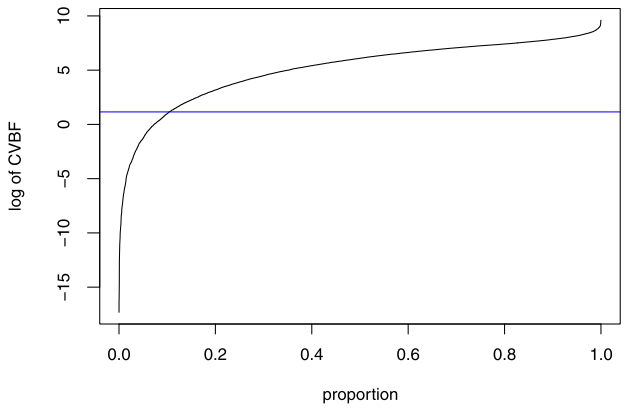

Figure 7: Quantile function of log Bayes factors corresponding to 10,000 randomly chosen splits of the luminosity data. The training and validation splits were of size 50 and 188, respectively. The flat line indicates the smallest value that is considered substantial evidence against H_0.

which Jeffreys (1961) deemed to be the smallest value indicating substantial evidence against the null hypothesis.

Using $k = 50$, we did a more detailed analysis. First we computed a Bayes factor for each of 10,000 randomly selected data splits. The geometric mean of these 10,000 Bayes factors was 164.23, a value generally regarded as strong evidence against the null hypothesis. An empirical quantile function for the logarithm of the 10,000 Bayes factors is shown in Figure 7. It is advisable to ensure that CVBF behaves appropriately when the null hypothesis is true. If it does, then one is justified in regarding a Bayes factor of 164.23 as evidence against the null hypothesis. We thus generated 1000 data sets, each of size 61, from density (3) with $\alpha = 31.48$, which is the maximum likelihood estimate of α for the luminosity data. Each data set was augmented with $177 = 238 - 61$ values larger than 22, and then 100 random splits with $k = 50$ were considered for each data set of size 238. (The values ascribed to data larger than 22 are inconsequential since only values less than 22 are used in computing CVBF.) The geometric mean of the 100 values of CVBF was computed for each data set. These 1000 means had a geometric mean of 0.0176, only 0.6% were larger than $\sqrt{10}$, and the largest was 17.95.

5.2 Wind direction data

Here we consider daily peak wind directions from Dallas, TX, USA for the year 2014. The available data consist of $n = 364$ directions in degrees, with 0 (or 360) degrees indicating a peak wind from due north. Defining W to be the reading in degrees, the data were transformed to radians as follows:

$$X = \begin{cases} -\frac{\pi}{2}\left(\frac{W}{90} - 1\right), & 0 \le W \le 270 \\ -\frac{\pi}{2}\left(\frac{W}{90} - 5\right), & 270 < W \le 360. \end{cases}$$

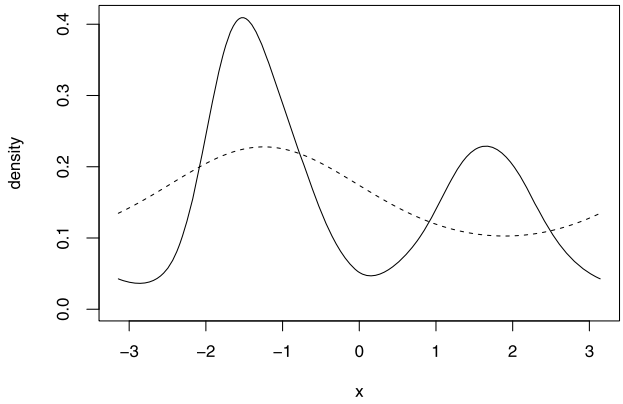

Figure 8: Kernel estimate (solid line) and fitted von Mises distribution (dashed line) for the wind data.

With this transformation northerly winds correspond to $x = \pi/2$ and southerly winds to $-\pi/2$.

The circularity of the data should be taken into account when estimating their distribution. One anticipates that the density of W is continuous at 270 degrees, which entails that the density, f, of X is such that $f(-\pi) = f(\pi)$. This condition is ensured by using a kernel estimate of the following form:

$$\hat{f}(x|h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \quad -\pi \le x \le \pi, \tag{4}$$

where

$$K_h(u) = \frac{1}{h}\sum_{j=-\infty}^{\infty} \phi\left(\frac{u - 2\pi j}{h}\right), \quad -\pi \le u \le \pi. \tag{5}$$

Kernel (5) is a *wrapped normal kernel* (Mardia and Jupp, 2000), and the estimate (4) integrates to 1 over the interval $(-\pi, \pi)$.

For descriptive purposes the bandwidth h of the kernel estimate was selected using the likelihood cross-validation method of van der Laan et al. (2004). The resulting kernel estimate is shown in Figure 8. The estimate shows that peak winds are usually from the north or south, with southerly peak winds being more frequent than northerly ones.

A commonly used model in circular data applications is the von Mises distribution, which is a circular analog of the normal distribution. The von Mises density is

$$f(x|\mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(x-\mu)}, \quad -\pi \le x \le \pi,$$

where $-\pi \le \mu \le \pi$, $\kappa > 0$ and I_0 is the modified Bessel function of order 0. The maximum likelihood estimates (MLEs) of μ and κ using all of the wind data are -1.245

and 0.3982, respectively. Figure 8 shows the MLE of the von Mises density along with the kernel estimate.

We shall compute CVBFs for comparing kernel estimates with the von Mises model. The prior for h was taken to be (2) with β equal to a circular data scale estimate computed from the validation data. The parameters μ and κ of the von Mises distribution were taken to be a priori independent, with μ being uniform on $(-\pi, \pi)$ and κ having an exponential density with mean equal to the MLE of κ from the validation data. The size k of each training set was taken to be 146. To ensure that this choice of k produces Bayes factors that behave appropriately under the null hypothesis, we generated 200 random samples of size 364 from a von Mises distribution with $\mu = -1.245$ and $\kappa = 0.3982$. For each data set, 50 random splits, each with $k = 146$, were considered. The geometric mean of the 50 CVBF values was computed for each of the 200 data sets. These 200 geometric means ranged between 0.080 and 0.738 and had a geometric mean of 0.237. We then considered 1000 random splits, each with $k = 146$, of the wind data and computed CVBF for each split. The smallest value of CVBF among the 1000 data splits was $3.48 \cdot 10^{11}$ and the geometric mean of the CVBF values was $7.51 \cdot 10^{20}$, providing overwhelming evidence against the von Mises model.

The von Mises model may seem like something of a straw man since the nonparametric estimate is strikingly bimodal. To make the goodness of fit problem more challenging, we thus consider testing the fit of a mixture of two von Mises distributions. Taking $\mu_1 \leq \mu_2$, the model considered is

$$f(x|\mu_1, \mu_2, \kappa_1, \kappa_2, w) = wf(x|\mu_1, \kappa_1) + (1 - w)f(x|\mu_2, \kappa_2), \quad -\pi \leq x \leq \pi.$$

We use a prior in which κ_1, κ_2 and w are mutually independent, the pair (μ_1, μ_2) is independent of (κ_1, κ_2, w), w is uniform on $(0, 1)$, each κ has a standard exponential density, and μ_1 and μ_2 are order statistics for a random sample of size 2 from the uniform density on $(-\pi, \pi)$. Initially we used the Markov chain Monte Carlo (MCMC) method to determine a good point estimate of the mixture model. A multivariate normal proposal distribution was used. Several iterations of MCMC were used to tweak the parameters of the proposal, and ultimately excellent mixing was obtained. The mixture density corresponding to the average of 10,000 values generated from the posterior is shown in Figure 9. The mixture density matches the kernel estimate reasonably well.

We computed values of CVBF corresponding to training set sizes (k) of 73, 109, 146, 182, 218, 255 and 291. At each k we computed CVBF for 200 random splits. The marginal for the mixture model was determined (for each data split) by using importance sampling. Denoting the prior by π and letting $\boldsymbol{\theta} = (\mu_1, \mu_2, \kappa_1, \kappa_2, w)$, the marginal for data $x_1, \ldots, x_m$ is approximated by

$$\frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{m} f(x_j|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)/g(\boldsymbol{\theta}_i),$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ is a random sample from g, and g is a multivariate normal distribution whose mean and covariance matrix are determined from our initial MCMC analysis. The geometric mean of the 200 CVBF values at each k is shown in Figure 10. These values show that the von Mises mixture model is strongly favored over the kernel model.

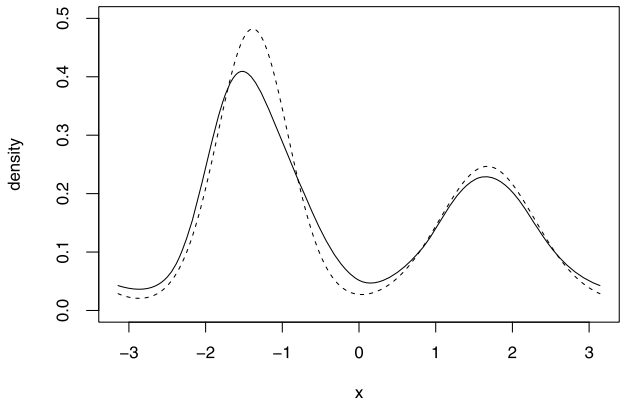

Figure 9: Kernel estimate (solid line) and fitted mixture of two von Mises distributions (dashed line) for the wind data.

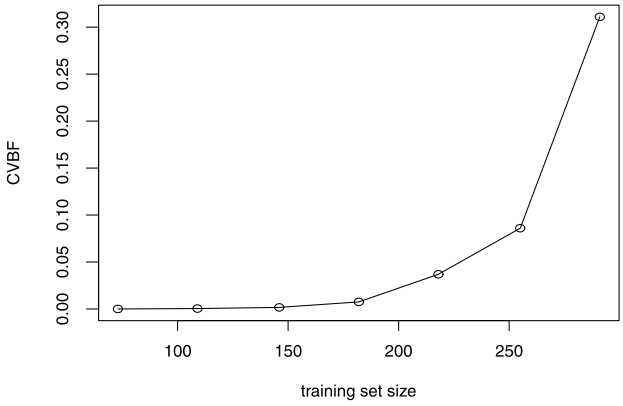

Figure 10: Values of CVBF for comparing kernel estimates with the von Mises mixture model. Each value is the geometric mean of CVBF over 200 random splits.

6 Bayes consistency

In our goodness-of-fit context, Bayes consistency occurs if the Bayes factor tends to 0 in probability when H_0 is true and to ∞ when H_0 is false. Here we provide a theorem on consistency in a case where the underlying density is defined on a known, finite interval. While the finite support assumption is less than ideal, we at least provide a first rigorous step towards a theoretical justification of our methodology.

Notation.

- Given observations $X_1, \ldots, X_n$, the training set is $\boldsymbol{X}_T = (X_1, \ldots, X_k)$ and the validation set $\boldsymbol{X}_V = (X_{k+1}, \ldots, X_n)$.

- The kernel K is a standard normal density.
- $I_{kh} = \int_a^b \frac{1}{kh} \sum_{i=1}^k K(\frac{x-X_i}{h})\, dx$.
- $\hat{f}(x|h, \boldsymbol{X}_T) = \frac{1}{kh} \sum_{i=1}^k K(\frac{x-X_i}{h}) \cdot I_{kh}^{-1}$.
- $J_{kn} = \sum_{i=k+1}^n \log f_0(X_i)$.
- $\ell_k(h) = \sum_{i=k+1}^n \log \hat{f}(X_i|h, \boldsymbol{X}_T)$.
- $\tilde{\ell}_k(h) = (n-k) \int_a^b \log \hat{f}(x|h, \boldsymbol{X}_T) f_0(x)\, dx$.
- $\ell(\boldsymbol{\theta}) = \sum_{i=k+1}^n \log \hat{f}(X_i|\boldsymbol{\theta})$.
- The priors for h and $\boldsymbol{\theta}$ are π and p, respectively.
- $h_M = k^{-\beta}$, $h_1 = k^{-\alpha}$ $(0 < \beta < \alpha)$, $h_j = h_1 + (h_M - h_1)(j-1)/(M-1)$, $j = 1, \ldots, M$.
- $\mathcal{H}_M = \{h_1, \ldots, h_M\}$.
- $\mathrm{BF}_{kn} = (h_M - h_1)(M-1)^{-1} \sum_{j=1}^M \exp(\ell_k(h_j))\pi(h_j) / \int_\Theta \exp(\ell(\boldsymbol{\theta}))p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$.

Assumptions.

1. $X_1, \ldots, X_n$ is a random sample from density f_0 that has known, finite support (a, b).
2. The density f_0 and its first derivative are bounded and continuous on (a, b), and f_0'' exists and is uniformly continuous on (a, b). In addition, f_0, f_0' and f_0'' are right and left continuous at a and b, respectively, with $f_0'(a+) = 0$ and $f_0'(b-) = 0$.
3. When H_0 is true, $f_0 \equiv f(\cdot|\boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}_0 \in \Theta$. The integral
$$\int_a^b \log f(x|\boldsymbol{\theta}) f_0(x)\, dx$$
exists for all $\boldsymbol{\theta}$ under both null and alternative hypotheses. When the alternative is true, there exists a $\boldsymbol{\theta}_0 \in \Theta$ such that $\boldsymbol{\theta}_0$ maximizes
$$\int_a^b \log f(x|\boldsymbol{\theta}) f_0(x)\, dx,$$
with respect to $\boldsymbol{\theta}$. Furthermore, the Kullback–Leibler discrepancy
$$\int_a^b \log[f_0(x)/f(x|\boldsymbol{\theta}_0)] f_0(x)\, dx$$
is strictly positive, and $\int_a^b [\log f_0(x)]^2 f_0(x)\, dx < \infty$.

4. Under both null and alternative hypotheses the marginal for the parametric model is asymptotic to the following Laplace approximation:

$$(2\pi)^{d/2} n^{-d/2} |\boldsymbol{I}(\hat{\boldsymbol{\theta}})|^{-1/2} p(\hat{\boldsymbol{\theta}}) \exp(\ell(\hat{\boldsymbol{\theta}})),$$

where $\hat{\boldsymbol{\theta}}$ is the MLE based on $\boldsymbol{X}_V$ and $\boldsymbol{I}(\boldsymbol{\theta})$ is the $d \times d$ matrix with ij element

$$\int_a^b \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(x|\boldsymbol{\theta}) f_0(x)\, dx.$$

5. The MLE $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_0$ under both null and alternative hypotheses as $n \to \infty$. Also, $\boldsymbol{I}$ and p are continuous at $\boldsymbol{\theta}_0$, and $p(\boldsymbol{\theta})$ is positive for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$.

6. Under both null and alternative hypotheses, $\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)$ converges in distribution to a positive random variable Y as $n \to \infty$.

7. The upper bound, $k^{-\beta}$, on the bandwidth is such that β can be arbitrarily close to 0. The lower bound, $k^{-\alpha}$, is such that $1/4 < \alpha < 1$. We assume that $k = n^c$ with $c < [8(\alpha + 1/5)]^{-1}$, and $M = O(k)$.

Theorem 1. *Suppose that Assumptions 1–7 hold. If the null hypothesis is true, then* BF_{kn} *is bounded above by a random variable* $\mathcal{B}_n$ *satisfying*

$$\log \mathcal{B}_n = -Pn^{1-c(2\alpha+4/5)} + o_p(n^{1-c(2\alpha+4/5)})$$

as $n \to \infty$*, where* P *is a positive constant and* $1 - c(2\alpha + 4/5) > 1/2$*. If the alternative is true*

$$\log BF_{kn} = Qn + o_p(n)$$

as $n \to \infty$*, where* Q *is a positive constant.*

Some remarks are in order concerning our theorem.

- Note that our proof is for the case where only one random split is used. Clearly, though, a Bayes factor based on the geometric mean of many random splits will perform even better.

- The Bayes factor analyzed actually uses a Riemann sum approximation to the marginal of the kernel model. This seems perfectly reasonable, however, since in practice numerical methods (such as a Riemann sum) must be used to approximate this marginal.

- It is noteworthy that under the null hypothesis our Bayes factor converges to 0 at a rate of $\exp(-Pn^\eta)$ for some $1/2 < \eta < 1$. This is in contrast to typical results when testing a parametric null hypothesis against a nonparametric alternative, where under the null hypothesis the Bayes factor converges to 0 at slower than an exponential rate. (See, for example, McVinish et al. (2009).) The reason for exponential convergence in our case is the inefficiency of the kernel estimate relative

to the parametric model under H_0. This creates a sharper contrast between null and alternative models when using our methodology. This is a tangible benefit of using our method as opposed to one in which the null is a special case of a larger alternative model.

- Exponential convergence of Bayes factors under alternatives is typical, and we note that our method also has this desirable property.
- Our proof requires that $k = n^c$, where c must be less than 5/18.
- Finally, it is worth noting that our proof of consistency does not require using any of the optimality properties of likelihood cross-validation as proven by van der Laan et al. (2004). In particular, we do not need to assume (as do van der Laan et al. (2004)) that the density is bounded away from 0.

Proof. We begin with a lemma and its proof.

Lemma 1. *The statistic* $\mathcal{I}_k = \min_{h \in \mathcal{H}_M} I_{kh}^2$ *converges in probability to 1 as* $n \to \infty$.

Proof. We have

$$I_{kh} = \frac{1}{k} \sum_{i=1}^{k} \left[\Phi\left(\frac{b - X_i}{h} \right) - \Phi\left(\frac{a - X_i}{h} \right) \right],$$

where Φ is the standard normal cdf, and hence I_{kh} is between 0 and 1 almost surely. For ϵ an arbitrarily small positive number,

$$\begin{aligned} P(|\mathcal{I}_k - 1| > \epsilon) &\leq P(\cup_{j=1}^{M} \{|I_{kh_j}^2 - 1| > \epsilon\}) \\ &\leq \sum_{j=1}^{M} P(|I_{kh_j}^2 - 1| > \epsilon) \\ &= \sum_{j=1}^{M} P\left(I_{kh_j} < (1-\epsilon)^{1/2} \right). \end{aligned} \tag{6}$$

Letting E_j denote $E(I_{kh_j})$, consider

$$P\left(I_{kh_j} < (1-\epsilon)^{1/2} \right) \leq P(I_{kh_j} - E_j < a_\epsilon) + P(E_j - 1 < a_\epsilon),$$

where $a_\epsilon = [(1-\epsilon)^{1/2} - 1]/2$.

We have

$$I_{kh} = 1 - \frac{1}{k} \sum_{i=1}^{k} \left[\Phi\left(\frac{a - X_i}{h} \right) + \Phi\left(\frac{X_i - b}{h} \right) \right],$$

and hence

$$E_j = 1 - \int_a^b \left[\Phi\left(\frac{a - x}{h_j} \right) + \Phi\left(\frac{x - b}{h_j} \right) \right] f_0(x)\, dx$$

$$= \quad 1 - h_j \int_{(a-b)/h_j}^{0} \Phi(y)\,[f_0(a - h_j y) + f_0(b + h_j y)]\; dy.$$

Using Assumption 2, it follows from the last expression that

$$E_j = 1 - h_j(f_0(a+) + f_0(b-))/\sqrt{2\pi} + h_j^2\gamma_j, \tag{7}$$

where $\gamma_1, \ldots, \gamma_M$ are all bounded by the same number. Assumption 7 then implies that, for some positive constant C_1, $|E_j - 1| \le C_1 k^{-\beta}$ for all j, and so for all k sufficiently large $P(E_j - 1 < a_\epsilon) = 0$ for $j = 1, \ldots, M$.

Theorem 1 of Hoeffding (1963) implies that for $0 < t < E_j$,

$$P(I_{kh_j} - E_j \le -t) \le e^{-2kt^2}.$$

But for all k sufficiently large $-a_\epsilon < E_j$ for $j = 1, \ldots, M$, and therefore

$$P(I_{kh_j} - E_j \le a_\epsilon) \le \exp(-2ka_\epsilon^2), \quad j = 1, \ldots, M.$$

Combining previous results and using (6) now implies that, for all k sufficiently large,

$$P(|\mathcal{I}_k - 1| > \epsilon) \le M \exp(-2ka_\epsilon^2),$$

from which Lemma 1 follows since $M = O(k)$ by Assumption 7. □

Null case. Using Assumptions 4–6, it is enough to prove that

$$\frac{1}{M}\sum_{j=1}^{M} \exp(\ell_k(h_j) - J_{kn})\pi(h_j)$$

is bounded by a quantity that converges in probability to 0 at an exponential rate. To this end, for some positive constant C_2, the last quantity is bounded by

$$C_2 \exp(\ell_k(\hat{h}) - J_{kn})$$

where $\hat{h}$ maximizes $\ell_k(h)$ for $h \in \mathcal{H}_M$.

Write

$$\ell_k(\hat{h}) - J_{kn} = \tilde{\ell}_k(\hat{h}) - J_{kn} + \ell_k(\hat{h}) - \tilde{\ell}_k(\hat{h}),$$

and consider

$$\begin{aligned}
\frac{1}{n-k}\tilde{\ell}_k(h) &= \int_a^b \log \hat{f}(x|h, \boldsymbol{X}_T) f_0(x)\, dx \\
&= \int_a^b \log f_0(x) f_0(x)\, dx + \int_a^b [\hat{f}(x|h, \boldsymbol{X}_T) - f_0(x)]\, dx \\
&\quad -\frac{1}{2}\int_a^b [\hat{f}(x|h, \boldsymbol{X}_T) - f_0(x)]^2 \frac{f_0(x)}{\tilde{f}_0(x)^2}\, dx
\end{aligned}$$

$$= \int_a^b \log f_0(x) f_0(x)\, dx - \frac{1}{2}\int_a^b [\hat{f}(x|h, \boldsymbol{X}_T) - f_0(x)]^2 \frac{f_0(x)}{\tilde{f}_0(x)^2}\, dx,$$

where $\tilde{f}_0(x)$ is between $\hat{f}(x|h, \boldsymbol{X}_T)$ and $f_0(x)$.

If D is a bound on f_0, then for all $x \in (a, b)$,

$$\tilde{f}_0(x) \le \max\left(D, \frac{K(0)}{hI_{kh}}\right) \le \max\left(D, \frac{k^\alpha K(0)}{\min_{h\in\mathcal{H}_M} I_{kh}}\right),$$

and hence

$$\frac{1}{\tilde{f}_0(x)^2} \ge \min\left(D^{-2}, \frac{k^{-2\alpha}\min_{h\in\mathcal{H}_M} I_{kh}^2}{K^2(0)}\right) = C_k.$$

Therefore,

$$\int_a^b [\hat{f}(x|\hat{h}, \boldsymbol{X}_T) - f_0(x)]^2 \frac{f_0(x)}{\tilde{f}_0(x)^2}\, dx \ge C_k ISE(\hat{h}_0), \tag{8}$$

where $ISE(h) = \int_a^b [\hat{f}(x|h, \boldsymbol{X}_T) - f_0(x)]^2 f_0(x)\, dx$ for all h, and $\hat{h}_0$ is the minimizer of $ISE(h)$ over all h in an appropriately large set of which $\mathcal{H}_M$ is a subset. It follows that

$$\begin{aligned}\frac{1}{n-k}\left(\tilde{\ell}_k(\hat{h}) - J_{kn}\right) &\le -\frac{1}{2}C_k ISE(\hat{h}_0) + \int_a^b \log f_0(x) f_0(x)\, dx - \frac{1}{n-k}J_{kn} \\ &= -\frac{1}{2}C_k ISE(\hat{h}_0) + O_p(n^{-1/2}).\end{aligned}$$

We may use results of Hall and Marron (1987) to argue that either $k^{3/4}ISE(\hat{h}_0)$ or $k^{4/5}ISE(\hat{h}_0)$ converge in probability to a positive constant, depending on how f_0 behaves at the boundary of its support. We take the more conservative approach and assume that $k^{4/5}ISE(\hat{h}_0) \xrightarrow{p} A_1 > 0$. From Lemma 1 $C_k \sim k^{-2\alpha}/K^2(0)$, in probability, as $k \to \infty$, and hence we may write

$$-\frac{1}{2}C_k ISE(\hat{h}_0) = -\frac{k^{-(2\alpha+4/5)}A_1}{2K^2(0)} + o_p(k^{-(2\alpha+4/5)}).$$

Combining previous results,

$$\begin{aligned}\exp(\ell_k(\hat{h}) - J_{kn}) &\le \exp\left(-\frac{nk^{-(2\alpha+4/5)}A_1}{2K^2(0)} + \ell_k(\hat{h}) - \tilde{\ell}_k(\hat{h})\right) \\ &\quad \times \exp\left(O_p(n^{1/2}) + o_p\left(nk^{-(2\alpha+4/5)}\right)\right).\end{aligned}$$

Since $k = n^c$, to guarantee consistency at an exponential rate, we need $1 - c(2\alpha + 4/5) > 1/2$, or $c < (1/2)(2\alpha + 4/5)^{-1}$, which holds by the condition on c in Assumption 7.

We next consider $\ell_k(\hat{h}) - \tilde{\ell}_k(\hat{h})$, which we must show is $o_p(nk^{-(2\alpha+4/5)})$. Let d be such that $1/2 < d < 1 - c(2\alpha + 4/5)$. We have

$$P(|\ell_k(\hat{h}) - \tilde{\ell}_k(\hat{h})| > n^d) \le P\left(\bigcup_{j=1}^M \left\{|\ell_k(h_j) - \tilde{\ell}_k(h_j)| > n^d\right\}\right)$$

$$\begin{aligned} &\leq \sum_{j=1}^{M} P\left(|\ell_k(h_j) - \tilde{\ell}_k(h_j)| > n^d\right) \\ &= \sum_{j=1}^{M} \int_a^b \cdots \int_a^b P\Big(|\ell_k(h_j) - \tilde{\ell}_k(h_j)| > n^d \Big| \\ &\quad \boldsymbol{X}_T = (x_1, \ldots, x_k)\Big) \times \prod_{i=1}^{k} f_0(x_i)\, dx_i. \end{aligned}$$

We have

$$P\left(\ell_k(h_j) - \tilde{\ell}_k(h_j) > n^d \Big| \boldsymbol{X}_T\right) =$$

$$P\left(\frac{1}{n-k}\sum_{i=k+1}^{n} \log \hat{f}(X_i|h_j, \boldsymbol{X}_T) - \mu_j > \frac{n^d}{n-k}\Big| \boldsymbol{X}_T\right), \tag{9}$$

where $\mu_j = \int_a^b \log \hat{f}(x|h_j, \boldsymbol{X}_T) f_0(x)\, dx$. Given $\boldsymbol{X}_T$, $\log \hat{f}(X_i|h_j, \boldsymbol{X}_T)$, $i = k+1, \ldots, n$, are i.i.d. with mean μ_j. In addition, $\log \hat{f}(X_i|h_j, \boldsymbol{X}_T)$ is bounded above and below almost surely. These properties allow us to use a result of Hoeffding (1963) to bound the probability (9).

It is easy to establish that for all $x \in (a, b)$

$$\frac{1}{h_j} K\left(\frac{b-a}{h_j}\right) \leq \hat{f}(X_i|h_j, \boldsymbol{X}_T) \leq K(0)\left[K\left(\frac{b-a}{h_j}\right)\right]^{-1},$$

and hence

$$A_j = -\log h_j + \log K\left(\frac{b-a}{h_j}\right) \leq \log \hat{f}(X_i|h_j, \boldsymbol{X}_T) \leq \log K(0) - \log K\left(\frac{b-a}{h_j}\right) = B_j.$$

We may now use Theorem 1 of Hoeffding (1963) to assert that

$$P\left(\ell_k(h_j) - \tilde{\ell}_k(h_j) > n^d \Big| \boldsymbol{X}_T\right) \leq \exp\left(-2(n-k)\left(\frac{n^d}{(n-k)(B_j - A_j)}\right)^2\right).$$

Importantly, the bound is free of $\boldsymbol{X}_T$, and hence

$$P(|\ell_k(\hat{h}) - \tilde{\ell}_k(\hat{h})| > n^d) \leq 2\sum_{j=1}^{M} \exp\left(-2n^{2d}(n-k)^{-1}(B_j - A_j)^{-2}\right).$$

Since $d > 1/2$, $n^{2d}/(n-k) \sim n^{(2d-1)} \to \infty$. Furthermore,

$$\begin{aligned} (B_j - A_j)^{-2} &= h_j^4\left[(b-a)^2 + h_j^2(\log\sqrt{2\pi} - \log(b-a) + \log h_j)\right]^{-2} \\ &\geq h_j^4 (b-a)^{-2}, \end{aligned}$$

with the last inequality holding for all n sufficiently large. Therefore,

$$\frac{n^{2d}}{n-k}(B_j - A_j)^{-2} \geq \frac{n^{2d-4c\alpha}}{n-k}(b-a)^{-2},$$

and

$$P(|\ell_k(\hat{h}) - \tilde{\ell}_k(\hat{h})| > n^d) \leq 2M \exp\left(-2n^{2d-4c\alpha}(n-k)^{-1}(b-a)^{-2}\right).$$

We need to choose c and d to satisfy $2d - 4c\alpha > 1$ and $1/2 < d < 1 - c(2\alpha + 4/5)$. These two inequalities imply that $1/2 + 2c\alpha < d < 1 - c(2\alpha + 4/5)$, which requires $1/2 + 2c\alpha < 1 - c(2\alpha + 4/5)$. The last inequality is true under the condition imposed on c in Assunption 7, thus establishing the existence of an appropriate d and completing the proof of consistency in the null case.

Alternative case. Because of Assumptions 4 and 5, the result will be established by showing that

$$\exp\left(-\ell(\boldsymbol{\theta}_0)\right) \frac{(h_M - h_1)}{M} \sum_{j=1}^{M} \exp\left(\ell_k(h_j)\right) \pi(h_j)$$

tends to ∞ at the stated exponential rate. The last quantity is at least

$$\frac{(h_M - h_1)}{M} \exp\left(\ell_k(h_1) - \ell(\boldsymbol{\theta}_0)\right) \pi(h_1).$$

By Assumptions 2 and 3 and the result of Ahmad and Lin (1976),

$$\ell_k(h_1) - \ell(\boldsymbol{\theta}_0) + \log \pi(h_1) = (n-k) \int_a^b \log f_0(x) f_0(x)\, dx - \ell(\boldsymbol{\theta}_0) - \beta/h_1^2 + o_p(n).$$

Therefore, $\ell_k(h_1) - \ell(\boldsymbol{\theta}_0) + \log \pi(h_1)$ equals

$$\begin{aligned}
&(n-k)\left[\int_a^b \log f_0(x) f_0(x)\, dx - \int_a^b \log f(x|\boldsymbol{\theta}_0) f_0(x)\, dx\right] \\
&+(n-k)\left[\int_a^b \log f(x|\boldsymbol{\theta}_0) f_0(x)\, dx - (n-k)^{-1} \sum_{i=m+1}^{n} \log f(X_i|\boldsymbol{\theta}_0)\right] \\
&+O(n^{2c\alpha}) + o_p(n),
\end{aligned}$$

which is equal to

$$(n-k) \int_a^b \log[f_0(x)/f(x|\boldsymbol{\theta}_0)] f_0(x)\, dx + o_p(n). \qquad \square$$

7 Concluding remarks

We have proposed a Bayesian goodness-of-fit test based on ordinary kernel estimators. The idea that makes this possible is that kernel estimates indexed by bandwidth comprise a parametric model for the distribution of data that *are not used in calculating the*

kernel estimates. This idea in conjunction with data splitting leads to straightforward calculation of Bayes factors that compare a specific parametric model with the kernel model. Prior specification for alternative models is straightforward since the only prior required is for the bandwidth of the kernel estimate. A particular prior for the bandwidth of a Gaussian-kernel estimator was proposed and shown by examples and simulations to perform very well. We also applied our method to a wrapped Gaussian-kernel estimator in a setting with circular data.

The ideas proposed herein for testing model fit easily extend to other settings. For example, the same method could be used to test the fit of a multivariate distribution, to test the fit of a parametric regression model, or to compare multiple curves in either k-sample goodness-of-fit or testing the equality of regression functions. A challenging but undoubtedly fascinating problem for future research would be to determine, at least asymptotically, optimal sizes for the training and validation sets.

References

Ahmad, I. and Lin, P.-E. (1976). "A nonparametric estimation of the entropy for absolutely continuous distributions." *IEEE Transactions on Information Theory*, 22: 372–375. MR0408987. 674

Alqallaf, F. and Gustafson, P. (2001). "On cross-validation of Bayes models." *Canadian Journal of Statistics*, 29: 333–340. MR1840713. doi: http://dx.doi.org/10.2307/3316081. 655

Berger, J. and Guglielmi, A. (2001). "Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives." *Journal of the American Statistical Association*, 96: 174–184. MR1952730. doi: http://dx.doi.org/10.1198/016214501750333045. 653

Berger, J. and Pericchi, L. (1996). "The intrinsic Bayes factor for model selection and prediction." *Journal of the American Statistical Association*, 91: 109–122. MR1394065. doi: http://dx.doi.org/10.2307/2291387. 656, 657

Bickel, P. and Rosenblatt, M. (1973). "On some global measures of the deviation of density function estimates." *Annals of Statistics*, 1: 1071–1095. MR0348906. 654

Ciardullo, R., Feldmeier, J. J., Jacoby, G. H., de Naray, R. K., Laychak, M. B., and Durrell, P. R. (2002). "Planetary Nebulae as Standard Candles. XII. Connecting the Population I and Population II Distance Scales." *Astrophysical Journal*, 577: 31–50. 661

Ciardullo, R., Jacoby, G. H., Ford, H. C., and Neil, J. D. (1989). "Planetary nebulae as standard candles. II. The calibration in M31 and its companions." *Astrophysical Journal*, 339: 53–69. 661

Ferguson, T. (1983). "Bayesian density estimation by mixtures of normal distributions." In: Rizvi, H. and Rustagi, J. (eds.), *Recent Advances in Statistics*, 287–302. New York: Academic Press. MR0736538. 654

Hall, P. and Marron, J. (1987). "Extent to which least-squares cross-validation minimises integrated squared error in nonparametric density estimation." *Probability Theory and Related Fields*, 74: 567–581. MR0876256. doi: http://dx.doi.org/10.1007/BF00363516. 672

Hoeffding, W. (1963). "Probability inequalities for sums of bounded random variables." *Journal of the American Statistical Association*, 58: 13–30. MR0144363. 671, 673

Jeffreys, H. (1961). *Theory of Probability*. Third edition. Clarendon Press, Oxford. MR0187257. 659, 664

Johnson, V. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests." *Journal of the Royal Statistical Society B*, 72: 143–170. MR2830762. doi: http://dx.doi.org/10.1111/j.1467-9868.2009.00730.x. 657

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. Wiley Series in Probability and Statistics. Chichester: Wiley. MR1828667. 665

McVinish, R., Rousseau, J., and Mengersen, K. (2009). "Bayesian goodness of fit testing with mixtures of triangular distributions." *Scandinavian Journal of Statistics*, 36: 337–354. MR2528988. doi: http://dx.doi.org/10.1111/j.1467-9469.2008.00620.x. 669

Müller, P. and Quintana, F. (2004). "Nonparametric Bayesian data analysis." *Statistical Science*, 19: 95–110. MR2082149. doi: http://dx.doi.org/10.1214/088342304000000017. 653

Rayner, J., Thas, O., and Best, D. (2009). *Smooth Tests of Goodness of Fit: Using R*. Wiley Series in Probability and Statistics. Chichester, UK: John Wiley & Sons, Ltd, second edition. 653

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. London: Chapman & Hall. MR0848134. doi: http://dx.doi.org/10.1007/978-1-4899-3324-9. 656, 663

Tokdar, S. and Martin, R. (2013). "Bayesian test of normality versus a Dirichlet process mixture alternative." Unpublished manuscript, arXiv:1108.2883v3 [math.ST]. 653

Tokdar, S. T., Chakrabarti, A., and Ghosh, J. K. (2010). "Bayesian nonparametric goodness of fit tests." In: Chen, M.-H., Dey, D. K., Müller, P., Sun, D., and Ye, K. (eds.), *Frontiers of statistical decision making and Bayesian analysis*. New York: Springer. In honor of James O. Berger. MR2766461. doi: http://dx.doi.org/10.1007/978-1-4419-6944-6. 653

van der Laan, M., Dudoit, S., and Keleş, S. (2004). "Asymptotic optimality of likelihood-based cross-validation." *Statistical Applications in Genetics and Molecular Biology*, 3: online publication. MR2101455. doi: http://dx.doi.org/10.2202/1544-6115.1036. 655, 665, 670

Xu, X., Lu, P., MacEachern, S., and Xu, R. (2011). "Calibrated Bayes factors for model comparison." Unpublished manuscript. 656

Acknowledgments

We greatly appreciate the AE and reviewers for their helpful comments that improved the substance of the paper. Research of Jeffrey D. Hart was supported by NSF grant DMS-0604801. Research of Taeryon Choi was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2013R1A1A2074463).