Bayesian Analysis (2016) **11**, Number 4, pp. 957–975

Bayes Factors for Smoothing Spline ANOVA

Chin-I. Cheng* and Paul L. Speckman†

Abstract. This paper describes an approach for variable selection and hypothesis testing in semiparametric additive models using Bayes factors in smoothing spline analysis of variance (SSANOVA) models. Effects can be linear or nonparametric (i.e., smooth or interactions between selected linear and smooth effects). To evaluate the importance of each term in the model, we develop Bayes factors for both linear and nonparametric terms. We compute approximate Bayes factors by Monte Carlo and Laplace integration. These Bayes factors can be computed to compare any two sub-models including one model nested in another. This permits formal tests of any portion or simultaneous portions of an SSANOVA model. We demonstrate this approach with an example.

Keywords: smoothing spline ANOVA, Bayes factor, laplace integration, reproducing kernel, semiparametric model.

1 Introduction

The introduction of semiparametric additive models (e.g., Speckman, 1988; Hastie and Tibshirani, 1990; Ruppert et al., 2003) has broadened the applications of regression to allow users to better describe real situations. One such approach is the smoothing spline ANOVA (SSANOVA) model, introduced by Wahba et al. (1995). In the basic context of linear models, SSANOVA provides a decomposition of the model space as a tensor sum of inner product spaces. This decomposition generalizes the usual linear model, in particular, analysis of covariance models, by allowing for nonparametric (i.e., "smooth") components as well as discrete and continuous factors and interactions between all types of factors. The result is a decomposition into main and interaction effects that provides not only the flexibility to fit complex models but also makes it natural to select a parsimonious model from a large class of semiparametric additive models. From a Bayesian perspective, the SSANOVA decomposition of a model tends to create terms whose posterior correlation is low, facilitating Bayesian computation. Almost all of the literature on additive models has been devoted to estimation. The purpose of this article is to describe Bayesian methodology for implementing Bayes factors, which can be used to compare models or applied to nested models to formally test any portion of the model. This paper complements Cheng and Speckman (2012), which introduced a Bayesian framework for SSANOVA. We indicate how the priors may be modified for implementing Bayes factors, and we present computational algorithms.

While the literature on estimation in additive models is extensive, much less attention has been given to testing. In the context of smoothing splines, the test of whether a specific continuous effect can be parametric or needs a nonparametric model has received the most research. Cox et al. (1988) proposed a locally most powerful test for

*Central Michigan University, cheng3c@cmich.edu
†University of Missouri-Columbia, speckmanp@missouri.edu

 DOI: 10.1214/15-BA974

generalized smoothing spline and partial spline models, and Wahba (1990) derived the generalized maximum likelihood ratio (GML) test and the generalized cross-validation (GCV) test to examine the smooth effect in a model. These tests have also been applied in the context of smoothing splines or SSANOVA for linear and generalized linear models by Guo (2002), Liu and Wang (2004) and Liu et al. (2005) among others. There is a considerable literature on lack-of-fit tests in the context of nonparametric regression; see, for example, Eubank et al. (1995), Bowman and Azzalini (1997), and Hart (1997). But the literature on general tests is sparse. Hastie and Tibshirani (1990) introduced approximate F-tests, and Fan and Jiang (2005) extended generalized likelihood ratio tests to certain classes of additive models using the backfitting estimator. Crainiceanu and Ruppert (2004) consider semiparametric additive models in the context of mixed linear models using P-splines. They developed test statistics for models with one variance component. In the context considered here, that corresponds to one nonparametric term. Finally, Kim (2013) has proposed bootstrap tests based on likelihood ratio and F-type statistics in a general context for SSANOVA models with a mixed model representation. Our results give a Bayesian analogue for such methods and help fill a gap in the literature of testing in SSANOVA models.

The methods presented here build on the Bayesian approach to SSANOVA models in Cheng and Speckman (2012). The work is comparable to Reich et al. (2009), whose focus in SSANOVA models was on prediction through Bayesian model averaging. The priors proposed in Cheng and Speckman (2012), based on classical frequentist SSANOVA models, are closely related to priors often recommended for Bayes factors in linear models. While noninformative priors are not possible for terms under consideration in Bayes factors, there is considerable interest in so-called "objective" priors suitable for linear models, especially priors related to the g-priors of Zellner and Siow (1980). Liang et al. (2008) discussed the theoretical properties of the mixture g-priors used in Bayes factors. Moreover, Shi and Dunson (2011), Morey et al. (2011) and Lodewyckx et al. (2011) provided alternatives to compute the Bayes factor efficiently.

Our work is also comparable to Basu and Chib (2003), Chib and Jeliazkov (2006) and Chib and Greenberg (2007), who all use Bayes factors for model selection in semiparametric models. We note that SSANOVA is based on reproducing kernels, so our approach does not depend on the number and location of knots as in Chib and Greenberg (2010).

In Section 2, the fully Bayesian SSANOVA model in Cheng and Speckman (2012) is briefly reviewed for completeness. Readers familiar with the model can skip to Section 2.5, which describes some necessary changes between the priors used for estimation and testing. Section 3 is devoted to the derivation and evaluation of Bayes factors. The method is illustrated by a reanalysis of the potassium measurement on dogs data in Section 4.

2 Bayesian smoothing spline ANOVA

In this section, we briefly review the smoothing spline ANOVA model, some useful reproducing kernels, and the Bayesian smoothing spline ANOVA model introduced by

Cheng and Speckman (2012). Readers interested in full details of the model should refer to that paper. In this section, we focus on providing a rationale for priors suitable for implementing Bayes factors and some subtle differences between testing and estimation.

2.1 Bayesian smoothing splines

Consider first the nonparametric regression model with one independent variable x_i and response variable y_i,

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where without loss of generality $x_i \in [0, 1]$. Assume further that the ε_i are independent $N(0, \delta)$ random variables and that f is an unknown smooth function. The smoothing spline $\widehat{f}$ is the minimizer of the penalized least squares equation,

$$\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_0^1 (f^{(m)}(x))^2 dx, \tag{1}$$

where $f^{(m)} = d^m f / dx^m$, λ is a smoothing parameter, and minimization is taken over the Sobolev space $\mathcal{H} = \{f : f, \ldots, f^{(m-1)}$ are absolutely continuous, $f^{(m)} \in L_2\}$. The exact solution, which can be found, for example, in Wahba (1990), can be derived using reproducing kernel Hilbert spaces. Briefly, it can be shown that $\widehat{f} \in \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ and $\mathcal{H}_1$ are Hilbert spaces, $\mathcal{H}_0$ is the null space of functions $\{f \in \mathcal{H} : f^{(m)} \equiv 0\}$, and $\mathcal{H}_1$ is the orthogonal complement of $\mathcal{H}_0$ in $\mathcal{H}$. If $\mathcal{H}_1$ has reproducing kernel $R_1(x_1, x_2)$, let

$$\boldsymbol{\Sigma} = [R_1(x_i, x_j)]_{n \times n}.$$

Suppose $\{\phi_1(x), \ldots, \phi_m(x)\}$ span the null space $\mathcal{H}_0$, and define $\boldsymbol{T} = [\phi_i(x_j)]_{n \times m}$. It can be shown that $\widehat{f}(x) = \sum_{i=1}^{m} d_i \phi_i(x) + \sum_{i=1}^{n} c_i R(x_i, x)$ for constants d_i and c_i. In fact, with $\boldsymbol{d} = (d_1, \ldots, d_m)'$ $\boldsymbol{c} = (c_1, \ldots, c_n)'$, and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, (1) can be written as $\|\boldsymbol{y} - \boldsymbol{T}\boldsymbol{d} - \boldsymbol{\Sigma}\boldsymbol{c}\|^2 + \lambda \boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c}$, so the minimizer of (1) is obtained by solving

$$\min_{\boldsymbol{c},\boldsymbol{d}} \|\boldsymbol{y} - \boldsymbol{T}\boldsymbol{d} - \boldsymbol{\Sigma}\boldsymbol{c}\|^2 + \lambda \boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c}.$$

Following Cheng and Speckman (2012), let $\boldsymbol{g} = \boldsymbol{\Sigma}\boldsymbol{c}$, so the penalized sum of squares is

$$\min_{\boldsymbol{d},\boldsymbol{g}} \|\boldsymbol{y} - \boldsymbol{T}\boldsymbol{d} - \boldsymbol{g}\|^2 + \lambda \boldsymbol{g}'\boldsymbol{\Sigma}^{-}\boldsymbol{g} \tag{2}$$

for any generalized inverse $\boldsymbol{\Sigma}^{-}$. If $(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{g}})$ solves (2), the minimizer of (1) satisfies

$$\widehat{\boldsymbol{f}} = \boldsymbol{T}\widehat{\boldsymbol{d}} + \widehat{\boldsymbol{g}}, \tag{3}$$

where $\widehat{\boldsymbol{f}} = (\widehat{f}(x_1), \ldots, \widehat{f}(x_n))'$. Although $\widehat{\boldsymbol{f}}$ is overparameterized in this formulation, the solution to (2) is unique. In this notation, smoothing splines have an obvious Bayesian interpretation. With the normal errors model and partially improper priors

$$\begin{aligned} \boldsymbol{y} \mid \boldsymbol{d}, \boldsymbol{g} &\sim N(\boldsymbol{T}\boldsymbol{d} + \boldsymbol{g}, \delta \boldsymbol{I}), \\ [\boldsymbol{d}] &\propto 1, \quad \boldsymbol{g} \sim N(\boldsymbol{0}, (\delta/\lambda)\boldsymbol{\Sigma}), \end{aligned} \tag{4}$$

$\hat{\boldsymbol{f}}$ in (3) is exactly the posterior mode. We suggest an improper prior, the Jeffrey's prior on δ, $[\delta] \propto 1/\delta$, $\delta > 0$.

The parameter λ, known as the "smoothing parameter" in the spline literature, controls the degree of smoothing in fitting the spline. It must be chosen judiciously to get a good fit. In frequentist applications, automatic methods such as GCV (Golub et al., 1979) are generally used. For a Bayesian, the solution is a suitable prior on λ. The Jeffreys prior on δ implies that the prior on the λ must be proper (see Sun and Speckman, 2008). Note also that λ is comparable to the parameter $1/g$ in the Zellner g-prior (Zellner, 1986) for linear models. We propose either scaled χ^2 priors following Zellner and Siow (1980) or scaled Pareto priors following Liang et al. (2008),

$$[\lambda \mid b] = \frac{\lambda^{-1/2} e^{-\lambda/(2b)}}{\Gamma(1/2)(2b)^{1/2}}, \quad \lambda > 0, \tag{5}$$

or

$$[\lambda \mid b] = \frac{b}{(b+\lambda)^2}, \quad \lambda > 0. \tag{6}$$

We believe this form of prior is suitable for both testing and estimation. In both cases, the choice of scale parameter b is necessarily subjective, but we argue that in practical problems the choice is not overly prescriptive, and it's relatively easy to elicit a reasonable prior. We follow White (2006) and Cheng and Speckman (2012) and recommend a procedure based on the "equivalent degrees of freedom" (Hastie and Tibshirani, 1990) of the smoother matrix associated with that term. It is well known that $\hat{\boldsymbol{f}} = \boldsymbol{S}_\lambda \boldsymbol{y}$ for a suitable matrix $\boldsymbol{S}_\lambda$. In Cheng and Speckman (2012), it is shown that

$$\boldsymbol{S}_\lambda = \boldsymbol{O}_1\boldsymbol{O}_1' + \boldsymbol{O}_2(\boldsymbol{I}_{(n-m)\times(n-m)} + \lambda(\boldsymbol{O}_2'\boldsymbol{\Sigma}\boldsymbol{O}_2)^-)^{-1}\boldsymbol{O}_2',$$

where $\boldsymbol{O} = (\boldsymbol{O}_1, \boldsymbol{O}_2)$ is an orthonormal matrix such that $\boldsymbol{O}_1$ spans the column space of $\boldsymbol{T}$. In ordinary linear regression, degrees of freedom is equal to the trace of the hat matrix. In analogy, Hastie and Tibshirani (1990) defined "equivalent degrees of fredom" to be

$$d(\lambda) = \mathrm{tr}(\boldsymbol{S}_\lambda).$$

It can be shown that $m < d(\lambda) < n$, where m is the dimension of the null space (the rank of $\boldsymbol{T}$). The interpretation is that a smoothing spline with $d(\lambda)$ equivalent degrees of freedom is roughly equivalent to linear regression using a parametric model with $d(\lambda)$ terms, e.g., a polynomial of degree $d(\lambda) - 1$.

We suggest choosing b so that the median of the distribution of $d(\lambda)$ under the prior on λ reflects the investigator's prior belief on the smoothness of that component, say d_0 equivalent degrees of freedom. Since $d(\lambda)$ is a monotonic function of λ, the median of the distribution of $d(\lambda)$ is $d(\tilde{m})$, where $\tilde{m}$ is the median of the prior on λ. We solve $d(\tilde{m}) = d_0$ for $\tilde{m}$ to obtain the median of the prior in (5) or (6). In practice, we generally take $m = 2$, so the null space is the space of linear functions. In many applications, having approximately five degrees of freedom represents reasonable prior belief. This corresponds to fitting a fourth degree polynomial. However, the choice of

prior equivalent degrees of freedom depends on the nature of the function being fitted. In a setting with low noise and prior belief that the function is quite complex, one may want to choose a much larger value of prior equivalent degrees of freedom. These ideas extend to higher dimensions such as fitting spatial components, where again a higher degree of prior complexity may be desired. As with any subjective prior, we recommend using a range of priors to ensure that the analysis is not overly dependent on the choice of a prior.

2.2 Additive models

Now extend the nonparametric regression model to have d independent variables, $x_1, \ldots, x_d$, which can be either continuous or discrete, and response variable y, so

$$y_i = f(x_{i1}, \ldots, x_{id}) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the ε_i are independent $N(0, \delta)$ random variables and f is an unknown function.

Following Wahba et al. (1995), the unknown mean function is decomposed as

$$\begin{aligned} E(y \mid x_1, x_2, \ldots, x_d) &= f(x_1, x_2, \ldots, x_d) \\ &= \mu + \sum_{k=1}^{d} f_k(x_k) + \sum_{i<j} f_{ij}(x_i, x_j) + \cdots, \end{aligned} \tag{7}$$

where $\mu = \mu(x_1, \ldots, x_d)$ represents linear effects in the null space that are not penalized, the f_k represent main effects, which can be linear effects or smooth effects, and the f_{ij} represent interaction effects, which can be between linear and linear effects, linear and smooth effects, or smooth and smooth effects. The result is an ANOVA-type decomposition and is called the smoothing spline ANOVA (SSANOVA) model. The SSANOVA decomposition creates interpretable main and interaction effects and provides flexibility for model fitting and variable selection. Assume p terms are retained in (7) and rewrite (7) as $f = \mu + f_1 + \cdots + f_p$. Each subspace in the decomposition has a reproducing kernel, say R_k, with corresponding matrix $\boldsymbol{\Sigma}_k = [R_k(x_{ik}, x_{jk})]$. We construct an $n \times m$ $(m < n)$ matrix $\boldsymbol{T}$ whose columns span the smallest space containing the null spaces of all the $\boldsymbol{\Sigma}_k$. With $\boldsymbol{g}_k = \boldsymbol{\Sigma}_k \boldsymbol{c}_k$ parameterizing the part of $\boldsymbol{f}_k$ not in the null space of $\boldsymbol{\Sigma}_k$, there is a vector additive representation of the model as

$$\boldsymbol{f} = \boldsymbol{T}\boldsymbol{d} + \sum_{k=1}^{p} \boldsymbol{g}_k.$$

The solution for the unknown function f in the SSANOVA model is obtained by solving

$$\min_{\boldsymbol{d}, \boldsymbol{g}_k} \left\| \boldsymbol{y} - \boldsymbol{T}\boldsymbol{d} - \sum_{k=1}^{p} \boldsymbol{g}_k \right\|^2 + \sum_{k=1}^{p} \lambda_k \boldsymbol{g}_k' \boldsymbol{\Sigma}_k^{-} \boldsymbol{g}_k. \tag{8}$$

For more detail, please refer to Gu (2002) or Cheng and Speckman (2012). The solution to (8) is again unique and has the form

$$\widehat{\boldsymbol{f}} = \boldsymbol{T}\widehat{\boldsymbol{d}} + \sum_{k=1}^{p} \widehat{\boldsymbol{g}}_k.$$

The Bayesian interpretation (4) carries over. With the normal errors model and independent priors,

$$\begin{aligned} \boldsymbol{y} \mid \boldsymbol{d}, \boldsymbol{g}_k, \delta &\sim N\left(\boldsymbol{T}\boldsymbol{d} + \sum_{k=1}^{p} \boldsymbol{g}_k, \delta \boldsymbol{I}\right), \\ [\boldsymbol{d}] &\propto 1, \\ \boldsymbol{g}_k \mid \lambda_k, \delta &\sim N(\boldsymbol{0}, (\delta/\lambda_k)\boldsymbol{\Sigma}_k), k = 1, \ldots, p, \end{aligned} \tag{9}$$

the SSANOVA solution $\hat{\boldsymbol{f}}$ is exactly the posterior mode.

2.3 Some reproducing kernels for SSANOVA

We must specify a reproducing kernel matrix $\boldsymbol{\Sigma}_k$ for each component in the model. We follow Gu (2002) for many useful ones. Throughout this section, we will denote *discrete* variables by t_1, t_2, ..., and *continuous* variables by x_1, x_2, ...

Discrete variable main effects. Suppose variable t_1 is a discrete factor with ℓ levels. The SSANOVA reproducing kernel is

$$R_D(t_1, t_2) = I_{\{t_1 = t_2\}} - \frac{1}{\ell},$$

where $I_{\{\cdot\}}$ is the indicator function. The corresponding reproducing kernel matrix for a discrete variable t_1 is defined as $\boldsymbol{\Sigma}_D = [\boldsymbol{\Sigma}_{ij}]_{n \times n}$, where

$$\boldsymbol{\Sigma}_{ij} = R_D(t_{i1}, t_{j1}) = I_{(t_{i1} = t_{j1})} - \frac{1}{\ell}. \tag{10}$$

This is equivalent to the following one-way ANOVA model. Suppose $y_i \sim N(\mu_0 + \sum_{j=1}^{\ell} x_{ij}\alpha_j, \delta)$, $i = 1, \ldots, n$, where, for each i, exactly one x_{ij} is one and the remaining are zero, $j = 1, \ldots, \ell$. Take the prior $\boldsymbol{\alpha} \sim N(\boldsymbol{0}, (\delta/\lambda)(\boldsymbol{I}_\ell - 1/\ell \boldsymbol{1}_\ell \boldsymbol{1}_\ell'))$. Then, with $\boldsymbol{X} = [x_{ij}]_{n \times \ell}$, the prior on effect $\boldsymbol{X}\boldsymbol{\alpha}$ is $N(\boldsymbol{0}, (\delta/\lambda)\boldsymbol{\Sigma}_D)$.

Continuous variable main effects. One special case is important for testing and model comparison. Consider an ordinary linear term of the form $\boldsymbol{g} = \boldsymbol{X}\boldsymbol{\beta}$, where we assume for simplicity that $\boldsymbol{X}$ is of full rank and $\boldsymbol{X}'\boldsymbol{1} = \boldsymbol{0}$. The reproducing kernel matrix for this term is

$$\boldsymbol{\Sigma}_L = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'. \tag{11}$$

If $\boldsymbol{\beta} \sim N(\boldsymbol{0}, (\delta/\lambda)(\boldsymbol{X}'\boldsymbol{X})^{-1})$, then $\boldsymbol{g} = \boldsymbol{X}\boldsymbol{\beta} \sim N(\boldsymbol{0}, (\delta/\lambda)\boldsymbol{\Sigma}_L)$. Thus this kernel for a linear term in continuous variables corresponds to a Zellner g-prior with g replaced by $1/\lambda$.

Nonparametric variable main effects. We will use thin-plate splines to model smooth effects in one or two dimensions. For $x, y \in \mathbb{R}^d$, $d = 1, 2, \ldots$, and $2m > d$, define the thin-plate spline semi-kernel

$$R_{STP}(x, y) = E(||x - y||),$$

where $||x-y||$ is Euclidean distance,

$$E(\mu) = \begin{cases} (-1)^{\frac{d}{2}+1+m}|\mu|^{2m-d}\log|\mu|, & d \text{ even}, \\ |\mu|^{2m-d}, & d \text{ odd}, \end{cases}$$

and m generalizes the order of the penalty term differential operator in (1) to higher dimensions. For $d = 1$ and $m = 2$, the result is a cubic spline on the real line. Because R_{STP} is not nonnegative definite, it is not a true kernel. However, it is nonnegative definite on the complement of the null space. Let $\phi_1, \ldots, \phi_m$ be a complete orthonormal system in $\mathcal{H}_0$ (i.e., polynomials of total degree $m-1$), and let P_0 be the projection operator onto $\mathcal{H}_0$ defined as $(P_0 f)(x) = \sum_{\nu=1}^{m}(f, \phi_\nu)_0 \phi_\nu(x)$. Then the genuine reproducing kernel is

$$R_{TP}(x, y) = (I - P_{0(x)})(I - P_{0(y)})E(||x - y||),$$

where I is the identity operator and $P_{0(x)}$ and $P_{0(y)}$ are the projection operators applied to the arguments x and y, respectively. In practice, we compute the matrix kernel used in (8) as follows. Again let $\boldsymbol{O} = (\boldsymbol{O}_1, \boldsymbol{O}_2)$ be an orthogonal matrix such that $\boldsymbol{O}_1$ spans the null space of $\boldsymbol{T}$. The genuine reproducing kernel matrix for the thin-plate spline is

$$\boldsymbol{\Sigma}_{TP} = \boldsymbol{O}_2\boldsymbol{O}_2'\boldsymbol{K}_{TP}\boldsymbol{O}_2\boldsymbol{O}_2',$$

where $\boldsymbol{K}_{TP} = [R_{TP}(x_i, x_j)]_{n\times n}$. (See Gu (2002) or Wang (2011) for details.)

For $m = 2$ and $d = 1$ (the usual cubic smoothing spline on the line), one can take $\phi_1(x) = 1$, $\phi_2(x) = x$, and the semi-kernel is $R_{STP}(x_i, x_j) = |x_i - x_j|^3$. For $m = 2$, $d = 2$ (the thin-plate spline on the plane) and $\boldsymbol{x} = (x_1, x_2)$, one can take $\phi_1(\boldsymbol{x}) = 1$, $\phi_2(\boldsymbol{x}) = x_1$, $\phi_3(\boldsymbol{x}) = x_2$, and $R_{STP}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \log\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$.

Interaction effects. In SSANOVA, an interaction effect belongs to a tensor product reproducing kernel Hilbert space. The products of reproducing kernels on the marginal domains form reproducing kernels on the product domain (see Gu, 2002). Thus the reproducing kernel for an interaction effect is the direct product of the reproducing kernels in the marginal spaces. For example, the reproducing kernel for the interaction between a thin-plate spline over $\boldsymbol{x} \in \mathbb{R}^d$ and a discrete factor indexed by t is

$$R_{TP,D}((\boldsymbol{x}_i, t_i), (\boldsymbol{x}_j, t_j)) = R_{TP}(\boldsymbol{x}_i, \boldsymbol{x}_j)R_D(t_i, t_j).$$

For two discrete spaces corresponding to variables t_1 and t_2, the interaction kernel is

$$\begin{aligned} R_{D,D}((t_{1i}, t_{2i}), (t_{1j}, t_{2j})) &= R_D(t_{i1}, t_{j1})R_D(t_{i2}, t_{j2}) \\ &= \left(I_{(t_{i1}=t_{j1})} - \frac{1}{\ell_1}\right) \times \left(I_{(t_{i2}=t_{j2})} - \frac{1}{\ell_2}\right), \end{aligned}$$

where ℓ_1 and ℓ_2 are the number of levels of t_1 and t_2, respectively. In computing interaction terms, the elements of an interaction kernel matrix are simply the products of the elements of the two main effects kernel matrices.

Many other reproducing kernels are possible. In general, the Bayesian SSANOVA framework presented here can be used with other reproducing kernels; see, for example, Wang (2011).

2.4 Priors for Bayesian SSANOVA

We follow the model of Cheng and Speckman (2012). Starting with the basic Bayesian additive model (9), we take the Jeffrey's prior on δ,

$$[\delta] \propto \frac{1}{\delta}, \quad \delta > 0. \tag{12}$$

which implies that the priors on the λ_k must be proper (see Sun and Speckman, 2008).

Much of the previous literature on Bayes factors in linear models begins with a Zellner–Siow kind of model for testing a single term (e.g., Zellner and Siow, 1980; Liang et al., 2008). More recently, Rouder et al. (2012) and Min and Sun (2015) have proposed more complicated ANOVA models with independent g-priors on many or all of the effects. The models here are in the same spirit. SSANOVA suggests a Bayesian framework with independent priors on the λ_k. We propose either scaled χ^2 priors (5) following Zellner and Siow (1980) or scaled Pareto priors (6) following Liang et al. (2008). The main problem is the choice of scale parameter b_k for the prior on λ_k.

For smooth terms, we use "effective degrees of freedom" as in Section 2.1. For effects terms for a discrete explanatory variable and kernel (10), we generally take $b_k = 1$. This follows the models in Rouder et al. (2012) and Min and Sun (2015). For linear regression terms, we use (11) with $b_k = 1/n$ following Zellner–Siow. In all cases, prior information can inform alternative choices for the b_k.

2.5 Testing vs. estimation

In mixed linear models, priors for testing and estimation may be different. It is customary to take flat priors on fixed effects and hierarchical priors on random effects for purposes of estimation (e.g., Gelman et al., 2014). However, when computing Bayes factors, priors must be proper on each effect except possibly for effects common to all models.

As seen in Cheng and Speckman (2012), the reproducing kernels from classical SSANOVA correspond to hierarchical Bayesian models. In particular, we followed the SSANOVA prescription as closely as possible, using a prior with covariance matrix (10) for both fixed and random effects. The prior for a main effects term corresponding to $\boldsymbol{\Sigma}_D$ is not a conventional prior for either fixed or random effects in standard Bayesian analysis since the coefficient vector $\boldsymbol{\alpha}$ is constrained to satisfy $\mathbf{1}'\boldsymbol{\alpha} = 0$. This is exactly the prior recommended by Rouder et al. (2012) for use in testing a fixed effects term. Moreover, in conventional Zellner–Siow models, the factor to be tested must satisfy $\mathbf{1}'\boldsymbol{X} = \mathbf{0}$. One solution for the main effects model is equivalent to the reproducing kernel matrix $\boldsymbol{\Sigma}_D$. One could model a random effects term $\boldsymbol{X}\boldsymbol{\alpha}$ with a standard conditionally independent prior $\boldsymbol{\alpha}_k \sim N(\mathbf{0}, (\delta/\lambda_k)\boldsymbol{I})$ in place of the restricted kernel $\boldsymbol{\Sigma}_D$. This has implications, especially for interaction terms, that we do not explore here.

The treatment in Cheng and Speckman (2012) modeled all terms in the SSANOVA model as random except for terms in the null space, which consists of a constant plus the null spaces from kernels corresponding to smooth terms. For example, for smoothing

in one dimension with a cubic spline using parameterization (3), the model has the form $f(x_i) = d_0 + d_1 x_i + g(x_i)$, where $g(x)$ is the nonparametric *smooth* part of the model and $d_1 x$ is the parametric part. For estimation, it is appropriate to put flat priors on d_0 and d_1. If one wants to test for the presence of the nonparametric part $g(x)$ of the model, a flat prior on d_0 and d_1 can be used. However, if one wants to test for the presence of an effect f including the linear term, then one must put a proper prior on d_1 as well. We use the Zellner–Siow prior with kernel (11) on the factor $d_1 x_i$ and prior (5) with $b_k = 1/n$ for models where we implement Bayes factors. Note that this prior may give poor performance for estimation. The shrinkage prior on the linear part of the model can introduce substantial of bias.

In retrospect, we believe that using $\boldsymbol{\Sigma}_D$ for a fixed ANOVA effect term in an estimation model as we did in Cheng and Speckman (2012) may also introduce unwanted bias. One could use a flat prior for a fixed effect as well. In practice, this means including such a term in the null space. For example, if the fixed effect term has the form $\boldsymbol{X}_k \boldsymbol{\beta}_k$ and the null space for the rest of the model is the span of $\boldsymbol{T}$, then the fixed effect could be included in the model by augmenting the null space to be $\boldsymbol{T}_1 = (\boldsymbol{T}, \boldsymbol{X}_k)$ with $\boldsymbol{d}_1' = (\boldsymbol{d}', \boldsymbol{\beta}_k')'$. Reparameterization may be necessary to insure that $\boldsymbol{T}_1$ has full rank.

3 Bayes factors

The significance of each term in the Bayes SSANOVA model will be evaluated by Bayes factors (Kass and Raftery, 1995). A Bayes factor is the ratio of marginal likelihoods for two competing models. With no closed form, we approximate the marginal likelihood by Monte Carlo or Laplace integration. The details are described in the following sections.

3.1 Marginal likelihood

The marginal likelihood function of $\boldsymbol{y}$ under model M_k is

$$p(\boldsymbol{y} \mid M_k) = \int_{\omega_k} f_k(\boldsymbol{y} \mid \omega_k) \pi_k(\omega_k)\, d\omega_k, \quad k = 1, 2,$$

where $f_k(\boldsymbol{y} \mid \omega_k)$ is the likelihood function of $\boldsymbol{y}$ given parameters ω_k, and $\pi_k(\omega_k)$ is the prior for ω_k. The Bayes factor (Kass and Raftery, 1995) between M_2 and M_1 is defined to be

$$BF_{21} = \frac{p(\boldsymbol{y} \mid M_2)}{p(\boldsymbol{y} \mid M_1)}. \tag{13}$$

Care must be in using improper priors with Bayes factors. Even if the posterior is proper, the normalizing constant for an improper prior is arbitrary, rendering (13) indeterminate. However, it is permissible to use an improper prior on one or more common terms in both models such as (9) and (12) here.

Consider now the Bayes SSANOVA model (9). Since $\boldsymbol{\Sigma}_k$ may be singular, we need a full rank parameterization. Let $\boldsymbol{\Sigma}_k = \boldsymbol{Q}_k \boldsymbol{D}_k \boldsymbol{Q}_k'$ be the spectral decomposition of $\boldsymbol{\Sigma}_k$,

where $\boldsymbol{Q}_k$ is a $n \times r_k$ suborthogonal matrix of eigenvectors corresponding to the r_k nonzero eigenvalues of $\boldsymbol{\Sigma}_k$, and $\boldsymbol{D}_k$ is the diagonal matrix with the nonzero eigenvalues in the diagonals. Setting $\boldsymbol{g}_k = \boldsymbol{Q}_k \boldsymbol{v}_k$,

$$\boldsymbol{v}_k \mid \delta, \lambda_k \quad \sim \quad N_{r_k}\left(0, \frac{\delta}{\lambda_k}\boldsymbol{D}_k\right), \quad k = 1, \ldots, p.$$

To write the likelihood function in matrix format, let $r = \sum_{k=1}^{p} r_k$ and define $\boldsymbol{X} = [\boldsymbol{Q}_1, \boldsymbol{Q}_2, \ldots, \boldsymbol{Q}_p]$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)$ and $\boldsymbol{\beta} = [\boldsymbol{v}_1', \boldsymbol{v}_2', \ldots, \boldsymbol{v}_p']'$. Thus $\boldsymbol{X}$ is an $n \times r$ matrix. The likelihood function for the Bayesian SSANOVA model is

$$h(\boldsymbol{d}, \boldsymbol{\beta}, \delta, \boldsymbol{\lambda} | \boldsymbol{y}) = \frac{\exp(-\frac{1}{2\delta}||\boldsymbol{y} - \boldsymbol{T}\boldsymbol{d} - \boldsymbol{X}\boldsymbol{\beta}||^2)}{(2\pi)^{n/2}\delta^{n/2}} \frac{\exp(-\frac{1}{2\delta}\boldsymbol{\beta}'\boldsymbol{D}^{-1}(\boldsymbol{\lambda})\boldsymbol{\beta})}{(2\pi)^{r/2}\delta^{r/2}|\boldsymbol{D}(\boldsymbol{\lambda})|^{1/2}} \frac{1}{\delta}\pi(\boldsymbol{\lambda}),$$

where

$$\boldsymbol{D}(\boldsymbol{\lambda}) = \begin{pmatrix} \lambda_1^{-1}\boldsymbol{D}_1 & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \lambda_2^{-1}\boldsymbol{D}_2 & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \lambda_p^{-1}\boldsymbol{D}_p \end{pmatrix},$$

and $\pi(\boldsymbol{\lambda})$ is the prior for $\boldsymbol{\lambda}$. To compute the Bayes factor in (13), we need the marginal likelihood for $\boldsymbol{y}$ under M_2 and M_1. One advantage of the Zellner–Siow prior in linear models is the computational efficiency due to the conditionally conjugate prior. Our model is closely related to the linear model structure. With $\boldsymbol{O}_2$ defined as before as projection onto the complement of the range of $\boldsymbol{T}$, define $\widetilde{\boldsymbol{y}} = \boldsymbol{O}_2'\boldsymbol{y}$, $\widetilde{\boldsymbol{X}} = \boldsymbol{O}_2'\boldsymbol{X}$ and $\boldsymbol{W}(\boldsymbol{\lambda}) = \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}} + \boldsymbol{D}^{-1}(\boldsymbol{\lambda})$. With this notation, integrating out $\boldsymbol{d}$, $\boldsymbol{\beta}$ and δ in $h(\boldsymbol{d}, \boldsymbol{\beta}, \delta, \boldsymbol{\lambda}|\boldsymbol{y})$ yields

$$\begin{aligned} h_1(\boldsymbol{\lambda} \mid \boldsymbol{y}) &= \int h(\boldsymbol{d}, \boldsymbol{\beta}, \delta, \boldsymbol{\lambda} \mid \boldsymbol{y})\, d\boldsymbol{d}\, d\boldsymbol{\beta}\, d\delta \\ &= z(\boldsymbol{\lambda} \mid \boldsymbol{y})\pi(\boldsymbol{\lambda}), \end{aligned}$$

where

$$\begin{aligned} z(\boldsymbol{\lambda} \mid \boldsymbol{y}) &= |(\boldsymbol{T}'\boldsymbol{T})|^{-1/2}|\boldsymbol{W}(\boldsymbol{\lambda})|^{-1/2}|\boldsymbol{D}(\boldsymbol{\lambda})|^{-1/2} \\ &\quad \times \frac{\boldsymbol{\Gamma}(\frac{n-m}{2})}{[\pi(\widetilde{\boldsymbol{y}}'\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{y}}'\widetilde{\boldsymbol{X}}\boldsymbol{W}^{-1}(\boldsymbol{\lambda})\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{y}})]^{(n-m)/2}}. \end{aligned} \tag{14}$$

There is an alternate form which is useful if $\widetilde{\boldsymbol{X}}$ has many more columns than rows. Using the identity $|\boldsymbol{A} + \boldsymbol{X}\boldsymbol{B}\boldsymbol{X}'| = |\boldsymbol{B}||\boldsymbol{A}||\boldsymbol{B}^{-1} + \boldsymbol{X}'\boldsymbol{A}^{-1}\boldsymbol{X}|$, we have $|\boldsymbol{W}(\boldsymbol{\lambda})||\boldsymbol{D}(\boldsymbol{\lambda})| = |\boldsymbol{I} + \widetilde{\boldsymbol{X}}\boldsymbol{D}(\boldsymbol{\lambda})\widetilde{\boldsymbol{X}}'|$. Similarly, the Sherman–Morrison–Woodbury formula implies $\boldsymbol{I} - \widetilde{\boldsymbol{X}}(\boldsymbol{D}(\boldsymbol{\lambda})^{-1} + \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}' = (\boldsymbol{I} + \widetilde{\boldsymbol{X}}\boldsymbol{D}(\boldsymbol{\lambda})\widetilde{\boldsymbol{X}}')^{-1}$, hence

$$\widetilde{\boldsymbol{y}}'\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{y}}'\widetilde{\boldsymbol{X}}\boldsymbol{W}^{-1}(\boldsymbol{\lambda})\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{y}}'[\boldsymbol{I} + \widetilde{\boldsymbol{X}}\boldsymbol{D}(\boldsymbol{\lambda})\widetilde{\boldsymbol{X}}']^{-1}\widetilde{\boldsymbol{y}}.$$

With $\boldsymbol{M}(\boldsymbol{\lambda}) = (\boldsymbol{I}_{n-m} + \widetilde{\boldsymbol{X}}\boldsymbol{D}(\boldsymbol{\lambda})\widetilde{\boldsymbol{X}}')$, we have

$$z(\boldsymbol{\lambda} \mid \boldsymbol{y}) \quad = \quad \frac{\boldsymbol{\Gamma}(\frac{n-m}{2})}{|(\boldsymbol{T}'\boldsymbol{T})|^{1/2}|\boldsymbol{M}(\boldsymbol{\lambda})|^{1/2}[\pi(\widetilde{\boldsymbol{y}}'\boldsymbol{M}^{-1}(\boldsymbol{\lambda})\widetilde{\boldsymbol{y}})]^{(n-m)/2}}. \tag{15}$$

3.2 Computing Bayes factors

Exact analytical evaluation of $h_2(\boldsymbol{y}) = \int h_1(\boldsymbol{\lambda} \mid \boldsymbol{y})\, d\boldsymbol{\lambda}$ is not possible. Evaluation by quadrature is possible if $\boldsymbol{\lambda}$ has dimension 1 or 2. One could use an MCMC method such as applying the Savage–Dickey density ratio (Dickey, 1971; Verdinelli and Wasserman, 1995) or bridge sampling (Meng and Wong, 1996). However, simpler methods appear to give excellent results, and it's much quicker to compute the Bayes factors directly than to run complete MCMC simulations for each model under consideration. Moreover, there are cases where the priors may differ between model selection and estimation, so there is no point in running an MCMC simulation. We approximate $h_2(\boldsymbol{y})$ in the example below by Monte Carlo or Laplace integration.

Following Kass and Raftery (1995), one method to approximate the marginal likelihood $h_2(\boldsymbol{y})$ is Laplace approximation. To improve this approximation, take the transformation, $\lambda_k = \exp(\phi_k), k = 1, \ldots, p$, with Jacobian $e^{\phi_1 + \cdots + \phi_p}$, and let $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_p)$. With the notation $e^{\boldsymbol{\phi}} = (e^{\phi_1}, \ldots, e^{\phi_p})$, define

$$s(\boldsymbol{\phi}) = \log\left[z(e^{\boldsymbol{\phi}} \mid \boldsymbol{y})\pi(e^{\boldsymbol{\phi}}) \exp(\phi_1 + \cdots + \phi_p)\right].$$

Then the marginal likelihood is $h_2(\boldsymbol{y}) = \int \exp(s(\boldsymbol{\phi}))\, d\boldsymbol{\phi}$. Let $\widehat{\boldsymbol{\phi}}$ maximize $s(\boldsymbol{\phi})$, and let $\boldsymbol{H}$ be the Hessian matrix of $s(\boldsymbol{\phi})$ evaluated at $\widehat{\boldsymbol{\phi}}$. Replacing $s(\boldsymbol{\phi})$ by the second order Taylor expansion

$$s(\boldsymbol{\phi}) \doteq s(\widehat{\boldsymbol{\phi}}) + \frac{1}{2}(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})'\boldsymbol{H}(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}}),$$

the Laplace approximation is

$$\begin{aligned} h_2(\boldsymbol{y}) &= \int \exp(s(\boldsymbol{\phi}))d\boldsymbol{\phi} \\ &\doteq \exp(s(\widehat{\boldsymbol{\phi}}))(2\pi)^{p/2}|\boldsymbol{H}|^{-1/2}. \end{aligned}$$

This approximation works well with moderate to large samples when the posterior distribution of $\boldsymbol{\lambda}$ has small variance.

Perhaps the simplest approximation is naive Monte Carlo (Kass and Raftery, 1995). If we generate a random sample $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}, \ldots, \boldsymbol{\lambda}^{(N)}$ from the prior, $\pi(\boldsymbol{\lambda})$, then the marginal likelihood is approximated by

$$h_2(\boldsymbol{y}) = \int z(\boldsymbol{\lambda} \mid \boldsymbol{y})\pi(\boldsymbol{\lambda})\, d\boldsymbol{\lambda} \doteq \frac{1}{N}\sum_{j=1}^{N} z(\boldsymbol{\lambda}^{(j)} \mid \boldsymbol{y}).$$

This approach is easy to compute. Either (14) or (15) can be used to evaluate $z(\boldsymbol{\lambda} \mid \boldsymbol{y})$. The former is more efficient if $\widetilde{\boldsymbol{X}}$ has more rows than columns. If $\widetilde{\boldsymbol{X}}$ has substantially more columns than rows, then (15) is preferable. However, if the posterior distribution for $\boldsymbol{\lambda}$ is highly concentrated in one or more coordinates, this form of Monte Carlo approximation is not very efficient.

The precision of Monte Carlo can often be improved by importance sampling (e.g., Gelman et al., 2014). We implement importance sampling in the following way. After reparameterization, the Laplace approximation suggests that the posterior of $\boldsymbol{\phi}$ is ap-

proximately $\mathrm{N}(\widehat{\boldsymbol{\phi}}, -\boldsymbol{H}^{-1})$. Let $\boldsymbol{\psi} \sim \mathrm{N}(\widehat{\boldsymbol{\phi}}, -\boldsymbol{H}^{-1})$, and let $n(\boldsymbol{\psi} \mid \widehat{\boldsymbol{\phi}}, -\boldsymbol{H}^{-1})$ denote the corresponding multivariate normal density. Then

$$h_2(\boldsymbol{y}) = \int \exp(s(\boldsymbol{\phi}))\, d\boldsymbol{\phi} = E_{\boldsymbol{\psi}} \left(\frac{\exp(s(\boldsymbol{\psi}))}{n(\boldsymbol{\psi} \mid \widehat{\boldsymbol{\phi}}, -\boldsymbol{H}^{-1})} \right).$$

(Here $n(\boldsymbol{\psi} \mid \widehat{\boldsymbol{\phi}}, -\boldsymbol{H}^{-1})$ is the importance density.) Thus if $\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(N)}$ is a random sample from $\boldsymbol{\psi}$,

$$h_2(\boldsymbol{y}) \doteq \frac{1}{N} \sum_{j=1}^{N} \frac{\exp(s(\boldsymbol{\psi}^{(j)}))}{n(\boldsymbol{\psi}^{(j)} \mid \widehat{\boldsymbol{\phi}}, \boldsymbol{H})}.$$

One advantage of Monte Carlo over Laplace approximation is that one can easily obtain the standard error of the Monte Carlo approximation to the marginal likelihood. Moreover, the delta method yields the approximate standard error for the Bayes factor. These approximations and standard errors are illustrated in the next example.

4 Application: potassium measurement on dogs

This data set was originally given in Grizzle and Allen (1969) and has been estimated with SSANOVA by Wang and Ke (2004) and Bayesian SSANOVA by Cheng and Speckman (2012). The data consists of measurements of coronary sinus potassium concentrations for 36 dogs that were assigned to four groups: control, extrinsic cardiac denervation three weeks prior to coronary occlusion, extrinsic cardiac denervation immediately prior to coronary occlusion, and bilateral thoratic sympathectomy and stellectomy three weeks prior to coronary occlusion. Coronary sinus potassium concentrations were measured on each dog every two minutes from 1 to 13 minutes after occlusion. The variable group is coded as 1 to 4 to represent the 4 groups, and dog is numbered 1 to 36. The variable time is transformed into $[0, 1]$. The design has the dog factor nested within the group factor. There are 9 dogs in each group. The goal is to examine the significance of the two categorical variables, group and dog, the continuous variable, time, and the interactions among those variables to the response variable, potassium, coronary sinus potassium concentrations level. Here we extend the work of Cheng and Speckman (2012) to model selection with Bayes factors.

Based on 95% credible sets, Cheng and Speckman (2012) informally concluded that the following terms should be included in the model: the smooth effect $s_3(time)$, the level effects $l_1(group)$ and $l_2(dog)$, and the interaction effects $l_{13}(group, time)$, $l_{23}(dog, time)$ and $ls_{23}(dog, time)$ to potassium concentrations level. Here $l_{13}(group, time)$ denotes the interaction between group and the linear time effect, $l_{23}(dog, time)$ denotes the interaction between dog and the linear time effect, and $ls_{23}(dog, time)$ denotes the interaction between dog and the smooth effect in time. In addition to those effects, Wang and Ke (2004) included the interaction between group and the smooth effect in time, $ls_{13}(group, time)$, in their final model based on AIC. As pointed out by Cheng and Speckman (2012), Wang and Ke (2004) derived their best model using AIC without any kind of testing procedure for individual terms. Note that AIC is generally believed

to be too liberal and may include a large proportion of irrelevant variables (e.g., George, 2000).

We conducted a formal model selection procedure using Bayes factors. The following nested sequence of models was considered:

$$\begin{aligned}
\text{Model 5} \quad &: \quad f(group, dog, time) = \mu + \beta_3 time + s_3(time) + l_1(group) \\
&\quad +l_2(dog) + l_{23}(dog, time) + l_{13}(group, time) \\
&\quad +ls_{23}(dog, time) + ls_{13}(group, time), \\
\text{Model 4} \quad &: \quad \text{Model 5} - ls_{13}(group, time), \\
\text{Model 3} \quad &: \quad \text{Model 4} - ls_{23}(dog, time), \\
\text{Model 2} \quad &: \quad \text{Model 4} - l_{13}(group, time), \\
\text{Model 1} \quad &: \quad \text{Model 2} - l_1(group).
\end{aligned}$$

Following Section 2.4, Zellner–Siow priors corresponding to (11) and (5) with $b = 1/n$ were used for the interaction between level and linear effects, $l_{13}(group, time)$ and $l_{23}(dog, time)$. We chose Gamma $(1/2, 2)$ priors for the level effects and Gamma $(1/2, 2b_k)$ priors for the smooth effects and the interaction effects involved curvature effects. As described in Section 2.1, the scale parameters b_k were selected to achieve desired median *a priori* effective degrees of freedom. We chose to set median prior effective degrees of freedom to approximately 3, 4 and 36 for $s_3(time)$, $ls_{13}(group, time)$ and $ls_{23}(dog, time)$, respectively, as suggested by Cheng and Speckman (2012). The corresponding b_k are 0.4241, 1.1944, and 0.2391.

The Bayes factors to evaluate the significance of each term in model 5 as approximated by importance sampling Monte Carlo and Laplace approximation are listed in Table 1. For this example, importance sampling was substantially much more efficient than basic Monte Carlo; results shown are based on a sample of 5,000. The Laplace approximations are quite good except for BF_{51}. However, this discrepancy did not affect our conclusions. Following the guidelines of Kass and Raftery (1995), there is overwhelming evidence for the dog–time nonparametric interaction effect $ls_{23}(dog, time)$. The evidence for nonparametric group–time interaction effect $ls_{13}(group, time)$ is rather weak, but there is some evidence for the linear interaction effect $l_{13}(group, time)$. Following customary practice in linear models, we did not examine main effects or lower order interaction effects included by significant interaction effects. However, to illustrate the method, we did test for an overall group effect by computing the Bayes factor comparing model 5 with model 1. The resulting $BF \approx 47.6$ is very strong evidence for a *group* effect. Note that the Bayes factor for the nonparametric group–time effect $ls_{13}(group, time)$ only gives very weak support. However, this term was included in the final model of Wang and Ke (2004).

To study the robustness of the analysis with respect to choice of prior, we reran the analysis using three alternate sets of priors. Following the discussion in Section 2.5, we put the Zellner–Siow prior rather than the flat prior on the linear term in *time*. As seen in Table 2, the effect is completely negligible. Next, we experimented with different equivalent degrees of freedom (edf) for the priors on the smooth terms $s_3(time)$, $ls_{13}(group, time)$ and $ls_{23}(dog, time)$. Note that the data set has only 7 time points. With cubic spline smoothing, there are only 5 degrees of freedom possible for the main

	Tested terms	Bayes Factors MC estimate (SE)	Laplace approx.
BF_{54}	$ls_{13}(group, time)$	3.11(0.07)	2.59
BF_{43}	$ls_{23}(dog, time)$	$1.1 \times 10^6(1.5 \times 10^4)$	1.1×10^6
BF_{42}	$l_{13}(group, time)$	10.05(0.11)	9.79
BF_{51}	$(ls_{13}(group, time)$ $+l_{13}(group, time)+l_1(group))$	47.60(1.08)	37.67

Table 1: The Bayes factors (BF) for testing the terms $ls_{13}(group, time)$, $ls_{23}(dog, time)$, $l_{13}(group, time)$ and $ls_{13}(group, time) + l_{13}(group, time) + l_1(group)$ for the potassium measurement on dogs data set. The Monte Carlo estimates, computed by importance sampling, include approximate standard errors (SE).

	Bayes Factors Z–S prior	(1, 2, 18) edf	(4, 12, 70) edf
BF_{54}	3.11	1.98	2.03
BF_{43}	1.1×10^6	4.27×10^5	2.35×10^6
BF_{42}	9.52	9.51	9.51
BF_{51}	45.17	29.03	28.96

Table 2: Robustness study for Bayes factors in Table 1: Column 1, Zellner–Siow prior on the linear term for *time*; Column 2, (1, 2, 18) edf priors; Column 3, (4, 12, 70) edf priors.

effect smooth part. Assuming very smooth prior belief, we tried (1, 2, 18) prior edf, respectively. Finally, to try a prior assuming complex relationships, we tried (4, 12, 70) edf, respectively. The results are also presented in Table 2. The choice of prior does affect certain Bayes factors, especially BF_{51} testing for a group effect. In our opinion, these results demonstrate that the Bayes factor analysis is quite satisfactory, especially if one believes that the extreme choices are not very realistic.

Another alternative for evaluating the significance of each term is the deviance information criteria (DIC) (Spiegelhalter et al., 2002). For comparison, Table 3 lists the terms selected based on smaller DIC. The preferred model contains $ls_{23}(dog, time)$, $ls_{13}(group, time)$ and all lower order terms such as $s_3(time)$, $l_1(group)$, $l_2(dog)$, $l_{23}(dog, time)$ and $l_{13}(group, time)$. There is some discrepancy in the models determined by the two methods. DIC supports $ls_{13}(group, time)$ but the Bayes factor is not convincing. We believe that this data set is an example of a situation where analysis using Bayes factors is more illuminating than DIC.

5 Conclusions

The proposed Bayes factors permit testing of all possible effects in SSANOVA models including linear effects, smooth effects, interaction effects or combinations of arbitrary terms in the model. This provides a flexible method for variable selection in semiparametric and nonparametric models comparable to hypothesis testing in ordinary linear

Terms	*Model(DIC)*	*Selected*
$ls_{13}(group, time)$	M5(278.2) vs M4(282.0)	M5
$ls_{23}(dog, time)$	M4(282.0) vs M3(443.2)	M4
$l_{13}(group, time)$	M4(282.0) vs M2(295.6)	M4
$ls_{13}(group, time)+l_{13}(group, time)+ l_1(group)$	M5(278.2) vs M1(284.1)	M5

Table 3: Model selection based on the Deviance Information Criterion for the potassium measurement on dogs data set.

and linear mixed models. Being Bayesian, our approach is different from the generalized likelihood statistics for nonparametric models proposed, for example, by Fan et al. (2001).

These Bayes factors can be computed reasonably efficiently in our set up. The Bayes factors give more precise conclusions than examining credible sets for variable selection in additive models. In particular, credible sets can be hard to interpret for nonparametric (smooth) effects, especially for interaction effects.

Frequentist variable selection for SSANOVA models is often performed by AIC. This is generally believed to be liberal, may tend to favor more complicated models, and does not give any guidance on the amount of statistical evidence for retaining or omitting terms from a model. In contrast, Bayesian SSANOVA can provide inference for individual terms of interest when desired. It would be interesting to compare the results obtained by Bayes factors with the resampling approach for hypothesis testing proposed in Kim (2013).

Finally, in our opinion, Bayes factors provide more precise guidance than DIC, which is a somewhat informal guide to model selection with no control for sampling variability.

References

Basu, S. and Chib, S. (2003). "Marginal Likelihood and Bayes Factors for Dirichlet Process Mixture Models." *Journal of the American Statistical Association*, 98(461): 224–235. MR1965688. doi: http://dx.doi.org/10.1198/01621450338861947. 958

Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press. 958

Cheng, C.-I. and Speckman, P. L. (2012). "Bayesian Smoothing Spline Analysis of Variance Models." *Computational Statistics and Data Analysis*, 56(12): 3945–3958. MR2957845. doi: http://dx.doi.org/10.1016/j.csda.2012.05.020. 957, 958, 959, 960, 961, 964, 965, 968, 969

Chib, S. and Greenberg, E. (2007). "Semiparametric Modeling and Estimation of Instrumental Variable Models." *Journal of Computational and Graphical Statistics*, 16(1): 86–114. MR2345749. doi: http://dx.doi.org/10.1198/106186007X180723. 958

Chib, S. and Greenberg, E. (2010). "Additive Cubic Spline Regression with Dirichlet Process Mixture Errors." *Journal of Econometrics*, 156(2): 322–336. MR2609935. doi: http://dx.doi.org/10.1016/j.jeconom.2009.11.002. 958

Chib, S. and Jeliazkov, I. (2006). "Inference in Semiparametric Dynamic Models for Binary Longitudinal Data." *Journal of the American Statistical Association*, 101(474): 685–700. MR2256181. doi: http://dx.doi.org/10.1198/016214505000000871. 958

Cox, D., Koh, E., Wahba, G., and Yandell, B. S. (1988). "Testing the (Parametric) Null Model Hypothesis in (Semiparametric) Partial and Generalized Spline Models." *The Annals of Statistics*, 16(1): 113–119. MR0924859. doi: http://dx.doi.org/10.1214/aos/1176350693. 957

Crainiceanu, C. M. and Ruppert, D. (2004). "Likelihood Ratio Tests in Linear Mixed Models with One Variance Component." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 66(1): 165–185. MR2035765. doi: http://dx.doi.org/10.1111/j.1467-9868.2004.00438.x. 958

Dickey, J. M. (1971). "The weighted likelihood ratio, linear hypotheses on normal location parameters." *The Annals of Mathematical Statistics*, 42: 204–223. MR0309225. 967

Eubank, R. L., Hart, J. D., Simpson, D. G., and Stefanski, L. A. (1995). "Testing for additivity in nonparametric regression." *The Annals of Statistics*, 23(6): 1896–1920. MR1389857. doi: http://dx.doi.org/10.1214/aos/1034713639. 958

Fan, J. and Jiang, J. (2005). "Nonparametric Inferences for Additive Models." *Journal of the American Statistical Association*, 100(471): 890–907. MR2201017. doi: http://dx.doi.org/10.1198/016214504000001439. 958

Fan, J., Zhang, C., and Zhang, J. (2001). "Generalized Likelihood Ratio Statistics and Wilks Phenomenon." *The Annals of Statistics*, 29(1): 153–193. MR1833962. doi: http://dx.doi.org/10.1214/aos/996986505. 971

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014). *Bayesian Data Analysis, Third Edition*. CRC Press, Boka Raton. MR3235677. 964, 967

George, E. I. (2000). "The Variable Selection Problem." *Journal of the American Statistical Association*, 95(452): 1304–1308. MR1825282. doi: http://dx.doi.org/10.2307/2669776. 969

Golub, G. H., Heath, M., and Wahba, G. (1979). "Generalized cross-validation as a method for choosing a good ridge parameter." *Technometrics*, 21: 215–223. MR0533250. doi: http://dx.doi.org/10.2307/1268518. 960

Grizzle, J. E. and Allen, D. M. (1969). "Analysis of growth and dose response curves." *Biometrics*, 25: 357–381. 968

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer-Verlag Inc. MR1876599. doi: http://dx.doi.org/10.1007/978-1-4757-3683-0. 961, 962, 963

Guo, W. (2002). "Inference in Smoothing Spline Analysis of Variance." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 64(4): 887–898. MR1979393. doi: http://dx.doi.org/10.1111/1467-9868.00367. 958

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-fit Tests*. Springer-Verlag Inc. MR1461272. doi: http://dx.doi.org/10.1007/978-1-4757-2722-7. 958

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall Ltd. MR1082147. 957, 958, 960

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90: 773–795. 965, 967, 969

Kim, M.-K. (2013). "Semi-Parametric Mixed-Effects Models for the Analysis of QT intervals." PhD, University of California at Santa Barbara. MR3153273. 958, 971

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of G Priors for Bayesian Variable Selection." *Journal of the American Statistical Association*, 103(481): 410–423. MR2420243. doi: http://dx.doi.org/10.1198/016214507000001337. 958, 960, 964

Liu, A., Meiring, W., and Wang, Y. (2005). "Testing Generalized Linear Models Using Smoothing Spline Methods." *Statistica Sinica*, 15(1): 235–256. MR2125730. 958

Liu, A. and Wang, Y. (2004). "Hypothesis Testing in Smoothing Spline Models." *Journal of Statistical Computation and Simulation*, 74(8): 581–597. MR2074614. doi: http://dx.doi.org/10.1080/00949650310001623416. 958

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P., and Wagemakers, E.-J. (2011). "A Tutorial on Bayes Factor Estimation with the Product Space Method." *Journal of Mathematical Psychology*, 55(5): 331–347. MR2838422. doi: http://dx.doi.org/10.1016/j.jmp.2011.06.001. 958

Meng, X.-L. and Wong, W. H. (1996). "Simulating Ratios of Normalizing Constants Via a Simple Identity: A Theoretical Exploration." *Statistica Sinica*, 6: 831–860. MR1422406. 967

Min, X. and Sun, D. (2015). "Bayesian model selection for a linear model with grouped covariates." *Annals of the Institute of Statistical Mathematics*, 1–27. doi: http://dx.doi.org/10.1007/s10463-015-0518-9. 964

Morey, R. D., Rouder, J. N., Pratte, M. S., and Speckman, P. L. (2011). "Using MCMC Chain Outputs to Efficiently Estimate Bayes Factors." *Journal of Mathematical Psychology*, 55(5): 368–378. MR2838425. doi: http://dx.doi.org/10.1016/j.jmp.2011.06.004. 958

Reich, B. J., Storlie, C. B., and Bondell, H. D. (2009). "Variable Selection in Bayesian Smoothing Spline ANOVA Models: Application to Deterministic Computer Codes." *Technometrics*, 51(2): 110–120. MR2668168. doi: http://dx.doi.org/10.1198/TECH.2009.0013. 958

Rouder, J. N., Morey, R. D., Speckman, P. L., and Province, J. M. (2012). "Default Bayes Factors for ANOVA Designs." *Journal of Mathematical Psychology*, 56(5): 356–374. MR2983394. doi: http://dx.doi.org/10.1016/j.jmp.2012.08.001. 964

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge. MR1998720. doi: http://dx.doi.org/10.1017/CBO9780511755453. 957

Shi, M. and Dunson, D. B. (2011). "Bayesian Variable Selection Via Particle Stochastic Search." *Statistics & Probability Letters*, 81(2): 283–291. MR2764295. doi: http://dx.doi.org/10.1016/j.spl.2010.10.011. 958

Speckman, P. (1988). "Kernel smoothing in partial linear models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 50(3): 413–436. MR0970977. 957

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). "Bayesian Measures of Model Complexity and Fit (Pkg: P583-639)." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 64(4): 583–616. MR1979380. doi: http://dx.doi.org/10.1111/1467-9868.00353. 970

Sun, D. and Speckman, P. L. (2008). "Bayesian Hierarchical Linear Mixed Models for Additive Smoothing Splines." *Annals of the Institute of Statistical Mathematics*, 60(3): 499–517. MR2434409. doi: http://dx.doi.org/10.1007/s10463-007-0127-3. 960, 964

Verdinelli, I. and Wasserman, L. (1995). "Computing Bayes factors using a generalization of the Savage–Dickey density ratio." *Journal of the American Statistical Association*, 90(430): 614–618. MR1340514. 967

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM [Society for Industrial and Applied Mathematics]. MR1045442. doi: http://dx.doi.org/10.1137/1.9781611970128. 958, 959

Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995). "Smoothing Spline ANOVA for Exponential Families, with Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy: The 1994 Neyman Memorial Lecture." *The Annals of Statistics*, 23(6): 1865–1895. MR1389856. doi: http://dx.doi.org/10.1214/aos/1034713638. 957, 961

Wang, Y. (2011). *Smoothing Splines: Methods and Applications*. CRC Press. MR2814838. doi: http://dx.doi.org/10.1201/b10954. 963

Wang, Y. and Ke, C. (2004). "ASSIST: A Suite of S functions Implementing Spline smoothing Techniques." 968, 969

White, G. A. (2006). "Bayesian Semiparametric Spatial And Joint Spatial Temporal Smoothing." Ph.D. dissertation, University of Missouri Columbia, Department of Statistics. MR2709925. 960

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions." In: Goel, P. and Zellner, A. (eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. Amsterdam: North-Holland/Elsevier. MR0881437. 960

Zellner, A. and Siow, A. (1980). "Posterior Odds Ratios for Selected Regression Hypotheses." In: Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.), *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, 585–603. University of Valencia. MR0638871. 958, 960, 964

Acknowledgments

The authors gratefully acknowledge partial support from National Science Foundation Grants SES-0720229 and SES-1024080.