Bayesian Testing of Linear Versus Nonlinear Effects Using Gaussian Process Priors

Joris Mulder

Department of Methodology and Statistics, Tilburg University, Tilburg, Netherlands

ABSTRACT
A Bayes factor is proposed for testing whether the effect of a key predictor variable on a dependent variable is linear or nonlinear, possibly while controlling for certain covariates. The test can be used (i) in substantive research for assessing the nature of the relationship between certain variables based on scientific expectations, and (ii) for statistical model building to infer whether a (transformed) variable should be added as a linear or nonlinear predictor in a regression model. Under the nonlinear model, a Gaussian process prior is employed using a parameterization similar to Zellner’s g prior resulting in a scale-invariant test. Unlike existing p-values, the proposed Bayes factor can be used for quantifying the relative evidence in the data in favor of linearity. Furthermore, the Bayes factor does not overestimate the evidence against the linear null model resulting in more parsimonious models. An extension is proposed for Bayesian one-sided testing of whether a nonlinear effect is consistently positive, consistently negative, or neither. Applications are provided from various fields including social network research and education.

1. Introduction

Linearity between explanatory and dependent variables is a key assumption in most statistical models. In linear regression models, the explanatory variables are assumed to affect the dependent variables in a linear manner, in logistic regression models it is assumed that the explanatory variables have a linear effect on the logit of the probability of a success on the outcome variable, in survival or event history analysis a log-linear effect is generally assumed between the explanatory variables and the event rate, etc. Sometimes nonlinear functions (e.g., polynomials) are included of certain explanatory variables (e.g., for modeling curvilinear effects), or interaction effects are included between explanatory variables, which, in turn, are assumed to affect the dependent variable(s) in a linear manner.

Despite the central role of linear effects in statistical models, statistical tests of linearity versus nonlinearity are hardly used. In practice researchers tend to eyeball the relationship between the variables based on a scatterplot. When a possible nonlinear relationship is observed, various linear transformations (e.g., polynomial, logarithmic, Box-Cox) are applied and significance tests are executed to see if the coefficients of the transformed variables are significant or not. Eventually, when the nonlinear trend results in a reasonable fit, standard statistical inferential methods are applied (such as testing whether certain effects are zero and/or evaluating interval estimates).

This procedure is problematic for several reasons. First, executing many different significance tests on different transformed variables may result in p-hacking and inflated Type I errors. Second, regardless of the outcome of a significance test, for example, when testing whether the coefficient of the square of the predictor variable, \( X^2 \), equals zero or not, \( H_0 : \beta_{X^2} = 0 \) versus \( H_1 : \beta_{X^2} \neq 0 \), we would not learn whether \( X^2 \) has a linear effect on \( Y \) or not; only whether an increase of \( X^2 \) results on average in an increase/decrease of \( Y \) or not. Third, nonlinear transformations (e.g., polynomials, logarithmic, Box-Cox) are only able to create approximate linearity for a limited set of nonlinear relationships. Fourth, polynomial regression results in nonlocality which implies that the fitted function for a given value of the predictor variable heavily depends on values far away from that point (Magee 1998). Fifth, eyeballing the relationship can be subjective, and instead a principled and probabilistic approach is needed.

To address these shortcomings this article proposes a Bayes factor for testing the following models:

\[ M_0 : \text{“}X\text{ has a linear effect on }Y\text{”} \]

versus

\[ M_1 : \text{“}X\text{ has a nonlinear effect on }Y\text{”} \]

(1)

possibly while controlling for certain covariates. The proposed test can be used to assess whether a variable should be added as a linear or as a nonlinear predictor when building statistical models. Besides model building, testing whether an effect has a linear or nonlinear effect on an outcome variable can also be of direct scientific interest in applied research. For example in neuropsychology, Kanai et al. (2012) was interested in the nature of the relationship between (the square root of) the number of hits and (the square root of) the number of false alarms.
Facebook friends and gray matter density in regions of the brain that are related to social perception and associative memory to better understand the reasons for people to participate in online social networking. The relationship was quantified using correlations. Correlational analysis however, is only meaningful if the relationships are linear. Furthermore, in longitudinal social network research, the volume of past social interactions between actors is often assumed to have a negative linear effect on the logarithm of the time of the next social interaction between actors (Butts 2008; Mulder and Leenders 2019). Formal statistical tests of arithm of the time of the next social interaction between actors correlations. Correlational analysis however, is only meaningful onlinesocialnetworking. Therelationshipwasquantifiedusing
to better understand the reasons for people to participate in
that are related to social perception and associative memory
2
to model nonlinear relationships and its clear Bayesian inter-
testing of strictly positive or strictly negative onlineareffects.
areneededtoanswersuchresearchquestions.
Theaboveexamplesalsoillustratethatscientificexpectations
are often formulated about the direction of the (nonlinear) effect
which can be either positive or negative. In linear modeling, such
directional expectations are evaluated using one-sided hypoth-
thesis testing. To our knowledge no tests are available for testing
whether anonlineareffectisconsistentlypositiveorconsistently
esis testing. To our knowledge no tests are available for testing
directional expectations are evaluated using one-sided hypoth-
doing mainly focus on estimating rather than testing nonlinear
properties for null hypothesis testing which dates back from Jef-
linearity are more likely a priori than large deviations (a key
for this parameter is chosen such that small deviations from
captures the expected prior deviation from linearity. The prior
for this parameter is chosen such that small deviations from
linearlyare more likely a priori than large deviations (a key
property for null hypothesis testing which dates back from Jef-
Jeffreys 1961). Moreover the prior is parameterized by extending
the well-known g prior for linear regression analysis (Zellner
1986; Liang et al. 2008) resulting in a scale-invariant Bayes factor
test. Furthermore, for the nonlinear one-sided test, truncated
Gaussian process priors are specified which assume that the
function is either consistently positive or consistently negative.
Current available techniques for nonlinear statistical modeling
mainly focus on estimating rather than testing nonlinear effects.
Flexible estimation procedures are implemented in the R
packages brms (Bürkner 2017) and CVEK (Deng and Liu 2020),
which rely on Bayesian Gaussian process regression. Moreover
there is a huge literature on generalized additive models (Hastie
and Tibshirani 1986) which have been implemented in R pack-
ages such as gam (Hastie and Hastie 2020) and mgcv (Wood
2015). For an overview of R packages using generalized additive
models, see Perperoglou et al. (2019). Generalized additive mod-
els are also available in commercial statistical software such as
SAS using PROC GAM. Other approaches for semi-parametric
statistical modeling for combinations of linear or nonlinear
predictors include partial linear models (Härdle, Liang, and Gao
2000) and partial index models (Wang et al. 2010).
Some of these software packages provide approximate p-
values for significance testing of nonlinear terms (Li 2009; Wood
2013; Liu and Coull 2017). Following the standard interpreta-
tion, these p-values reflect the probability of observing at least as
extreme nonlinear effects as the observed nonlinear effect under
the assumption that the nonlinear effect is zero. The proposed
Bayes factor on the other hand quantifies the probability of
the observed data under the linear model $M_0$ relative to the
probability of the observed data under the nonlinear alternative
model $M_1$, and therefore, the outcome can be interpreted as a
relative measure of evidence in favor of linearity while p-values
cannot. For this reason, Bayes factors are more suitable for model
building than p-values as they quantify the relative evidence in
the data between linearity and nonlinearity. Fisherian p-values
on the other hand can only be used for falsifying a linear effect:
If the p-value does not fall below a significance threshold, there
is insufficient evidence to reject the linear effect but we also cannot
claim there is evidence in favor of linearity. Moreover p-values
tend to result in an overestimation of the evidence against the
null model (Benjamin and Berger 2019). For the current test,
this implies that Bayes factors result in a more parsimonious
model than p-values where predictor variables are not unnec-
essarily modeled as nonlinear because of a seemingly significant
nonlinear term. Finally Bayes factors have the appealing prop-
erty of being independent of the (possibly unknown) sampling
plan (Wagenmakers 2007) and being able to simultaneously test
multiple (one-sided) models (Mulder et al. 2021).
The article is organized as follows. Section 2 describes the lin-
ear and nonlinear Bayesian models and the corresponding Bayes
factor. Its behavior is also explored and compared with a p-value
in numerical simulations. Section 3 describes the nonlinear
one-sided Bayesian test. Subsequently, Section 4 presents four
applications of the proposed methodology in different research
fields. We end the article with a short discussion in Section 5.

2. A Bayes Factor for Testing (Non)linearity

2.1. Model Specification

Under the standard linear regression model, denoted by $M_0$,
we assume that the mean of the dependent variable $Y$ depends
proportionally on the key predictor variable $X$, possibly while
correcting for certain covariates. Mathematically, this implies
that the predictor variable is multiplied with the same coeffi-
cient, denoted by $\beta$, to compute the (corrected) mean of the
dependent variable for all values of $X$. The linear model can then
be written as

$$M_0 : y \sim N(\beta x + \gamma Z, \sigma^2 I_n),$$

where $y$ is a vector containing the $n$ observations of the depen-
dent variable, $x$ contains the $n$ observations of the predictor
variable, $Z$ is a $n \times k$ matrix of covariates (which are assumed
to be orthogonal to the key predictor variable) with correspond-
ing coefficients $\gamma$, and $\sigma^2$ denotes the error variance which is
multiplied with the identity matrix of size $n$, denoted by $I_n$. To
complete the Bayesian model, we adopt the standard g prior
approach (Zellner 1986) by setting a Gaussian prior on $\beta$ where
the variance is scaled based on the error variance, the scale of the
predictor variable, and the sample size, together with a flat prior
for the nuisance regression coefficients and the independence
Jeffreys’ prior for the error variance, that is,

$$\beta | \sigma^2 \sim N(0, \sigma^2 g(x'x)^{-1})$$

$$p(\gamma) \propto 1$$

$$p(\sigma^2) \propto \sigma^{-2}.$$
The prior mean is set to the default value of 0 so that, a priori, small effects in absolute value are more likely than large effects (as is common in applied research) and positive effects are equally likely as negative effects (an objective choice in Bayesian one-sided testing Jeffreys 1961 and Mulder, Hoijtink, and Klugkist 2010). By setting \( g = n \) we obtain a unit-information prior (Kass and Wasserman 1995; Liang et al. 2008) which will be adopted throughout this article.\(^1\)

Under the alternative nonlinear model, denoted by \( M_1 \), we assume that the mean of the dependent variable does not depend proportionally on the predictor variable. This implies that the observations of the predictor variable can be multiplied with different values for different values of the predictor variable \( X \). This can be written as follows:

\[
M_1 : y \sim \mathcal{N}(\beta(x) \circ x + \gamma Z, \sigma^2 \mathcal{I}_n),
\]

where \( \beta(x) \) denotes a vector of length \( n \) containing the coefficients of the corresponding \( n \) observations of the predictor variable \( x \), and \( \circ \) denotes the Hadamard product. The vector \( \beta(x) \) can be viewed as the \( n \) realizations when plugging the different values of \( x \) in a unknown theoretical function \( \beta(x) \). Thus, in the special case where \( \beta(x) \) is a constant function, say, \( \beta(x) = \beta \), model \( M_1 \) would be equivalent to the linear model \( M_0 \).

Next we specify a prior distribution for the function of the coefficients. Because we are testing for linearity, it would be likely to expect relatively smooth changes between different values under the alternative. This implies that the difference between \( \beta_i(x_i) \) and \( \beta_j(x_j) \) is expected to be small when the distance between \( x_i \) and \( x_j \) is small, and vice versa. A Gaussian process prior for the function \( \beta(x) \) has this property which is defined by

\[
\beta(x) | \tau^2, \xi, x \sim \mathcal{GP}(0, \tau^2 k(x, x'|\xi)),
\]

which has a zero mean function and a kernel function \( k(\cdot, \cdot) \) which defines the covariance of the coefficients as a function of the distance between values of the predictor variable. A squared exponential kernel will be used which is given by

\[
k(x_i, x_j|\xi) = \exp \left\{ -\frac{1}{2} \xi^2 (x_i - x_j)^2 \right\},
\]

for \( i, j = 1, \ldots, n \). As can be seen, predictor variables \( x_i \) and \( x_j \) that are close to (far away from) each other have a larger (smaller) covariance, and thus, are on average closer to (further away from) each other. The hyperparameter \( \xi \) controls the smoothness of the function where values close to 0 imply very smooth function shapes and large values imply highly irregular shapes (as will be illustrated later). Note that typically the smoothness is parameterized via the reciprocal of \( \xi \), which is referred to as the length-scale parameter. Here we use the current parameterization so that the special value \( \xi = 0 \) would come down to a constant function, say \( \beta(x) = \beta \), which would correspond to a linear relationship between the predictor and the outcome variable.

The hyperparameter \( \tau^2 \) controls the prior magnitude of the coefficients, that is, the overall prior variance for the coefficients. We extend the \( g \) prior formulation to the alternative model by setting \( \tau^2 = \sigma^2 g(x'|x)^{-1} \) and specify the same priors for \( \gamma \) and \( \sigma^2 \) as under \( M_0 \). Furthermore, by taking into account that the Gaussian process prior implies that the coefficients for the observed predictor variables follow a multivariate normal distribution, the priors under \( M_1 \) given the predictor variables can be formulated as

\[
\beta(x) | \sigma^2, \xi, x \sim \mathcal{N}(0, \tau^2 g(x'|x)^{-1} k(x, x'|\xi))
\]

\[
p(\gamma) \propto 1
\]

\[
p(\sigma^2) \propto \sigma^{-2}.
\]

To complete the model a half-Cauchy prior is specified for the key parameter \( \xi \) having prior scale \( s_\xi \), that is,

\[
\xi \sim \text{half-C}(s_\xi).
\]

The motivation for this prior is based on one of Jeffreys (1961) desiderata which states that small deviations from the null value are generally more likely a priori than large deviations otherwise there would be no point in testing the null value. In the current setting this would imply that small deviations from linearity are more likely to be expected than large deviations. This would imply that values of \( \xi \) close to 0 are more likely a priori than large values, and thus, that the prior distribution for \( \xi \) should be a decreasing function. The half-Cauchy distribution satisfies this property. Further note that the half-Cauchy prior is becoming increasingly popular for scale parameters in Bayesian analyses (Gelman 2006; Polson and Scott 2012; Mulder and Pericchi 2018).

The prior scale for the key parameter \( \xi \) under \( M_1 \) should be carefully specified as it defines which deviations from linearity are most plausible. To give the reader more insight about how \( \xi \) affects the distribution of the slopes of \( y \) as function of \( x \), Figure 1 displays 10 random draws of the function of slopes when setting \( \xi_1 = 0 \) (Figure 1(a)), \( \xi_1 = \exp(-2) \) (Figure 1(b)), \( \xi_1 = \exp(-1) = 1 \) (Figure 1(c)), \( \xi_1 = \exp(0) \) (Figure 1(d)) while fixing \( \tau^2 = \sigma^2 g(x'|x)^{-1} = 1 \), where the slope function is defined by

\[
\eta(x) = \frac{d}{dx} [\beta(x) \circ x] = \beta(x) + \frac{d}{dx} [\beta(x)] \circ x.
\]

The figure shows that by increasing \( \xi \) we get larger deviations from a constant slope. Based on these plots we qualify the choices \( \xi_1 = \exp(-2) \), \( \exp(-1) \), and 1 as small deviations, medium deviations, and large deviations from linearity, respectively.

Because the median of a half-Cauchy distribution is equal to the scale parameter \( s_\xi \), the scale parameter could be set based on the expected deviation from linearity. Furthermore, it is important to note that the expected deviation from linearity depends on the range of the predictor variable: In a very small range it may be expected that the effect is close to linear but in a wide range of the predictor variable, large deviations from linearity may be expected. Given the plots in Figure 1, one could set the prior scale equal to \( s_\xi = 6c/\text{range}(x) \), where \( c \) can be interpreted as a standardized measure for the deviation from linearity such that setting \( c = \exp(-2) \), \( \exp(-1) \), or \( \exp(0) \) would imply small,
medium, or large deviations from linearity, respectively. Thus, if the range of \( x \) would be equal to 6 (as in the plots in Figure 1), the median of \( \xi \) would be equal to \( \exp(-2) \), \( \exp(-1) \), and \( \exp(0) \) as plotted in Figure 1.

2.2. Bayes Factor Computation

The Bayes factor is defined as the ratio of the marginal (or integrated) likelihoods under the respective models. For this reason it is useful to integrate out the coefficient \( \beta \) under \( \mathcal{M}_0 \) and the coefficients \( \beta(x) \) under \( \mathcal{M}_1 \), which are in fact nuisance parameters in the test. This yields the following integrated models

\[
\mathcal{M}_0 : \begin{cases} 
  y|x, \gamma, \sigma^2 \sim N(Z\gamma, \sigma^2 g(x'x)^{-1}xx' + \sigma^2 I_n) \\
  p(\gamma) \propto 1 \\
  p(\sigma^2) \propto \sigma^{-2}
\end{cases}
\]

\[
\mathcal{M}_1 : \begin{cases} 
  y|x, \gamma, \sigma^2, \xi \sim N(Z\gamma, \sigma^2 g(x'x)^{-1}k(x, x'|\xi) o xx' + \sigma^2 I_n) \\
  p(\gamma) \propto 1 \\
  p(\sigma^2) \propto \sigma^{-2} \\
  \xi \sim \text{half-C}(\xi_0)
\end{cases}
\]

As can be seen \( \sigma^2 \) is a common factor in all (co)variances of \( y \) under both models. This makes inferences about \( \xi \) invariant to the scale of the outcome variable. Finally note that the integrated models clearly show that the model selection problem can concisely be written as

\[
\mathcal{M}_0 : \xi = 0 \\
\mathcal{M}_1 : \xi > 0
\]

because \( k(x, x'|\xi) = 1 \) when setting \( \xi = 0 \).

Using the above integrated models, the Bayes factor can be written as

\[
B_{01} = \frac{p(y|\mathcal{M}_0)}{p(y|\mathcal{M}_1)} = \frac{\int \int p(y|x, \gamma, \sigma^2, \xi = 0)p(\gamma)p(\sigma^2) d\gamma d\sigma^2}{\int \int \int p(y|x, \gamma, \sigma^2, \xi)p(\gamma)p(\sigma^2)p(\xi) d\gamma d\sigma^2 d\xi},
\]

which can be interpreted as the relative evidence in the data between the linear model \( \mathcal{M}_0 \) and the nonlinear model \( \mathcal{M}_1 \). Different methods can be used for computing marginal likelihoods. Throughout this article we use an importance sample estimate. The R code for the computation of the marginal likelihoods and the sampler from the posterior predictive distribution can be found in the supplementary material.

2.3. Numerical Behavior

2.3.1. Smooth Nonlinear Alternative

Numerical simulations were performed to evaluate the performance of the proposed Bayes factor. The nonlinear function was set equal to \( \beta(x) = 3h(\phi(x)) \), for \( h = 0, \ldots, 5 \), where \( \phi \) is the standard normal probability density function (Figure 2; first row, left panel). In the case where \( h = 0 \), the effect is linear, and as \( h \) increase, the effect becomes increasingly nonlinear. The dependent variable was computed as \( \beta(x) o x + \epsilon \), where \( \epsilon \) was sampled from a normal distribution with mean 0 and \( \sigma = .1 \).
The logarithm of the Bayes factor, denoted by \( \log(B_{01}) \), was computed between the linear model \( M_0 \) and the nonlinear model \( M_1 \) (Figure 2; second row, left panel) while setting the prior scale equal to \( s_\xi = \exp(-2) \) (small prior scale; solid line), \( \exp(-1) \) (medium prior scale; dashed line), and \( \exp(0) \) (large prior scale; dotted line) for sample size \( n = 20 \) (black lines), 50 (red lines), and 200 (green lines) for equally distant predictor values in the interval \((-3, 3)\). Overall we see the expected trend where we obtain evidence in favor of \( M_0 \) in the case \( h \) is close to zero and evidence in favor of \( M_1 \) for larger values of \( h \). Moreover, as \( h \) increases, we see that the evidence for \( M_1 \) increases, which is an illustration of information consistent behavior (Mulder et al. 2020). We also see that the evidence for \( M_0 \) (\( M_1 \)) is larger for larger sample sizes and larger prior scale when \( h = 0 \).
(h ≫ 0) as anticipated given the consistent behavior of the Bayes factor.

### 2.3.2. Discontinuous Stepwise Alternative

Next we investigated the robustness of the test for nonlinear relationships that are not smooth (unlike the Gaussian processes with a squared exponential kernel under $M_1$). A similar analysis was performed as above but using a discontinuous step function $\beta(x) = h 1(x > 0)$, where $1(\cdot)$ is the indicator function, for $h = 0, \ldots, .5$ (Figure 2; first row, right panel). Again the dependent variable was computed as $\beta(x) \circ x + \epsilon$ and the logarithm of the Bayes factor was computed (Figure 2; second row, right panel).

The Bayes factor shows a similar behavior as earlier where the data were generated using a smooth nonlinear alternative. The similarity of the results can be explained by the fact that even though the step function cannot be generated using a Gaussian process with a squared exponential kernel, the closest approximation of the step function is still nonlinear, and thus, evidence is found against the linear model $M_0$ in the case $h > 0$. This illustrates that the proposed Bayes factor is robust for testing nonsmooth nonlinear effects.

### 2.3.3. Comparison with $p$-Values

Finally we investigated how the proposed Bayesian test relates to existing classical tests for nonlinearity under a generalized additive model. Because the $p$-value would be different for different specifications of the nonlinear term (e.g., degrees of freedom, basis function, etc.), we kept the comparison simple and only considered a $p$-value for a nonlinear smoothing spline under the default setting of a fitted generalized additive model from the 	exttt{gam} package (Hastie and Hastie 2020). The behavior was similar for other nonlinear term (such as a loess smooth terms). The $p$-value was computed for the above two scenario’s where the alternative was either a smooth nonlinear function or a discontinuous stepwise function. Figure 2 (fourth row) shows the $p$-value as a function of the effect size $h$ for different sample sizes. As a comparison we also computed the posterior probability of the linear model $M_0$ via $P(M_0|y) = \frac{B_{01}}{B_{01} + 1}$, by assuming equal prior model probabilities (Figure 2; third row). These plots show that the $p$-value depends on the effect size in the same manner as the posterior probability. The posterior probability of the null model however, is generally larger than the $p$-value which is an illustration that $p$-values tend to overestimate the evidence against the null (Sellke, Bayarri, and Berger 2001; Benjamin et al. 2017).

To get more insights about this, we also explored how the proposed Bayes factor (using a medium prior scale) behaves as a function of the sample size while keeping the $p$-value fixed at 0.05 (using the default smoothing spline $s$ from the 	exttt{gam} package). We achieved this by generating datasets under a nonlinear model having standard normal pdf shape having a $p$-value that is approximately equal to 0.05 (the data were generated so that $p$ lied between 0.0499 and 0.0501). For each sample size, $n = 20, 30, 40, \ldots, 500$, 50 datasets were generated (each resulting in $p \approx 0.05$), and the median of the Bayes factors was computed. The results can be found in Figure 3. The figure shows that on average the evidence for the null model increases as the sample size grows. This behavior can be explained by the fact that the effect size that results in a $p$-value of 0.05 decreases when the sample size increases, and thus, the evidence against the null decreases when the sample size grows when keeping $p$ fixed. As the sample size grows the evidence eventually increases toward the null model, which conflicts with the conclusion based on the $p$-value. This conflicting behavior is related to the Jeffreys-Lindley paradox (Lindley 1957; Jeffreys 1961). See also Berger and Delampady (1987), Berger and Mortera (1999), and Wagenmakers (2007), for more illustrations of this phenomenon.

### 3. Extension to One-Sided Nonlinear Testing

When testing linear effects, the interest is often in whether the effect is either positive or negative if the null does not hold. Equivalently in the case of nonlinear effects the interest would be whether the effect is consistently increasing or consistently decreasing over the range of $X$. To model this we divide the parameter space under the nonlinear model $M_1$ in three subspaces:

- $M_{1, positive}$: “the nonlinear effect of $X$ on $Y$ is consistently positive”
- $M_{1, negative}$: “the nonlinear effect of $X$ on $Y$ is consistently negative”
- $M_{1, complement}$: “the nonlinear effect of $X$ on $Y$ is neither consistently positive, nor consistently negative”

Note that the first model implies that the slope function is consistently positive, that is, $\eta(x) > 0$, the second implies that the slope is consistently negative, that is, $\eta(x) < 0$, while the third complement model assumes that the slope function is neither consistently positive nor negative.

Following standard Bayesian methodology (Klugkist, Laudy, and Hoijtink 2005; Mulder, Hoijtink, and Klugkist 2021), we consider a Gaussian process prior under the unconstrained nonlinear model $M_1$, and specify truncations in the constrained

---

1 The R code for obtaining the $p$-value of a default nonlinear smoothing spline using the 	exttt{gam} package was: 
\[
\text{gaml} <- \text{gam}(y \sim 1 + x + s(x), \\
data = \text{data.frame}(x=x, y=y)); \text{summary(gaml)}\text{anova}[3].
\]
subspaces under $\mathcal{M}_{1,\text{positive}}$, $\mathcal{M}_{1,\text{negative}}$, and $\mathcal{M}_{1,\text{complement}}$. For model $\mathcal{M}_{1,\text{positive}}$, the truncated Gaussian process prior in the subspace where the derivative of the function is positive can then be written as

$$
\pi_{1,\text{pos}}(\beta(x)|\xi_1, \sigma_1^2, x) = \pi_1(\beta(x)|\xi_1, \sigma_1^2, x) \times f(\eta(x) > 0(\beta(x)) \times \Pr(\eta(x) > 0|\mathcal{M}_1, \xi_1, \sigma_1^2, x)^{-1},
$$

where $\pi_1(\beta(x)|\xi_1, \sigma_1^2, x)$ denotes the unconstrained Gaussian process prior, $f(\cdot)$ denotes the indicator function, and the prior probability, which serves as normalizing constant, equals

$$
\Pr(\eta(x) > 0|\mathcal{M}_1, \xi_1, \sigma_1^2, x) = \int_{\eta(x) > 0} \pi_1(\beta(x)|\xi_1, \sigma_1^2, x) d\beta(x),
$$

where the derivative $\eta(x)$ was given in Equation (6). Note that the prior probability for a consistently negative effect is equal because the prior mean of $\beta(x)$ equals 0. The truncated priors under $\mathcal{M}_{1,\text{negative}}$ and $\mathcal{M}_{1,\text{complement}}$ can be specified in a similar manner. Consequently, the Bayes factor of each constrained model against the unconstrained model $\mathcal{M}_1$ is then given by the ratio of the posterior and prior probabilities that the constraints hold under $\mathcal{M}_1$, for example,

$$
B_{(1,\text{pos})}^{1,\text{comp}} = \frac{\Pr(\eta(x) > 0|\mathcal{M}_1, y)}{\Pr(\eta(x) > 0|\mathcal{M}_1)}.
$$

Bayes factors between the above three models can then be computed using the transitive property of the Bayes factor, for example, $B_{(1,\text{pos})}(1,\text{comp}) = B_{(1,\text{pos})}/B_{(1,\text{comp})}$.

4. Empirical Applications

4.1. Neuroscience: Facebook Friends Versus Gray Matter

Kanai et al. (2012) studied the relationship between the number of Facebook friends and the gray matter density in regions of the brain that are related to social perception and associative memory to better understand the reasons for people to participate in online social networking. Here we analyze the data from the right entorhinal cortex ($n = 41$). Due to the nature of the variables a positive relationship was expected. Based on a significance test (Kanai et al. 2012) and a Bayes factor (Wetzels and Wagenmakers 2012) on a sample of size 41, it was concluded that there is evidence for a nonzero correlation between the square root of the number of Facebook friends and the gray matter density. In order for a correlation to be meaningful however, it is important that the relationship is (approximately) linear.

Here we test whether the relationship is linear or nonlinear. Furthermore, in the case of a nonlinear relationship, we test whether the relationship is consistently positive, consistently negative, or neither. Besides the predictor variable, the employed model has an intercept. The predictor variable is shifted to have a mean of 0 so that it is independent of the vector of ones for the intercept.

The Bayes factor between the linear model against the nonlinear model when using a prior scale of $\exp(-1)$ (medium effect) was equal to $B_{01} = 2.50$ (with $\log(B_{01}) = 0.917$). This implies very mild evidence in favor of a linear relationship between the square root of the number of Facebook friends and gray matter density in this region of the predictor variable. When assuming equal prior model probabilities, this would result in posterior model probabilities of 0.714 and 0.286 for $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively. Thus, if we would conclude that the relation is linear, there would be a conditional error probability of drawing the wrong conclusion of 0.286. Table 1 presents the Bayes factors also for the other prior scales as well as the posterior probabilities. The table also gives the $p$-values for a nonlinear smoothing spline and a loess smooth term, which are equal to 0.287 and 0.233, respectively. Because these $p$-values do not fall below the threshold value of 0.05, we would be in a state of indifference where there is not enough evidence to reject the linear model but we also cannot claim that there is evidence in favor of the linear model. Based on the Bayes factors and posterior probabilities on the other hand, we do observe mild evidence in favor of linearity.

Figure 4 (upper left panel) displays the data (circles; replotted from Kanai et al. 2012) and 50 draws of the posterior distribution density for the mean function under the nonlinear model at the observed values of the predictor variable. As can be seen most draws are approximately linear, and because the Bayes factor

| Table 1. Log Bayes factors for the linear model versus the nonlinear model using different prior scales $s_\xi$ (e$^{-2} =$ small, e$^{-1} =$ medium, 1 = large), and corresponding posterior model probabilities, and $p$-values for a smoothing spline (ss) and a loess smooth term using the gam package. |
|-----------------|------|-------------|-------------|-------------|-------------|
|                 | n    | Small       | Medium      | Large       | Small       | Medium      | Large       | p-value     |
| Age and knowing gay | 41   | 0.508       | 0.917       | 1.45        | 0.624       | 0.714       | 0.810       | 0.287       | 0.233       |
| Past activity and waiting time | 63   | -37.7       | -38.3       | -38.1       | 0.000       | 0.000       | 0.000       | 0.000       | 0.000       |
| Mother’s IQ and child test scores | 500  | -0.776      | -0.361      | 0.394       | 0.315       | 0.412       | 0.597       | 0.011       | 0.033       |
| Fb friends and gray matter | 434  | -2.46       | -2.07       | -1.38       | 0.079       | 0.112       | 0.201       | 0.002       | 0.001       |
Figures (circles) and 50 draws of the mean function (lines) under the nonlinear model $M_1$ for the four different applications. In the lower right panel, draws are given in the case the mother finished her high school (blue lines) or not (green lines).

Even though we found evidence for a linear effect, there is still posterior model uncertainty and therefore, we computed the Bayes factors between the one-sided models (9) under the nonlinear model $M_1$. This resulted in Bayes factors for a consistently positive effect, a consistently negative effect, and the complement model against the unconstrained model of $B_{(1,\text{pos})}(1) = 5.894$, $B_{(1,\text{neg})}(1) = 0.000$, and $B_{(1,\text{comp})}(1) = 0.242$, implying clear evidence for a consistently positive effect with $B_{(1,\text{pos})}(1,\text{neg}) \approx \infty$ and $B_{(1,\text{pos})}(1,\text{comp}) \approx 24.28$. These results are confirmed when checking the slopes of the posterior draws of the nonlinear mean function in Figure 4 (upper left panel).

4.2. Sociology: Age and Attitude Toward Gay

We consider data presented in Gelman et al. (2014) from the 2004 National Annenberg Election Survey containing respondents’ age, sex, race, and attitude on three gay-related questions from the 2004 National Annenberg Election Survey. Here we are interested in the relationship between age and the proportion of people who know someone who is gay ($n = 63$). It may be expected that older people know less people who are gay and thus, a negative relationship may be expected. Here we test whether the relationship between these variables is linear or not. In the case of a nonlinear relationship we also perform the one-sided test whether the relationship is consistently positive, negative, or neither. Again an intercept is included in the employed model.

When setting the prior scale to a medium deviation from linearity, the logarithm of the Bayes factor between the linear model against the nonlinear model was approximately equal to $-38.3$, which corresponds to a Bayes factor of 0. This implies convincing evidence for a nonlinear effect. When using a small or large prior scale, the Bayesian tests result in the same conclusion, as well as when using $p$-values (Table 1). Figure 4 (upper right panel) displays the data (black circles) and 50 posterior draws of the mean function, which have clear nonlinear curves which fit the observed data.

Next we computed the Bayes factors for the multiple one-sided test which results in decisive evidence for the complement model that the relationship is neither consistently positive nor consistently negative, with $B_{(1,\text{comp})}(1,\text{pos}) = \infty$ and $B_{(1,\text{comp})}(1,\text{neg}) = \infty$. This is confirmed when checking the posterior draws of the mean function in Figure 4 (upper right panel). We see a slight increase of the proportion of respondents who know someone who’s gay toward the age of 45, and a decrease afterwards.
4.3. Social Networks: Inertia and Dyadic Waiting Times

In longitudinal social networks, it is often assumed that actors in a network have a tendency to continue to initiate social interactions with each other as a function of the volume of past interactions. This is also called inertia. In the application of the relational event model (Butts 2008; Mulder and Leenders 2019), it is often assumed that the expected value of the logarithm of the waiting time between social interactions depends linearly on the number of past social interactions between actors. Here we consider relational (email) data from the Enron e-mail corpus (Cohen 2009). We consider a subset of the last \( n = 500 \) emails (excluding four outliers) in a network of 156 employees in the Enron data (Cohen 2009). We use a model with an intercept.

Based on a medium prior scale under the nonlinear model, the logarithm of the Bayes factor between the linear model against the nonlinear model equals \( \log(B_{10}) = -0.361 \), which corresponds to \( B_{10} = 1.43 \), implying approximately equal evidence for both models. The posterior probabilities are equal to 0.412 and 0.588 for models \( M_0 \) and \( M_1 \), respectively. When using the small and large prior scale the Bayes factors are similar (Table 1), where the direction of the evidence flips toward the null model when using a large prior scale. This suggests that a large deviation from linearity is least likely. Based on the posterior draws of the mean function in Figure 4 (lower left panel) we also see an approximate linear relationship. The nonlinearity seems to be mainly caused by larger observations of the predictor variable. The evidence is inconclusive about the nature of the relationship (linear or nonlinear), and therefore, the use of a linear inertia effect would be justifiable. Based on the \( p \)-values, which fall below 0.05 however, the linear model would be rejected and a nonlinear inertia effect would be used making the model possibly unnecessary complex and resulting in a decrease of statistical power.

For the one-sided tests, the Bayes factors yield most evidence for a consistent decrease but the evidence against the complement model is mild with \( B_{11,\text{neg}}(1,\text{pos}) = \infty \) and \( B_{(1,\text{neg})(1,\text{comp})} = 5.14 \). This suggests that there is most evidence that dyads (i.e., pairs of actors) that have been more active in the past are likely to communicate more frequently in the future.

5. Discussion

In order to make inferences about the nature of the relationship between stochastic variables, principled statistical tests are needed. In this article a Bayes factor was proposed that allows one to quantify the relative evidence in the data between a linear relationship and a nonlinear relationship, possibly while controlling for certain covariates. The test can be used (i) when building (generalized) linear models to check whether a (transformed) predictor affects the dependent variable in a linear manner, and (ii) when testing scientific theories about the nature of a relationship between two variables. The proposed Bayes factor adds to the increasing call for using statistical tests in applied research that can provide evidence in favor of a null hypothesis (e.g., Gönen et al. 2005; Schönbrodt and Wagenmakers 2018; Keysers, Gazzola, and Wagenmakers 2020; Mulder et al. 2021).

A Gaussian process prior with a square exponential kernel was used to model the nonlinear relationship under the alternative (nonlinear) model. The model was parameterized similar as Zellner’s \( g \) prior to make inferences that are invariant of the scale of the dependent variable and predictor variable. Moreover the Gaussian process was parameterized using the reciprocal of the length-scale parameter which controls the smoothness of the nonlinear trend so that the linear model would be obtained when setting this parameter equal to 0. A standardized scale for this parameter was proposed to quantify the deviation from linearity under the alternative model.

In the case of a nonlinear effect a Bayes factor was proposed for testing whether the effect was consistently positive, consistently negative, or neither. This test can be seen as a nonlinear extension of Bayesian one-sided testing. Unlike the linear one-sided test, the Bayes factor depends on the prior scale for the nonlinear one-sided test. Thus, the prior scale also needs to be carefully chosen for the one-sided test depending on the expected deviation from linearity.

The proposed Bayes factor contributes to the existing literature on testing nonlinear effects which, so far, has been dominated by Fisherian \( p \)-values. Unlike \( p \)-values, the proposed Bayes factor can be used for quantifying statistical evidence in favor of linearity. The Bayes factor also does not depend on data that

---

3As discussed by Gelman and Hill (2007) an interaction effect could also be reasonable to consider. Here we did not add the interaction effect for illustrative purposes. We come back to this in the Discussion.
where not observed (such as hypothetical data that deviate more from linearity than the observed data). Thereby the Bayes factor does not overestimate the evidence against the null resulting in more parsimonious statistical models without unnecessary nonlinear terms based on seemingly significant p-values. As a next step it would be useful to extend the methodology to correct for covariates that have a nonlinear effect on the outcome variable (e.g., using additive Gaussian processes; Duvenaud, Nickisch, and Rasmussen 2017; Cheng et al. 2019), to test nonlinear interaction effects, or to allow other kernels under the alternative model to capture different nonlinear shapes. We leave this for future work.

Supplementary Materials

The supplementary material contains the R code for the computation of the marginal likelihoods and the sampler from the posterior predictive distribution.

Acknowledgments

The author is grateful to Andrew Gelman for sharing the data for the second and fourth empirical application. Furthermore the author would like to thank Joran Jongerling, two anonymous reviewers, and the editorial board for their feedback which resulted in important improvements of the article.

Funding

This work was supported by an ERC starting grant (758791).

References

Bayarri, M. J., and García-Donato, G. (2007), “Extending Conventional Priors for Testing General Hypotheses in Linear Models,” *Biometrika*, 95, 135–152. [3]

Benjamin, D. J., and Berger, J. O. (2019), “Three Recommendations for Improving the Use of p-Values,” *The American Statistician*, 73, 186–191. [2]

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K.A, Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C.D., Clyde, M., Cook, T.D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A.P., Forster, M., George, E.I., Gonzalez, R., Goodman, S., Green, E., Green, D.P., Greenwald, A.G., Hadfield, J.D., Hedge, L.V., Held, L., Ho, T.H., Hoijtingk, H., Hruschka, D.J., Imai, K., Imbens, G., Ioannidis, J.P.A., Jeon, M., Holland Jones, J., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S.E., McCarthy, M., Moore, D.A., Morgan, S.L., Munafò, M., Nakagawa, S., Nyhan, B., Parker, T.H., Perichci, L., Perugini, M., Rouder, J., Roussou, J., Savalei, V., Schönbrodt, F.D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D.J., Winship, C., Wolpert, R.L., Xie, Y., Young, C., Zinman, J., Johnson, V.E. (2017), “Robust Hypothesis Test for Nested Hypotheses and its Relationship to the Schwarz Criterion,” *Journal of the American Statistical Association*, 90, 928–934. [3]

Keysers, C., Gazzola, V., and Wagenmakers, E.-J. (2020), “Using Bayes Factor Hypothesis Testing in Neuroscience to Establish Evidence of Absence,” *Nature Neuroscience*, 23, 788–799. [9]

Klugkist, I., Laudy, O., and Hoijtingk, H. (2005), “Inequality Constrained Analysis of Variance: A Bayesian Approach,” *Psychological Methods*, 10, 477–493. [6]

Li, C.-S. (2009), “Using p-Splines to Test the Linearity of Partially Linear Models,” *Statistical Modelling*, 6, 542–552. [2]

Li, C.-S., Xu, J., and Yang, C. (2009), “Using p-Splines to Test the Linearity of Partially Linear Models,” *Statistical Modelling*, 9, 155–200. [2,9]

Perugini, M., Rouder, J., and Wagenmakers, E.-J. (2020), “Using Constrained Posterior Priors,” *Journal of Mathematical Psychology*, 93, 159–174. [3]

The Matrix-F Prior for Estimating and Testing Covariance Matrices,” *Bayesian Analysis*, 13, 1193–1214. [3]

Mulder, J., Williams, D. R., Gu, X., Tomarken, A., Böing-Messing, F., Olsson-Collentine, A., Meijerink-Bosman, M., Menke, J., van Aert, R., Fox, J.-P., Hoijtingk, H., Rosseel, R., Wagenmakers, E. J., and van Lissa, C. (2021), “Bfpack: Flexible Bayes Factor Testing of Scientific Theories in R,” *Journal of Statistical Software*, 100, 1–63. [2,9]

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., and Schmid, M. (2019), “A Review of Spline Function Procedures in R,” *BMC Medical Research Methodology*, 19, 1–16. [2]
Polson, N. G., and Scott, J. G. (2012), "On the Half-Cauchy Prior for a Global Scale Parameter," *Bayesian Analysis*, 7, 887–902. [3]

Rasmussen, C., and Williams, C. (2007), *Gaussian Processes for Machine Learning*, Cambridge, MA: MIT Press. [2]

Rouder, J. N., Speckman, P. L., Sun, D. D. M., and Iverson, G. (2009), "Bayesian t Tests for Accepting and Rejecting the Null Hypothesis," *Psychonomic Bulletin & Review*, 16, 225–237. [3]

Schönbrodt, F. D., and Wagenmakers, E.-J. (2018), "Bayes Factor Design Analysis: Planning for Compelling Evidence," *Psychonomic Bulletin & Review*, 25, 128–142. [9]

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001), "Calibration of p Values for Testing Precise Null Hypotheses," *The American Statistician*, 55, 62–71. [6]

Wagenmakers, E.-J. (2007), "A Practical Solution to the Pervasive Problem of p Values," *Psychonomic Bulletin and Review*, 14, 779–804. [2,6]

Wang, J.-L., Xue, L., Zhu, L., and Chong, Y. S. (2010), "Estimation for a Partial-Linear Single-Index Model," *The Annals of Statistics*, 38, 246–274. [2]

Wetzels, R., and Wagenmakers, E. J. (2012), "A Default Bayesian Hypothesis Test for Correlations and Partial Correlations," *Psychonomic Bulletin and Review*, 19, 1057–1064. [7]

Wood, S. (2013), "On p-Values for Smooth Components of an Extended Generalized Additive Model," *Biometrika*, 100, 221–228. [2]

(2015), "Package ‘mgcv’," *R package version*, 1, 29. [2]

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, eds. P. K. Goel and A. Zellner, pp. 233–243. Amsterdam: North-Holland/Elsevier. [2]