Parametric and semiparametric hypotheses in the linear model

Steven N. MacEachern and Subharup Guha*

August, 2008

*Steven MacEachern is Professor, Department of Statistics, Ohio State University, Cockins Hall, 1958 Neil Avenue, Columbus, OH 43210 (snm@stat.ohio-state.edu); Subharup Guha is Assistant Professor, Department of Statistics, University of Missouri-Columbia, 209C Middlebush Hall, Columbia, MO 65211 (email: GuhaSu@missouri.edu). Thanks to NSF grant numbers SES-0437251 and DMS-0605041.
Abstract

The independent additive errors linear model consists of a structure for the mean and a separate structure for the error distribution. The error structure may be parametric or it may be semiparametric. Under alternative values of the mean structure, the best fitting additive errors model has an error distribution which can be represented as the convolution of the actual error distribution and the marginal distribution of a misspecification term. The model misspecification term results from the covariates’ distribution. Conditions are developed to distinguish when the semiparametric model yields sharper inference than the parametric model and vice-versa. The main conditions concern the actual error distribution and the covariates’ distribution. The theoretical results explain a paradoxical finding in semiparametric Bayesian modelling, where the posterior distribution under a semiparametric model is found to be more concentrated than is the posterior distribution under a corresponding parametric model. The paradox is illustrated on a set of allometric data.

Keywords: Bayes factor, Dirichlet process, Kullback-Liebler divergence, nonparametric Bayes.

1 INTRODUCTION

Recent advances in modelling and computation have enabled us to fit more realistic models in a wide variety of settings. One modelling strategy which has proven enormously successful is to modify a standard Bayesian hierarchical model by replacing a distribution of parametric form with one which has a nearly arbitrary form. This
results in a semiparametric or nonparametric Bayesian model which we fit with a Markov chain Monte Carlo simulation. Dey et al. (1998), Walker et al. (1999) and Müller and Quintana (2004) provide excellent overviews of these techniques. Further developments are reported in the more recent literature.

This paper contrasts inference for parametric and semiparametric linear models. To ensure that the models are comparable, we demand that the two versions of the model imply mutually absolutely continuous distributions for each finite set of observables. Specifically, we focus on the independent errors linear model, specified by

\[ Y = X\beta + \epsilon \]  

(1)

where, with a departure from convention, the intercept is swept into the error term. That is, the \( n \) components of \( \epsilon \) are independent draws from some distribution, \( F \) which may have non-zero mean. In the normal theory linear model, \( F \) is presumed to be a Gaussian distribution. For our semiparametric models, we presume that \( F \) has a density with respect to Lebesgue measure and that this density has full support on the real line.

One of the main motivations for fitting a semiparametric model is that its support contains the support of the parametric model. Thus, for large samples and with additional mild conditions, when \( F \) does follow the specified parametric form, the predictive distributions will be nearly equivalent. However, when \( F \) does not follow the specified parametric form, we obtain more accurate predictive distributions for \( Y|X \) under the semiparametric model.

Setting this important predictive advantage of the semiparametric model aside, we
consider inference for the regression coefficients, $\beta$. When contemplating a semiparame-
metric model, one often also fits a corresponding parametric model. A phenomenon
which occurs on a regular basis is that the posterior distribution for $\beta$ is tighter under
the semiparametric model than it is under the parametric model. This phenomenon
is driven by the marginal likelihood for $\beta$. Figure 1 shows estimated marginal log-
likelihoods for two models fit to a set of allometric data. The semiparametric model
fits the data better, as indicated by the higher log-likelihood. The sharper curvature
of the log-likelihood for the semiparametric model leads to a smaller difference be-
tween the two models for extreme $\beta$, leading to a more concentrated posterior under
the semiparametric model.

Figure 1: Estimated log-marginal likelihoods $\pm$ 2 standard errors. The upper and
lower curves correspond, respectively, to the semiparametric and the parametric mod-
els. The data, models and the method of estimation are described in Section 3

This phenomenon is surprising since the prior distribution for the semiparametric
model is far more diffuse than is the prior distribution for the parametric model. The typical expectation is that, when updated with the same data, a more diffuse prior distribution will lead to a more diffuse posterior distribution. For other data sets, the converse of the phenomenon occurs, with the posterior distribution for $\beta$ more diffuse under the semiparametric model than under the parametric model.

In this paper, we explain this paradoxical behavior. The key ingredients of the explanation are marginal distribution of the covariate, $X$, and the Kullback-Liebler divergences between particular distributions. The next section presents analytical results for the simple versus simple setting. The results extend to the case where the regression coefficients have a continuous distribution. Section 3 illustrates the behavior of the posterior for the allometric data relating brain weight to body weight and provides further evidence through simulation. The results are used to draw the distinction between parametric and semiparametric hypotheses. This highlights the importance of specifying a complete probability model for the data under null and alternative hypotheses. We believe that it also establishes the need for routine use of semiparametric models.

2 THEORETICAL RESULTS

Consider an analysis of the mean structure given in Equation (1) at two distinct values for $\beta$. We presume that one of the values is the true value of $\beta$, denoted by $\beta_0$. A Bayesian hypothesis test of $H_0 : \beta = \beta_0$ against the alternative value $H_1 : \beta = \beta_1$ can be accomplished by computing the posterior odds of the two hypotheses. An
equivalent procedure, if the prior odds are even, is to compute the Bayes factor comparing the two hypotheses.

The Bayes factor is the ratio of the marginal likelihoods of the data under the two parameter values, \( B = m_0(Y, X)/m_1(Y, X) \). Since the hypotheses are simple and the distribution of \( X \) is assumed to be identical under the two models, we can compute the Bayes factor from the adjusted data \( m_j(Y|X) = \prod_{i=1}^{n} f_j(Y_i - X_i^T \beta_j) \), \( j = 1, 2 \), where \( f_j(\cdot) \) represents the error density under model \( j \).

Let \( H \) be the distribution of \( X^T(\beta_0 - \beta_1) \). Given a constant \( h \), let \( F_1 \ast h \) denote the distribution \( F_1 \) shifted by \( h \). The log-Bayes factor of the models in the hypothesis test relies on the log-marginal likelihoods, and defining \( h_i = X_i^T(\beta_0 - \beta_1) \) we have

\[
\log(B) = \log(m_0(Y, X)/m_1(Y, X)) = \sum_{i=1}^{n} \log(f_0(Y_i - X_i^T \beta_0)/f_1(Y_i - X_i^T \beta_0 - h_i)).
\]

Thus \( \log(B) \) is a sum of independent and identically distributed variates. Under \( H_0 \), the mean of a single one of these variates is the expected Kullback-Liebler divergence, \( E^H KL(F_0, F_1 \ast h) \), where \( h \sim H \). The large sample properties of the comparison between the two hypotheses is driven by the expected Kullback-Liebler divergence, with mild conditions ensuring that a central limit theorem holds, so that \( \log(B)/n = E^H KL(F_0, F_1 \ast h) + O(1/\sqrt{n}) \).

The distributions \( F_0 \) and \( F_1 \) are essential to the computation of \( B \). When they both fall in a parametric family, say the family of normal distributions, we refer to \( H_0 \) and \( H_1 \) as parametric hypotheses. When \( F_0 \) and \( F_1 \) do not fall in the parametric family, we refer to \( H_0 \) and \( H_1 \) as semiparametric hypotheses. In the sequel, we reserve \( m_0 \) and \( m_1 \) for the models and marginal densities under the semiparametric
hypotheses. We use $m_2$ and $m_3$ for the models and marginal densities under the parametric hypotheses.

### 2.1 Kullback-Leibler divergences

For a pair of distributions $F$ and $G$ which both have densities with respect to Lebesgue measure, the Kullback-Liebler divergence from $F$ to $G$ is defined by

$$KL(F,G) = \int \log(f(x)/g(x))f(x)dx.$$  \hspace{1cm} (2)

The Kullback-Liebler divergence satisfies a number of properties which it will prove helpful to recall.

1. **Zero information.** $KL(F,G) = 0$ if and only if $F = G$. If $F \neq G$, then $KL(F,G) > 0$.

2. **Independent experiments.** Suppose that two independent experiments are performed, where a random vector may be split into two parts, $X = (X_1, X_2)$. Under the model $F$, $X_i \sim F_i$; under the model $G$, $X_i \sim G_i$. Under each model, $X_1$ and $X_2$ are independent. Then $KL(F,G) = KL(F_1,G_1) + KL(F_2,G_2)$.

3. **Dependent experiments and non-negativity.** Suppose that a random vector may be split into two parts, $X = (X_1, X_2)$. Let $F_1$ denote the distribution of $X_1$ and $F_{2|1}$ denote the conditional distribution of $X_2|X_1$ under the model $F$. Use similar notation for the model $G$. Then $KL(F,G) \geq KL(F_1,G_1)$.

4. **1-1 transformations.** Let $F_1$ and $G_1$ denote the distributions of a random vector $X$ under models $F$ and $G$. Suppose $Y = \psi(X)$, where $\psi$ is a 1-1 trans-
formation. Let $F_2$ and $G_2$ denote the corresponding distributions of $Y$ under the two models. Then $KL(F_1, G_1) = KL(F_2, G_2)$.

These properties are easily established through examination of the expected log-likelihoods.

### 2.2 Log-Bayes factors - random covariates

The Bayes factor comparing the two semiparametric models is driven by $E^H KL(F_0, F_1 \ast h)$. The fits of the parametric models are based on possible approximations to the actual error distributions. We will use $G_0$ and $G_1$ to represent the best fitting normal approximations to $F_0$ and to $F_1$, respectively. The best approximation $G$ to $F_0$ is the normal distribution which minimizes $KL(F_0, G)$. The minimizer, $G_0$, has the same mean and variance as $F_0$. Similarly, $G_1$ has the same mean and variance as $F_1$.

To contrast the comparison of the hypotheses under the two models, we examine the ratio of Bayes factors under the semiparametric and parametric models. Equivalently, we compute the difference in log-Bayes factors.

$$\log\left( \frac{m_0(Y,X)}{m_1(Y,X)} \right) - \log\left( \frac{m_2(Y,X)}{m_3(Y,X)} \right) = \log\left( \frac{m_0(Y,X)}{m_2(Y,X)} \right) - \log\left( \frac{m_1(Y,X)}{m_3(Y,X)} \right). \quad (3)$$

A positive difference in (3) indicates greater support for $\beta_0$ under the semiparametric model; a negative difference indicates greater support under the parametric model.

We take expectations of (3) with respect to the model $m_0$. Provided the divergences are finite, this becomes

$$KL(F_0, G_0) + KL(F_0 \ast H, F_1) - KL(F_0 \ast H, G_1). \quad (4)$$
The proof is presented in the Appendix. With finite second moments for the divergences, we immediately have

$$\log(B_0)/n - \log(B_1)/n = KL(F_0, G_0) + KL(F_0 \ast H, F_1) - KL(F_0 \ast H, G_1) + O(1/\sqrt{n}).$$

As a precursor to the theorems, we note the expression

$$Y_i - X_i^T \beta_1 = Y_i - X_i^T \beta_0 + X_i^T (\beta_0 - \beta_1). \tag{5}$$

When $\beta_0$ provides the true regression coefficients and the semiparametric model in Equation (1) holds, we note that $Y_i - X_i^T \beta_0$ follows the distribution $F_0$ and that $Y_i - X_i^T \beta_1$ is therefore distributed as $F_0 \ast H$. The expression in (4) with $F_1 = F_0 \ast H$, therefore simplifies to $KL(F_0, G_0) - KL(F_1, G_1)$. This simple expression enables us to describe the impact of the covariates’ distribution on inference.

Suppose that the covariate is normally distributed with mean $\mu$ and variance $\Sigma$. Then $H$ is a normal distribution with mean $\mu^T(\beta_0 - \beta_1)$ and variance $(\beta_0 - \beta_1)^T \Sigma (\beta_0 - \beta_1)$. In a loose sense, we expect that, when the covariate follows a multivariate normal distribution, the convolution of $F_0$ and $H$ will result in an $F_1$ which is closer to normal than is $F_0$. The following theorem provides a precise statement in terms of the Kullback-Liebler divergence between parametric and semiparametric models.

**Theorem 1** Assume that $H_0 : \beta = \beta_0$ holds for the semiparametric model in Equation (1). Further assume that $X_i \sim N(\mu, \Sigma), i = 1, \ldots, n$. Then $KL(F_0, G_0) \geq KL(F_1, G_1)$.

**Proof:** As we remarked earlier, the assumption that the covariate $X$ is normally distributed implies that $X^T(\beta_0 - \beta_1)$ is normally distributed with its distribution
denoted by $H$. The distribution $G_1$ is the best normal approximation to $F_1 = F_0 * H$. This best normal approximation has the same mean and variance as $F_1$, and so $G_1 = G_0 * H$.

Let the symbol $(X \perp Y)$ denote the bivariate random variable $(X, Y)$ having mutually independent components $X$ and $Y$. The annotations below refer to properties of the Kullback-Liebler divergence given earlier.

\[
KL(F_0, G_0) = KL(F_0, G_0) + KL(H, H) \quad \text{zero information}
\]
\[
= KL((F_0 \perp H), (G_0 \perp H)) \quad \text{independent experiments}
\]
\[
= KL((F_0 * H, H), (G_0 * H, H)) \quad \text{1-1 transformation}
\]
\[
\geq KL(F_0 * H, G_0 * H) \quad \text{non-negativity}
\]
\[
= KL(F_1, G_1).
\]

The consequence of Theorem 1 is that the expected log-Bayes factor comparing $F_0$ to $F_1$ exceeds the expected log-Bayes factor comparing $G_0$ to $G_1$. Theorem 1 also implies that, as $\beta_1$ moves away from $\beta_0$, the gap in expected log-Bayes factors narrows. This suggests that the posterior distribution for $\beta$ will be more concentrated under the semiparametric model, in agreement with empirical observation (see Figure 1).

Specifically,

**Corollary 1** Assume that $H_0 : \beta = \beta_0$ holds for the semiparametric model in Equation (1). Further assume that $X_i \sim N(\mu, \Sigma)$, $i = 1, \ldots, n$. Let $\beta_2 = \beta_0 + c(\beta_1 - \beta_0)$ for some $c > 1$. Then, with evident notation, $KL(F_1, G_1) \geq KL(F_2, G_2)$.

The corollary is proved in the same fashion as Theorem 1.
As a converse to Theorem 1, when the parametric model holds, so that \( F_0 = G_0 \), but the covariate is not normally distributed, \( F_1 \) will no longer be normally distributed. The next theorem clarifies the impact that this has on the difference in log-Bayes factors.

**Theorem 2** Assume that \( H_0 : \beta = \beta_0 \) holds for the model in Equation (1). Further assume that \( F_0 \) is a normal distribution, so that \( G_0 = F_0 \). Then, if \( X \sim G_X \) where \( X^T(\beta_0 - \beta_1) \sim H \), and \( H \) is not a normal distribution, \( KL(F_0, G_0) < KL(F_1, G_1) \).

**Proof:** The proof of this statement for a weak inequality would be immediate since \( KL(F_0, G_0) = 0 \). Strict inequality follows from the zero-information property of the Kullback-Liebler divergence, and because \( G_1 \) is different than \( F_1 = G_0 \ast H \) when \( H \) is not normal.

Theorems 1 and 2 present results for simple versus simple comparisons of models, describing the difference in expected log-likelihoods at \( \beta_0 \) and \( \beta_1 \). The theorems apply to each individual value of \( \beta_1 \), and so extend to arbitrary sets for \( \beta_1 \). Specifically, Theorem 1 ensures that, with a normally distributed covariate, the difference in expected log-likelihood surfaces between semiparametric and parametric models is always positive. This indicates greater support for \( \beta_0 \) under the semiparametric model. It tends to lead to a sharper likelihood surface for \( \beta \) under the semiparametric model and hence to a sharper posterior distribution. Importantly, these results do not require knowledge of \( \beta_0 \). They are a direct consequence of the model in Equation (1).

Theorem 2 provides a counterpoint to this result. When the parametric model
holds at $\beta_0$ and the covariate is not normally distributed, the parametric model tends to yield a sharper likelihood for $\beta$.

As noted earlier, the difference $\log(B_0)/n - \log(B_1)/n$ equals $KL(F_0, G_0) - KL(F_1, G_1) + O(1/\sqrt{n})$, provided a central limit theorem holds. The following theorem, proved in the Appendix, gives sufficient conditions for a central limit theorem.

**Theorem 3** Under the conditions of Theorem 1 or Theorem 2, the difference
\[ \sqrt{n} \left( \log(B_0)/n - \log(B_1)/n \right) \] is asymptotically normal if the following three conditions hold:

(A) The density $f_0$ is bounded.

(B) The distributions $F_0$ and $H$ have finite fourth moments.

(C) The vectors $(X_i, Y_i)$ $i = 1, 2, \ldots$, are i.i.d. such that the vectors $(X_i, Y_i - X_i^T \beta_0)$ form a random sample from the distribution $G \perp F_0$

### 2.3 Log-Bayes factors - fixed covariates

Given a fixed value of the covariate $x$ and a random observation $Y$, define the random variable
\[
\Delta(H_n, h^*, \epsilon) = \log \left( \frac{f_{0}(\epsilon)}{g_0(\epsilon)} \right) - \log \left( \frac{f_{1,n}(\epsilon + h^*)}{g_{1,n}(\epsilon + h^*)} \right)
\]

where $h^* = x^T \beta_0$, $\epsilon = Y - x^T \beta_0$, $f_{1,n}$ is the density of $F_0 \ast H_n$ and $g_{1,n}$ is the density of its nearest normal approximation. For every integer $n$, let $x_{n,1}, \ldots, x_{n,n}$ be a given set of non-random covariates. Let $Y_{n,1}, \ldots, Y_{n,n}$ be the corresponding random responses. Setting $h_{n,k} = x_{n,k}^T (\beta_0 - \beta_1)$, we denote by $H_n$ the empirical distribution of $h_{n,k}$, where
The log-Bayes factor corresponding to a random observation $Y_{n,k}$ is then

$$
\Delta_{n,k} = \log\left(\frac{m_0(Y_{n,k}|x_{n,k})}{m_2(Y_{n,k}|x_{n,k})}\right) - \log\left(\frac{m_1(Y_{n,k}|x_{n,k})}{m_3(Y_{n,k}|x_{n,k})}\right)
$$

$$
= \log\left(\frac{f_0(Y_{n,k} - x_{n,k}^T\beta_0)}{g_0(Y_{n,k} - x_{n,k}^T\beta_0)}\right) - \log\left(\frac{f_{1,n}(Y_{n,k} - x_{n,k}^T\beta_0 + h_{n,k})}{g_{1,n}(Y_{n,k} - x_{n,k}^T\beta_0 + h_{n,k})}\right)
$$

where $\epsilon_{n,k} = Y_{n,k} - x_{n,k}^T\beta_0$. Analogous to the mean difference in log-Bayes factors, $\log(B_0)/n - \log(B_1)/n$, we define the random variable $\Gamma_n$ in the case of non-random covariates as

$$
\Gamma_n = \frac{1}{n} \sum_{k=1}^{n} \Delta_{n,k}.
$$

The following lemma provides sufficient conditions which guarantee that $\Gamma_n$ is asymptotically normal and equals $M_n + O_p(n^{-1/2})$, where $\lim_n M_n = KL(F_0, G_0) - KL(F_1, G_1)$. We assume that the empirical distributions $H_n$ have support lying in the interval $[-M, M]$, for some large number $M$. The Appendix contains an outline of the proof.

**Theorem 4** Suppose that all of the following conditions hold:

(A) The density $f_0$ is bounded, continuous and has no zeros.

(B) The distribution $F_0$ has more than four finite moments. That is, there exist $m > 0$ and $\delta > 0$ such that the $(2 + \delta)(m + 2)$-th moment of $F_0$ is finite.

(C) For $m$ as defined above, there exists $x_0 > 0$ such that for all $|x| > x_0$ we have $f_0(x) \geq c_1 \cdot \exp(-c_2|x|^{m+2})$.

(D) The covariates $h_{n,k}$ are uniformly bounded in absolute value by $M > 0$.

(E) The sequence of distributions $H_n$, where $n = 1, 2, \ldots$, converges to $H$. 

13
Let \( M_n = \frac{1}{n} \sum_{k=1}^{n} E(\Delta_{n,k}) \) be the expectation of \( \Gamma_n \). Then
\[
\sqrt{n}(\Gamma_n - M_n) \Rightarrow N(0, \sigma^2),
\]
where \( \lim_n M_n = KL(F_0, G_0) - KL(F_1, G_1) \) and \( \sigma^2 = E^{H}Var^{F_0}\Delta(H, h, \epsilon) \).

**Remark:** Condition (C) is a mild tail condition on the error density \( f_0 \). It bounds \( f_0 \) away from zero with a sub-density whose tails that are thinner than normal. When \( m \) is large, as it is when \( F_0 \) is a short-tailed distribution, the sub-density has tails that decay extremely quickly.

**Remark:** Suppose for each \( n \), that the covariates correspond to the quantiles \( \frac{1}{n+1}, \ldots, \frac{n}{n+1} \) of a distribution \( H \) having bounded support. Thus \( H_n \) converges in distribution to \( H \). Applying Theorem 4, we have that the variable \( \Gamma_n \) is asymptotically normal and that \( \Gamma_n = M_n + O_p(\sqrt{n}) \) and \( \lim_n M_n = KL(F_0, G_0) - KL(F_1, G_1) \).

## 3 APPLICATIONS

Peters (1983) gives a thorough overview of allometric studies that relate the body size of an animal, \( X \), to some other physiological characteristic of interest, \( Y \). He provides a biological explanation of empirical power laws of the form \( Y = e^{\alpha}X^{\beta} \), where \( \alpha \) and \( \beta \) represent constants. In this example, we study the relationship between the body mass of an animal, \( X \), and its brain mass, \( Y \). One empirical theory in allometry states that the power law holds in this case with \( \beta \) approximately equal to \( 3/4 \). The rationale is that the brain regulates the body metabolism, which is “known” to vary as \( X^{3/4} \). Another theory, the so-called surface area law, states that the brain mass \( Y \)
is proportional to body surface area since it serves as the end-point for nerve channels. With body densities being approximately equal, body volume is also a measure of body mass, $X$. If, further, body shape is stable across a range of body sizes, body surface area is proportional to body mass to the $2/3$ power. Hence the brain mass $Y$ is proportional to $X^{2/3}$.

We consider the data given in Weisberg (1985) relating the body mass of 62 mammals to their brain mass. The variables are transformed to the logarithmic scale to achieve linearity and constant variance. The transformed covariate is then recentered to justify the specification of independent priors on the regression coefficients. The sample quantiles of the centered covariate are plotted against the theoretical values for the normal distribution in Figure 3. The plot suggests that the covariate is nearly normally distributed. For the transformed set of variables, the power law becomes $Y = \alpha + X\beta$. The model in Equation (1) applies, with the regression slope $\beta$ being the parameter of interest.

The traditional analysis of allometric data relies on a simple linear regression model (e.g., Weisberg, 1985). Standard regression diagnostics do not indicate any departures from the assumptions of linearity and constant variance. The errors appear to be approximately normal, although there is some evidence of skewness to the right. The assumption of independence does appear to be violated (see MacEachern and Peruggia, 2001, for an explanation and specialized diagnostics for independence), though for our purposes we focus on the traditional analysis which assumes independence.

The coincidence of an approximately normally distributed covariate and non-normal error distribution at the null value of $\beta$ suggests that Theorem 1 applies. The
Theorem suggests, in turn, that inference about the slope will be sharpened with a semiparametric specification of the model. Pursuing this line, we contrast the performance of parametric and semiparametric Bayesian models. We describe distributions on the error distribution, $F$, and the regression coefficient separately.

The semiparametric model which we investigate is a mixture of Dirichlet processes (MDP) model. The model relies on the Dirichlet process, a nonparametric model that serves as a prior distribution for an unknown distribution. The MDP model pushes the Dirichlet process up one stage in a hierarchy that effectively creates a semiparametric model.

$$
\begin{align*}
\gamma &\sim N(\mu_\gamma, \sigma_\gamma^2); \\
\tau^{-2} &\sim IG(c, d); \\
\sigma^{-2} &\sim IG(a, b) \\
P|\tau &\sim Dir(N(0, \tau^2)); \\
\theta_i|P &\sim P
\end{align*}
$$
Figure 3: A normal probability plot of the log-transformed, recentered covariates for the Weisberg data.

\[ \eta_i \mid \sigma \sim N(0, \sigma^2) \]

\[ \epsilon_i = \gamma + \theta_i + \eta_i \]

\[ Y_i = X_i^T \beta + \epsilon_i, \]

where \( IG(a, b) \) represents the inverse gamma distribution with mean \( b(a - 1)^{-1} \) and variance \( b^2(a - 1)^{-2}(a - 2)^{-1} \), \( N(a, b) \) represents the normal distribution with mean \( a \) and variance \( b \), and \( Dir(\alpha) \) represents the Dirichlet process with base measure \( \alpha \). With our choice of \( \alpha \), the prior distribution for \( P \) has full support among all distributions on the real line. The term \( \gamma \) is included to allow some control over the mean of the distribution for \( \epsilon \). The mean of \( P \) makes an additional contribution to \( \epsilon \)'s mean. The distribution \( P \) is a discrete distribution. The term \( \eta \) yields a continuous distribution for \( \epsilon \). The distribution for \( \epsilon \) may be described as a countable

The parametric model is matched to the above model to ensure that our results are not due to differences in prior specification. The model is chosen to provide the same prior predictive distribution for a single $\epsilon_i$.

\[
\begin{align*}
\gamma & \sim N(\mu_\gamma, \sigma_\gamma^2); \quad \tau^{-2} \sim IG(c,d); \quad \sigma^{-2} \sim IG(a,b) \\
\theta_i | \tau & \sim N(0, \tau^2); \quad \eta_i | \sigma \sim N(0, \sigma^2) \\
\epsilon_i & = \gamma + \theta_i + \eta_i \\
Y_i & = X_i^T \beta + \epsilon_i.
\end{align*}
\]

Conditional on the higher order parameters, the $\epsilon_i$ form a random sample from the $N(\gamma, \sigma^2 + \tau^2)$ distribution.

The prior variance $\sigma_\gamma^2$ was set equal to 100 to reflect vague prior knowledge. Taking a peek at the data to select reasonable parameter values, we set $a = b = c = d = 2$ and $\mu_\gamma = 3.14$. Focusing on the theory that $\beta = 3/4$, when working with a continuous distribution for $\beta$, we took $\beta \sim N(0.75, 1/16)$. The prior variance was chosen to guarantee a negligible probability of negative values of $\beta$. A least squares analysis yields an estimate of $\beta$ of 0.752 with a standard error of 0.028, supporting the belief that $\beta$ is approximately $3/4$.

The semiparametric and parametric models are compared in two fashions. First, the performance of the models is compared for simple versus simple and for simple
versus complex tests. Then the impact of the models on the posterior distribution for $\beta$ are investigated. Log-marginal likelihoods, either at specific values of $\beta$ or averaged across a distribution for $\beta$, underlie these tests and posterior distributions. Since these are not available in closed form, we estimated them with numerical methods. The MDP model was fit with the MCMC methods developed in Bush and MacEachern (1996), and its log-marginal likelihood was estimated with the methods of Basu and Chib (2003). The log-marginal likelihood of the parametric model was estimated by Monte Carlo integration. The MCMC runs and Monte Carlo sample sizes were long enough and large enough, respectively, to yield small standard errors for the log-marginal likelihoods.

Figure 1 presents the estimated log-marginal likelihoods and associated standard errors for the parametric and semiparametric models for a grid of 21 evenly spaced points in [.65,.85]. The Bayes factor test of $H_0 : \beta = 0.75$ against $H_1 : \beta = \beta_1$ is accomplished by exponentiating the difference in log-marginal likelihoods. In all cases, the test based on the semiparametric model shows stronger evidence in favor of $H_0$ than does the test based on the parametric model. The visual narrowing of the difference between the semiparametric (top) curve and the parametric (bottom) curve in the figure show this difference. For the specific alternative with $\beta_1 = 0.67$, the two tests yield Bayes factors of 34.6 and 17.9, providing strong evidence in favor of $\beta = 0.75$. The evidence in the semiparametric analysis is noticeably stronger than that in the parametric analysis.

The parametric and semiparametric analyses also enable us to make comparisons across the types of models. The comparison of the simple, semiparametric hypothesis $H_0 : \beta = 0.75$ against the simple, parametric hypothesis $H_1 : \beta = 0.75$ favors
the semiparametric model with a Bayes factor of 45.42. A test of the complex, semiparametric hypothesis \( H_0 : \beta \sim N(0.75, 1/16) \) against the simple, parametric hypothesis, \( H_1 : \beta = 0.75 \) favors the semiparametric model with a Bayes factor of 5.37. The first of these results provides strong evidence in favor of a semiparametric model, while the second indicates that the semiparametric model with unknown \( \beta \) fits better than the best of the parametric models.

Finally, the posterior distributions for \( \beta \) under the two complex models were investigated. The MDP model is supported more strongly by the data than is the parametric model. The log-marginal likelihoods were estimated to be \(-74.5\) and \(-78.3\), respectively, leading to a Bayes factor of 44.8 when comparing the two models. For each model, we obtained 25 independent replications of 90% and 95% central posterior probability regions for \( \beta \). All of the intervals were centered near 0.75, and all of the intervals excluded \( 2/3 \). At both confidence levels, all 25 of the semiparametric intervals were entirely contained in each of the 25 corresponding parametric intervals. As noted in the introduction, the sharper marginal likelihood under the semiparametric model leads to narrower posterior probability intervals. Since the prior distribution is centered very near the peak of the likelihood, the intervals under the semiparametric model are contained inside those of the parametric model. Again, this is exactly the behavior suggested by Theorem 1.

### 3.1 Simulations

The applicability of the theoretical results can be further demonstrated by simulation. The following two examples illustrate the applicability of Theorems 1 and 2.

**Example 1: MDP model.** Suppose the distribution \( F_0 \) arises from the MDP model
described earlier. The density is given by

\[ f_0(\epsilon) = \frac{1}{\sigma} \sum_{i=1}^{\infty} p_i \phi\left(\frac{\epsilon - \alpha - \theta_i}{\sigma}\right). \]

(7)

Being a mixture of normals having the same scale parameter, \( f_0 \) is bounded above by \((2\pi\sigma^2)^{-1/2}\), implying that assumption (A) of Theorem 3 is satisfied.

For all positive \( j \), the \( j^{th} \) moment of the distribution \( P = \sum_{i=1}^{\infty} p_i \delta_{\theta_i} \) is finite almost surely. This follows from the fact that all positive moments of the normal base measure are finite (Ferguson, 1973). Since

\[
\int \epsilon^4 f_0(\epsilon) d\epsilon = \sum_{i=1}^{\infty} p_i \int \epsilon^4 \frac{1}{\sigma} \phi\left(\frac{\epsilon - \theta_i}{\sigma}\right) d\epsilon \\
= 3\sigma^4 + 6\sigma^2 \sum_{i=1}^{\infty} p_i (\theta_i + \alpha)^2 + \sum_{i=1}^{\infty} p_i (\theta_i + \alpha)^4,
\]

which depends only on the first four moments of \( P \), the distribution \( F_0 \) has a finite fourth moment. Thus, if the covariate \( X \) is normally distributed, the Lemma applies and together with Theorem 1 indicates that the difference \( \log(B_0)/n - \log(B_1)/n \) is positive with high probability.

Setting \( \sigma = 0.1 \) and \( \tau = \alpha = 1 \) in the MDP model, we generated 1000 independent realizations of the error distribution \( F_0 \). As Muliere and Tardella (1998) note, the countable mixture for \( P \) must be approximated by a finite mixture. Following Ishwaran and James (2000), the distribution \( P \) was accurately approximated by a distribution restricted to 100 points. The covariates’ distribution was assumed to be \( N(2,4) \) and the true slope was \( \beta_0 = 1 \). For each of the 1000 realizations of \( F_0 \): (i) The model in Equation (1) was used to generate \( n = 1000 \) i.i.d. \((X, Y)\) pairs. (ii) Taking \( \beta_1 = 1.5 \) in the hypotheses \( H_1 \) and \( H_3 \), the difference \( \log(B_0)/n - \log(B_1)/n \) was computed and (iii) the standard error of the difference was estimated.
Figure 4 presents a histogram of the 1000 differences computed in (ii). All of the differences were positive, and all except 27 of the differences exceeded zero by more than three standard errors. This demonstrates the validity of Theorem 1. The average difference equals 0.837.

![Histogram of log(B0/n - log(B1)/n in Example 1.](image)

**Example 2: Normal model.** If the true error density is normal and the covariate is non-normal, Theorem 2 and the Lemma state that the difference log(B0)/n − log(B1)/n is close to −KL(F1,G1), a negative value, with high probability. To demonstrate this, we set σ = 0.1 and τ = α = 1 as in Example 1, and suppose that the true error F0 ∼ N(α,σ^2 + τ^2). The covariate X is assumed to be exponentially distributed with mean 2. Thus, Theorem 2 and the Lemma are formally applicable. We let β0 = 1 be the true regression coefficient in Equation (1). We test this against β1 = 1.5 in the hypotheses H1 and H3. The limiting difference −KL(F1,G1) can also be numerically computed and equals -0.040.

Since the (X,Y) pairs are unconditionally i.i.d. in this example, we generated 100 i.i.d. (X,Y) pairs from the above true model. The difference log(B0)/n − log(B1)/n
was -0.06 with a standard error of 0.013. The small standard error indicates the high probability that the difference will be negative.

4 CONCLUSIONS

The development of the previous sections places the focus of inference both on the mean structure of the linear model and on the form of the actual error distribution. The implication is that it is essential to consider whether the hypotheses under investigation specify both mean and error structure, or whether they specify only the mean structure.

We distinguish three cases of simple versus complex tests. The first case is purely parametric. The null hypothesis, given by \((\beta_0, G)\), specifies both the mean structure and the parametric form for \(G\). The alternative allows a different mean structure but retains the parametric form for the error distribution, as denoted by \((\beta, G)\). Most applied statistical work that we have seen, either classical or Bayesian, presumes the same parametric form under the simple and complex models. When the assumption of a particular parametric form is relaxed, as it is to create classical nonparametric tests (e.g., Randles and Wolfe, 1979), it is often replaced by the assumption that the additive model holds—implying that the error distribution holds the same form for different values of \(\beta\), whatever that form may be.

In the second case, the null hypothesis is parametric, given by \((\beta_0, G)\), while the complex, alternative hypothesis allows departures in both the mean structure and the form of the error distribution. In the notation of this paper, the alternative would
be specified as \((\beta, F)\). This formulation of the problem allows the form of the error distribution to differ when the null hypothesis does not hold. It is appropriate, for example, when a treatment affects only some of a population or where the impact of a treatment is different on different units in a population.

In the third case, both hypotheses are semiparametric. The null hypothesis, given by \((\beta_0, F)\), specifies the mean structure. The alternative hypothesis specifies a departure from the mean structure and would be written \((\beta, F)\).

The allometric study of Section 3 provides an example which we may use to select from the three pairs of hypotheses. We rule out the first pair, since the simple hypothesis provides no description of the data’s distribution in the event that the simple hypothesis fails to hold. The choice between the second and third pairs of hypotheses depend on whether one believes that the theories which determine \(\beta_0\) also specify a form for the error distribution. Our reading of the literature and conversations with evolutionary biologists lead us to the view that the theories do not specify any form for the error distribution, although a presumed normal form has been used both to produce estimates, via linear regression, and for inferential purposes (e.g., Weisberg, 1985). This leads us to select the third case, of semiparametric against semiparametric hypotheses.

We believe that the second or third pairs of hypotheses will generally be appropriate for selection, and that the first pair only rarely reflects the theories that generate the hypotheses. In keeping with this belief, we recommend the routine use of semiparametric models for the alternatives, even when the point null is parametric.

The analysis in Section 3 illustrates the price of misspecifying the hypotheses. For the allometric data, inference about the mean structure is less precise than it should
be. As the simulations show, inference about the mean structure can be artificially overprecise in some settings if parametric hypotheses are selected.

This work has been motivated by observed differences in posterior distributions under semiparametric Bayesian models and parametric Bayesian models. The theoretical development of Section 2 tells us that the results are not due to the specifics of the models we have used. The theory also suggests that we will observe similar behavior with non-Bayesian likelihood based methods.

5 APPENDIX

Proof of (4): Let us use $f^*_0$ to denote the density of the distribution $F_0 \ast H$. Taking expectations in (3) with respect to the model $m_0$:

$$
E^m \log\left(\frac{m_0(Y;X)}{m_2(Y;X)}\right) - E^m \log\left(\frac{m_1(Y;X)}{m_3(Y;X)}\right)
$$

$$
= E^m \log\left(\frac{m_0(\cdot)}{m_2(\cdot|X;\beta_0)}\right) - E^m \log\left(\frac{m_1(\cdot)}{m_3(\cdot|X;\beta_1)}\right)
$$

$$
= \int \log\left(\frac{f^*_0(\epsilon)}{g^*_0(\epsilon)}\right) f_0(\epsilon) d\epsilon - \int \log\left(\frac{f_1(\epsilon')}{g_1(\epsilon')}\right) f^*_0(\epsilon') d\epsilon'
$$

$$
= KL(F_0, G_0) + \int \log\left(\frac{f_0(\epsilon)}{f_1(\epsilon')}\right) f^*_0(\epsilon') d\epsilon' - \int \log\left(\frac{f^*_0(\epsilon')}{g^*_0(\epsilon')}\right) f^*_0(\epsilon') d\epsilon'
$$

$$
= KL(F_0, G_0) + KL(F_0 \ast H, F_1) - KL(F_0 \ast H, G_1).
$$

The following lemma is useful for the proof of Theorems 3 and 4.

**Lemma 1:** For a bounded density $\psi(x)$, the integral $\int |\log \psi(x)|^q \psi(x) dx$ is finite for all positive integers $q$.

**Proof of Lemma 1:** For any density $\psi(x)$, there exists a $\delta \in (0, 1)$ such that $\int \psi^{1-\delta}(x) dx$ is finite. Since $\psi(x)$ is bounded, given any positive integer $q$ and $\gamma > 0$,
the function $|\log \psi|^q \cdot \psi^\gamma$ is bounded because $\lim_{\psi \to 0} |\log \psi|^q \cdot \psi^\gamma$ equals zero. Since $|\log \psi|^q \psi$ equals $|\log \psi|^q \psi^{\delta} \cdot \psi^{1-\delta}$, we have that $\int |\log \psi(x)|^q \psi(x)dx$ is finite.

**Proof of Theorem 3:** The difference in log-Bayes factor in (3) equals

$$\log \left( \frac{f_0(\epsilon)}{g_0(\epsilon)} \right) - \log \left( \frac{f_1(\epsilon')}{g_1(\epsilon')} \right)$$

where $\epsilon \sim F_0$ and $\epsilon' \sim F_1 = F_0 \ast H$ under hypothesis $H_0$. The (i.i.d.) central limit theorem holds if the second moment exists, which is implied by the finiteness of the following integrals: (i) $\int \log^2 f_0(x) f_0(x) dx$, (ii) $\int \log^2 g_0(x) f_0(x) dx$, (iii) $\int \log^2 f_1(x) f_1(x) dx$, and (iv) $\int \log^2 g_1(x) f_1(x) dx$.

Assumption (A) and Lemma 1 above imply that the integral (i) is finite. Now

$$f_1(x) = \int f_0(x - t) \cdot dH(t). \quad (8)$$

This implies that the density $f_1$ is bounded, so that the integral (iii) is also finite.

Assumption (B) implies that the fourth moments of $F_0$ and $F_1 = F_0 \ast H$ are finite. The expressions $\log^2 g_0(x)$ and $\log^2 g_1(x)$ are fourth degree polynomials in $x$, so the integrals (ii) and (iv) are both finite. This proves the result.

The following Lemma is a useful tool:

**Lemma 2:** Assume that the conditions of Theorem 4 are satisfied. Let the random variable $\Delta(H_n, h^*, \epsilon)$ be defined as in section 2.3 and $\delta > 0$ be defined as in Theorem 4. For all $0 \leq \gamma \leq 2 + \delta$,

$$\lim_{n} \frac{1}{n} \sum_{k=1}^{n} E^{F_0} \left| \Delta_{n,k} - E^{F_0} \left( \Delta_{n,k} \right) \right|^\gamma = E^H E^{F_0} \left| \Delta(H, h^*, \epsilon) - E^{F_0} \Delta(H, h^*, \epsilon) \right|^\gamma < \infty.$$

**Proof of Lemma 2:** As a first step to the proof, we observe that

$$\frac{1}{n} \sum_{k=1}^{n} E^{F_0} \left| \Delta_{n,k} - E^{F_0} \left( \Delta_{n,k} \right) \right|^\gamma = E^H_n E^{F_0} \left| \Delta(H_n, h, \epsilon) - E^{F_0} \Delta(H_n, h, \epsilon) \right|^\gamma,$$
where the expectation \( E^{H_n} \) is taken over the argument \( h \). We prove first that for \( 0 \leq \gamma \leq 2 + \delta \),

\[
\sup_{h^*, n, \gamma} E^{F_0} \left| \Delta(H_n, h^*, \epsilon) - E^{F_0} \Delta(H_n, h^*, \epsilon) \right|^{\gamma} < \infty \tag{9}
\]

where \( h^* \) is constrained to lie in \([-M, M]\). Since \( E|X|^p \leq E|X|^q + 1 \) for all random variables \( X \) and \( 0 \leq p \leq q \), it is sufficient to prove the result for \( \gamma = 2 + \delta \). This is equivalent to \( \sup_{h^*, n} E^{F_0} |\Delta(H_n, h^*, \epsilon)|^{2+\delta} < \infty \). It is sufficient, therefore, that the following integrals are uniformly bounded for all \( n \) and \( h^* \in [-M, M] \):

(i) \( \int |\log f_0(\epsilon)|^{2+\delta} f_0(\epsilon) d\epsilon \),
(ii) \( \int |\log g_0(\epsilon)|^{2+\delta} f_0(\epsilon) d\epsilon \),
(iii) \( \int |\log f_{1,n}(\epsilon + h^*)|^{2+\delta} f_0(\epsilon) d\epsilon \),
(iv) \( \int |\log g_{1,n}(\epsilon + h^*)|^{2+\delta} f_0(\epsilon) dx \).

Lemma 1 guarantees that integral (i) is finite. The expressions \( \log g_0(\epsilon) \) and \( \log g_{1,n}(\epsilon + h^*) \) are second degree polynomials in \( \epsilon \). The integral (ii) is therefore finite by assumption (B) of Theorem 4. Since \( h^* \in [-M, M] \), the integral (iv) is uniformly bounded for all \( H_n \) and \( h^* \). If \( \epsilon \) belongs to the set \( A_1 = (-\infty, -x_0 - 2M] \), assumption (C) implies that \( \inf_{h^*} f_{1,n}(\epsilon + h^*) \) is bounded below by \( c_1 \cdot \exp(-c_2|\epsilon - 2M|^{m+2}) \).

Similarly, if \( \epsilon \) belongs to the set \( A_2 = [x_0 + 2M, \infty) \), \( \inf_{h^*} f_{1,n}(\epsilon + h^*) \) is bounded below by \( c_1 \cdot \exp(-c_2|\epsilon + 2M|^{m+2}) \). If \( \epsilon \) belongs to \( A_3 = (-x_0 - 2M, x_0 + 2M) \), \( \inf_{h^*} f_{1,n}(\epsilon + h^*) \) exceeds \( c_0 \) for some \( c_0 > 0 \). This follows from assumption (A) of Theorem 4. Splitting the integral (iii) over the sets \( A_1 \), \( A_2 \) and \( A_3 \), we note that \( |\log f_{1,n}(\epsilon + h^*)|^{2+\delta} \) is uniformly bounded above by \( a_1|\epsilon + 2M|^{(2+\delta)(m+2)} + a_2 \), where \( a_1 \) and \( a_2 \) are constants that do not depend on \( H_n \) or \( h^* \). The supremum of the integral (iii) is therefore finite, by assumption (B).

Notice that \( f_{1,n} \) is bounded and continuous because \( f_0 \) is bounded and continuous. Furthermore, since \( H_n \to H \), \( f_{1,n} \) tends pointwise to the density \( f_1 \). Thus,
the random variable $\Delta(H_n, h^*, \epsilon)$ is continuous in $h^*$. The bounded convergence theorem and (9) applied with $\gamma = 2$, imply that $E^{F_0} \Delta(H_n, h^*, \epsilon)$ is continuous in $h^*$, and also that $\lim_{n} E^{F_0} \Delta(H_n, h^*, \epsilon)$ equals $E^{F_0} \Delta(H, h^*, \epsilon)$. For $0 \leq \gamma < 2 + \delta$ and any $h^*$, the limit, as $n$ grows, of $E^{F_0} \left| \Delta(H_n, h^*, \epsilon) - E^{F_0} \Delta(H_n, h^*, \epsilon) \right|^\gamma$ is $E^{F_0} \left| \Delta(H, h^*, \epsilon) - E^{F_0} \Delta(H, h^*, \epsilon) \right|^\gamma$. Thus, for $\gamma < 2 + \delta$, the bounded convergence theorem gives us

$$
\lim_n E^{H} E^{F_0} \left| \Delta(H_n, h^*, \epsilon) - E^{F_0} \Delta(H_n, h^*, \epsilon) \right|^\gamma = E^{H} E^{F_0} \left| \Delta(H, h^*, \epsilon) - E^{F_0} \Delta(H, h^*, \epsilon) \right|^\gamma
$$

(10)

Again applying the bounded convergence theorem and relation (9) with $\gamma = 2 + \delta$, we find that $E^{F_0} \left| \Delta(H_n, h^*, \epsilon) - E^{F_0} \Delta(H_n, h^*, \epsilon) \right|^{\gamma'}$ is a bounded and continuous function of $h^*$ for $\gamma' < 2 + \delta$. Fixing $n$, we obtain for $\gamma' < 2 + \delta$,

$$
\lim_j E^{H_j} E^{F_0} \left| \Delta(H_n, h^*, \epsilon) - E^{F_0} \Delta(H_n, h^*, \epsilon) \right|^{\gamma'} = E^{H} E^{F_0} \left| \Delta(H_n, h^*, \epsilon) - E^{F_0} \Delta(H_n, h^*, \epsilon) \right|^{\gamma'}
$$

(11)

where the outer expectation runs over $h^*$. Equations (10) and (11) give us the stated result. ■

**Proof of Theorem 4:** The $\Delta_{n,k}$ form a triangular array. To obtain the central limit theorem, we verify the Lyapounov condition. Consider

$$
R_n = \frac{\frac{1}{n^{\delta/2}} \sum_{k=1}^{n} E^{F_0} \left| \Delta_{n,k} - E^{F_0}(\Delta_{n,k}) \right|^{2+\delta}}{\left( \frac{1}{n} \sum_{k=1}^{n} Var^{F_0}(\Delta_{n,k}) \right)^{1+\delta/2}} = \frac{1}{n^{\delta/2}} \cdot \frac{\frac{1}{n} \sum_{k=1}^{n} E^{F_0} \left| \Delta_{n,k} - E^{F_0}(\Delta_{n,k}) \right|^{2+\delta}}{\left( \frac{1}{n} \sum_{k=1}^{n} Var^{F_0}(\Delta_{n,k}) \right)^{1+\delta/2}}
$$

(12)

with $\delta$ defined as in assumption (B) of Theorem 4. Lemma 2 implies that the numerator and denominator of $n^{\delta/2} R_n$ have finite limits. Thus $R_n$ has the limit zero, verifying the Lyapounov condition. Writing the mean and variance of $\Gamma_n$ as $M_n$ and
$V_n$, this implies
\[ \sqrt{n} \left( \frac{\Gamma_n - M_n}{\sqrt{n}V_n} \right) \Rightarrow N(0, 1). \]

Now $nV_n = \frac{1}{n} \sum_{k=1}^{n} Var^{F_0}(\Delta_{n,k})$ which tends to $E^H Var^{F_0} \Delta(H, h, \epsilon)$, by Lemma 2.

Writing this quantity as $\sigma^2$, we have that
\[ \sqrt{n} (\Gamma_n - M_n) \Rightarrow N(0, \sigma^2). \]

Finally, also by Lemma 2, $M_n = \frac{1}{n} \sum_{k=1}^{n} E^{F_0}(\Delta_{n,k})$ has the limit $E^H E^{F_0} \Delta(H, h, \epsilon)$, which equals $KL(F_0, G_0) - KL(F_1, G_1)$. 

6 REFERENCES


