Selecting the derivative of a functional covariate in scalar-on-function regression

Giles Hooker 1,2 · Han Lin Shang 3

Received: 6 July 2021 / Accepted: 11 March 2022 / Published online: 23 April 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
This paper presents tests to formally choose between regression models using different derivatives of a functional covariate in scalar-on-function regression. We demonstrate that for linear regression, models using different derivatives can be nested within a model that includes point-impact effects at the end-points of the observed functions. Contrasts can then be employed to test the specification of different derivatives. When nonlinear regression models are employed, we apply a C test to determine the statistical significance of the nonlinear structure between a functional covariate and a scalar response. The finite-sample performance of these methods is verified in simulation, and their practical application is demonstrated using both chemometric and environmental data sets.

Keywords Model selection · Variable selection · Likelihood ratio test · C test

1 Introduction
Recent advances in computer recording and storing technology facilitate functional data sets, which motivated many researchers to consider various functional regression models for estimating the relationship between predictor and response variables, where at least one variable is function-valued. The functional formulation of a linear model dates back to a discussion by Hastie and Mallows (1993), Dalzell and Ramsay (1993); see Ramsay and Silverman (2005) for a detailed overview and Ramsay et al. (2009) for software implementation.

Since then, models to incorporate functional variable have been extended to include generalized linear models (Aguilera et al. 2008; Müller and Stadmüller 2005), additive regression (Febrero-Bande and González-Manteiga 2013; McLean et al. 2014), polynomial models (Yao and Müller 2010), non-parametric functional regression models (Ferraty and Vieu 2006), semi-functional partial linear models (Aneiros-Pérez and Vieu 2006, 2008) and many more. Because the fast development in functional regression models, it has received increasing popularity in various fields of application, such as age-specific mortality and fertility forecasting in demography (Hyndman and Shang 2009), analysis of spectroscopy data in chemometrics (Ferraty and Vieu 2002), earthquake modeling (Quintela-del-Río et al. 2011) and ozone-level prediction (Quintela-del-Río and Francisco-Fernández 2011).

Despite relatively mature literature on functional models, there has been little attention to selecting which derivative of observed functional data $X(t)$ to use as a covariate. One distinguishing feature of functional data is access to multiple derivatives $X^{(k)}(t)$ of $X(t)$. It is therefore natural to consider using one or more of these as a covariate. Indeed Ferraty and Vieu (2002) discusses the use of semi-metrics based on derivatives for non-parametric regression, and Ferraty and Vieu (2009) empirically finds that the use of second derivatives provides significant performance improvement in the example data set we use below. The choice of derivative can be empirically motivated or derived from domain knowledge; in the example of particulate matter emissions studied below, we may expect vehicular acceleration, hence engine workload, to be the relevant predictor of emissions, although this should be empirically assessed. This paper examines formal methods of comparing models that use different derivatives of $X(t)$. 

---

1 Department of Statistics, University of California, Berkeley, California 94720, USA
2 Research School of Finance, Actuarial Studies and Statistics Australian National University, Canberra, Australia
3 Department of Actuarial Studies and Business Analytics, Macquarie University, Sydney, Australia

Giles Hooker
ghooker@berkeley.edu

© Springer
Below we distinguish between two forms of the model. When $X(t)$ or its derivatives enter the model linearly, integration by parts provides a means of embedding smooth linear functionals of both $X(t)$ and $X^{(1)}(t)$ within a larger space:

$$
\int_0^1 \gamma(t)X^{(1)}(t)dt = \gamma(1)X(1)
$$

$$
-\gamma(0)X(0) - \int_0^1 \gamma^{(1)}(t)X(t)dt.
$$

Here $\gamma(t)$ is a weight function that generalizes linear coefficients to functional covariates. Starting from using $X$, we can assess whether $X^{(1)}$ is more appropriate by first testing the expansion of a model using $X(t)$ to include separate point-impact effects for the endpoints $X(0)$ and $X(1)$. This will not necessarily correspond to a functional linear model of $X^{(1)}$ unless the estimated coefficients also correspond to some coefficient function, i.e., $\alpha(1) = \alpha(0) + \int_0^1 \alpha^{(1)}(t)dt$.

We thus formulate a contrast to test whether the reduction of the expanded model corresponds to the left-hand side of (1). These can be done via $F$ tests formulated for a penalized linear regression. The same formulation allows us to reverse the inference – to start with $X^{(1)}(t)$ and test for $X(t)$ – and to consider changes of more than one derivative as well.

This analysis illustrates the geometry of the space of models on different derivatives: their subspaces intersect apart from one dimension of variation in each case. We can already observe from (1) that models employing both $X(t)$ and $X^{(1)}(t)$ are equivalent to using only one along with point impact terms for $X(1)$ and $X(0)$.

When $X^{(k)}(t)$ does not enter the model through a smooth linear operator, the formulation in (1) cannot be applied generically. Instead, we propose a form of the $C$ test of Davidson and MacKinnon (1981) to allow us to form nested models after estimating each separately. While this approach or the use of model selection methods and information criteria could be employed for functional linear models, the approach we take here elucidates the geometry of the decision space and allows for better control and reporting of error probabilities.

Our paper proceeds as follows. In Sect. 2, we develop linear contrasts to test the adequacy of different derivatives within a linear model specification. Although not investigated here, these methods can be readily extended to generalized linear models. Section 3 examines a $C$ test for explicitly nonlinear models that can also be used in conjunction with functional principal components regression. Section 4 provides some simulation results for our methods and Sections 5 and 6 illustrate these tests on the Tecator data (see also Ferraty and Vieu 2006) and on the particulate matter data explored in Asencio et al. (2014).

Below, we distinguish between a data generating process and a fitted model. Here we use the notation $y_i = f(X_i) + \epsilon_i$ to indicate a data generating process or a hypothesized model (where $f$ may be given by a linear model in terms of functional parameters), and $y_i \sim g(X_i)$ to indicate that we fit the model $g(X_i)$ (including any parameters) to the data. For this paper, we will only consider fitting by penalized least squares, i.e., minimizing $\sum |y_i - g(X_i)|^2 + P(g)$ where $P(g)$ is quadratic in any parameters that are used to fit. However, extensions to other likelihoods are fairly immediate.

### 2 Contrasts for tests between derivatives

This section develops formalized tests between functional derivatives used as covariates. Without loss of generality, we assume a collection of data $[y_i, X_i(t)]$ for $i = 1, \ldots, n$ with each $X_i$ a function of the interval $[0, 1]$. We consider a model of the form

$$
y_i = \alpha + \int_0^1 \beta(t)X_i^{(k)}(t)dt + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),
$$

where $\alpha$ provides an intercept and $\beta(t)$ is a weight function over $t$. Here $k$ indicates the order of the derivative to use, typically $k \in \{0, 1, 2\}$, and a central challenge is the choice of $k$.

A nested test can always be constructed by estimating a model that includes multiple derivatives:

$$
y_i \sim \beta_0 + \sum_k \int_0^1 \beta_k(t)X_i^{(k)}(t)dt.
$$

We explore this framework below, but we note that this test is complicated by the association between the derivatives of $X_i(t)$. These derivatives do not cover the same linear space, but they overlap considerably. In this paper, we apply integration by parts to compare models with different derivatives; we can embed both models into a common space by adding a finite number of point impacts and constructing a set of contrasts to distinguish different derivatives.

We begin by setting up contrasts to produce tests between derivatives defined through integration by parts and then discuss the numerical implementation of tests of these contrasts within common functional data packages. We will do this for three specific tests: taking $X_i$ as a baseline and testing whether $X_i^{(1)}$ is more appropriate, the reverse procedure starting from $X_i^{(1)}$ and testing whether $X_i$ is better, and testing a change of two derivatives from $X_i$ to $X_i^{(2)}$. While these tests can be given as special cases of a more general procedure, we expect they cover all the cases that are likely to be practically relevant.
2.1 Taking one more derivative

We start by considering a null hypothesis functional linear model

\[ H_0 : y_i = \alpha + \int_0^1 \beta_0(t)X(t)dt + \epsilon_i, \quad (2) \]

and wish to test in the direction of a model that is based on the first derivative of \(X(t)\):

\[ H_a : y_i = \alpha + \int_0^1 \beta_1(t)X^{(1)}(t)dt + \epsilon_i. \quad (3) \]

To carry such a test out, we need to formulate a contrast that we obtain via integration by parts in (1). Here we can conduct a nested test for the adequacy of the model at \(k = 0\) by estimating the functional linear model augmented with point impacts at the endpoints:

\[ y_i \sim \alpha + \gamma_0 X_i(0) + \gamma_1 X_i(1) + \int_0^1 \gamma(t) X_i(t)dt, \quad (4) \]

in which scalar coefficients \((\alpha, \gamma_0, \gamma_1)\) are estimated along with the functional coefficient \(\gamma(t)\). This model adds two degrees of freedom to the original functional linear model (2) and can thus be represented as a nested test with an appropriate contrast matrix. However, this model is over-specified and corresponds to (3) only under the constraint:

\[ \gamma_0 + \gamma_1 + \int_0^1 \gamma(t)dt = 0. \quad (5) \]

The left-hand side of this equation represents a linear contrast that can be tested via a Wald-type procedure or equivalently by solving for \(\gamma_1\) and fitting the model

\[ y_i \sim \alpha + \gamma_0 [X_i(0) - X_i(1)] \]
\[ + \int_0^1 \gamma(t)[X_i(t) - X_i(1)]dt, \quad (6) \]

from which it should be clear that (2) cannot be expressed as being nested within (6). Because these are equivalent tests, we will take the first approach and test the agreement with model (3) through the contrast (5).

We can thus define a two-stage procedure:

1) Test the significance of \(\gamma_0\) and \(\gamma_1\) in (4) to assess the adequacy of using \(X(t)\) as a covariate relative to the alternative \(X^{(1)}(t)\).

2) Test the significance of the contrast (5) as a goodness of fit assessment of the functional linear model using \(X^{(1)}(t)\) as a covariate.

As we discuss below, the test of the contrast should, in theory, be equivalent to a comparison of (3) with (4). However, the numerical implementation of these tests in commonly-used functional data analysis software may render this correspondence inexact in practice. We recommend assessing both the estimated (3) and (4) under constraint (5) when choosing a model.

2.2 Taking one less derivative

Using similar arguments, we can also consider testing a lower-order derivative as an alternative. To illustrate this, we swap the hypotheses (3) and (2) so that we start with a model for \(X^{(1)}(t)\) and consider \(X(t)\) as an alternative. Here again, integration by parts yields

\[ \int_0^1 \beta_0(t)X_i(t)dt = \beta_0^{(-1)}(1)X_i(1) - \beta_0^{(-1)}(0)X_i(0) \]
\[ - \int_0^1 \beta_0^{(-1)}(t)X_i^{(1)}(t)dt, \]

where we have used the anti-derivative

\[ \beta_0^{(-1)}(t) = \int_0^t \beta_0(s)ds, \]

and we will set \(\beta_0^{(-1)}(0) = 0\) since a constant can be added arbitrarily. Thus we can test the adequacy of null hypothesis (3) with (2) as a potential alternative by fitting a model

\[ y_i \sim \delta_1 X_i(1) + \int_0^1 \delta(t)X_i^{(1)}(t)dt, \]

and testing \(H_0 : \delta_1 = 0\). As above, the additional degrees of freedom over-specify the model and an exact agreement with (3) requires the constraint:

\[ \delta_1 - \delta(1) = 0. \]

These can again be tested in a two-stage procedure.

2.3 Moving more than one derivative

The same arguments can be extended to test moves of more than one derivative. For example, we may test between the null (2) and alternative

\[ H_a : y = \alpha + \int_0^1 \beta_2(t)X^{(2)}(t)dt + \epsilon \quad (7) \]

by iterating integration by parts:

\[ \int_0^1 \beta_2(t)X_i^{(2)}(t)dt = \beta_2(1)X_i^{(1)}(1) - \beta_2(0)X_i^{(1)}(0) \]
which can be assessed by including end-point impacts for \( X \) and \( X^{(1)} \):

\[
y_i \sim \zeta_{00} X_i(0) + \zeta_{01} X_i(1) + \zeta_{10} X_i^{(1)}(0) + \zeta_{11} X_i^{(1)}(1) + \int_0^1 \zeta(t) X_i(t)dt,
\]

with the alternative model based on \( X_i^{(2)}(t) \) corresponding to the contrasts:

\[
\begin{align*}
\zeta_{00} + \zeta_{01} - \int_0^1 \left( \zeta_{10} + \int_0^t \zeta(s)ds \right) dt &= 0,
\end{align*}
\]

\[
\zeta_{10} + \zeta_{11} + \int_0^1 \zeta(t) dt = 0.
\]

The same two-step procedure can then be used to assess the fit of both models.

Beyond providing a framework for constructing nested tests to move between derivatives, the results above also allow us to understand the power that we have to distinguish between potential models. In particular, we observe from (4) that these models will be indistinguishable if \( \beta_1(0) = \beta_1(1) = 0 \). If the \( X_i(t) \) are periodic with \( X_i(0) = X_i(1) \) — if represented by Fourier components, for example — (4) is not estimable, but we also have no power if \( \beta_1(0) = \beta_1(1) \). Testing in the converse direction will similarly have no power if \( \int_0^1 \beta_0(t) dt = 0 \).

### 2.4 Some numerical comments

We provide a numerical implementation of the test above through the penalized basis expansion framework described in Ramsay and Silverman (2005) and taken up in several software packages, see fda package of Ramsay et al. (2020), fda.usc package of Febrero-Bande and Oviedo de la Fuente (2012) and refund package of Goldsmith et al. (2020). Before describing the calculation that we undertake, we first note a number of numerical issues that may make the correspondence between fitting the model (4) under constraints (5) inexact and some consequences of this.

The first observation is that derivatives are not necessarily represented exactly in many popular FDA software packages. For example, the \( \hat{\mathcal{f}} \hat{\mathcal{d}} \)a package represents the functions \( X_i \) via a basis expansion. The derivatives of such functions need not themselves be within the span of this basis, but \( \text{deriv.} \hat{\mathcal{f}}\hat{\mathcal{d}} \) will create \( X_i^{(k)} \) as functional data by projecting the derivative onto the basis expansion for \( X_i \). This introduces a numerical error into the integration by parts formula (1) whose severity depends on the basis used and the smoothness of the \( X_i \).

Additionally, any estimate for \( \beta_k(t) \) is subject to bias associated with the basis expansion or the smoothing penalty. Thus in (4) it may be easier, with finite data, to estimate \( \gamma(t) \) with target \( -\beta_1^{(1)}(t) \) than to estimate \( \beta_1(t) \) directly in (3) or \textit{vice versa}. To account for this, we have introduced a noncentrality parameter in the tests we describe in Sect. 2.5. However, these biases can still affect the level or power of the test, particularly when one representation of the relationship is significantly smoother than another.

Both of these observations mean that the observed squared error from fitting (4) under constraints (5) may be different from that for fitting (3) despite these being theoretically equivalent. If, as we find in the Tecator data in Sect. 5, our contrasts both conclude that (2) is inadequate but (3) is not, the choice of using (4) versus re-fitting (3) depends on their predictive performance and the purpose of the modeling exercise.

In this paper, we have not examined the use of functional principal components regression (Yao et al. 2005). If we use the eigenfunctions for some derivative \( X^{(k)} \) as a basis expansion (that we hold fixed even when examining a different derivative), the calculations below remain unchanged. However, we would expect a strong bias towards using the derivative that produced the eigenfunctions. It may be more natural to project onto a different eigenbasis for each derivative, in which case the \( C \) test detailed in Sect. 3 can be employed, but the choice of representation from \( \beta_0 \) to \( \beta_1 \) is much harder to account for mathematically.

### 2.5 Implementation

All the models described above can be fit through functions in one of several software packages that carry out functional data analysis. Our discussion centers on using a single functional covariate, but the extension to an additional scalar or functional covariates is straightforward.

We use a basis expansion \( \Phi(t) = [\phi_1(t), \ldots, \phi_K(t)] \) to represent \( \beta^k(t) \) and define the design matrix

\[
\left[ Z^k \right]_{ij} = \int_0^1 X_i^{(k)}(t) \phi_j(t) dt,
\]

so that \( \int_0^1 X_i^{(k)}(t) \gamma(t) dt = Z_i^k g \) for a vector of coefficients \( g \). We write \( X_0 \) and \( X_1 \) for the vectors containing the \( X_i(0) \) and \( X_i(1) \) respectively and we assume that a quadratic smoothing penalty is applied to \( \gamma(t) \) that can be represented as \( g^\top P g \) for some matrix \( P \) (e.g. see Ramsay et al. 2009).

We then estimate parameters in (4) by minimizing

\[
\left\| Y - \alpha \gamma_0 X_0 - \gamma_1 X_1 - Z^0 g \right\|^2 + \lambda g^\top P g,
\]
which gives

\[ \hat{g} = \left( \hat{Z}^\top \hat{Z} + \lambda \hat{P} \right)^{-1} \hat{Z}^\top Y, \]

using the augmented objects \( \hat{g} = (\alpha, \gamma_0, \gamma_1, g) \), \( \hat{Z} = [1, X_0, X_1, Z^0] \) and \( \hat{P} \) contains \( P \) preceded by three rows and columns of 0’s.

We can estimate \( \sigma^2 \) from

\[ \hat{\sigma}^2 = \frac{1}{n - df} \left| Y - \hat{Z} \hat{g} \right|^2, \quad df = \text{tr} \left( \hat{Z} \left( \hat{Z}^\top \hat{Z} + \lambda \hat{P} \right)^{-1} \hat{Z}^\top \right). \]

Using the sandwich matrix

\[ V = \left( \hat{Z}^\top \hat{Z} + \lambda \hat{P} \right)^{-1} \hat{Z}^\top \hat{Z} \left( \hat{Z}^\top \hat{Z} + \lambda \hat{P} \right)^{-1}, \]

we can obtain an \( F \) statistic for the contrast \( C \hat{g} \)

\[ F = \frac{1}{pC \hat{\sigma}^2} \hat{g}^\top C^\top \left( C^\top V C \right)^{-1} C \hat{g}, \]

where we are interested in the contrast matrices

\[ C_1 = [0_{2 \times 1} I_{2 \times 2} 0_{2 \times k}] \]

to assess the significance of \( \gamma_0 \) and \( \gamma_1 \) and

\[ C_2 = [0 \ 1 \ 1 \ m] \]

to test (5) where \( m_j = f_0^1 \phi_j(t) dt \) and \( p_C \) is the dimension of the column space of \( C \).

Under the null hypothesis, and ignoring smoothing and numerical biases, \( F \) should be distributed as an \( F \) statistic with degrees of freedom corresponding to \( p_C \) and \( n - df \). However, smoothing can generate a significant bias in favor of the alternative, compromising the level of the test. To account for this, we introduce a noncentrality parameter. Recall that for \( \nu_j \sim N(\mu_j, \sigma^2) \), \( j = 1, \ldots, p \) and \( \epsilon_i \sim N(0, \sigma^2) \), \( i = 1, \ldots, n \), then \( \sum_j \nu_j^2 / \sum_i \epsilon_i^2 \) has a noncentral \( F \) distribution with degrees of freedom \( p \) and \( n \) and noncentrality parameter \( \sum_j \mu_j^2 / \sigma^2 \). Here, for each \( \lambda \), we treat the smoothing bias in \( \hat{g} \) as an offset which we estimate from the effect of applying the smoothing parameter \( \lambda \) to the fitted values of a minimally-smoothed model, where the fitting is expected to asymptotically remove the noise associated with measurement error. Specifically, our calculations proceed as follows:

1. Obtain fitted values \( \hat{Y}_0 \) and an estimate of residual variance \( \hat{\sigma}_a^2 \) using the alternative model (4), using the smallest smoothing parameters that allow for model identifiability; in our implementation we used \( \lambda = 10^{-11} \).

2. Project \( \hat{Y}_0 \) onto the null hypothesis space of our test as follows:

\[ \hat{Y}_0 = \hat{Z}(I - C^\top (CC^\top)^{-1}C)(\hat{Z}^\top \hat{Z})^{-1} \hat{Z}^\top \hat{Y}_a. \]

This first represents \( \hat{Y}_a \) in terms of the coefficients in the model (4), and then projects into the null space of \( C \), being equivalent in this case to setting \( \gamma_0 = \gamma_1 = 0 \).

3. Re-obtain coefficients from the projected \( \hat{Y}_0 \)

\[ \hat{g}_0 = \left( \hat{Z}^\top \hat{Z} + \lambda \hat{P} \right)^{-1} \hat{Z}^\top \hat{Y}_0, \]

and form the noncentrality parameter

\[ \eta = \frac{1}{\hat{\sigma}_a^2} \hat{g}_0^\top C^\top \left( C^\top V C \right)^{-1} C \hat{g}_0. \]

We now test \( F \) against the relevant quantile of an \( F \) distribution with \( p_C \) and \( n - df \) degrees of freedom and noncentrality parameter \( \eta \). The noncentrality parameter corrects the level of the test for smoothing bias; from (1), if \( \beta_1(0) = \beta_1(1) = 0 \) using either \( X(t) \) or \( X^{(1)}(t) \) is equivalent and a distinction between them depends on whether \( \beta_1(t) \) or \( \beta_1^{(1)}(t) \) incurs more bias. In the context of our test, the point impacts at 0 and 1 can be correlated with an over-smoothed linear functional, thereby affecting the level of the test. We base the noncentrality parameter on the orthogonal projection of an under-smoothed model onto the null hypothesis space to estimate this bias as well as possible.

Notice that \( \hat{Y}_a \) and \( \hat{\sigma}_a \) are calculated as part of a search over smoothing parameter values. Thus the calculation of the noncentrality parameter only requires a second penalized regression at each value of \( \lambda \).

In a similar fashion, we may test (2) as an alternative to a null hypothesis (3) by applying the same structure as above, forming \( \tilde{Z} = [1, X_1, Z^1] \) and the same contrast \( C_1 \) to reject (3) and then assessing

\[ C_2 = [0 \ 0 \ -1 \ \Phi(1)] \]

using the analogous statistic and noncentrality parameter as above.

Similarly a comparison of the 0th with the 2nd derivative as an alternative can be made by letting \( X_0' = X^{(1)}(0) \) and \( X_1' = X^{(1)}(1) \) and setting \( \tilde{Z} = [1, X_0, X_1, X_0', X_1', Z^0] \) and assessing

\[ C_1 = [0_{4 \times 1} I_{4 \times 4} 0_{4 \times k}] \]

and

\[ C_2 = \begin{bmatrix} 0 \ 0 \ 0 \ 1 \ 1 \ m \\ 0 \ 1 \ 1 \ -1 \ 0 \ -m \end{bmatrix} \]
with \( \bar{m}_j = \int_0^1 \int_0^t \phi_j(s)dsdt \). When not available analytically, we can obtain this vector by observing that

\[
\int_0^1 \int_0^t \phi_j(s)dsdt = \int_0^1 \phi_j(t)dt - \int_0^1 t \phi(t)dt
\]

by a further integration by parts argument. The expressions above can be evaluated by, for example, the function inprod in the fda package. See the code in the supplementary materials for details.

### 3 C Test for more general models

An analysis using integration by parts as described above requires \( X^{(i)}(t) \) to enter the model via a smooth linear operator \( \int_0^1 \beta^i(t)X^{(i)}(t)dt \). When this is not the case, for example, in the non-parametric regression methods of Ferraty and Vieu (2002) or the additive models in McLean et al. (2014), we cannot generically embed models based on \( X^{(k)}(t) \) and \( X^{(j)}(t) \) for \( j \neq k \) within a common space and thereby allow a nested hypothesis test. Here we explore a naïve alternative given by the \( C \) test to assess non-nested models in Davidson and MacKinnon (1981) but modified by sample splitting to account for using pre-estimated models.

Specifically, we consider hypotheses

\[
H_0 : \ y = m(X) + \epsilon \\
H_1 : \ y = s(X^{(1)}) + \epsilon,
\]

where \( \epsilon \) denotes independent normally distributed error term with mean 0 and variance \( \sigma^2 \). We then employ the fitted values after fitting each of \( \hat{m} \) and \( \hat{s} \) in linear regression:

\[
y - \hat{m}(X) = \alpha + \hat{\theta}_0 \hat{m}(X) + \hat{\theta}_1 \hat{s}(X^{(1)}) + \epsilon,
\]

and our test becomes that of

\[
H_0 : \ \theta_1 = 0
\]

in the linear regression.

The form of the model in (10) allows us to exactly represent both \( H_0 \) as \( \theta_0 = \theta_1 = 0 \) and \( H_1 \) by \( \theta_0 = -1 \) and \( \theta_1 = 1 \). In addition to testing \( \theta_1 = 0 \), we also conduct a goodness of fit test of \( \theta_0 = 0 \) resulting in two tests, of which both should be rejected. We expect, in general, that we should have \( \theta_0 \approx -\theta_1 \) with correspondingly similar \( p \)-values except under \( H_0 \) and indeed find this is the case.

The framework of this test is, in fact, very general and would allow us to test the structure of a model (e.g., linear versus nonparametric) as well as between derivatives. In our experiments below, we consider both \( m \) and \( s \) as functional linear models and nonparametric kernel regression.

Our model (10) is a slight elaboration on the original test in Davidson and MacKinnon (1981) which corresponds to fixing \( \alpha = 0 \) and \( \theta_0 = -\theta_1 \). For nonlinear models, Davidson and MacKinnon (1981) proposed a \( J \) test that estimated any parameters in \( m \) along with \( (\alpha, \theta_0, \theta_1) \) jointly, and a \( P \) test following our scheme but with an added linear effect for the Jacobian of \( m \) with respect to parameters in order to account for their estimation. Here we examine possibly non-parametric models for both \( m \) and \( s \) which might not be trainable jointly with \( (\alpha, \theta_0, \theta_1) \) (for example the kernel methods in Ferraty and Vieu (2002), that we examine here), or may not have a Jacobian. Davidson and MacKinnon (1981) advocate the use of the \( C \) test only as a preliminary assessment because it involves first estimating \( \hat{m} \) and \( \hat{s} \) and then fitting (10) on the same data, rendering the standard linear regression standard errors inconsistent. Here, we avoid this issue by employing subsample-splitting (see, e.g., Jarque 1987,) to estimate \( \hat{m} \) and \( \hat{s} \) on half of the data (selected at random) and once these are estimated, conduct a test for \( \theta_0 \) and \( \theta_1 \) using their fitted values on the second half. This makes the test statistic for \( \theta_1 \) independent of \( \hat{m} \) and \( \hat{s} \) and the resulting test at least conservative.

Our procedure is summarized below, based on first separating the data into disjoint sets \( S_1 \) and \( S_2 \):

1. Fit separates nonparametric scalar-on-function regressions using \( S_1 \), to obtain both \( \hat{m} \) and \( \hat{s} \) along with their fitted values on \( S_2 \).
2. Conduct a \( t \) test for the significance of \( \theta_1 \) in (10) using fitted values from \( S_2 \). Assess the test for \( \theta_1 \) only if we do not reject \( H_0 : \theta_0 = 0 \).

### 4 Simulation examples

We explore the properties of the tests described above through a simulated framework. For this, we use a generative model

\[
y_i = \int_0^1 \beta(t)X_i^{(1)}(t)dt + \epsilon_i, \quad \epsilon_i \sim N(0, 0.01),
\]

where we set

\[
\beta_1(t) = b_0 + 0.5 \sin(2\pi t) + 0.3 \sin(4\pi t) + 0.1 \sin(6\pi t),
\]

and we note that when \( b_0 = 0 \), we have \( \beta_1(0) = \beta_1(1) = 0 \) and the following models also hold

\[
y_i = \int_0^1 \beta_1^{(1)}(t)X_i(t)dt + \epsilon_i,
\]

\[
y_i = \int_0^1 \beta_1^{(-1)}(t)X_i^{(2)}(t)dt + \epsilon_i,
\]
both of which we will use as null hypotheses below. In this framework, varying $b_0$ allows us to test power.

We generated functional covariates $X_i$ by generating random coefficients for a Fourier basis with 25 basis functions plus linear and exponential terms. Specifically

$$X_i(t) = d_0 + d_1 \frac{(t - 1/2)}{2} + d_2 e^{t(1/2)} + \sum_{k=1}^{12} (f_k e^{\sin(-k(3/2)0)} \sin(2\pi k t) + g_k e^{-(k-1) \cos(2\pi k t)}),$$

where all coefficients $d_j, f_j, g_j$ are independently normally distributed, and we have included scaling factors as part of the basis. These are then projected onto a B-spline basis of order 6 with 21 knots (see Ramsay et al. 2009). The projected functions are then used as the covariates when generating the $y_i$ as in (11).

Throughout the following, we represent coefficient functions via a basis comprising of the functions $\{1, t, \sin(2\pi k t), \cos(2\pi k t)\}_{k=1}^{12}$. This is a Fourier basis augmented with a linear term. The linear term is necessary to represent $p_1^{(c-1)}(t) = \int_0^t \beta_1(s) ds$ when $b_0 \neq 0$ and we have included it in all estimation procedures. We also take $P$ to be derived from a second derivative penalty $\int_0^1 \hat{\gamma}^{(2)}(t)^2 dt$ in each of the models.

Figure 1 demonstrates the power of the three tests detailed in Sect. 2 as a function of $b_0$ along with $C$ tests based on linear and nonparametric functional models. We generated a sample of 250 covariates as above and reproduced the responses $y_i$ 1000 times to obtain power. Our procedure tests two contrasts, only the first of which should be rejected, and we therefore do not apply a multiple testing correction. The power curves that we report give the probability of rejecting each contrast and the combined probability of arriving at the correct model: rejecting the first contrast but failing to reject the second. We report these values both when using a small but fixed value of $\lambda$ — enough to ensure that matrix inverses are well defined — and for $\lambda$ chosen by ordinary cross-validation.

We also applied the $C$ test to each data set where we record the rejection rate of both $b_0$ and $\theta_1$ where we have used functional linear models. Both with a small value of $\lambda$ and when chosen by OCV, a functional kernel regression as implemented using the default parameters from fregre.np in the fda.usc package (Febrero-Bande and Oviedo de la Fuente 2012).

We find that the power of testing $X^{(2)}$ versus $X^{(1)}$ increases much more rapidly than testing $X$ versus $X^{(1)}$ (note two orders of magnitude difference in the range of the x-axes in Fig. 1). Using $\lambda$ chosen by cross-validation improves power relative to $\lambda = 10^{-9}$ for testing $X$ versus $X^{(1)}$, but the opposite is true when testing $X$ versus $X^{(2)}$. This is likely due to smoothing effects; the relationship for $X^{(2)}$ is considerably smoother than for $X$, thus introducing significant smoothing bias and, therefore, a large noncentrality parameter. Without the noncentrality parameter, $Z$ can have a very high correlation with the point impacts $\{X(0), X(1), X^{(1)}(0), X^{(1)}(1)\}$ which then compensates for the bias induced by smoothing penalties and results in the level of the uncorrected test exceeding 0.5. By contrast, the point impacts in the test of $X^{(2)}$ versus $X^{(1)}$ are nearly uncorrelated with the functional component of the model, partly accounting for its higher power. This result suggests that including point impacts may be useful to alleviate smoothing bias in functional linear regression, whether or not they necessarily imply the use of a different derivative.

The $C$ test results reinforce these conclusions. We observe that inflated $\alpha$-levels at $b_0 = 0$ for functional linear models, likely due to smoothing bias and lower power than the test based on contrasts. A noncentrality correction along the lines of the contrast test may be possible, but we have not pursued it here. Under-smoothing results in a larger power reduction, likely due to over-fitting. We also observe very little power
in the $X$ versus $X^{(2)}$ test – in this case, and the generating model is not contained within either model class. We observe power near one for the test between nonparametric models, which we attribute to smoothing bias.

Figure 2 presents further analysis of these simulation results. We confine ourselves only to tests of the null hypothesis since tests of alternative hypotheses maintain the alpha-level for contrasts and are poorly behaved for C tests. The upper row plots the power of each of the tests we examine as the smoothing parameter $\lambda$ is varied. Here we observe similar patterns between tests of different derivatives and that the optimum lambda tends to be small among the values that we investigated. The bottom row examines the power of tests with cross-validated values of $\lambda$ and $C$ tests for sample sizes 50, 100, and 250. Here we see power increasing with sample size, most noticeably for tests of $X^{(2)}$ versus $X^{(1)}$, but these do not improve the level of $C$ tests.

5 Tecator data

We illustrate the proposed tests using an example that focuses on estimating the fat content of meat samples based on near-infrared (NIR) absorbance spectra. These data were obtained from http://lib.stat.cmu.edu/datasets/tecatore, and have been studied by Ferraty and Vieu (2006), and Aneiros-Pérez and Vieu (2006), among many others. Each sample contains finely chopped pure meat with different percentages of fat, protein, and moisture contents. For each unit $i$ (among 215 pieces of finely chopped meat), we observe one spectrometric curve, denoted by $T_i$, which corresponds to the absorbance measured at a grid of 100 wavelengths (i.e., $T_i = [T_i(t_1), \ldots, T_i(t_{100})]$). We also observe its fat, protein, and moisture contents $X \in R^3$, obtained by chemical processing. Graphical displays of the original spectrometric curves and their first and second derivatives are shown in Fig. 3.

The ability to consider derivatives, as a by-product of conceiving the data as functions, is of great advantage for inference, modeling, and forecasting (see, e.g., Ramsay and Hooker 2017) and data visualization (see, e.g., Shang 2019). In chemometrics, derivative spectroscopy uses first or higher derivatives of absorbance with respect to wavelength for qualitative analysis and quantification. The use of derivatives of spectral data was introduced in the 1960s when it was shown to have many advantages (see, e.g., Savitzky and Golay 1964), corroborated on this data set by the performance improvement relative to other derivatives found in Ferraty and Vieu (2009). While these data are well-studied, we have chosen these data to illustrate the long-standing agreement that the second derivative of the spectrum is most predictive of fat content, which we test formally here.

Here, we extend the functional linear models described above to include protein and moisture content as scalar covariates and apply the procedures described in Sect. 2 to determine the optimal derivative use based on a linear specification:

$$y_i = \beta_0 + Z_i \beta + \int_{850}^{1050} \beta^{(k)}(t)X^{(k)}_i(t)dt + \epsilon_i,$$

(12)

where $Z_i$ represents the linear effect of covariates, which can be incorporated naturally into the tests above. Applying our linear-specification test, we find that tests of $X$ versus $X^{(1)}$, $X^{(1)}$ versus $X^{(2)}$ and $X$ versus $X^{(2)}$ all reject at the cross-validated value of $\lambda$; $X$ versus $X^{(1)}$ rejects the second test indicating possible further model elaborations, but no others do. We also tested the converse $X^{(2)}$ versus $X^{(1)}$ without rejecting. These results are displayed graphically in the bottom panel of Fig. 3, where we have plotted p-values of each test as a function of $\lambda$ and indicated values minimizing cross-validation with vertical lines.

The second test for $X$ versus $X^{(2)}$ also rejects, suggesting that simply using the second derivative may not be adequate.
We assess the robustness of this conclusion by employing a C test of both parametric and non-parametric models. To do this, we selected 160 observations at random and used these to obtain models, with the remaining 65 observations used to carry out the test based on fitted values. C tests based on functional linear models with smoothing parameters chosen by ordinary cross validation repeated the patterns found using our contrast tests: tests rejected $X^{(0)}$ and $X^{(1)}$ in favor of $X^{(2)}$ but failed to distinguish between the first and second derivative or to find any improvement in fit by adding either when $X^{(2)}$ is used as the null hypothesis. Tests based on nonparametric models without accounting for protein and moisture obtained from default settings in \texttt{fregre.np} in the \texttt{fda.usc} package found significant results in all tests except when $X^{(2)}$ is used as the null hypothesis. Summaries of all tests are given in Table 1.

We also ran a C test in which a functional linear model based on $X^{(2)}$ was taken as a null hypothesis, and we used the corresponding nonparametric model as an alternative. This test did not reject (p-value of 0.975); while our nonparametric model did not include protein and moisture covariates, we would expect large departures from linearity to be still detectable. This gives us confidence that a functional linear model based on the second derivative of the spectrum is indeed optimal for predicting fat content.
Table 1  Results of tests of derivatives for the tecator data. Note that in the cast of contrast-based tests, p-values assess the goodness of fit of the null hypothesis and alternative hypothesis model, whereas C tests assess the significance of the coefficient of each model when used to predict out-of-sample residuals from the null hypothesis. In the case of C tests, the similarity of p-values is expected when the alternative substitutes for the null model resulting in coefficients near 1 and −1 respectively.

| Hypothesis | Contrasts | Linear C test | Nonparametric C test |
|------------|-----------|---------------|----------------------|
|            | Null      | Alt           | Null                 | Alt           |
| $X^{(0)} v X^{(1)}$ | 0.016     | 0748          | 0.411                | 0.416         | 3.31e-13 | < 2e-16 |
| $X^{(0)} v X^{(2)}$ | 3.09e-9   | 0.004         | 0.032                | 0.032         | < 2e-16 | < 2e-16 |
| $X^{(1)} v X^{(2)}$ | 0.003     | 0.524         | 0.022                | 0.022         | 8.80e-07 | 7.08e-09 |
| $X^{(2)} v X^{(1)}$ | 0.424     | 0.718         | 0.315                | 0.310         | 0.553   | 0.351   |

Fig. 4  Left: plots of velocity and acceleration covariates, colored by whether log PM$_{10}$ > 0. Right: p-values for functional linear models based on contrasts as a function of λ.

6 Particulate matter data

We repeat our analysis using the particulate matter data first examined in Asencio et al. (2014) and re-examined in McLean et al. (2014) and Hall and Hooker (2016). These data relate log base ten particulate matter measured at the truck’s tailpipe to the recent history of driving conditions. Here we follow Hall and Hooker (2016) and examine the relationship between log PM$_{10}$ and the past 30 seconds of vehicle velocity as the conservative upper end of potential dependence. The papers above all made use of acceleration (i.e., the first derivative of measured velocity) as a functional covariate; the tests employed here check for the adequacy of velocity, acceleration, and using a further derivative. To avoid smoothing biases at the endpoints of the functional data, we initially obtain covariate functions from the range [-10, 50] and project the resulting smooth and its derivatives to a basis expansion covering the range [0, 30] before proceeding with our analysis. Figure 4 presents plots of both acceleration and velocity along with the p-values from our contrasts tests as a function of smoothing parameters.

Table 2 presents the results of the tests described above, and here our evidence is less clear than in the tecator data. Tests based on contrasts only suggest that using $X^{(3)}$ is inadequate. $C$ tests based on both linear models also prefer $X^{(1)}$ to $X^{(0)}$ but re-scale $X^{(1)}$ when it occurs in the model, possibly as a result of smoothing bias. $C$ tests using nonparametric specifications prefer $X^{(1)}$ to either alternative, although they still re-scale significantly. McLean et al. (2014) found evidence of a nonlinear relationship between log PM$_{10}$ and $X^{(1)}$; a $C$ test between the linear and nonparametric specifications agrees with this, with a p-value of 3.70e-05 for the coefficient of the nonparametric fitted values.

We note that under the assumption that $X(t)$ (or its derivatives) cease affecting PM$_{10}$ past some time-point $\theta < 30$, we should have $\beta(30) = 0$, meaning that we could change the contrast tests to reduce the number of point impacts to consider. However, this would also require explicitly con-
straining $\beta(30) = 0$ for the relevant models, and we have not pursued this here.

### 7 Discussion

The use of derivatives is a feature that distinguishes functional from multivariate data. The selection of which derivative to use can make a substantial difference to the performance of functional regression. Despite the observation that derivatives can be important, there has been relatively little formal attention to this problem.

Within a linear model, derivatives of functional covariates cover non-nested function spaces. However, we have shown that a simple integration by parts analysis allows us to embed models based on two different derivatives within a common space by adding a finite number of point impacts. This allows us to construct finite-dimensional contrasts to assess the fit of each derivative which can be tested using standard procedures. In contrast to linear models, more general functional regression models cannot be as readily embedded in a common space. However, we have suggested adapting the $C$ test of Davidson and MacKinnon (1981) with subsample splitting to distinguish between two models that have already been fit.

While we have shown that tests based on contrasts perform well in simulation, there is clear scope for further development. An important component of our tests is a correction for the bias, since $\beta(t)$ or $\beta^{(1)}(t)$ may be easier to estimate, and this can affect the conclusions that we draw; while our noncentrality parameter appears to work well, a better theoretical grounding for it would give useful guidance. Similarly, the finite-dimensional representation of $X^{(k)}(t)$ can make the implicit function theorem inexact if its derivative is not in the span of this representation, yielding a further potential source of bias that should be better understood. In non-parametric models, we have proposed a generic framework, but this comes at the cost of sample splitting and will likely be inefficient for any given non-parametric model and poor Type-I error control. The use of noncentrality parameters here may help control the latter, but the detailed analysis will need to focus on the particular model at hand.

### References

- Aguilera, A., Escabias, M., Valderrama, M.: Discussion of different logistic models with functional data. Application to systemic lupus erythematosus. Comput. Statistics and Data Anal. 53(1), 151–163 (2008)
- Aneiros-Pérez, G., Vieu, P.: Semi-functional partial linear regression. Statistics & Prob. Lett. 76(11), 1102–1110 (2006)
- Aneiros-Pérez, G., Vieu, P.: Nonparametric time series prediction: A semi-functional partial linear modeling. J. Multivar. Anal. 99, 834–857 (2008)
- Asencio, M., Hooker, G., Gao, H.O.: Functional convolution models. Stat. Model. 14(4), 315–335 (2014)
- Dalzell, C.J., Ramsay, J.O.: Computing reproducing kernels with arbitrary boundary constraints. SIAM J. Sci. Comput. 14(3), 511–518 (1993)
- Davidson, R., MacKinnon, J.G.: Several tests for model specification in the presence of alternative hypotheses. Econometrica 49(3), 781–793 (1981)
- Febrero-Bande, M., González-Manteiga, W.: Regression additive models for functional data. TEST 22(2), 278–292 (2013)
- Febrero-Bande, M., Oviedo de la Fuente, M.: Statistical computing in functional data analysis: The R package fda.usc. J. Stat. Softw. 51(4), 1–28 (2012)
- Ferraty, F., Vieu, P.: The functional nonparametric model and applications to spectrometric data. Comput. Statist. 17(4), 545–564 (2002)
- Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis. Springer, New York (2006)
- Ferraty, F., Vieu, P.: Additive prediction and boosting for functional data. Computational Statistics & Data Analysis 53(4), 1400–1413 (2009)
- Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M.W., Swihart, B., Xiao, L., Crainiceanu, C. and Reiss, P.T.: refund: Regression with Functional Data. R package version 0.1-23. (2020) https://cran.r-project.org/web/packages/refund/index.html
Hall, P., Hooker, G.: Truncated linear models for functional data. J. Royal Statistical Soc.: Series B (Statistical Methodology) 78(3), 637–653 (2016)

Hastie, T., Mallows, C.: A statistical view of some chemometrics regression tools (discussion). Technometrics 35(2), 140–143 (1993)

Hyndman, R.J., Shang, H.L.: Forecasting functional time series (with discussions). J. Korean Statistical Soc. 38(3), 199–211 (2009)

Jarque, C.M.: Sample splitting and applied econometric modeling. J. Bus. & Econom. Statistics 5(2), 267–274 (1987)

McLean, M.W., Hooker, G., Staicu, A.M., Scheipl, F., Ruppert, D.: Functional generalized additive models. J. Comput. Graph. Stat. 23(1), 249–269 (2014)

Müller, H.-G., Stadmüller, U.: Generalized functional linear models. Ann. Stat. 33(2), 774–805 (2005)

Quintela-del-Río, A., Ferraty, F., Vieu, P.: Analysis of time of occurrence of earthquakes: A functional data approach. Math. Geosci. 43(6), 695–719 (2011)

Quintela-del-Río, A., Francisco-Fernández, M.: Nonparametric functional data estimation applied to ozone data: Prediction and extreme value analysis. Chemosphere 82(6), 800–808 (2011)

Ramsay, J., Hooker, G.: Dynamic Data Analysis: Modeling Data with Differential Equations. Springer, New York (2017)

Ramsay, J., Hooker, G., Graves, S.: Functional Data Analysis with R and MATLAB. Springer, Dordrecht (2009)

Ramsay, J. O., Graves, S. and Hooker, G.: fda: Functional Data Analysis. R package version 5.1.5.1. (2020) https://CRAN.R-project.org/package=fda

Ramsay, J., Silverman, B.: Functional Data Analysis, 2nd edn. Springer, New York (2005)

Savitzky, A., Golay, M.J.E.: Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36(8), 1627–1639 (1964)

Shang, H.L.: Visualizing rate of change: An application to age-specific fertility rates. J. R. Stat. Soc. Ser. A 182(1), 249–262 (2019)

Yao, F., Müller, H.-G.: Functional quadratic regression. Biometrika 97(1), 49–64 (2010)

Yao, F., Müller, H.-G., Wang, J.-L.: Functional data analysis for sparse longitudinal data. J. Am. Stat. Assoc. 100(470), 577–590 (2005)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.