Bayes Factor Consistency

Siddhartha Chib
John M. Olin School of Business, Washington University in St. Louis
and
Todd A. Kuffner
Department of Mathematics, Washington University in St. Louis

July 1, 2016

Abstract

Good large sample performance is typically a minimum requirement of any model selection criterion. This article focuses on the consistency property of the Bayes factor, a commonly used model comparison tool, which has experienced a recent surge of attention in the literature. We thoroughly review existing results. As there exists such a wide variety of settings to be considered, e.g. parametric vs. nonparametric, nested vs. non-nested, etc., we adopt the view that a unified framework has didactic value. Using the basic marginal likelihood identity of Chib (1995), we study Bayes factor asymptotics by decomposing the natural logarithm of the ratio of marginal likelihoods into three components. These are, respectively, log ratios of likelihoods, prior densities, and posterior densities. This yields an interpretation of the log ratio of posteriors as a penalty term, and emphasizes that to understand Bayes factor consistency, the prior support conditions driving posterior consistency in each respective model under comparison should be contrasted in terms of the rates of posterior contraction they imply.

Keywords: Bayes factor; consistency; marginal likelihood; asymptotics; model selection; nonparametric Bayes; semiparametric regression.

1 Introduction

Bayes factors have long held a special place in the Bayesian inferential paradigm, being the criterion of choice in model comparison problems for such Bayesian stalwarts as Jeffreys, Good, Jaynes, and others. An excellent introduction to Bayes factors is given by Kass & Raftery (1995). The meritorious reputation of the Bayes factor derives from its relatively good performance across key inference desiderata, including interpretability, Occam's razor, and an ability to choose the best model among those under comparison in large samples. One concrete definition of the latter property is consistency.

Informally, let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the only two candidate statistical models, each specifying a set of distributions for the data and a prior distribution on this set. Assume *a priori* that the models are assigned equal odds. The posterior under model $\mathcal{M}_k$, $k = 1, 2$, is given by

$$\{\text{posterior } |\mathcal{M}_k\} = \frac{\{\text{likelihood } |\mathcal{M}_k\} \times \{\text{prior } |\mathcal{M}_k\}}{\text{normalizing constant}}. \tag{1}$$

The normalizing constant, which is simply the integral of $\{\text{likelihood } |\mathcal{M}_k\} \times \{\text{prior } |\mathcal{M}_k\}$ over the parameter space (which may be finite- or infinite-dimensional), is called the marginal likelihood under $\mathcal{M}_k$, denoted by $m(\text{data}|\mathcal{M}_k)$.

The Bayes factor for comparing model $\mathcal{M}_1$ to model $\mathcal{M}_2$ is

$$BF_{12} = \frac{m(\text{data}|\mathcal{M}_1)}{m(\text{data}|\mathcal{M}_2)}.$$

A 'large' value of BF_{12} indicates support for $\mathcal{M}_1$ relative to $\mathcal{M}_2$, and a 'small' value (> 0) indicates support for $\mathcal{M}_2$ relative to $\mathcal{M}_1$. Bayes factor consistency refers to the stochastic convergence of BF_{12}, under the true probability distribution, such that $BF_{12} \to \infty$ if $\mathcal{M}_1$ is the best model, and $BF_{12} \to 0$ if $\mathcal{M}_2$ is the best model.

The marginal likelihood is generally not analytically tractable to compute, and hence must be approximated in practice. The conventional approaches to studying the large-sample properties of the Bayes factor involve studying the large-sample properties of the marginal likelihood, or a suitable approximation, under each model, or to derive suitable bounds for these quantities.

The literature has witnessed a surge in interest regarding Bayes factors for model comparison since the development of accurate asymptotic approximations in the Bayesian setting, and the advent of Markov Chain Monte Carlo (MCMC) methods for estimating the marginal likelihood. These breakthroughs enabled practitioners to overcome analytically intractable problems and brought the Bayes factor into everyday use. Key references include Tierney & Kadane (1986), Gelfand & Smith (1990), Kass & Vaidyanathan (1992), Carlin & Chib (1995), Chib (1995), Green (1995), Verdinelli & Wasserman (1995), Meng & Wong (1996), Chen & Shao (1997), DiCiccio et al. (1997), Chib & Jeliazkov (2001), Han & Carlin (2001) and Basu & Chib (2003).

In this article, we review existing consistency results through the lens of a simple decomposition. Chib (1995) developed a convenient approach for estimating the marginal likelihood using an identity found by a rearrangement of the posterior (1). Namely,

$$m(\text{data}|\mathcal{M}_k) = \frac{\{\text{likelihood }|\mathcal{M}_k\} \times \{\text{prior }|\mathcal{M}_k\}}{\{\text{posterior }|\mathcal{M}_k\}}.$$

We use this identity to write the natural logarithm of the Bayes factor as the sum of three log ratios, each of which can be studied in terms of their large sample behavior when evaluated at the same sequence of points. Specifically,

$$\log BF_{12} = \log \frac{\{\text{likelihood }|\mathcal{M}_1\}}{\{\text{likelihood }|\mathcal{M}_2\}} + \log \frac{\{\text{prior }|\mathcal{M}_1\}}{\{\text{prior }|\mathcal{M}_2\}} - \log \frac{\{\text{posterior }|\mathcal{M}_1\}}{\{\text{posterior }|\mathcal{M}_2\}}.$$

Whatever the target of inference may be in each model, such as a finite-dimensional parameter or a density, we argue that looking at the sample-size-dependent sequence of values of a consistent estimator for the target in each model, respectively, facilitates the application of a rich literature concerning likelihood and posterior asymptotics.

We first formally define the Bayes factor and its consistency property in the most general setting, and in the sequel make the concepts in each setting more precise. Throughout, we try to avoid technical details which, while nontrivial, are not essential to follow the arguments. We also focus only on the usual Bayes factor, rather than its variants. The practical aspects of the parametric and nonparametric settings are sufficiently different that they must be discussed separately, though we may accommodate both settings in our generic description of the problem below. For the technical details, the interested reader is directed to the relevant articles as indicated throughout. Recent monographs include Ghosh & Ramamoorthi (2003), Hjort et al. (2010) and Giné & Nickl (2015).

2 Problem Setting and Basic Argument

2.1 Notation and Definitions

We borrow from notational conventions in the nonparametric literature, c.f. Ghosal et al. (2008); Hjort et al. (2010); Walker (2004a). Let $\mathbf{y}^{(n)} \equiv \mathbf{y}$ be the observed values of a sequence of n random variables $\mathbf{Y}^{(n)} = \{Y_1, \ldots, Y_n\}$ which are independent and identically distributed according to some distribution P_0 on a measurable space $(\mathcal{Y}, \mathcal{A})$, having density p_0 with respect to a suitable dominating measure μ on $(\mathcal{Y}, \mathcal{A})$. The joint distribution of $\mathbf{Y}^{(n)}$, which is the n-fold copy of P_0 denoted by P_0^n, is absolutely continuous with respect to a common measure μ^n on the sample space $\mathcal{Y}^n$. It is desired to choose the best model among two or more candidates, where 'best' intuitively means that it most closely resembles the unknown true distribution of the data generating process, P_0^n. The 'best' model may be within some class of parametric models, semiparametric models, or fully nonparametric models. The simplest case is when the 'best' model corresponds to the true model, i.e. when the truth is contained in one of the models under consideration.

For sample size n, a generic model is denoted by $M_{n,\alpha} = \{\mathcal{F}_{n,\alpha}, \Pi_{n,\alpha}, \lambda_{n,\alpha}\}$, where $\mathcal{F}$ is a set of μ-probability densities on $(\mathcal{Y}, \mathcal{A})$ equipped with a σ-algebra ensuring that the maps $(y, p) \mapsto p(y)$ are measurable. The models are indexed by $\alpha \in A_n$, where for each $n \in \mathbb{N}$, the index set A_n is countable. The prior distribution $\Pi_{n,\alpha}$ is a probability measure on $\mathcal{F}_{n,\alpha}$, and $\lambda_{n,\alpha}$ is a probability measure on A_n. The assumption that the index set is countable for every n essentially removes some technical problems which would arise if all models had zero prior probability, which would occur if there were infinitely many possible models.

We allow that the sets $\mathcal{F}_k$ and/or $\mathcal{F}_l$ may be of any general form. That is, M_k and M_l could be any combination of parametric, semiparametric or nonparametric models. For example, in a parametric model M, $\mathcal{F} = \{p_{\boldsymbol{\theta}}^n(\mathbf{y}), \boldsymbol{\theta} \in \boldsymbol{\Omega} \subset \mathbb{R}^d, \pi(\boldsymbol{\theta})\}$, where $p_{\boldsymbol{\theta}}^n(\mathbf{y})$ denotes the density of $\mathbf{y}$ with respect to μ^n under model M prescribing the set $\mathcal{F}$.

For a generic model M, the overall prior is a probability measure on the set of probability densities,

$$\Pi_n = \sum_{\alpha \in A_n} \lambda_{n,\alpha} \Pi_{n,\alpha}. \tag{2}$$

The posterior distribution of a model index $B \subset A_n$, given this prior distribution, is the random measure

$$\begin{aligned}\Pi_n(B|y_1, \ldots, y_n) &= \frac{\int_B \prod_{i=1}^n p(y_i) d\Pi_n(p)}{\int \prod_{i=1}^n p(y_i) d\Pi_n(p)} \\ &= \frac{\sum_{\alpha \in A_n} \lambda_{n,\alpha} \int_{p \in \mathcal{P}_{n,\alpha}: p \in B} \prod_{i=1}^n p(y_i) d\Pi_{n,\alpha}(p)}{\sum_{\alpha \in A_n} \lambda_{n,\alpha} \int_{p \in \mathcal{P}_{n,\alpha}} \prod_{i=1}^n p(y_i) d\Pi_{n,\alpha}(p)},\end{aligned}$$

and the marginal likelihood is defined as

$$m(\mathbf{y}|M) = \frac{\sum_{\alpha \in A_n} \lambda_{n,\alpha} \int_{p \in \mathcal{P}_{n,\alpha}: p \in B} \prod_{i=1}^n p(y_i) d\Pi_{n,\alpha}(p)}{\Pi_n(B|y_1, \ldots, y_n)}. \tag{3}$$

For clarity, note that the given, fixed true density p_0 corresponds to a 'best' index element, say $\beta_n \in A_n$, in the sense that β_n refers to the model which is closest to the true model according to a chosen measure. In the above, the set $B \subset A_n$ of indices could simply be the single element β_n, or a collection of models.

Consider two candidate models M_k and M_l, prescribing sets of density functions $\mathcal{F}_k$ and $\mathcal{F}_l$, with associated prior distributions on these sets of density functions Π_k and Π_l, and let $m(\mathbf{y}|M_k)$ and $m(\mathbf{y}|M_l)$ denote the respective marginal likelihoods. The Bayes factor is defined by

$$BF_{kl} = \frac{m(\mathbf{y}|M_k)}{m(\mathbf{y}|M_l)} = \frac{\lambda_{n,k} \int \prod_{i=1}^n p(y_i) \Pi_{n,k}(p)}{\lambda_{n,l} \int \prod_{i=1}^n p(y_i) \Pi_{n,l}(p)}. \tag{4}$$

The prior probabilities on the models, $\lambda_{n,\alpha}$, do not affect the consistency arguments, and thus there is no loss of generality in assuming $\lambda_{n,k} = \lambda_{n,l}$, so that these quantities may be ignored in what follows.

Consistency of the Bayes factor refers to stochastic convergence of the quantity (4).

Definition 1 (Conventional Bayes Factor Consistency)**.** The Bayes factor for comparing M_k and M_l, $BF_{kl} = m(\mathbf{y}|M_k)/m(\mathbf{y}|M_l)$, is consistent if:

(i) $BF_{kl} \to_p 0$ (or $\log BF_{kl} \to_p -\infty$) when M_l contains the true model ($p_0^n \in \mathcal{F}_{n,l}$); and

(ii) $BF_{kl} \to_p \infty$ (or $\log BF_{kl} \to_p \infty$) when M_k contains the true model ($p_0^n \in \mathcal{F}_{n,k}$).

The probability measure associated with these convergence results is the one associated with the infinite-dimensional product measure corresponding to the true distribution, the n-fold product measure P_0^n as $n \to \infty$. When both relations hold with probability one, the

Bayes factor is said to be almost surely consistent. In all stochastic convergence statements in this paper, we are considering $n \to \infty$. We further note that such convergence statements are pointwise and not uniform.

It will often be convenient to work with the natural logarithm of the Bayes factor, referred to by I. J. Good as the *weight of evidence*. As noted in the definition, and pointed out by Dutta et al. (2012) and Chakrabarti & Ghosh (2011), we must be careful about which probability measure is associated with the stochastic convergence statement. Moreover, consistency of the Bayes factor in the sense of Definition 1 requires that one of the models being considered contains the true model. Definition 3 accommodates the more general setting.

Bayes factor consistency is not the same as model selection consistency. The reason this is often referred to as model selection consistency is that this ensures the sequence of posterior probabilities of the true model will converge to one, at least in fixed-dimensional settings; see Liang et al. (2008), Casella et al. (2009) and Shang & Clayton (2011). This is what is more conventionally thought of as model selection consistency (Fernández et al., 2001). It has been emphasized (Moreno et al., 2015) that the Bayesian and frequentist notions of model selection consistency do not necessarily agree; c.f. Shao (1997) where model selection consistency means convergence in probability of the selected model to the submodel which minimizes the mean squared prediction error.

Note that if a prior is improper, then the marginal likelihood is also improper. In that case, the ratio (4) cannot be interpreted. We consider here only proper priors for which the marginal likelihoods and Bayes factors are well-defined. There is a growing literature concerning alternative measures of evidence which allow for improper priors (Berger & Pericchi, 2001), e.g. posterior Bayes factors (Aitkin, 1991), fractional Bayes factors (O'Hagan, 1995) and intrinsic Bayes factors (Berger & Pericchi, 1996). Other notable proposals for measuring evidence, as alternatives to the usual Bayes factor considered here, include posterior likelihood ratios (Dempster, 1973; Smith & Ferrari, 2014), test martingales (Shafer et al., 2011) and relative belief ratios (Evans, 2015). A recent overview of objective Bayesian approaches can be found in Bayarri et al. (2012).

2.2 A Unified Framework for Analysis

In this general setup, following the basic marginal likelihood identity (BMI) of Chib (1995), we have that

$$m(\mathbf{y}|M_{n,\alpha}) = \frac{p^n(\mathbf{y}|\mathcal{F}_{n,\alpha})\Pi_{n,\alpha}(\mathcal{F}_{n,\alpha})}{\Pi_{n,\alpha}(\mathcal{F}_{n,\alpha}|\mathbf{y})}. \tag{5}$$

Using this identity, the natural logarithm of the Bayes factor for comparing M_k and M_l may be expressed as

$$\log BF_{kl} = \log \frac{p^n(\mathbf{y}|\mathcal{F}_{n,k})}{p^n(\mathbf{y}|\mathcal{F}_{n,l})} + \log \frac{\Pi_{n,k}(\mathcal{F}_{n,k})}{\Pi_{n,l}(\mathcal{F}_{n,l})} - \log \frac{\Pi_{n,k}(\mathcal{F}_{n,k}|\mathbf{y})}{\Pi_{n,l}(\mathcal{F}_{n,l}|\mathbf{y})}. \tag{6}$$

To establish stochastic convergence as $n \to \infty$ under probability law P_0^n, when $p_0^n \in \mathcal{F}_{n,k}$ or $p_0^n \in \mathcal{F}_{n,l}$, we examine each of the three terms appearing on the r.h.s. of (6) separately. We emphasize that no single technique will work across all model comparison problems; the tools needed for analysis of each term in this decomposition will depend on the nature of the model comparison problem.

The first term is a type of log likelihood ratio. It is convenient for this term to be bounded in P_0^n-probability and, hence, to be $O_p(1)$. This happens, for example, if this quantity converges in distribution. The arsenal of tools for this term includes everything from classical likelihood theory to generalized likelihood ratio tests (Fan et al., 2001) and other variants for the semiparametric setting. However, we emphasize that the goal is not simply to make assumptions which ensure convergence in distribution, but to also understand settings for which this would fail to happen, as such failure will impact Bayes factor consistency.

The second term is desired to be $O(1)$, as we want the prior in each model to have a negligible effect on the posterior for large samples. If the priors are continuous and bounded, as they will be when proper, this term will be bounded. This term can be problematic in some cases, however, including when the dimension of the parameter space grows with the sample size.

The final term involves a log ratio of posteriors. Consistency requires that this converge in P_0^n-probability to $-\infty$ if $p_0 \in \mathcal{F}_{n,l}$ and to $\to \infty$ in P_0^n-probability if $p_0^n \in \mathcal{F}_{n,k}$. Essentially this just means that the correct model has a faster rate of posterior contraction, as a function of the available sample size, when compared to an incorrect or more complex alternative. This can often be established by existing theorems, provided one is willing to make suitable assumptions on the concentration of the priors, and the size of the model space. We review such conditions below.

An attractive feature of this framework is that the third term acts like a penalty term, and it is this term which typically drives the consistency result. This observation leads to interesting questions about when the penalty is sufficiently large to prevent the selection of an incorrect model, and to prevent overfitting. Intuitively, if two models are close in some sense but only one of them is correct, consistency requires some condition ensuring that the posterior in the correct model contracts faster, and that it is possible to distinguish the models asymptotically. The latter condition can be roughly thought of as saying the frequentist notion of the 'null' and 'alternative' hypotheses are well-separated in an appropriate sense. This is related to the existence of uniformly exponentially consistent tests. Moreover, if the models are nested, and both contain the truth, the penalty term must be sufficiently large to ensure the smaller model is chosen. In the classical nested parametric setting, the larger model, by virtue of having a higher-dimensional parameter space, will have prior mass that is more spread out. Thus if the true density is contained in a smaller model, with a more concentrated prior mass around the true density, the posterior in the smaller model will contract faster than the more complex model. In more general, non-nested settings with nonparametric models, a similar intuition will hold.

If the true P_0 is not contained in the set of candidate models, it will certainly not be possible for any model selection procedure to select the correct model. Many authors argue that it is unrealistic to assume the true distribution of the data generating process is exactly specified by one of the candidate models, and instead use the qualifier 'pseudo-true' to mean the 'best' approximating model among those being considered. While we sympathize with such authors and use this convention when appropriate, we often make no distinction in what follows between true and pseudo-true models (and parameter values). This issue is most relevant when both models are misspecified and not well-separated, so that it is difficult to construct a consistent sequence of tests. Moreover, while consistency in the usual sense is not possible when none of the candidate models contain the true distribution, it will still be possible to define some notion of consistency in the sense of choosing the model which is closest to the true distribution in a relative entropy sense; c.f. Definition 3.

2.3 Advantages of This Approach

Since (5) is an identity, then it holds for any density in $\mathcal{F}$, the set of densities being considered. It will also hold for every sequence of estimates of the density (or parameter, in the parametric setting). Thus, to establish consistency of the Bayes factor, it is sufficient to show convergence in P_0^n-probability of the Bayes factor for any sequence of estimates which are the values of consistent estimators for the unknown densities or parameters in the respective models being compared.

A second advantage is that we examine the asymptotic behavior of each component, rather than the aggregate asymptotic behavior of the marginal likelihood, and in this sense we learn more about the asymptotic behavior of important quantities. We are able to exploit a richer literature to study likelihood ratios and posterior contraction rates separately, and this can help identify important unanswered questions or problems for which there is scope to improve existing results. Compared to existing proofs, which rely on results which relate the prior support conditions to bounds on the marginal likelihood, we are directly considering how the prior support conditions affect the behavior of the posterior in each model, which is conceptually appealing.

2.4 Model Comparison Types

When comparing two models, it is possible that both, only one or neither of the models contains the true distribution. When a model contains the true distribution, it is said to be correctly specified; otherwise it is said to be misspecified. We can further classify the types of model comparison as: (a) both models are parametric (§ 4); (b) one model is parametric, the other is nonparametric (§ 5); (c) one model is parametric, the other is semiparametric (§ 6); (d) both models are nonparametric (§ 7); (e) both models are semiparametric (§ 7); (f) one model is nonparametric, the other is semiparametric (§ 7).

Thus there are six frameworks of model comparison and, within any of these, one could consider misspecified models, non i.i.d. observations, or other 'non-regular' settings. Much of the Bayes factor literature makes distinctions between the situations where the models are nested, overlapping or non-nested. These classifications may be roughly understood in the parametric setting as follows. Model 1 is nested in Model 2 if it is possible to obtain Model 1 by some restriction on the parameters of Model 2. This is the case most commonly studied in classical frequentist hypothesis testing for linear models; the relationship between the parameter spaces and likelihoods facilitates the derivation of the asymptotic distribution of the likelihood ratio statistic under the null model. Non-nested models can be either strictly non-nested or overlapping. Overlapping models are those for which neither model is nested in the other, but that there still exists some choice of parameter values for each model such that they yield the same joint distribution for the data. Strictly non-nested models do not allow for this latter possibility. In the nonparametric setting, borrowing from the literature on adaptation, the distinction is in terms of coarser and smoother models. More recently, Ghosal et al. (2000) and Ghosal et al. (2008) have adopted the terminology of bigger and smaller models, with a precise definition of model complexity given in Ghosal et al. (2000). We also note that, while the classical large sample evaluation of Bayes factors considers the dimension of the model to be fixed, there is a growing literature concerning the asymptotic behavior of Bayes factors as the model dimension grows with the sample size. We reference this literature throughout.

We focus on the 'regular' versions of each of the above frameworks, with comments and references in appropriate places regarding extensions to common alternative settings. While we have tried to incorporate as many of the recent developments as possible, the size of the Bayes factors literature renders any attempt at being comprehensive as an exercise in futility, and our omission of certain elements should not be interpreted as a judgment that these contributions are less important.

We caution the reader that in any particular setting, there will be model-specific assumptions required to ensure that the model comparison problem is well-defined. For example, in the regression settings we mention below, there would be some additional assumptions on the matrix of predictor variables, the error distribution and the joint distribution of the predictors and errors to ensure the consistency of a suitable estimator for the unknown mean regression function or its parameters, and hence also for a consistent estimator of the density of interest. These assumptions are crucial, but we do not dwell on them here. Instead, we focus only on the aspects of the problem which are unique to the study of Bayes factor asymptotics.

3 Frameworks, Concepts and Connections

We review some key definitions and concepts needed for large-sample model comparison across the parametric, semiparametric and nonparametric frameworks. Posterior consistency and rates of contraction are also considered. A running example of mean regression modeling is introduced.

3.1 Parametric Framework

Let $\mathbf{Y}^{(n)} = \{Y_1, \ldots, Y_n\}$ be a sequence of random variables from a population with some density indexed by parameter vector $\boldsymbol{\theta} \in \Omega$; the Y_i are assumed to be conditionally independent and identically distributed, given $\boldsymbol{\theta}$. The parameter sets considered are subsets of $\mathbb{R}^d$, where d is finite and fixed, in particular, $d < n$ whenever estimation is considered for fixed n. A model, M_k, $k = 1, \ldots, K$ (where $K \in \mathbb{N}$, hence a countably infinite set) consists of a parameter space Ω_k, a density $f_k(\mathbf{y}|\boldsymbol{\theta}_k \in \Omega_k, M_k)$ and a prior density $\pi_k(\boldsymbol{\theta}_k|M_k)$ for $\boldsymbol{\theta}_k \in \Omega_k$. The vector of observed values $\mathbf{y}^{(n)} = (y_1, \ldots, y_n)$ are the realizations of $\mathbf{Y}^{(n)}$, though we suppress the superscript on $\mathbf{y}^{(n)} \equiv \mathbf{y}$ for notational convenience. The model space we will consider is $\mathcal{M} = \cup_{k=1}^{K} M_k$, a countable union of models.

A Bayesian constructs a prior distribution Π, which expresses beliefs about the parameter. Combining the prior with the observations yields the posterior distribution

$$\Pi_n(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})}{\int_{\boldsymbol{\Omega}} \pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})}{m(\mathbf{y})} \tag{7}$$

with the prior $\pi(\boldsymbol{\theta})$ representing the density corresponding to the prior probability measure Π. The marginal likelihood under model M_k is

$$m(\mathbf{y}|M_k) = \int f(\mathbf{y}|\boldsymbol{\theta}_k, M_k)\pi(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k. \tag{8}$$

Example 1 (Linear Regression)**.** *Let* $\{Y_i\}_{i=1}^n = \{(Z_i, X_{ij})\}_{i=1}^n$, $j = 1, \ldots, p$, *so that* $\mathbf{Y}$ *is a matrix of response-covariate pairs, where* $\mathbf{Z} = \{Z_1, \ldots, Z_n\}$ *is an* n*-dimensional vector of* *response variables and $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix. It is assumed that*

$$Z_i = X_{ij}\boldsymbol{\beta} + \varepsilon_i \tag{9}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter vector. We typically assume that the ϵ_i are i.i.d. $\mathcal{N}(0, \sigma^2)$ with $\sigma > 0$, and thus $\mathbf{Z}|\mathbf{X} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n)$. It is typically assumed that an intercept term is included, so that the first column of $\mathbf{X}$ is a vector of 1s. It is also conventional for illustrative purposes to assume that the X_{ij} are fixed. A Bayesian model specifies a prior for $\boldsymbol{\beta} \sim \Pi_1$, independent of σ^2, and could either treat σ^2 as known or specify a prior $\sigma^2 \sim \Pi_2$.

Complete specification of the Bayesian model comparison procedure would also require a prior distribution over the space of models; this would be necessary to calculate the posterior odds from the Bayes factor. This will not be necessary for our purposes. The priors for the parameter are assumed to be proper; use of improper priors in conjunction with Bayes factors for model selection can be problematic (e.g. Lindley's paradox). We refer readers to Robert (1993) and Villa & Walker (2015) for discussion of Lindley's paradox. Ghosh et al. (2005) point out that the paradox disappears if the Bayesian and frequentist asymptotic frameworks are the same, i.e. both using Bahadur asymptotics (fixed alternatives) or Pitman (contiguous) alternatives. When improper priors are used, such as those arising from objective Bayesian analysis, there have been several proposals to alter the standard model comparison framework. Dawid (2011), suggested that using the posterior odds instead of the Bayes factor can overcome the problem, while O'Hagan (1995), Dawid & Musio (2015) and others have proposed variants such as fractional or intrinsic Bayes factors.

3.2 Nonparametric Framework

Let $\mathbf{Y}^{(n)} = \{Y_1, \ldots, Y_n\}$ be a sequence of random variables, which are assumed to be independent and identically distributed according to a true distribution F_0, having true density f_0 with respect to Lebesgue measure. In the parametric setting, estimation is concerned with the parameter $\boldsymbol{\theta}$ of an (assumed) known distribution, while in the nonparametric setting, estimation is concerned with a density function. A Bayesian constructs a prior distribution Π on the set of densities $\mathcal{F}$ prescribed by the model, which expresses beliefs about the location of the true density. The posterior distribution of a set of densities A in the set Ω of all densities with respect to Lebesgue measure is given by

$$\Pi_n(A) = \frac{\int_A f(\mathbf{y})\Pi(df)}{\int_\Omega f(\mathbf{y})\Pi(df)} = \frac{\int_A f(\mathbf{y})\Pi(df)}{m(\mathbf{y})}, \tag{10}$$

where $m(\mathbf{y})$ again denotes the marginal likelihood.

Given two densities f and g, which are both absolutely continuous with respect to the same dominating measure μ over a set $\mathcal{S}$, define the Kullback-Leibler divergence of g from f as $d_{KL}(f,g) \equiv d_{KL}(g\|f) = \int_{\mathcal{S}} g \log(g/f) d\mu$. Conventionally the 'best' density would be g, so that one is speaking of the divergence of some proposed distribution f from the 'best' density. The simplest case is when $g \equiv p_0$. A Kullback-Leibler neighborhood of the density g, of size $\epsilon > 0$, is defined by

$$N_g(\epsilon) = \left\{ f : \int g(x) \log \frac{g(x)}{f(x)} dx < \epsilon \right\}. \tag{11}$$

Definition 2 (Kullback-Leibler Property)**.** A prior distribution Π over the space of densities is said to possess the Kullback-Leibler property if

$$\Pi\{f : d_{KL}(f,g) < \epsilon\} > 0 \tag{12}$$

for all $\epsilon > 0$ and for all g.

Expositions of the Kullback-Leibler property, with examples, are found in Wu & Ghosal (2008), Walker et al. (2004), Petrone & Wasserman (2002), Ghosal et al. (1999a), and Barron et al. (1999).

Each model M_α under consideration contains an element f_α which is such that

$$f_\alpha := \arg\min_{f \in \mathcal{F}_\alpha} d_{KL}(f, p_0).$$

The element f_α is the closest member of model M_α to the true density p_0. If the model contains the truth, then $d_{KL}(f_\alpha, p_0) = 0$. Otherwise the 'best' density in M_α is simply defined as the closest density to the true density in the Kullback-Leibler sense. Consider two models M_k and M_l. Each model contains a density, respectively f_k and f_l, defined in the sense above.

Definition 3 (Pragmatic Bayes Factor Consistency (Walker et al., 2004))**.** A pragmatic version of Bayes factor consistency is said to hold if

(i) $BF_{kl} \to 0$ almost surely (or $\log BF_{kl} \to -\infty$ a.s.) when $d_{KL}(f_l, p_0) < d_{KL}(f_\alpha, p_0)$ for all $\alpha \neq l$; and

(ii) $BF_{kl} \to \infty$ almost surely (or $\log BF_{kl} \to \infty$ a.s.) when $d_{KL}(f_k, p_0) < d_{KL}(f_\alpha, p_0)$ for all $\alpha \neq k$.

Again, the probability measure associated with the stochastic convergence statement is the infinite product measure P_0^∞. This says that the Bayes factor will asymptotically choose the model containing the density which is closest, in the Kullback-Leibler sense, to the true model, among all possible densities contained in the models under comparison.

A fundamentally important result related to Definition 3 is given in Theorem 1 of Walker et al. (2004), which concerns the general setting where it is possible that neither model contains the true density. Formally, Bayesian models are characterized by a prior Π, and associated with a value $\delta_\Pi \geq 0$, which is such that $\Pi\{f : d_{KL}(f, p_0) < d\} > 0$ only and for all $d > \delta_\Pi$. Let $\delta_\alpha \geq 0$ be the value associated with Π_α. Walker et al. (2004) show that $BF_{KL} \to \infty$ almost surely if and only if $\delta_k < \delta_l$. Moreover, if one model has the Kullback-Leibler property while the other does not, then the Bayes factor will asymptotically prefer the model possessing the Kullback-Leibler property. This was also shown in Dass & Lee (2004) for the setting of a testing a point null hypothesis against a non-parametric alternative with a prior possessing the Kullback-Leibler property. The proof in the latter paper, however, does not generalize to general null models; c.f. Ghosal et al. (2008). A related result due to Chib et al. (2016) is that the less misspecified model, in the Kullback-Leibler sense, will asymptotically yield a larger marginal likelihood with probability tending to one. This is a result about model selection consistency when using the relative (pairwise) marginal likelihood as a selection criterion.

When two or more models under consideration have the Kullback-Leibler property, stronger conditions are needed to ensure Bayes factor consistency. These are explicitly discussed in Ghosal et al. (2008) and McVinish et al. (2009). We discuss these conditions in § 5.

Example 2 (Nonparametric Mean Regression)**.** *A nonparametric regression model specifies*

$$Z_i = r(X_{ij}) + \varepsilon_i, \tag{13}$$

where $r(\cdot)$ is an unknown function, and the ε_i are i.i.d. with mean zero. Again, for illustrative purposes it is typically assumed that the X_{ij} are fixed.

3.3 Semiparametric Framework

In the semiparametric framework, it is assumed that a model has both parametric and nonparametric components. Then the semiparametric prior has two ingredients, Π_{Par} for the parametric part and Π_{NP} for the nonparametric part. The overall prior is given by $\Pi_{\text{Par}} \times \Pi_{\text{NP}}$.

Within the mean regression modeling setting, there are two common manifestations of semiparametric models. The first is a partially linear model, and the second is found in Kundu & Dunson (2014).

Example 3 (Partially Linear Model with Known Error Distribution)**.**

$$Z_i = X_{ij}\boldsymbol{\beta} + r(X_{ij}) + \varepsilon_i, \tag{14}$$

where $r(\cdot)$ is an unknown function in an infinite-dimensional parameter space, and $X_{ij}\boldsymbol{\beta}$ is a linear component.

A semiparametric prior is $\Pi = \Pi_{\boldsymbol{\beta}} \times \Pi_r$. See Example 5 in § 6 for more details. A simple version of the above example would be to modify Example 2. Assume $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ given σ, that $r(\cdot) \sim \Pi_1$ independent of σ and $\boldsymbol{\varepsilon}$, and that $\sigma \sim \Pi_2$ where Π_1 and Π_2 are prior distributions. Then, formally, this would be a semiparametric model rather than fully nonparametric.

Example 4 (Linear Model with Unknown Error Distribution)**.**

$$Z_i = X_{ij}\boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim q(\cdot) \tag{15}$$

where $q(\cdot)$ is an unknown residual density, about which parametric assumptions are avoided.

3.4 Posterior Consistency

The random measure $\Pi_n(\cdot|\mathbf{y})$, which is the posterior distribution, is said to be consistent for some fixed measure P_0 if it concentrates in arbitrarily small neighborhoods of P_0, either with probability tending to 1 or almost surely, as $n \to \infty$. Posterior consistency implies consistency of Bayesian estimates in the frequentist sense. See Barron (1986) for more on the definitions of weak, strong and intermediate forms of consistency. Parametric Bayesian asymptotics are reviewed in Ghosh & Ramamoorthi (2003, Ch. 1). More recent contributions in Bayesian nonparametric asymptotics, where one may find other important references, include Castillo (2014) and Giné & Nickl (2015). We review here some essential results.

Schwartz (1965) established weak consistency of the posterior distribution under the condition that the prior has the Kullback-Leibler property.

Definition 4 (Weak Consistency)**.** Let F_0 and F be the cumulative distribution functions for the true density p_0 and a generic density f. Take any metric w on the cumulative distribution functions F_0 and F for which convergence in w is equivalent to convergence in distribution. Define a weak neighborhood of p_0 to be $W = \{f : w(F_0, F) < \delta\}$ for $\delta > 0$. A posterior distribution $\Pi_n(\cdot|\mathbf{y})$ is said to be weakly consistent if for almost all sequences under p_0, $\Pi_n(W|\mathbf{y}) \to 1$ for all weak neighborhoods of p_0.

After Doob (1949) pioneered the notion of posterior consistency in the weak topology on the space of densities, as in the definition above, Schwartz (1965) established the Kullback-Leibler property as an important criterion for demonstrating weak consistency. However, Diaconis & Freedman (1986a,b) have shown that priors satisfying this property in weak neighborhoods may not yield weakly consistent posteriors. Freedman (1963) and Kim & Lee (2001) have also given examples of inconsistency. After the Diaconis-Freedman critique and insightful analysis by Andrew Barron, among others, the focus in the literature then shifted to establishing sufficient conditions for strong consistency, specifically Hellinger consistency; see Wasserman (1998), Ghosal et al. (2000), Shen & Wasserman (2001), Walker & Hjort (2001), Walker (2003) and Walker (2004a,b). Another commonly used distance is the L_1 distance. By Hellinger consistency, we mean convergence according to the Hellinger metric on the set of densities. This metric induces the Hellinger topology on the set of densities.

Definition 5 (Hellinger Consistency)**.** For any pair of probability measures P and Q on a measurable space $(\mathcal{Y}, \mathcal{A})$ with corresponding densities p and q with respect to a dominating measure μ, the Hellinger distance is defined as

$$d_H(p,q) = \left\{ \int \left(\sqrt{p} - \sqrt{q}\right)^2 d\mu \right\}^{1/2}. \tag{16}$$

Furthermore, define a Hellinger neighborhood of the true density to be $S_\delta = \{f : d_H(p_0, f) < \delta\}$. A sequence of posteriors $\{\Pi_n(\cdot|\mathbf{y})\}$ is Hellinger consistent for p_0, often called strongly consistent, if $\Pi_n(S_\delta|\mathbf{y}) \to 1$ almost surely in P_0^∞ probability as $n \to \infty$.

Posterior consistency in nonparametric and semiparametric problems is still an active research area, though some important questions have been answered. Some relatively recent and important contributions include Barron et al. (1999), Walker (2004b), Shen & Wasserman (2001). Particularly lucid overviews of nonparametric asymptotics are given by Ghosal (2010), Martin & Hong (2012) and Giné & Nickl (2015, §7.4). Pseudo-posterior consistency was considered in Walker & Hjort (2001).

In the semiparametric setting, letting $\boldsymbol{\theta}$ and q respectively denote the finite-dimensional parametric and inifinite-dimensional nonparametric components, posterior consistency can refer to (i) consistency at the pair $(\boldsymbol{\theta}_0, q_0)$; (ii) consistency for the marginal posterior of the parametric component $\boldsymbol{\theta}_0$, after marginalizing out the nonparametric component q; or (iii) consistency for the marginal posterior of the nonparametric component q_0 after marginalizing out the parametric component $\boldsymbol{\theta}_0$. It is often the case that the nonparametric component is not of interest for inference, and is treated as a nuisance parameter. The remarkable results of Cheng & Kosorok (2008) on the posterior profile distribution are of interest in that case.

Barron (1986), Ghosal et al. (1999b) and Amewou-Atisso et al. (2003) noted the importance of the Kullback-Leibler property in establishing consistency for the marginal posterior of the parametric component in semiparametric regression models. There have been many recent developments in Bayesian asymptotics for semiparametric models, a representative sample of which are summarized in Castillo & Rousseau (2015), Rousseau et al. (2014) and Chib et al. (2016).

3.5 Rate of Posterior Contraction

An implication of posterior consistency is the existence of a *rate* of consistency, i.e. a sequence ϵ_n indexed by the sample size, corresponding to the size of a shrinking ball around the true density whose posterior probability tends to 1 as $n \to \infty$. We use the terms rate of *contraction*, rate of *convergence*, and rate of *consistency* interchangeably. It is especially important in the study of Bayes factor asymptotics that one can establish the rate of convergence of the posterior distribution at the true value. The convergence rate is the size ϵ_n of the smallest ball, centered at the true value, such that the posterior probability of this ball tends to one.

Definition 6 (Ghosal et al. (2008))**.** Suppose $\mathbf{Y}^{(n)}$ are an i.i.d. random sample from p_0, d is a distance on the set of densities. The convergence rate of the posterior distribution $\Pi_n(\cdot|\mathbf{Y}^{(n)})$ at the true density p_0 is at least ϵ_n, if as $n \to \infty$, $\epsilon_n \to 0$, and for every sufficiently large constant M,

$$\Pi_n(f : d(f, p_0) > M\epsilon_n|\mathbf{Y}^{(n)}) \to 0$$

in P_0^n-probability.

The sequence $\epsilon_n \to 0$ is often referred to as the targeted rate, where the target is the known optimal rate for the estimation problem at hand. Note that this definition utilizes the notion that complements of neighborhoods of the true density have posterior probability tending to zero.

We point out that, in fact, it is the *higher-order asymptotic* properties of Bayesian methods which play a key role in Bayes factor consistency. That is to say, if we consider consistency to be a first-order property, then the study of rates of consistency can be viewed as a first step towards Bayesian nonparametric higher-order asymptotics. Such analysis will undoubtedly be very different from Bayesian parametric higher-order asymptotics, such as the posterior expansions reviewed in Ghosh (1994).

The prior will completely determine the *attainable* rate of posterior contraction, which can be understood intuitively by imagining how quickly the posterior will contract as the thickness of the prior tails is increased or decreased (Martin & Walker, 2016). However, the *optimal* rate depends on the smoothness of the underlying true density, which is in general not known. In the regression setting, if the smoothness of the mean regression function is known, the prior can be suitably adjusted so that the attainable and optimal contraction rates coincide. When the smoothness is unknown, the prior should be more flexible so that it can adapt to the unknown optimal contraction rate. Of primary importance for Bayes factor asymptotics are the *relative attainable* rates of posterior contraction in each model under comparison.

The core ideas regarding posterior convergence rates in the general (not necessarily parametric) setting were developed in Ghosal et al. (2000) and Shen & Wasserman (2001). We also mention Castillo (2014); Ghosal & van der Vaart (2007); Ghosal et al. (2000); Giné & Nickl (2011) and Hoffmann et al. (2015). Informative and didactic treatments are found in Giné & Nickl (2015, §7.3), Ghosal (2010, §2.5), Ghosal et al. (2008) and Walker et al. (2007). Recent extensions include misspecified models (Kleijn & van der Vaart, 2006; Lian, 2009; Shalizi, 2009), non-Euclidean sample spaces (Bhattacharya & Dunson, 2010), and conditional distribution estimation (Pati et al., 2013). Pseudo-posterior convergence rates have recently been considered in Martin et al. (2013).

3.6 Connections with Other Literature

Arguably the most closely-related strand of literature to Bayes factor asymptotics is that concerning Bayesian adaptation. Foundational work in this area is due to Belitser & Ghosal (2003) and Huang (2004). We also mention Ghosal et al. (2003), Scricciolo (2006), Lember & van der Vaart (2007) and van der Vaart & van Zanten (2009). The convergence of the ratio of posteriors, under the probability law of the true model, will turn out to be determined by two factors which are well-studied in the Bayesian adaptation literature. First, the smoothness of the true density determines the minimax rate of convergence that any estimator can achieve. Second, the smoothness properties of the models under consideration will determine how closely their respective posteriors track the optimal minimax convergence rate. For recent progress on, respectively, adaptive minimax density estimation and Bayesian adaptation, see Goldenshluger & Lepski (2014) and Scricciolo (2015).

Parallel literatures concerning goodness-of-fit tests, prequential analysis, code length, and proper scoring rules contain many ideas relevant to Bayes factor asymptotics. There has been some cross-fertilization with the Bayes factor consistency literature. The interested reader is referred to Tokdar et al. (2010), Dawid (1992), Grünwald (2007) and Dawid & Musio (2015), respectively, for key ideas and references. Many of the asymptotic arguments presented in these parallel literatures are similar to those found in the statistics literature, though there is perhaps more emphasis in the former on information theory. Clarke & Barron (1990) and Barron (1998) are great resources for Bayesian asymptotics and the relevant information theory. Those authors derive the distribution of the relative entropy (Kullback-Leibler divergence), and in the process find bounds for the ratio of marginal likelihoods. They also give useful decomposition identities involving the Kullback-Leibler divergence and ratio of marginal likelihoods. Zhang (2006) further demonstrates how ideas from information theory can simplify existing results on posterior consistency.

4 Comparing Two Parametric Models

The study of Bayes factor consistency in the parametric setting is more appropriately dealt with using different tools, rather than folding it into the nonparametric setting utilizing the Kullback-Leibler property of the priors. As noted by Walker et al. (2004), a finite-dimensional parametric model will not possess the Kullback-Leibler property unless p_0 is known to belong to the assumed parametric family. Moreover, if the parameter space Θ_l is nested in Θ_k, then if $p_0^n \in \mathcal{M}_l$, the Kullback-Leibler property holds for both prior distributions prescribed by $\mathcal{M}_k$ and $\mathcal{M}_l$; c.f. Rousseau & Choi (2012).

Consider the Bayes factor,$BF_{kl} = m(\mathbf{y}|M_k)/m(\mathbf{y}|M_l)$, for comparing any two models $\mathcal{M}_k$ and $\mathcal{M}_l$. From the *basic marginal likelihood identity* (BMI) of Chib (1995), and momentarily suppressing subscripts, we have for each particular model that

$$m(\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|\mathbf{y})}. \tag{17}$$

This equation is an identity because the l.h.s. does not depend on $\boldsymbol{\theta}$. Since this holds for any $\boldsymbol{\theta}$, if we consider any two particular sequences of $\boldsymbol{\theta}$ values indexed by the sample size n, the resulting sequences of marginal likelihoods will be the same. This simplification will allow us to focus on the maximum likelihood estimator for simplicity. Taking logs of both sides of this identity, we have

$$\log m(\mathbf{y}) = \log f(\mathbf{y}|\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta}|\mathbf{y}), \tag{18}$$

which yields

$$BF_{kl} = \frac{f_k(\mathbf{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k)}{\pi(\boldsymbol{\theta}_k|\mathbf{y})} \Big/ \frac{f_l(\mathbf{y}|\boldsymbol{\theta}_l)\pi_l(\boldsymbol{\theta}_l)}{\pi(\boldsymbol{\theta}_l|\mathbf{y})} = \exp\{\log \frac{f_k(\mathbf{y}|\boldsymbol{\theta}_k)}{f_l(\mathbf{y}|\boldsymbol{\theta}_l)} + \log \frac{\pi_k(\boldsymbol{\theta}_k)}{\pi_l(\boldsymbol{\theta}_l)} - \log \frac{\pi(\boldsymbol{\theta}_k|\mathbf{y})}{\pi(\boldsymbol{\theta}_l|\mathbf{y})}\}, \tag{19}$$

or

$$\log BF_{kl} = \log \frac{f_k(\mathbf{y}|\boldsymbol{\theta}_k)}{f_l(\mathbf{y}|\boldsymbol{\theta}_l)} + \log \frac{\pi_k(\boldsymbol{\theta}_k)}{\pi_l(\boldsymbol{\theta}_l)} - \log \frac{\pi(\boldsymbol{\theta}_k|\mathbf{y})}{\pi(\boldsymbol{\theta}_l|\mathbf{y})}. \tag{20}$$

Note that (20) is also an identity. Thus, if we can find any two sequences $\tilde{\boldsymbol{\theta}}_{k,n}$ and $\tilde{\boldsymbol{\theta}}_{l,n}$ of consistent estimators of $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_l$, respectively, such that (i) $\log BF_{kl} \rightarrow_p -\infty$ when M_l is the true model, and (ii) $\log BF_{kl} \rightarrow_p \infty$ when M_k is the true model, then this will establish Bayes factor consistency.

As discussed below, one may also be interested in what happens when the model dimension grows with the sample size. Thorough treatment of such scenarios would entail a major extension of the proposed framework, and the priors could no longer be viewed as independent of n.

4.1 Nested Models

We say that M_l is nested in M_k if $\Omega_l \subset \Omega_k$, and for $\boldsymbol{\theta} \in \Omega_l$, $f_l(\mathbf{y}|\boldsymbol{\theta}) = f_k(\mathbf{y}|\boldsymbol{\theta})$. Mathematically this means that Ω_l is isomorphic to a subset of Ω_k.

It is conceptually convenient to adopt a narrow interpretation of nested linear regression models, in which

$$M_k: \ \mathbf{y} = \mathbf{X}_k\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}, \ \ \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma_k^2 I_n), \quad M_l: \ \mathbf{y} = \mathbf{X}_l\boldsymbol{\beta}_l + \boldsymbol{\varepsilon}, \ \ \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma_l^2 I_n)$$

and M_l is nested in model M_k in the sense that columns of $\mathbf{X}_l$ are a proper subset of the columns of $\mathbf{X}_k$.

We examine (20) term-by-term, with $\tilde{\boldsymbol{\theta}}_{k,n}$ and $\tilde{\boldsymbol{\theta}}_{l,n}$ equal to the respective maximum likelihood estimators of $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_l$.

Consider the first term evaluated at the maximum likelihood estimators under each model. Depending on which model is correct, this converges to a central or non-central chi-square random variable. It is thus bounded in probability, i.e. $O_p(1)$. A thorough treatment of the asymptotic distribution of the likelihood ratio statistic in the parametric setting, including nested, strictly non-nested, non-nested but overlapping, as well as misspecified models, is given by Vuong (1989). In particular, this term is asymptotically distributed as either a weighted sum of chi-square random variables or standard normal, depending on whether the models are nested, overlapping or strictly non-nested. The Wilks phenomenon occurs when the asymptotic null distributions of test statistics are independent of nuisance parameters (or nuisance functions). The Wilks phenomenon arises here in the special case that the models are nested. When the true parameter is on the boundary of the parameter space, some additional complications arise (Self & Liang, 1987).

Next, consider the middle term. As $n \to \infty$, the prior effect on the Bayes factor is of order $O(1)$. The prior should not depend on the sample size unless the dimension of the parameter space changes (grows) with the sample size.

Evaluating the final term at the maximum likelihood estimators under each model, we see a potential problem in that if the maximum likelihood estimator is consistent in both models, with the same rate of consistency, then the sequence of ratios of posteriors will diverge, as each posterior density becomes concentrated around the true parameter value (and hence the numerator and denominator both diverge). We therefore require that the two models are asymptotically distinguishable in some sense. When both models are correctly specified, i.e. correctly specified and nested, some condition is needed to prevent the larger model from being chosen over the smaller model. The conventional argument achieves this by approximation of the Bayes factor using the Bayesian Information Criterion (BIC) suggested by Schwarz (1978). This involves the Laplace approximation for the marginal likelihood, and the ratio of these marginal likelihood approximations in the Bayes factor yields a penalty term for model dimension. It should be noted, however, that it is possible for the Bayes factor to be consistent even when the BIC is not, as shown by Berger et al. (2003) in the setting that the model dimension grows with the sample size. Key references for the standard BIC-approximation-based argument for consistency of the

Bayes factor include Ando (2010); Gelfand & Dey (1994); Kass & Vaidyanathan (1992); O'Hagan & Forster (2004) and Dawid (2011). We sketch a proof through the lens of our decomposition.

In (20), we deal with the log ratio of posteriors by using a Laplace approximation for each posterior, which is naturally similar to using the Laplace approximation for the marginal likelihood. The Laplace approximation for the posterior given by Davison (1986) is

$$\pi(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})n^{-p/2}|I(\boldsymbol{\theta}^*)|^{1/2}}{f(y|\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)(2\pi)^{p/2}}\{1 + O_p(n^{-1})\}, \tag{21}$$

where $I(\boldsymbol{\theta}^*)$ is minus the $p \times p$ matrix of second partial derivatives of $\log f(\mathbf{y}|\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta})$ evaluated at the posterior mode $\boldsymbol{\theta}^*$.

The posterior mode can be replaced by the MLE since, asymptotically, the likelihood and its derivatives are negligible outside of a small neighborhood of the MLE. Plugging (21) into (20) yields

$$\begin{aligned}\log BF_{kl} &= \log \frac{f_k(\mathbf{y}|\boldsymbol{\theta}_k)}{f_l(\mathbf{y}|\boldsymbol{\theta}_l)} + \log \frac{\pi_k(\boldsymbol{\theta}_k)}{\pi_l(\boldsymbol{\theta}_l)} - \log \frac{\pi(\boldsymbol{\theta}_k|\mathbf{y})}{\pi(\boldsymbol{\theta}_l|\mathbf{y})} \\ &= \log \frac{f_k(\mathbf{y}|\boldsymbol{\theta}_k)}{f_l(\mathbf{y}|\boldsymbol{\theta}_l)} + \log \frac{\pi_k(\boldsymbol{\theta}_k)}{\pi_l(\boldsymbol{\theta}_l)} - \log \frac{f(\mathbf{y}|\boldsymbol{\theta}_k)}{f(\mathbf{y}|\boldsymbol{\theta}_k^*)} - \log \frac{\pi(\boldsymbol{\theta}_k)}{\pi(\boldsymbol{\theta}_k^*)} + \frac{(p_k - p_l)}{2}\log n \\ &\quad - \log \frac{|I(\boldsymbol{\theta}_l^*)|^{1/2}}{|I(\boldsymbol{\theta}_k^*)|^{1/2}} + \frac{(p_k - p_l)}{2}\log(2\pi) + \log \frac{f(\mathbf{y}|\boldsymbol{\theta}_l)}{f(\mathbf{y}|\boldsymbol{\theta}_l^*)} + \log \frac{\pi(\boldsymbol{\theta}_l)}{\pi(\boldsymbol{\theta}_l^*)}.\end{aligned}$$

Recall that $p_l < p_k$ since M_l is nested in M_k. Now, simply take advantage of the identity and use the sequence of MLEs, $\hat{\boldsymbol{\theta}}_{k,n} = \boldsymbol{\theta}_{k,n}^*$. In this case the third term on the second line and the third term on the last line are both zero. Also the fourth terms on each line would be zero. Multiply what remains by minus two; the first term will be the usual $\chi^2_{p_l - p_k}$ statistic, and so is bounded in probability. The second term will be $O(1)$ (plug in the MLEs for both). The ratio of the determinants and the term involving π are also $O(1)$. All that remains is $(p_k - p_l)\log n$. It is clear that for fixed-dimensional models, $-2\log BF_{kl} \to -\infty$ in P_0^n-probability as $n \to \infty$.

Moreno et al. (1998) consider both nested and non-nested models using intrinsic Bayes factors, and establish consistency for nested models by arguing that the intrinsic Bayes factor approaches the usual Bayes factor in the limit. Wang & Sun (2014) consider consistency of the Bayes factor for nested linear models when the model dimension grows with the sample size. Moreno et al. (2010, 2015) study Bayes factor consistency in the same setting, with emphasis on objective Bayes factors, such as those using intrinsic priors. The intrinsic Bayes factors are shown to be asymptotically equivalent to the BIC, though not using the Laplace approximation. When the number of parameters is $O(n^{\upsilon})$ for $\upsilon < 1$, the Bayes factor is consistent. The case $\upsilon = 1$ is termed *almost consistent* in the sense that consistency holds for all but a small set of alternative models. This small set of alternatives is expressed in terms of the pseudo-distance from the alternative model to the null model, as defined by Moreno et al. (2008). Shang & Clayton (2011) consider models with growing p in high-dimensional settings, and elucidate the connections with model selection consistency. Johnson & Rossell (2012) considered Bayesian variable selection in high-dimensional linear models and found that the posterior probability of the true model tends to zero when the number of covariates is $O(n^{1/2})$, local priors (which assign prior density zero to null values) are used for regression coefficients, and the relative prior probabilities assigned to all models are strictly positive. This does not contradict Bayes factor consistency, which is based on pairwise comparisons.

4.2 Non-Nested Models

In the non-nested case, models may be either overlapping or strictly non-nested. Overlapping models are those for which neither model is nested in the other, but that there is at least one set of parameter values such that the conditional density of the observations is the same in the two models. In strictly non-nested models, there are no parameter values for which the conditional density of the observations is the same in two models. It is known that the BIC is generally not a consistent model selection criterion when selecting among non-nested models (Hong & Preston, 2012; Sin & White, 1996).

The standard treatment of non-nested linear regression models involves rival sets of predictor variables. The arguments for the non-nested linear regression model setting mirror those from the nested setting, with some important differences. The first term in (20) will converge to a different distribution; see Vuong (1989). Provided that the two models are asymptotically distinguishable, only one will be correctly specified. The misspecified model posterior will not be consistent in the conventional sense, though there will still be a sort of pseudo-consistency, i.e. consistency for all pseudo-true parameter values, under certain conditions. See § 7 for key references. The log ratio of posterior densities will thus converge as desired, depending on which model is correctly specified. When both models are misspecified, some additional prior support conditions would be needed to ensure that the rate of posterior contraction is faster in the 'best' model.

Few papers have considered this setting. Casella et al. (2009, p.1208) state, "As far as we know, a general consistency result for the Bayesian model selection procedure for non-nested models has not yet been established." Typically authors deal with the non-nested setting by specifying an encompassing model for which both models being compared are nested in the larger encompassing model. Let $\mathcal{M}_{\text{full}}$ be the full model including all predictors. This can be used as a base model to which $\mathcal{M}_k$ and $\mathcal{M}_l$ may be compared using

$$BF_{kl} = \frac{BF_{k,\text{full}}}{BF_{l,\text{full}}},$$

under a suitable choice of priors; see Liang et al. (2008) and Guo & Speckman (2009). Casella et al. (2009) and Guo & Speckman (2009) establish consistency of Bayes factors for the special setting of normal linear models and a wide class of prior distributions, including intrinsic priors. The method of proof also utilizes the BIC approximation, though not the usual Laplace approximation (due to a property of intrinsic priors). Wang & Maruyama (2015) study consistency in linear models when the model dimension grows with the sample size, and argue that use of Zellner's g-prior is crucially important to establishing consistency. Girón et al. (2010) study objective Bayes factor consistency for non-nested linear models.

5 Parametric vs. Nonparametric

5.1 Nested Models

A parametric model can be nested in a nonparametric model in several ways, of which the most common are: (i) the parametric model is a finite-dimensional restriction of the infinite-dimensional parametric model, (ii) in a regression setting, the set of predictors in the parametric model is a subset of the predictors in the nonparametric model. A non-nested example is the regression setting where the set of predictors is not the same in each model.

5.1.1 The First Term

The first term can be studied in at least two ways. The first is to consider a suitable probability inequality, perhaps derived from the results of Wong & Shen (1995), under which this term can be shown to be exponentially small. A second approach is to use the generalized likelihood ratio (GLR) results of Fan et al. (2001) where the Wilks phenomenon is seen to arise for suitable choices of the nonparametric density estimator. In particular, the local linear density estimator of Fan (1993). Other important results with this theme include Portnoy (1988) and Murphy (1993).

Generalized likelihood ratios were first studied by Severini & Wong (1992). More recently, a detailed theory has been developed in Fan et al. (2001), Fan & Zhang (2004), and Fan & Jiang (2005, 2007). Consider a vector of functions, $\mathbf{p}$ and $\boldsymbol{\zeta}$, which are the parameters of a semiparametric or nonparametric model. Given $\boldsymbol{\zeta}$, a nonparametric estimator of $\mathbf{p}$ is given by $\hat{\mathbf{p}}_{\boldsymbol{\zeta}}$. The parameters $\boldsymbol{\zeta}$ are regarded as nuisance parameters and are estimated using the profile likelihood, i.e. by finding $\boldsymbol{\zeta}$ to maximize $\ell(\hat{\mathbf{p}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta})$ with respect to $\boldsymbol{\zeta}$. This yields the maximum profile likelihood $\ell(\hat{\mathbf{p}}_{\hat{\boldsymbol{\zeta}}}, \hat{\boldsymbol{\zeta}})$. Fan & Jiang (2007) emphasize that this is not a maximum likelihood, since $\hat{\mathbf{p}}_{\boldsymbol{\zeta}}$ is not an MLE. Suppose we are testing a parametric null hypothesis against a nonparametric alternative, i.e.

$$H_0 : p = p_\theta,\ \theta \in \Theta.$$

Denote the MLE under the null model as $(\hat{\theta}_0, \hat{\boldsymbol{\zeta}}_0)$, which maximizes the log-likelihood $\ell(\mathbf{p}_\theta, \boldsymbol{\zeta})$. Then $\ell(\mathbf{p}_{\hat{\theta}_0}, \hat{\boldsymbol{\zeta}}_0)$ is the MLE under the null. The GLR statistic is

$$GLR_n = \ell(\hat{\mathbf{p}}_{\hat{\boldsymbol{\zeta}}}, \hat{\boldsymbol{\zeta}}) - \ell(\mathbf{p}_{\hat{\theta}_0}, \hat{\boldsymbol{\zeta}}_0).$$

What is amazing about the GLR theory is that it is not necessary to use a genuine likelihood; quasilikelihoods may also be employed. In broad generality, the asymptotic null distribution of the GLR statistic is approximately χ^2.

A particularly attractive feature of the GLR framework in our study is that it allows for flexibility in the choice of estimator used for the nonparametric model. The method of estimation and the smoothing parameters used will in general affect μ_n and r, so that the relevant asymptotic distribution would need to take the estimation method into account.

However, the basic result of convergence in distribution is broadly applicable, and this alone suffices for our purposes. In particular, we can choose any convenient nonparametric estimator to achieve the desired rate of posterior contraction needed for Bayes factor consistency, while remaining confident that any reasonable choice will ensure convergence in distribution of the GLR statistic.

5.1.2 The Third Term

We need some additional assumptions for this problem to be well-defined. Here we mention only the assumptions which are most important for intuition, but warn the reader that other technical conditions can be found in the cited papers that are essential for proving Bayes factor consistency. In particular, suppose that (Ghosal et al., 2008; McVinish et al., 2009) $\mathcal{M}_k$ prescribes a finite-dimensional parametric set of densities $\mathcal{F}_k$.

Assumption 1. *Let ϵ_n be a sequence of numbers such that $\epsilon_n \to 0$ as $n \to \infty$. The nonparametric posterior $\Pi_l(\cdot|\mathbf{Y}^{(n)})$ is strongly consistent at p_0 with rate ϵ_n in the sense that*

$$\Pi_l(f : d(f, p_0) > \epsilon_n|\mathbf{Y}^{(n)}) \to 0$$

almost surely, in P_0^∞-probability, where d is some distance on the space of densities such as Hellinger or L_1.

This assumption is satisfied by most commonly-used nonparametric priors. If the posterior is not consistent, the model comparison problem would not be well-defined. In existing methods of proof, it is common to also make an assumption which ensures that the marginal likelihood under the parametric model is bounded from below.

Assumption 2. *Let f_θ denote a generic element of the parametric family prescribed by $\mathcal{M}_k$. For any $\theta \in \Theta \subset \mathbb{R}^s$, with s finite,*

$$\Pi_k(\theta' : d_{KL}(f_\theta, f_{\theta'}) < cn^{-1},\ V(f_\theta, f_{\theta'}) < cn^{-1}) > Cn^{-s/2},$$

with $V(p,q) = \int p \log(p/q)^2 d\mu$, where p, q are both absolutely continuous with respect to μ, c and C are positive constants.

Most models which for which a formal Laplace expansion is valid will satisfy this assumption. If both models are correctly specified, then to ensure that the simpler, parametric model is chosen, the rate of posterior contraction in the parametric model must be sufficiently faster than the corresponding rate of posterior contraction in the nonparametric model.

Assumption 3. *Let $A_{\epsilon_n}(\theta) = \{f : d(f, f_\theta) < C\epsilon_n\}$, where ϵ_n is the rate of consistency of the nonparametric posterior from Assumption 1. Then*

$$\sup_\theta \Pi_l(A_{\epsilon_n}(\theta)) = o(n^{-d/2}).$$

This controls the amount of mass that the nonparametric prior assigns to such neighborhoods of the parametric family of densities, for all values of $\theta \in \Theta$. Given this assumption and the rate of consistency for the nonparametric posterior, then the Bayes factor is consistent.

5.2 Non-Nested Models

When the parametric model is not nested in the alternative, the GLR framework does not help with the first term. In that case it may be easiest to utilize relevant probability inequalities to bound the log-likelihood ratio statistic so that it is exponentially small with probability exponentially close to 1. The third term can be dealt with similarly to the nested case, as discussed in detail in Ghosal et al. (2008)[§4].

5.3 Connections to Literature

Bayes factor consistency in finite-dimensional versus infinite-dimensional model comparison problems is addressed for i.i.d. data in Ghosal et al. (2008) and McVinish et al. (2009). The assumptions in the former paper are similar in spirit to those in the latter, which are given above. The results discussed below regarding the independent but not identically distributed setting for comparing a linear and partially linear model (Choi & Rousseau, 2015) are also relevant here.

Much of the existing literature in this area considers such model comparisons as a goodness of fit testing problem. An excellent review is found in Tokdar et al. (2010).

Dass & Lee (2004) considered testing a point null versus a nonparametric alternative. A common approach is to embed a parametric model into a nonparametric alternative using, for example, a mixture of Dirichlet processes (Carota & Parmigiani, 1996), a mixture of Gaussian processes (Verdinelli & Wasserman, 1998), or a mixture of Polya tree processes (Berger & Guglielmi, 2001). These methods are connected to the approach of Florens et al. (1996), in which a Dirichlet process prior was used for the alternative. Recent efforts in this area are discussed in Almeida & Mouchart (2014).

We mention that one well-studied example in this setting is the comparison of a nonparametric logspline model with a parametric alternative. The asymptotic properties of the nonparametric logspline model have been studied in Stone (1990, 1991). Joo et al. (2010) study the consistency property of information-based model selection criteria for this model comparison problem. Ghosal et al. (2008) use the logspline model as an example to illustrate that the prior support conditions are satisfied, in order to prove Bayes factor consistency in this setting.

6 Parametric vs. Semiparametric

It is increasingly common in regression modeling to compare parametric linear models against semiparametric models. A typical proof of Bayes factor consistency in this setting is found in Kundu & Dunson (2014). The proof involves bounding the marginal likelihood (upper and lower bounds) in each model and then considering what happens as $n \to \infty$. The arguments closely follow those of Guo & Speckman (2009), and can also accommodate increasing-dimensional parameter spaces, subject to a condition on the rate of growth in terms of the available sample size.

As in the parametric versus nonparametric comparison problem, existing results regarding posterior contraction rates in specific problems can be adapted to the study of the log ratio of posteriors in our framework. It is the log ratio of likelihoods which requires some additional care.

The log likelihood ratio can often be handled by some variant of semiparametric likelihood ratio tests. The results of Murphy & van der Vaart (1997) are particularly useful when the finite dimensional parametric component is of interest, and the infinite-dimensional nonparametric component is treated as a nuisance parameter. In some cases, the restricted likelihood ratio test may be used, as described in Ruppert et al. (2003)[Ch. 6]. That particular formulation relies on modeling the mean regression function as a penalized spline having a mixed effects representation, and is potentially useful when comparing a parametric linear mean regression model against an encompassing semiparametric model. This approach can also handle some types of dependent data; see the pseudo-likelihood ratio test proposed by Staicu et al. (2014).

In the semiparametric regression setting when one is interested in inference for the nonparametric component, a promising approach is described by Li & Liang (2008). Those authors extend the generalized likelihood ratio test described in § 5 to semiparametric models, which they use for inference about the nonparametric component. Remarkably, the Wilks phenomenon is again seen to arise for the generalized varying-coefficient partially linear model, which includes partially linear models and varying coefficient models as special cases.

Dealing with the log ratio of posteriors is of similar difficulty. When a parametric null model is being compared to a partially linear model, such that the null model is nested in the alternative, and if the true density is contained in the null model, it is clear that both models will possess the Kullback-Leibler property. This situation bears resemblance to the parametric versus nonparametric model comparison problem already discussed. For i.i.d. observations, conditions for Bayes factor consistency are studied in Rousseau (2008), Ghosal et al. (2008) and McVinish et al. (2009). The case of independent but not identically distributed observations is studied in Choi & Rousseau (2015).

Example 5.

$$\mathcal{M}_1 :\ Z_i = W_{ij}\boldsymbol{\beta} + \sigma\varepsilon_i, \quad \mathcal{M}_2 :\ Z_i = W_{ij}\boldsymbol{\beta} + r(X_i) + \sigma\varepsilon_i, \tag{22}$$

where $r(\cdot)$ is an unknown function in an infinite-dimensional parameter space, and $W_{ij}\boldsymbol{\beta}$ is a linear component.

A particular version of this comparison problem is studied by Choi & Rousseau (2015), where $\{W_{ij}\} \in [-1, 1]$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, with p finite, and $\{X_i\} \in [0, 1]$, $i = 1, \ldots, n$. Recall that a semiparametric prior is $\Pi = \Pi_{\boldsymbol{\beta}} \times \Pi_r$. Choi & Rousseau (2015) assume a Gaussian process prior for Π_r with a reproducing kernel Hilbert space (RKHS) $\mathbb{H} = L^2([0,1])$, and concentration function ϕ_r given by $\phi_r(\epsilon) = \inf_{h \in \mathbb{H}: \|h - r_0\|_\infty < \epsilon} \|h\|_{\mathbb{H}}^2 - \log \Pi_r[\|r\|_\infty < \epsilon]$, where $\|\cdot\|_\infty$ is the supremum norm, $\|r\|_\infty = \sup\{|r(x)| : x \in [0,1]\}$. Concentration functions and RKHS are discussed in van der Vaart & van Zanten (2008a,b). They show that the support of Π_r is the closure of $\mathbb{H}$ and, therefore, for any r_0 contained in the support of Π_r, there exists $\epsilon_n \downarrow 0$, and $c > 0$ such that $\phi_{r_0}(\epsilon_n) \leq n\epsilon_n^2$, and the following small ball probability holds

$$\Pi_r[\|r - r_0\|_\infty \leq \epsilon_n] \geq e^{-cn\epsilon_n^2}.$$

This property ultimately controls the rate of posterior contraction under the partially linear model, and is essential to the proof of Bayes factor consistency in Choi & Rousseau (2015).

Choi & Rousseau (2015) and Kundu & Dunson (2014) give regularity conditions for Bayes factor consistency in the semiparametric setting, both for convergence in probability and convergence almost surely. Lenk (2003) discusses testing a parametric model against a semiparametric alternative by embedding an exponential family in a semiparametric model, and develops an MCMC procedure for evaluating the adequacy of the parametric model. Choi et al. (2009) study the asymptotic properties of the Bayes factor when comparing a parametric null model to a semiparametric alternative. They impose strong conditions on the structure of the model, namely Gaussian errors with known variance, and an orthogonal design matrix, and a particular variance structure for the coefficients in a trigonometric series expansion representation of the nonparametric component of the mean function in the regression model. An interesting example of this model comparison setting is found in MacEachern & Guha (2011), which illustrates a seemingly paradoxical result that under certain prior support assumptions, the posterior under a semiparametric model can be more concentrated than under a finite-dimensional restriction of that model.

A different strand of semiparametric Bayesian literature exists for models defined by moment conditions. Empirical likelihood methods have become popular for frequentist estimation and inference in such models. Empirical likelihoods have many properties of parametric likelihoods, but until recently there was no formal probabilistic interpretation, which meant there was no Bayesian justification for their use. Schennach (2005, 2007) introduced the Bayesian exponentially-tilted empirical likelihood (ETEL) to study the marginal posterior of a finite-dimensional interest parameter in the presence of an infinite dimensional nuisance parameter. This analysis is extended by Chib et al. (2016) to the problem of selecting valid and relevant moments among different moment condition models. Using a semiparametric approach, it is shown that the ETEL satisfies a Bernstein-von Mises theorem in misspecified moment models. Moreover, the marginal likelihood-based moment selection procedure, based on the approach of (Chib, 1995), is proven to be consistent in the sense of selecting only correctly specified and relevant moment conditions. This significant new work opens up the Bayes factor computation and analysis for a large array of Bayesian semiparametric models that are specified only through the moments of the underlying unknown distribution. It also connects to the analysis of similar problems from a frequentist perspective such as the proportional likelihood ratio model (Luo & Tsai, 2012), which can be seen as a generalization of exponential tilt regression models, and its adaptation to mean regression modeling using empirical likelihood (Huang & Rathouz, 2011). These frequentist developments have spawned new results concerning likelihood ratio tests. For example, in the context of high-dimensional semiparametric generalized linear models, the reader is referred to Ning et al. (2015).

7 Directions for Future Research

We briefly discuss some potential avenues for future research which coincide with recent advances in the literature.

Other Model Comparison Problems. Some model comparison problems are less well-studied than the three model-comparison frameworks already discussed: (i) two nonparametric models; (ii) two semiparametric models; and (iii) a nonparametric and a semiparametric model. In principle, the concepts we have discussed in other cases continue to guide our intuition and direct our approach to the problem. However, we are unaware of existing theoretical results which are general enough to be broadly applicable in these settings. An ongoing area of research is that of testing nonparametric hypotheses when they are not well-separated; see Salomond (2015).

Rate of Convergence of Bayes Factors. McVinish et al. (2009) actually obtain an upper bound on the convergence rate of the Bayes factor when the parametric model is correct and the comparison is with an encompassing nonparametric model. Such results would be of great use in applied settings. In particular, it would be helpful to be able to state more precisely the notion that under certain prior support conditions, the Bayes factor cannot distinguish between a correct and incorrect model, or may even prefer an incorrect model, despite having very large sample sizes.

Empirical Bayes. Replacing some of the hyperparameters in the prior distribution by data dependent quantities is part of an empirical Bayesian analysis. The asymptotic properties of empirical Bayes procedures have recently been studied in nonparametric models, where the maximum marginal posterior likelihood estimator is used for the hyperparameters; see Petrone et al. (2014) and Rousseau & Szabó (2015). Consistency of the posterior in empirical Bayesian analysis is complicated by the need to also consider the sequence of estimators for the hyperparameter. Posterior concentration rates with empirical priors are studied in Martin & Walker (2016).

Summary Statistics. An exciting new approach to Bayes factor asymptotics by Marin et al. (2014) shows how Bayes factor consistency can hinge on whether or not the expectations of suitably chosen summary statistics are asymptotically different under the models being compared. It will be interesting to learn if those conditions, which are necessary and sufficient, can be related in a meaningful way to conditions on the log ratio of posterior densities.

Non-i.i.d. Observations and Conditional Densities. A natural extension to existing results concerns non-i.i.d. observations. Consistency and asymptotic normality for posterior densities in such settings are active research areas. See Ghosal & van der Vaart (2007). A related extension is the estimation of conditional densities. For example, in a semiparametric regression model, one might assume a linear model in the predictors and a nonparametric error distribution. One may not wish to assume that the predictors and errors are independent, and then could consider inference for the nonparametric component, i.e. the conditional error density, given the predictors. See e.g. Pelenis (2014). Pati et al. (2013) investigate the possibility of recovering the full conditional distribution of the response given the covariates.

Misspecified Models. Another important question is the large-sample behavior of the posterior when the model is misspecified. It is shown in (van Ommen, 2015, p.39-40) that the posterior is inconsistent in simple nested model settings with misspecification (the truth is not in the set of models being considered) despite all the relevant consistency theorems still holding, e.g. Kleijn & van der Vaart (2006), De Blasi & Walker (2013), and Ramamoorthi et al. (2015). An enlightening discussion of some of these issues, with more references, is given by van Erven et al. (2015). Other key references include Bunke & Milhaud (1998), Grünwald & Langford (2007), Shalizi (2009), Lee & MacEachern (2011), Müller (2013), and Chib et al. (2016). Section 4.3.2 of a recent Ph.D. thesis (van Ommen, 2015) contains a current discussion of Bayesian consistency under misspecification. Owhadi et al. (2015) consider the lack of robustness of Bayesian procedures to misspecification, which they call 'brittleness'. Kleijn & van der Vaart (2012) study the Bernstein-von Mises theorem under misspecification as do Chib et al. (2016).

References

Ando, T. (2010). *Bayesian Model Selection and Statistical Modeling.* Boca Raton: Chapman & Hall/CRC.

Aitkin, M. (1991). Posterior Bayes factors (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **53**, 111–142.

Almeida, C. & Mouchart, M. (2014). Bayesian testing for nested hypotheses under partial observability. *Sankhya* A **76**(2), 305–327.

Amemwou-Atisso, M., Ghosal, S., Ghosh, J. K. & Ramamoorthi, R. V. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli* **9**, 291–312.

Barron, A. R. (1986). Discussion of the paper by P. Diaconis and D. Freedman. *Annals of Statistics* **14**, 26–30.

Barron, A. R. (1998). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith, editors. *Bayesian Statistics 6*, 27–52. Oxford: Oxford University Press.

BARRON, A., SCHERVISH, M. J. & WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Annals of Statistics* **27**, 536–561.

BASU, S. & CHIB, S. (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *J. Amer. Statist. Assoc.* **98**, 224–235.

BAYARRI, M. J., BERGER, J. O., FORTE, A. & GARCÍA-DONATO, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Annals of Statistics* **40**, 1550–1577.

BELITSER, E. & GHOSAL, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Annals of Statistics* **31**, 536–559.

BERGER, J. O. & GUGLIELMI, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *J. Amer. Statist. Assoc.* **96**, 174–184.

BERGER, J. O. & PERICCHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91**, 109–122.

BERGER, J. O. & PERICCHI, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In: *Model Selection. Institute of Mathematical Statistics Lecture Notes – Monograph Series* **38**, 135–207. Beachwood: IMS.

BERGER, J. O., GHOSH, J. K. & MUKHOPADHYAY, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *J. Statist. Plann. Inference* **112**, 241–258.

BHATTACHARYA, A. & DUNSON, D. B. (2010). Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. *Biometrika* **97**(4), 851–865.

BUNKE, O. & MILHAUD, X. (1998). Asymptotic behavior of Bayes estimates under possibly incorrect models. *Annals of Statistics* **26**, 617–644.

CARLIN, B. P. & CHIB, S. (1995). Bayesian model choice via Markov Chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **57**, 473–484.

Carota, C. & Parmigiani, G. (1996). On Bayes factors for nonparametric alternatives. In *Bayesian Statistics* 5, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds., 507–511. Oxford University Press.

Casella, G., Giron, F. J., Martinez, M. L. & Moreno, E. (2009). Consistency of Bayesian procedures for variable selection. *Annals of Statistics* **37**, 1207–1228.

Castillo, I. (2014). On Bayesian supremum norm contraction rates. *Annals of Statistics* **42**, 2058–2091.

Castillo, I. & Rousseau, J. (2015). A Bernstein-von Mises theorem for smooth functionals in semiparametric models. *Annals of Statistics* **43**, 2353–2383.

Chakrabarti, A. & Ghosh, J. K. (2011). AIC, BIC and recent advances in model selection. In *Handbook of the Philosophy of Science Vol. 7 Philosophy of Statistics*, P.S. Bandyopadhyay & M.R. Forster, eds. Elsevier.

Chen, M.-H. & Shao, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *Annals of Statistics* **25**, 1563–1594.

Cheng, G. & Kosorok, M. R. (2008). General frequentist properties of the posterior profile distribution. *Annals of Statistics* **36**, 1819–1853.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90**, 1313–1321.

Chib, S. & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *J. Amer. Statist. Assoc.* **96**, 270–281.

Chib, S., Shin, M. & Simoni, A. (2016). Bayesian empirical likelihood estimation and comparison of moment condition models. Submitted.

Choi, T. & Rousseau, J. (2015). A note on Bayes factor consistency in partial linear models. *J. Statist. Plann. Inference* **166**, 158–170.

Choi, T., Lee, J. & Roy, A. (2009). A note on the Bayes factor in a semiparametric regression model. *J. Multivariate Anal.* **100**, 1316–1327.

Clarke, B. S. & Barron, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory* **36**, 453–471.

Dass, S. C. & Lee, J. (2004). A note on the consistency of Bayes factors for testing point null versus non-parametric alternatives. *J. Statist. Plann. Inference* **119**, 143–152.

Davison, A. C. (1986). Approximate predictive likelihood. *Biometrika* **73**, 323–332.

Dawid, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. In: J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, editors. *Bayesian Statistics 4*, p. 109–125. Oxford: Oxford University Press.

Dawid, A. P. (2011). Posterior model probabilities. In: P.S. Bandyopadhyay, M. Forster, editors. *Philosophy of Statistics*, 607–630. New York: Elsevier.

Dawid, A. P. & Musio, M. (2015). Bayesian model selection based on proper scoring rules. *Bayesian Analysis* **10**, 479–499.

De Blasi, P. & Walker, S. G. (2013). Bayesian asymptotics with misspecified models. *Statistica Sinica* **23**, 169–187.

Dempster, A. P. (1973). The direct use of likelihood for significance testing. In *Proceedings of Conference on Foundational Questions in Statistical Inference*, 335–354. Aarhus, Denmark.

Diaconis, P. & Freedman, D. (1986a). On the consistency of Bayes estimates (with discussion). *Annals of Statistics* **14**, 1–67.

Diaconis, P. & Freedman, D. (1986b). On inconsistent Bayes estimates of location. *Annals of Statistics* **14**, 68–87.

DiCiccio, T. J., Kass, R. E., Raftery, A. E. & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.* **92**, 903–915.

Doob, J. L. (1949). Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, 23–27. Paris: Colloques Internationaux du Centre National de la Recherche Scientifique.

Dutta, R., Bogdan, M. & Ghosh, J. K. (2012). Model selection and multiple testing - a Bayes and empirical Bayes overview and some new results. *J. Indian Statist. Assoc.* **50**, 105–142.

Evans, M. (2015). *Measuring Statistical Evidence Using Relative Belief.* Boca Raton: CRC Press.

Fan, J. (1993). Local linear regression smoothers and minimax efficiencies. *Annals of Statistics* **21**, 196–216.

Fan, J. & Jiang, J. (2005). Nonparametric inferences for additive models. *J. Amer. Statist. Assoc.* **100**, 890–907.

Fan, J. & Jiang, J. (2007). Nonparametric inference with generalized likelihood ratio tests (with discussion). *TEST* **16**, 409–478.

Fan, J. & Zhang, J. (2004). Sieve empirical likelihood ratio tests for nonparametric functions. *Annals of Statistics* **32**, 1858–1907.

Fan, J., Zhang, C. & Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of Statistics* **29**, 153–193.

Fernández, C., Ley, E. & Steel, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics* **100**, 381–427.

Florens, J. P., Richard, J. F. & Rolin, J. M. (1996). Bayesian encompassing specification tests of a parametric model against a non parametric alternative. Discussion Paper 96-08, Institut de statistique, Université Catholique de Louvain, `http://sites.uclouvain.be/IAP-Stat-Phase-V-VI/ISBApub/ISBAdp.html`.

Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Annals of Mathematical Statistics* **34**, 1386–1403.

Gelfand, A. E. & Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *J. R. Statist. Soc.* B **56**, 501–514.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398–409.

Ghosh, J. K. & Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics.* New York: Springer-Verlag.

Ghosal, S. (2010). The Dirichlet process, related priors and posterior asymptotics. In N.L. Hjort, C. Holmes, P. Müller, S.G. Walker, editors. *Bayesian Nonparametrics*, 35–79. Cambridge: Cambridge University Press.

Ghosal, S. & van der Vaart, A. (2007). Posterior convergence rates of posterior distributions for non-i.i.d. observations. *Annals of Statistics* **35**, 192–223.

Ghosal, S., Ghosh, J. K. & Ramamoorthi, R. V. (1999a). Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics* **27**, 143–158.

Ghosal, S., Ghosh, J. K. & Ramamoorthi, R. V. (1999b). Consistent semiparametric Bayesian inference about a location parameter. *J. Statist. Plann. Inference* **77**, 181–193.

Ghosal, S., Ghosh, J. K. & van der Vaart, A. (2000). Convergence rates of posterior distributions. *Annals of Statistics* **28**, 500–531.

Ghosal, S., Lember, J. & van der Vaart, A. (2003). On Bayesian adaptation. Acta Applicandae Mathematicae **79**, 165–175.

Ghosal, S., Lember, J. & van der Vaart, A. (2008). Nonparametric Bayesian model selection and averaging. *Elec. J. Statist.* **2**, 63–89.

Ghosh, J. K. (1994). *Higher-Order Asymptotics.* Institute of Mathematical Statistics, NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 4.

Ghosh, J., Pukayastha, S. & Samanta, T. (2005). Role of p-values and other measures of evidence in Bayesian analysis. In: D. Dey and C.R. Rao, editors. *Handbook of Statistics* Vol. 25, Ch. 5, p. 151–170. Amsterdam: Elsevier.

Giné, E. & Nickl, R. (2011). Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Annals of Statistics* **39**, 2883–2911.

Giné, E. & Nickl, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models.* Cambridge: Cambridge University Press.

Girón, F. J., Moreno, E., Casella, G. & Martínez, M. L. (2010). Consistency of objective Bayes factors for nonnested linear models and increasing model dimension. *RACSAM* **1**, 57–67.

Goldenshluger, A. & Lepski, O. (2014). On adaptive minimax density estimation on R^d. *Probab. Theory Relat. Fields* **159**, 479–543.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Grünwald, P. D. (2007). *The minimum description length principle.* Cambridge: MIT Press.

Grünwald, P. & Langford, J. (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. *Mach. Learn.* **66**, 119–149.

Guo, R. & Speckman, P. (2009). Bayes factor consistency in linear models. In *2009 International Workshop on Objective Bayes Methodology*, Philadelphia.

Han, C. & Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors. *J. Amer. Statist. Assoc.* **96**, 1122–1132.

Hjort, N. L., Holmes, C., Müller, P. & Walker, S. G., editors (2010). *Bayesian Nonparametrics.* Cambridge: Cambridge University Press.

Hoffmann, M., Rousseau, J. & Schmidt-Hieber, J. (2015). On adaptive posterior concentration rates. *Annals of Statistics* **43**, 2259–2295.

Hong, H. & Preston, B. (2012). Bayesian averaging, prediction and nonnested model selection. *J. Econometrics* **167**, 358–369.

Huang, T.-M. (2004). Convergence rates for posterior distributions and adaptive estimation. *Annals of Statistics* **32**, 1556–1593.

Huang, A. & Rathouz, P. J. (2011). Proportional likelihood ratio models for mean regression. *Biometrika* **99**(1), 223–229.

Johnson, V. E. & Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107**, 649–660.

Joo, Y., Wells, M. T. & Casella, G. (2010). Model selection error rates in nonparametric and parametric model comparisons. In *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown*, 166–183. Institute of Mathematical Statistics.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773–795.

Kass, R. E. & Vaidyanathan, S. K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *J. R. Statist. Soc.* B **54**, 129–144.

Kim, Y. & Lee, J. (2001). On posterior consistency of survival models. *Annals of Statistics* **29**(3), 666–686.

Kleijn, B. J. K. & van der Vaart, A. W. (2006). Misspecification in infinite dimensional Bayesian statistics. *Annals of Statistics* **34**, 837–877.

Kleijn, B. J. K. & van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic J. Statist.* **6**, 354–381.

Kundu, S. & Dunson, D. B. (2014). Bayes variable selection in semiparametric linear models. *J. Amer. Statist. Assoc.* **109**, 437–447.

Lee, J. & MacEachern, S. N. (2011). Consistency of Bayes estimators without the assumption that the model is correct. *J. Statist. Plann. Inference* **141**, 748–757.

Lember, J. & van der Vaart, A. (2007). On universal Bayesian adaptation. *Statistics & Decisions* **25**, 127–152.

Lenk, P. J. (2003). Bayesian semiparametric density estimation and model verification using a logistic-Gaussian process. *J. Comput. Graph. Statist.* **12**, 548–565.

Li, R. & Liang, H. (2008). Variable selection in semiparametric regression modeling. *Annals of Statistics* **36**, 261–286.

Lian, H. (2009). On rates of convergence for posterior distributions under misspecification. *Comm. Statist. Theory Methods* **38**, 1893–1900.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103**, 410–423.

Luo, X. & Tsai, W. Y. (2012). A proportional likelihood ratio model. *Biometrika* **99**(1), 211–222.

MacEachern, S. N. & Guha, S. (2011). Parametric and semiparametric hypotheses in the linear model. *Canadian J. Statistics* **39**(1), 165–180.

Marin, J.-M., Pillai, N. S., Robert, C. P. & Rousseau, J. (2014). Relevant statistics for Bayesian model choice. *J. Roy. Statist. Soc. Ser. B* **76**, 833–859.

Martin, R. & Hong, L. (2012). On convergence rates of Bayesian predictive densities and posterior distributions. Unpublished report, arXiv:1210.0103.

Martin, R. & Walker, S. G. (2016). Optimal Bayesian posterior concentration rates with empirical priors. Technical report, arXiv:1604.05734.

Martin, R., Hong, L. & Walker, S. G. (2013). A note on Bayesian convergence rates under local prior support conditions. Unpublished report, arXiv:1201.3102.

McVinish, R., Rousseau, J. & Mengersen, K. (2009). Bayesian goodness of fit testing with mixtures of triangular distributions. *Scand. J. Statist.* **36**, 337–354.

Meng, X.-L. & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica* **6**, 831–860.

Moreno, E. & Girón, F. J. (2008). Comparison of Bayesian objective procedures for variable selection in linear regression. *TEST* **17**, 472–490.

Moreno, E., Berolino, F. & Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypothesis testing. J. Amer. Statist. Assoc. **93**, 1451–1460.

Moreno, E., Girón, F. J. & Casella, G. (2010). Consistency of objective Bayes factors as the model dimension grows. *Annals of Statistics* **38**, 1937–1952.

Moreno, E., Girón, J. & Casella, G. (2015). Posterior model consistency in variable selection as the model dimension grows. *Statistical Science* **30**, 228–241.

Müller, U. K. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica* **81**, 1805–1849.

Murphy, S. A. (1993). Testing for a time dependent coefficient in Cox's regression model. *Scand. J. Statist.* **20**, 35–50.

Murphy, S. A. & van der Vaart, A. W. (1997). Semiparametric likelihood ratio inference. *Annals of Statistics* **25**, 1471–1509.

Ning, Y., Zhao, T. & Liu, H. (2015). A likelihood ratio framework for high dimensional semiparametric regression. arXiv:1412.2295.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. R. Statist. Soc.* B **57**, 99–138.

O'Hagan, A. & Forster, J. (2004). *Bayesian Inference*, 2nd edition, volume 2B of "Kendall's Advanced Theory of Statistics". London: Wiley.

Owhadi, H., Scovel, C. & Sullivan, T. (2015). On the brittleness of Bayesian inference. arXiv:1308.6306

Pati, D., Dunson, D. B. & Tokdar, S. T. (2013). Posterior consistency in conditional distribution estimation. *J. Multivariate Analysis* **116**, 456–472.

Pelinis, J. (2014). Bayesian regression with heteroscedastic error density and parametric mean function. *J. Econometrics* **178**, 624–638.

Petrone, S., Rousseau, J. & Scricciolo, C. (2014). Bayes and empirical Bayes: do they merge? *Biometrika* **101**, 285–302.

Petrone, S. & Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *J. Roy. Statist. Soc. Ser. B* **64**, 79–100.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Annals of Statistics* **16**, 356–366.

Ramamoorthi, R. V., Sriram, K. & Martin, R. (2015). On posterior concentration in misspecified models. *Bayesian Analysis* **10**, 759–789.

Robert, C. P. (1993). A note on Jeffreys-Lindley paradox. *Statistica Sinica* **3**, 601–608.

Rousseau, J. (2008). Approximating interval hypothesis: p-values and Bayes factors. In: J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith, editors. *Bayesian Statistics* 8, p. 417–452. Oxford: Oxford University Press.

Rousseau, J. & Choi, T. (2012). Bayes factor consistency in regression problems. Unpublished report.

Rousseau, J. & Szabó, B. (2015). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. arXiv:1504.04814

Rousseau, J., Salomond, J.-B. & Scricciolo, C. (2014). On some aspects of the asymptotic properties of Bayesian approaches in nonparametric and semiparametric models. *ESAIM: Proceedings* **44**, 159–171.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge: Cambridge University Press.

Salomond, J.-B. (2015). Bayesian testing for embedded hypotheses with application to shape constraints. arXiv: 1303.6466 .

Schennach, S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika* **92**(1), 31–46.

SCHENNACH, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *Annals of Statistics* **35**(2), 634–672.

SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Ver. Geb.* **4**, 10–26.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

SCRICCIOLO, C. (2006). Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *Annals of Statistics* **34**, 2897–2920.

SCRICCIOLO, C. (2015). Bayesian adaptation. *J. Statist. Plann. Inference* **166**, 87–101.

SELF, S. G. & LIANG, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82**(398), 605–610.

SEVERINI, T. & WONG, W. H. (1992). Profile likelihood and conditionally parametric models. *Annals of Statistics* **20**, 1768–1802.

SHAFER, G., SHEN, A., VERESHCHAGIN, N. & VOVK, V. (2011). Test martingales, Bayes factors and $p-$values. *Statistical Science* **26**, 84–101.

SHALIZI, C. R. (2009). Dynamics of Bayesian updating with dependent data and misspecified models. *Electronic J. Statist.* **3**, 1039–1074.

SHANG, Z. & CLAYTON, M. K. (2011). Consistency of Bayesian linear model selection with a growing number of parameters. *J. Statist. Plann. Inference* **141**, 3463–3474.

SHAO, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statistica Sinica* **7**, 221–264.

SHEN, X. & WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Annals of Statistics* **29**, 687–714.

SIN, C.-Y. & WHITE, H. (1996). Information criteria for selecting possibly misspecified parametric models. *J. Econometrics* **71**, 207–225.

Smith, I. & Ferrari, A. (2014). Equivalence between the posterior distribution of the likelihood ratio and a $p-$value in an invariant frame. *Bayesian Analysis* **9**, 939–962.

Staicu, A.-M., Li, Y., Crainiceanu, C.M. & Ruppert, D. (2014). Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis. *Scandinavian J. Statist.* **41**, 932–949.

Stone, C. (1990). Large sample inference for log-Spline models. *Annals of Statistics* **18**, 717–741.

Stone, C. (1991). Asymptotics for doubly flexible logspline response models. *Annals of Statistics* **19**, 1832–1854.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81**(393), 82–86.

Tokdar, S. T., Chakrabarti, A. & Ghosh, J. K. (2010). Bayesian nonparametric goodness of fit tests. In M.-H. Chen, D.K. Dey, P. Müller, D. Sun, K. Ye, editors. *Frontiers of Statistical Decision Making and Bayesian Analysis in Honor of James O. Berger*, 185–194. New York: Springer.

van der Vaart, A. W. & van Zanten, J. H. (2008a). Rates of contraction of posterior distributions based on Gaussian process priors. *Annals of Statistics* **36**(3), 1435–1463.

van der Vaart, A. W. & van Zanten, J. Y. (2008b). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh* 3, 200–222.

van der Vaart, A. W. & van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Annals of Statistics* **37**, 2655–2675.

van Erven, T., Grünwald, P. D., Mehta, N. A., Reid, M. D. & Williamson, R. C. (2015). Fast Rates in Statistical and Online Learning. *J. Machine Learning Research* **16**, 1793–1861.

van Ommen, T. (2015). Better predictions when models are wrong or underspecified. Ph.D. thesis, Universiteit Leiden.

Verdinelli, I. & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J. Amer. Statist. Assoc.* **90**, 614–618.

Verdinelli, I. & Wasserman, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Annals of Statistics* **26**, 1215–1241.

Villa, C. & Walker, S. (2015). On the mathematics of the Jeffreys-Lindley paradox. arXiv:1503.04098

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307–333.

Walker, S. G. (2003). On sufficient conditions for Bayesian consistency. *Biometrika* **90**, 482–488.

Walker, S. G. (2004). Modern Bayesian asymptotics. *Statistical Science* **19**, 111–117.

Walker, S. G. (2004). New approaches to Bayesian consistency. *Annals of Statistics* **32**, 2028–2043.

Walker, S., Damien, P. & Lenk, P. (2004). On priors with a Kullback-Leibler property. *J. Amer. Statist. Assoc.* **99**, 404–408.

Walker, S. & Hjort, N. L. (2001). On Bayesian consistency. *J. Roy. Statist. Soc. Ser. B* **63**, 811–821.

Walker, S. G., Lijoi, A. & Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Annals of Statistics* **35**, 738–746.

Wang, M. & Maruyama, Y. (2015). Consistency of Bayes factor for nonnested model selection when the model dimension grows. arXiv:1503.06155

Wang, M. & Sun, X. (2014). Bayes factor consistency for nested linear models with a growing number of parameters. *J. Statist. Plann. Inference* **147**, 95–105.

Wasserman, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. In D. Dey, P. Müller, D. Sinha, editors. *Practical Nonparametric and Semiparametric Bayesian Statistics.* New York: Springer.

Wong, W. H. & Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Annals of Statistics* **23**, 339–362.

Wu, Y. & Ghosal, S. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic J. Statistics* **2**, 298–331.

Zhang, T. (2006). From ε-entropy to KL-entropy: analysis of minimum information complexity density estimation. *Annals of Statistics* **34**, 2180–2210.