Understanding regression shape changes through nonparametric testing*

Tatiana Komarova[†] Kamila Nowakowicz [‡]

December 11, 2024

Latest version

Abstract

We propose a procedure for testing whether a nonparametric regression mean satisfies a shape restriction that varies within the domain of the regressor. Notably, the change points of these shape restrictions are unknown and must be estimated. Our test statistic is based on the empirical process, drawing inspiration from Khmaladze (1982). This paper extends the nonparametric methodology of Komarova and Hidalgo (2023) by proposing a method to estimate the shape change points and consequently addressing the additional estimation errors introduced by that stage. We analyze strategies for managing these errors and adapting the testing approach accordingly. Our framework accommodates various common shapes, such as (inverse) U-shapes, S-shapes, and W-shapes. Furthermore, our method is applicable to partial linear models, thereby encompassing a broad spectrum of applications. We demonstrate the efficacy of our approach through application to several economic problems and data.

Keywords: U-shape, inverted U-shape, S-shape, B-splines, P-splines, Khmaladze transformation

JEL Classification:

*We are grateful to Javier Hidalgo for helpful discussions.
[†]University of Cambridge, Faculty of Economics, tk670@cam.ac.uk
[‡]London School of Economics, Department of Economics, k.nowakowicz@lse.ac.uk

1 Introduction

This paper suggests a nonparametric procedure for testing shape constraints of the regression mean when the shape changes within the domain of the regressor. We allow the shape to potentially change not just once but multiple times. In economics and other disciplines changing shape patterns are encountered frequently. There is sizable empirical literature analyzing or attempting to establish U-shaped or hump-shaped relations. S-shape relations are also frequently encountered in economics in the context of poverty traps or innovations. If U-shapes can be described as changing a shape pattern only once in the domain – from decreasing to increasing, the S-shapes in a certain interpretation may involve a more complicated characterization as there are not just three different monotonicity patterns (down, then up, then down again) but also a switch from convex to concave.

A common empirical practice in cases of U-shapes or hump-shapes has often relied on quadratics despite some recognizing a potential need for nonparametric approaches[1]. The use of quadratic specifications, however, remain a widespread practice. It is intuitive why – quadratics may be appealing to researchers due to the simplicity of interpretation and ability to model both convex and concave responses. It is evident, however, that just using quadratics is a restrictive way to model nonlinearities, as: (a) it imposes symmetry around the turning point; (b) it has to be concave (convex) everywhere; (c) if it is concave (convex), it first has to decrease (increase) and then increase (decrease). In reality, nonlinear relationships may be much more complicated than that. We refer the reader to Appendix A, where we present examples of nonlinear relationship which will be either completely missed or largely misrepresented by the use of quadratics. This will be further confirmed and illustrated in our applications. We believe that the main reason why applied researchers have relied extensively on quadratic specifications has been the absence of a unifying nonparametric methodology that can be used for estimation and testing of nonlinear shapes in a variety of models, including partially linear models. This is precisely where our paper makes its contribution.

Some important and welcome developments in the context of U-shaped or inverse U-shaped relations have been made in the work of Lind and Mehlum (2010), Simonsohn (2018), Kostyshak (2017) and Ganz (2024). These works are discussed in more detail below. Komarova and Hidalgo

1. E.g., nonparametric fits for some of their specification were explored in Aghion, Van Reenen, and Zingales (2013) through their Lowess smoother in Figure 1, or Ashraf and Galor (2013) nonparametric fit in Figure 3, or Aghion et al. (2005) spline fit in Figure II.

(2023) discuss U-shapes and S-shapes as some of the applications of their method but they rely on the shape changing point(s) to be known which may be unrealistic in practice. Even though this paper builds to a certain extent on the methodology in Komarova and Hidalgo (2023), it makes an important step forward in allowing the turning points to be unknown. This implies the needs to estimate them and incorporate the extra estimation steps into the testing methodology, which is a nutshell description of the contributions of our paper. Our extension of the method in Komarova and Hidalgo (2023) is far from trivial. Theoretically, we have to make sure that the turning points are estimated in a way that their interference with the statistical properties of the test statistics is limited in a sample and is also asymptotically negligible. Practically, allowing the turning points to be found adaptively from the data adds robustness to the hypothesis testing and credibility to the conclusions from the test. It is worth emphasizing that our testing method applies to a very general class of regression function changing shapes which goes beyond just U-shapes and S-shapes. Additionally, our paper allows we make it explicit how to apply the testing methodology to partially linear model where the shape constraints enter through the nonparametric part and the effect of other variables is allowed through a linear index.

The paper proceeds as follows. Section 2 gives a literature review. Section 3 discusses the basic setting with one change in shape and, thus, one turning point and gives a brief overview of our testing methodology. The case of just one change in the shape of the function will be our leading examples throughout sections 3-5. Section 5 describes our testing approach in detail. Section 6 discussed extension such as (a) allowing for many changes in the shape of the function and, thus, for multiple turning point (b) controlling for other regressors in a linear way. Section 7 presents Monte Carlo simulations. Section 8 contains applications. Section 10 concludes.

2 Literature review

Even though there is no general approach in the literature to estimate and test regression shape changes for various shapes, there is some literature that attempts to address some special cases of this, such as U-shapes and hump-shapes.

The existing literature on testing U-shape constraints is relatively small. Although the shape appears in many settings in economics and social sciences, researchers usually use tests based on quadratic specification. Lind and Mehlum (2010), Simonsohn (2018) and Kostyshak (2017) all give compelling arguments for why tests based on quadratic approximations are not appropriate when testing for a U-shaped or hump-shaped relationship.

Lind and Mehlum (2010) was the first to explicitly highlight the problems with using "U-shaped" and "quadratic" as synonyms. It proposes a joint inequality test on the signs of first derivatives estimated at two points in the support. It's a parametric test and it relies on knowing the true functional form. As pointed out by Simonsohn (2018), that test is only valid if the correct functional form is used and is likely to suffer from a high false-positive rate when the model is misspecified.

Simonsohn (2018) proposes a simple test based on estimating two regression lines: for low and high values. He does not assume any functional form but instead tests if the average slope on either side of a switch point is significant, and if the slopes have opposite signs. The switch point is estimated from the data and is chosen to maximise the power of the test (instead of getting the best fit for the data like in our paper) using what the paper calls a "Robin Hood" algorithm, for it takes away observations from the more powerful line and assigns them to the less powerful one. This test is simple to use but it does have some drawbacks: it does not distinguish between single and multiple changes in the sign of derivative (would classify a W-shape or an N-shape as a U-shape) and its asymptotic properties have not been analysed. In particular, the implication of estimating the switch point to maximize power rather than fit the data are not clear (potentially, this may result in high Type I error).

Kostyshak (2017) uses a non-parametric test, where the test statistic is the smallest bandwidth such that a local polynomial regression is quasi-convex (i.e. U-shaped or monotone), followed by a test for monotonicity. This specification allows for the switch point to be unknown and for the presence of covariates, just like in our model. The test statistic is consistent but further asymptotic theory of the test is not provided. The testing algorithm relies on bootstrap (our test has a nice asymptotic distribution, but to improve the finite sample performance of our test we also resort to bootstrap in this paper). Kostyshak (2017) applies his test to life satisfaction in age and finds that much of the U-shape can be explained by the increase in financial satisfaction typically occurring later in life. A very interesting aspect of that application is that this finding would be completely missed by quadratic specifications. This resonates with our applications too, where we show that a quadratic specification may completely miss a U-shaped or hump-shaped relation. It appears that the idea of the test in Kostyshak (2017) may be extended to other shapes and multiple switch points by relying on more general tests for the number of peaks and valleys in the regression function and its derivatives, but it would require running a series of tests instead of a single test, and would give a researcher less control over the exact choice of a shape than our method.

An approach in a recent work Ganz (2024) is also designed towards testing for U-shape/inverse U-shape relations. The regression function is modeled using linear (first-degree) splines or quadratic I-splines and a candidate switch point is taken as one of the knots (in our approach the switch point is adaptively found first and then the system of knots is driven by the estimated switch point). The idea of Ganz (2024) is to estimate three models – one model is very flexible (not enforcing any constraints), the second one estimates a monotonic relationship, and the third one permits one switch point in line with an (inverse) U-shape relationship. If the fit of the first model is close to that of the third and better than the second, we conclude the relation is (inverse) U-shaped, otherwise we reject the (inverse) U-shaped relation. While the procedure performs well in simulations, the formal statistical properties of this test have not yet been established. It seems that, to ensure a flexible choice of switch point, the number of knots must increase with the sample size, but it is not clear how this affects the asymptotic behaviour of the test.

For more complex changing shapes – those beyond (inverse) U-shapes – to the best of our knowledge, there are no existing statistical testing procedures that allow unknown switch points (Komarova and Hidalgo (2023) can be used when such points are known).

The theoretical and empirical literature has, of course, dealt with non-linear shapes. For examples of U-shaped relationships in economics and other disciplines see e.g.Weiman (1977), Goldin (1995), Calabrese and Baldwin (2001), Groes, Kircher, and Manovskii (2014), Sutton and Trefler (2016) (also see discussions in Lind and Mehlum (2010), Simonsohn (2018) and Kostyshak (2017)). Inverse U-shaped relationships include the case of the so-called single-peaked preferences, which is an important class of preferences in psychology and economics. In empirical research, U-shaped functions e.g. often appear in environmental economics, particularly in studies relating electricity consumption to temperature. Typically, the switch between heating and cooling is set around 18.3°C (65°F). Traditionally, this non-linear relationship between temperature and electricity consumption is modeled using heating degree-days (HDD) and cooling degree-days (CDD) in least squares regressions, as in Pardo, Meneu, and Valor (2002). More advanced techniques, like panel threshold regression (Bessec and Fouquau (2008)) or semiparametric spline models (Engle, Granger, Rice, and Weiss (1986)), have also been used. Another area where U-shaped relationships are found is in the study of happiness across the lifespan. Research by Blanchflower (2020) and others (e.g., Clark (2007)) shows that, after controlling for factors like gender, education, marital status, and employment, happiness follows a U-shape, with a minimum around age 48. This pattern has been observed in both developed and developing countries, and similar findings have been confirmed for apes (Weiss et al. (2012)). However, these studies rely on quadratic specifications in age to test the relationship even when graphical evidence (like in Blanchflower (2020))) a more consistent with an asymmetric U-shape relationship. More advanced techniques, like semiparametric splines (Wunder, Wiencierz, Schwarze, and Küchenhoff (2013)), show a U-shape below age 60 but a downward trend beyond that. In happiness research, identifying the turning point in age is key and, thus, techniques like our would be most suitable also for that reason.

Some literature in accounting documented S-shaped relationships – when e.g. stock price response to unexpected earnings is first convex and then becomes concave after a switch point (and is monotonic throughout the domain). For specific examples see Freeman and Tse (1992) or Das and Lev (1994), among others. S-shaped growth curves of the adopted population in a large society is a generally accepted empirical feature of innovation diffusion (see discussions in Utterback (1996), Rogers (2003)). Thus, testing for an S-shape in this case would allow one to conclude whether technology evolves as one would expect. Newell, Genschel, and Zhang (2014) uses S-shaped curves to model decays in the availability or usage of traditional media. We have not been able to find a formal statistical test in the literature for this type of shape.

An important part of our analysis is estimating switch points between different shape patters (this is formally defined later). There are a few papers using a kernel approximation to estimate a minimum (or maximum) of an unknown function, starting with Parzen (1962) which describes a procedure for finding a mode of a probability density function. Eddy (1980) improves his method to achieve better convergence rate, he shows that the mean squared error of the mode estimator can converge to zero at rate $N^{-1-\varepsilon}$ for any $\varepsilon > 0$. Muller (1989) describes a similar procedure for finding a peak of a regression function. We are not aware of similar procedures using splines.

There is also a large literature on identifying break points in regression functions, i.e. points at which the function is either discontinuous or has a discontinuity in one of its derivatives. For example Feder (1975) develops asymptotic theory for linear estimators of segmented regressions, where the parameters of interest are both the parameters in each segment and points at which the behaviour of the function changes. Estimates of these kinds of break points are typically faster than $N^{-1/2}$ (see e.g. Muller (1992)), making them very attractive, but in this paper we avoid making any assumptions about a level of discontinuity (if any) at the switch points so we don not rely on any result of that kind. Other important papers in this strand of literature on structural breaks include Delgado and Hidalgo (2000), which suggests estimators of location and size of structural breaks in a nonparametric regression model and is applicable in both cross sectional and time series models. Hidalgo, Lee, and Seo (2019) gives robust inference in threshold regression models when it is not known a priori whether at the threshold point the true specification has a kink or a jump and the threshold itself is unknown. In a related work, Hidalgo, Lee, Lee, and Seo (2023) proposes a continuity test for the threshold regression model based on the findings about a risk lower bound in estimating the threshold parameter without knowing whether the threshold regression model is continuous or not.

Our model also involves a semi-parametric specification combining *B-spline* approximation with linear components, which has been analysed in a number of papers, e.g. Speckman (1988), Heckman (1986). Rice (1986) analyses convergence rates for semiparametric model combining splines and linear terms, under particular assumptions on variables. The main difference between his approach and ours is that he uses an N-dimensional space of splines with all the observations of the regressor treated as knots, whereas we use a space of splines smaller than the number of observations and we can define the knots independently of the data. The basis splines in his case are orthogonal, simplifying derivations, but because of the more dense spline system his model is more prone to overfitting, which he avoids by adding a penalty term. In our case the number of splines grows slower than the number of observations, allowing us to achieve consistency without adding a smoothing penalty term. In his model he shows that the estimate of the linear component is biased, with rate depending on the size of the penalty, and that to decrease bias one needs to use lower penalty than optimal. Under our assumptions we can use the results from Newey (1997) which show that the parametric component achieves root N consistency and is asymptotically unbiased.

3 Setting and a brief outline of main ideas

Our leading case in Sections 3-5 can be described by the following setting:

$$y = m(x) + z'\gamma_0 + u, \tag{1}$$

$$E[u|x, z] = 0, \tag{2}$$

where $m \in C^1[\underline{x}, \overline{x}]$ where C^1 denotes the class of smooth functions. The function $m(\cdot)$ and parameter γ_0 is unknown.[2]

To characterize the property of $m(\cdot)$ as that of a *changing shape*, we start with an illustration of a function that changes shape *once*.

Let us, first, denote $m|_{[a,b]}$ as $m(\cdot)$ restricted to the interval $[a, b]$ and, second, suppose that for some $s_1^0 \in [\underline{x}, \overline{x}]$,

$$m|_{[\underline{x},s_1^0]} \in \mathcal{M}_1\left([\underline{x}, s_1^0]\right), \quad m|_{[s_1^0,\overline{x}]} \in \mathcal{M}_2\left([s_1^0, \overline{x}]\right), \tag{3}$$

where $\mathcal{M}_1$ and $\mathcal{M}_2$ are two classes of functions that describe functional properties that can be localized in the sense that

$$m|_{[a,b]} \in \mathcal{M}_j([a, b]) \quad \Rightarrow \quad m|_{[c,d]} \in \mathcal{M}_j([c, d]) \quad \forall [c, d] \subseteq [a, b], \quad j = 1, 2. \tag{4}$$

We also assume that

$$\mathcal{M}_1([a, b]) \cap \mathcal{M}_2([a, b]) = \emptyset \quad \forall [a, b]. \tag{5}$$

We interpret s_1^0 as the *turning* or *switch* point as at that point the regression function changes its pattern from class $\mathcal{M}_1$ to class $\mathcal{M}_2$. We are ultimately interested in a scenario where s_1^0 is not known and has to be estimated from the data.

Consider the following two examples.

Example 1 (U-shape, inverse U-shape, quasi-convexity, quasi-concavity)**.** *To the best of our knowledge, there is no general agreement in the literature on how to define U-shaped relationships mathematically. On of the most common definitions is that the function first decreases till some switch point and then increases. However, some authors would also incorporate convexity requirements into this property. To avoid any ambiguity, below we state explicitly what mean by U-shape. Our testing procedures can, of course, additional convexity requirements.*

A (strict) U-shaped function $m(\cdot)$ is first (strictly) decreasing on some interval $[\underline{x}, s_1^0]$ and then

2. This setting can be easily extended to allow for $\psi(z, \gamma_0)$ instead of $z'\gamma_0$ for a known nonlinear function $\psi(z, \cdot)$ and unknown γ_0.

on $[s_1^0, \overline{x}]$ *it is (strictly) increasing. Clearly then*

$$\mathcal{M}_1([a,b]) = \left\{ m|_{[a,b]} : m'(x) < 0 \text{ a.e. on } [a,b] \right\},$$
$$\mathcal{M}_2([a,b]) = \left\{ m|_{[a,b]} : m'(x) > 0 \text{ a.e. on } [a,b] \right\}.$$

It is easy to see that $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *satisfy conditions (4) and (5) above. In the case of an inverse U-shape (also often called hump-shape), the roles of* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *are reversed.*

A non-strict version of U-shape may involve intervals of constancy and can be formulated as non-strict inequalities on the signs of the derivatives.

Related to U-shape is the class of quasi-convex functions which is defined as

$$\left\{ m(\cdot) : \ \forall x_1, x_2 \in [a,b] \ \ \forall \lambda \in [0,1] \ \ m(\lambda x_1 + (1-\lambda) x_2) \leq \max \left\{ m(x_1), m(x_2) \right\} \right\}.$$

Function m *is quasi-concave if and only if* $-m$ *is quasi-convex. A smooth function is quasi-convex (-concave) if and only if it first decreases (increases) up to some point and then increases (decreases) incorporating a special case of monotonicity when a switch point is located at one of the boundary points of the interval. For quasi-convex (-concave) functions this switch point may not be known a priori, and thus, it would have to be estimated. This description can be changed to a strict version.*

When considering a U-shape property a researcher may want to make further restriction on the function being convex. It is easy to do by adding an inequality $m''(x) > 0$ *to the definition of classes* $\mathcal{M}_1$ *and* $\mathcal{M}_2$.

Example 2 (S-shape)**.** *There is no generally agreed on definition of S-shape. E.g. one interpretation defines a (strict) S-shaped as* $m(\cdot)$ *which is first (strictly) convex and increasing on some interval* $[\underline{x}, s_1^0]$ *and then on* $[s_1^0, \overline{x}]$ *it is (strictly) concave and increasing. In our setting this means*

$$\mathcal{M}_1([a,b]) = \left\{ m|_{[a,b]} : m''(x) > 0 \text{ and } m'(x) > 0 \text{ a.e. on } [a,b] \right\},$$
$$\mathcal{M}_2([a,b]) = \left\{ m|_{[a,b]} : m''(x) < 0 \text{ and } m'(x) > 0 \text{ a.e. on } [a,b] \right\},$$

if $m(\cdot)$ *is twice differentiable (if not, convexity and concavity can be formulated without involving the derivatives). It is easy to see that* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *satisfy conditions (4) and (5) above. This interpretation of S-shape is close to prospect theory in behavioral economics (see Kahneman and*

Tversky (1979).)

Other fields may understand S-shape differently. E.g., another way to interpret it would be as the regression function first strictly decreasing then strictly increasing and then strictly decreasing again. This interpretation would require two interiors switch points $s_1^0 < s_2^0$ and three classes $\mathcal{M}_1([\underline{x}, s_1^0])$, $\mathcal{M}_2([s_1^0, s_2^0])$ and $\mathcal{M}_3([s_2^0, \overline{x}])$ with

$$\mathcal{M}_1([a,b]) = \mathcal{M}_3([a,b]) = \left\{ m|_{[a,b]} : m'(x) < 0 \text{ a.e. on } [a,b] \right\},$$
$$\mathcal{M}_2([a,b]) = \left\{ m|_{[a,b]} : m'(x) > 0 \text{ a.e. on } [a,b] \right\}.$$

More generally, we have an *ordered* sequence of interior switch points $s_1^0, s_2^0 \ldots, s_J^0$ such as

$$s_0^0 \equiv \underline{x} < s_1^0 < s_2^0 \ldots < s_J^0 < \overline{x} \equiv s_{J+1}^0$$

(where the support boundaries are denoted as s_0^0 and s_{J+1}^0 for notational convenience) and a sequence of properties $\mathcal{M}_j$, $j = 1, \ldots, J+1$, such that

$$m|_{[s_j^0, s_{j+1}^0]} \in \mathcal{M}_{j+1}\left([s_j^0, s_{j+1}^0]\right), j = 0, \ldots, J, \tag{6}$$

It is important that the ordering of $\mathcal{M}_j$, $j = 1, \ldots, J$ is predetermined – that is, we know the order in which the properties of the regression function change.

Condition C1. *(a) Classes $\mathcal{M}_j$, $j = 1, \ldots, J+1$, describe functional properties that can be localized in the sense that*

$$m|_{[a,b]} \in \mathcal{M}_j([a,b]) \quad \Rightarrow \quad m|_{[c,d]} \in \mathcal{M}_j([c,d]) \quad \forall [c,d] \subseteq [a,b], \quad j = 1, 2. \tag{7}$$

(b) We also assume that

$$\mathcal{M}_j([a,b]) \cap \mathcal{M}_{j+1}([a,b]) = \emptyset \quad \forall [a,b], \quad j = 1, \ldots, J. \tag{8}$$

Part (a) of Condition C1 refines the notion of what it means for a class to capture *shape* – this is a property that extends to subintervals. Part (b) gives a general condition for a *change in shape* that is formulated for any two consecutive classes.

We can establish the identification of s_j^0, $j = 1, \ldots, J$. Henceforth, $s_1 < s_2 < \ldots < s_J$ will denote a generic ordered sequence of switch points located in the interior of $[\underline{x}, \overline{x}]$.

Proposition 1 (Identification)**.** *In the model (6) with a given ordering $s_1^0 < s_2^0 < \ldots < s_J^0$ of switch points, the switch points s_j^0, $j = 1, \ldots, J$, are identified under Condition C1.*

Below is an example of a situation with multiple switch points.

Example 3 (two local regression peaks)**.** *Consider the case when the smooth regression function has two local regression peaks. Then, in addition to estimating the two locations of local regression peaks we have to estimate another point between them where the regression function has a local minimum and turns from the decreasing pattern to the increasing one.*

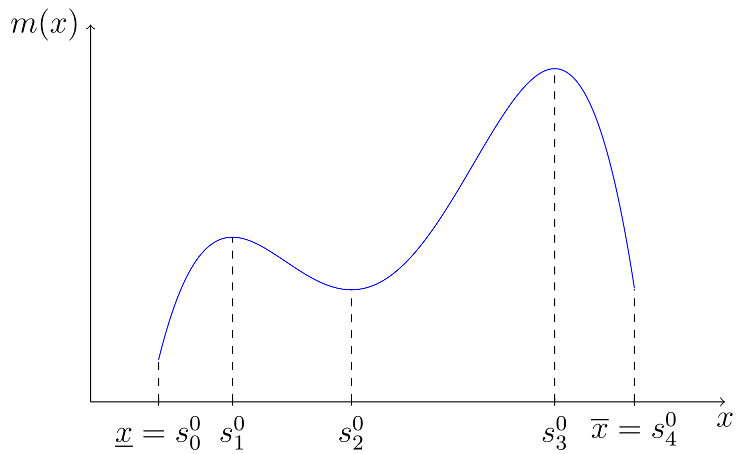


Figure 1: Two local regression peaks

Formally we have three interior switch points s_1^0, s_2^0, s_3^0 such that $s_0^0 \equiv \underline{x} < s_1^0 < s_2^0 < s_3^0 < \overline{x} \equiv s_4^0$, with the corresponding sets

$$\mathcal{M}_1([a,b]) = \mathcal{M}_3([a,b]) := \left\{ m|_{[a,b]} : m'(x) > 0 \text{ a.e. on } [a,b] \right\},$$
$$\mathcal{M}_2([a,b]) = \mathcal{M}_4([a,b]) := \left\{ m|_{[a,b]} : m'(x) < 0 \text{ a.e. on } [a,b] \right\}.$$

Points s_1^0 and s_3^0 are locations of the two local regression peaks whereas s_2^0 describes the location of the inevitable local minimum between s_1^0 and s_3^0.

It is easy to see that $\mathcal{M}_j$, $j = 1, \ldots, 4$, satisfy conditions (7) and (8).

Let $\mathcal{M}_0$ denote the class of all smooth regression functions m that satisfy (6):

$$\mathcal{M}_0 = \left\{ m \; : \; m \text{ satisfies (6) for some } s_1^0 < s_2^0 < \ldots < s_J^0 \right\}.$$

Our null hypothesis is

$$H_0 : m \in \mathcal{M}_0 \quad vs. \quad H_1 : m \notin \mathcal{M}_0 \tag{9}$$

(with the smoothness of functions in $\mathcal{M}_0$ being the maintained hypothesis).

The first step of our testing procedure will be to estimate m by a smooth join of *B-splines* of degree q_j defined on each estimated shape interval $[\hat{s}_j, \hat{s}_{j+1}]$, with $\hat{s}_0 = \underline{x}$, $\hat{s}_{J+1} = \overline{x}$. Suppose that the *B-spline* on $[\hat{s}_j, \hat{s}_{j+1}]$ is build on L_{j+1} *base B-spline*s (further details are in the next section). Our estimation will guarantee that as the sample size increases and all L_{j+1}, $j = 0, \ldots, J$, increase with it our estimator will be a consistent estimator of m under H_0.

Before advancing to the detailed technical description of that step, as well as the subsequent steps in testing, let us indicate what will differentiate our method from some other methods available in the literature.

From a big picture perspective, our methodology, just as in Komarova and Hidalgo (2023), is related to methods used in goodness of fit tests. Following Stute (1997a) or Andrews (1997) and Komarova and Hidalgo (2023), we base the testing procedure on functionals of the partial sums empirical process

$$\mathcal{K}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \widehat{u}_i \mathbb{1}(x_i < x), \qquad x \in [\underline{x}, \overline{x}] \tag{10}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Here

$$\widehat{u}_i = y_i - \widehat{m}_{\mathcal{B}}(x_i; \hat{s}) - z_i' \widehat{\gamma}, \quad i = 1, \ldots, n,$$

are the residuals obtained after m has been estimated by the nonparametric estimator $\widehat{m}_{\mathcal{B}}(x_i; \hat{s})$ by means of *B-splines* briefly described above and γ_0 has been estimated by $\widehat{\gamma}$ found simultaneously with $\widehat{m}_{\mathcal{B}}(x_i; \hat{s})$, see Section 4 for more detail (in a nutshell, $m_{\mathcal{B}}(x; \hat{s}) + z'\widehat{\gamma}$ denotes the best approximation of $m(x) + z'\gamma_0$ using the sum of the join of *B-splines* based on estimated switch points $\hat{s}$ for m and an additive separable linear function in z.) .

Unfortunately, after normalization, the limit covariance structure of $\mathcal{K}_n(x)$ depends on $\mathcal{M}_0$, making inferences based on $\mathcal{K}_n(x)$ very difficult to perform, if at all possible. For the simplicity of an illustration, consider the case of having no z_i on the right-hand side. Then

$$\mathcal{K}_n(x) = \frac{1}{n}\sum_{i=1}^{n} u_i \mathbb{1}(x_i < x) + \frac{1}{n}\sum_{i=1}^{n} \left(m_{\mathcal{B}}(x_i;\hat{s}) - \widehat{m}_{\mathcal{B}}(x_i;\hat{s})\right) \mathbb{1}(x_i < x) + \frac{1}{n}\sum_{i=1}^{n} \left(m(x_i) - m_{\mathcal{B}}(x_i;\hat{s})\right) \mathbb{1}(x_i < x). \quad (11)$$

In this decomposition the first term can be shown to be $\sqrt{n}$-convergent in distribution to the standard Brownian motion. The second term is also $O_p\left(\frac{1}{\sqrt{n}}\right)$, which means that the asymptotic distribution of $\mathcal{K}_n(x)$ might not be Gaussian and is difficult to characterize, making inferences very cumbersome. The third term in its turn can be represented as

$$\frac{1}{n}\sum_{i=1}^{n} \left(m(x_i) - m_{\mathcal{B}}(x_i;\hat{s})\right) \mathbb{1}(x_i < x) = \frac{1}{n}\sum_{i=1}^{n} \left(m(x_i) - m_{\mathcal{B}}(x_i;s^0)\right) \mathbb{1}(x_i < x) + \frac{1}{n}\sum_{i=1}^{n} \left(m_{\mathcal{B}}(x_i;s^0) - m_{\mathcal{B}}(x_i;\hat{s})\right) \mathbb{1}(x_i < x), \quad (12)$$

where the asymptotic behavior of first sub-term can be made asymptotically negligible with the choice of rates of L_j, $j = 1,\dots,J$, relative to n and the asymptotic behavior of the second sub-term depends on the rate of convergence of $\hat{s}$ to s^0.

In contrast to our method, the approach in Komarova and Hidalgo (2023) would assume the turning points in s^0 to be known (effectively making $\hat{s} = s^0$ in the decomposition above) and, thus, the right-hand side of (12) would only have the first sub-term significantly simplifying the ability to control the asymptotic behavior of that whole term. Our setting is more realistic as the turning points s^0 are taken to be *unknown*. This is a fundamental difference between this paper and Komarova and Hidalgo (2023) which results in very non-trivial theoretical and empirical challenges.

With estimates $\hat{s}$ and $m_{\mathcal{B}}(x_i;\hat{s})$ in hand we apply the transformation of $\mathcal{K}_n(x)$ analogous to the one used in Komarova and Hidalgo (2023) and based on ideas of Khmaladze (1982) as well as related to the *CUSUM* of recursive residuals proposed by Brown, Durbin, and Evans (1975). This leads to the asymptotic behavior of the transformation to be $\sqrt{n}$-convergent to a standard Brownian motion. Then testing is implemented using standard functionals such as Kolmogorov-Smirnov, Cramér -von-Mises or Anderson-Darling. In the next section we give the details of the estimation and testing procedure.

4 Modified null hypothesis and estimation methodology

We start with the discussion of estimating m under the null in line with our outline in Section 3.

For a given collection of switch points in the vector s, we can consider individual intervals $[s_{j-1}, s_j]$. On each of these intervals we consider a *B-spline* of degree q_j with knots that split $[s_{j-1}, s_j]$ into L'_j equally spaced intervals:[3]

$$m_{\mathcal{B};j}(x; s) \equiv \sum_{\ell=1}^{L_j} \beta_{\ell,j} p_{\ell,L_j,[s_{j-1},s_j],q_j}(x), \quad \text{where } L_j = L'_j + q_j, \tag{13}$$

and $\left\{p_{\ell,L_j,[s_{j-1},s_j],q_j}(\cdot)\right\}_{\ell=1}^{L_j}$ is the collection of the base *B-splines* base for the chosen system of knots and the chosen degree q_j (will be described shortly).

Then we can define

$$m_{\mathcal{B}}(x; s) = \sum_{j=1}^{J} m_{\mathcal{B};j}(x; s) \cdot \mathbb{1}[s_{j-1}, s_j) + m_{\mathcal{B};J+1}(x) \cdot \mathbb{1}[s_J, s_{J+1}], \quad x \in [\underline{x}, \overline{x}]. \tag{14}$$

Now we want to delve in more detail in the properties of *B-splines* in (13). These *B-splines* are constructed from polynomial pieces joined at some specific points called knots. In (14) we use *B-splines* whose domain and the system of knots differ on different sides of switch points. Generally, let q be the degree of a spline, L' be the number of subintervals of $[\underline{s}, \overline{s}]$ on which we define the spline (i.e. the number of polynomial pieces), then $L = L' + q$ is the number of *B-splines* in the basis.

We define the system of knots which split $[\underline{s}, \overline{s}]$ into L' equally spaced intervals. When defining *B-spline* of degree q we repeat the knots at the end points of the domain $q+1$ times. To be precise, we let

$$t = (t_\ell)_{\ell=1}^{L+2q+1} = \left(\underbrace{\underline{s}, \ldots, \underline{s}}_{q+1 \text{ times}}, \underline{s} + \frac{\overline{s} - \underline{s}}{L'}, \underline{s} + 2\frac{\overline{s} - \underline{s}}{L'}, \ldots, \underbrace{\overline{s}, \ldots, \overline{s}}_{q+1 \text{ times}} \right).$$

be the knot sequence. Then the ℓth *B-spline* of degree q defined on the knots t is a function of

3. The condition that these intervals are equally spaced is not important and is only imposed for the simplicity of the exposition. We only need that the system of knots has to become increasingly dense in $[s_{j-1}, s_j]$.

x we denote by $p_{\ell,L,[\underline{s},\overline{s}],q}(x)$. *B-splines* are defined recursively (see De Boor 1978) as follows:

$$p_{\ell-q,L-q,[\underline{s},\overline{s}],0}(x) = \mathbb{1}\left(x \in [t_\ell, t_{\ell+1})\right) = \begin{cases} 1 & \text{if } t_\ell \leq x < t_{\ell+1} \\ 0 & \text{otherwise} \end{cases}$$

and for $0 < k \leq q-1$:

$$p_{\ell,L-k,[\underline{s},\overline{s}],q-k} = \frac{x - t_\ell}{t_{\ell+q} - t_\ell} p_{\ell-1,L-k-1,[\underline{s},\overline{s}],q-k-1}(x) + \frac{t_{\ell+q+1} - x}{t_{\ell+q+1} - t_{\ell+1}} p_{\ell,L-k-1,[\underline{s},\overline{s}],q-k-1}(x).$$

By convention, anything divided by zero is zero.

An example of the steps in the construction of base *B-splines* for $q = 3$, $L = 8$, $[\underline{s}, \overline{s}] = [0, 1]$ is given in Figure 2.

Below is the list of some properties of *base B-splines*.

- $p_{\ell,L,[\underline{s},\overline{s}],q}(x)$ is non-negative and is positive over a domain spanned by $q+2$ adjacent knots, and is zero everywhere else;
- between each pair of consecutive knots $p_{\ell,L,[\underline{s},\overline{s}],q}(x)$ is a polynomial of degree q;
- at a knot which is repeated m times $p_{\ell,L,[\underline{s},\overline{s}],q}(x)$ has $q - m$ continuous derivatives;
- at any given x, at most $q + 1$ *B-splines* are non-zero;
- at any given x, the values of all *B-splines* sum to 1: $\forall x \in [\underline{s}, \overline{s}] \quad \sum_{\ell=1}^{L} p_{\ell,L,[\underline{s},\overline{s}],q}(x) = 1$.

The derivative of a *B-spline* is composed of polynomial sections of degree $q - 1$ defined over the same set of knots (with boundary knots having one less multiplicity), and is itself a *B-spline* of degree one lower. In particular, one can show, e.g. by induction (see e.g. De Boor (1978) or Procházková (2005)), that for a base *B-spline*,

$$\frac{\partial p_{\ell,L,[s_{j-1},s_j],q}}{\partial x} = \frac{q}{t_{\ell+q} - t_\ell} p_{\ell-1,L-1,[s_{j-1},s_j],q-1}(x) - \frac{q}{t_{\ell+q+1} - t_{\ell+1}} p_{\ell,L-1,[s_{j-1},s_j],q-1}(x), \tag{15}$$

which means that the derivative of the spline $m_{\mathcal{B};j}(x; s) \equiv \sum_{\ell=1}^{L} \beta_\ell p_{\ell,L,[s_{j-1},s_j],q}(x)$ is

$$\frac{\partial m_{\mathcal{B};j}(x; s)}{\partial x} = q \sum_{\ell=2}^{L} \frac{\Delta \beta_\ell}{t_{\ell+q} - t_\ell} p_{\ell-1,L-1,[s_{j-1},s_j],q-1}(x). \tag{16}$$

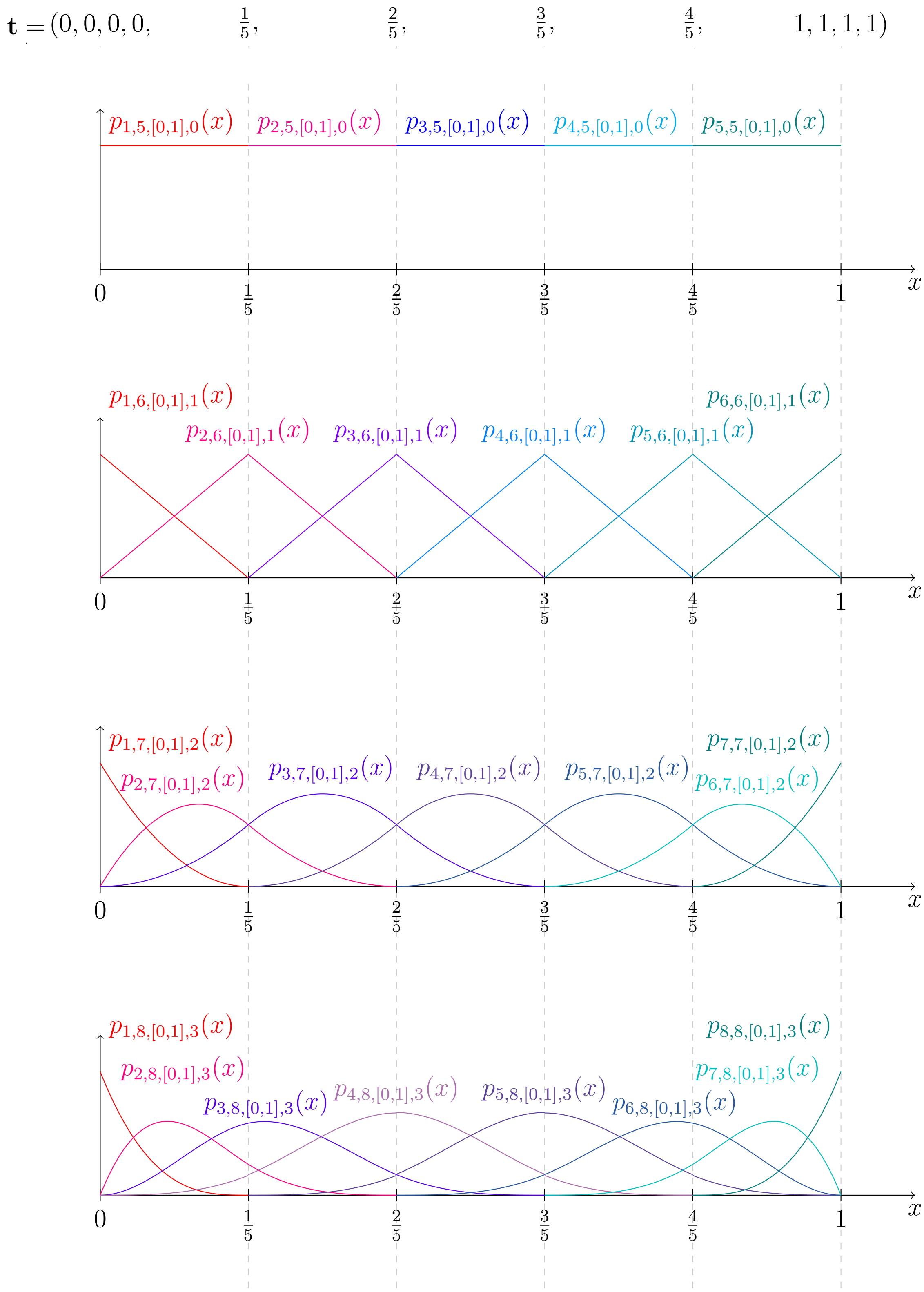


FIGURE 2: An example of *base B-spline* functions construction for $q = 3$.

Note that in the final expression the knots t are still based on the original q, not $q-1$.

Approximation and estimation of the regression mean $m(\cdot)$ by *B-splines* are appealing due to a convenient way to capture shape properties of interest, particularly those based on the derivatives of the regression function (such as U-shape, S-shape, etc.). In other words, the use of *B-splines* helps us to write the class $\mathcal{M}_0$ in (9) in terms of restrictions on the coefficients of the *base B-splines* in an approximation to $m(\cdot)$ (this requirement captured formally in Condition C2 below). With $L_j \to \infty$ as $n \to \infty$, $j = 1, \ldots, J+1$, the number of coefficients of the *B-splines* and the number of constraints will increase to infinity.

It is well understood that the choice of the number of knots determines the trade-off between overfitting and underfitting when there are respectively too many or too few knots. The main difference between *B-splines* and *P-splines* is that the latter tend to employ a large number of knots but to avoid oversmoothing they incorporate a penalty function based on the τ-th difference $\triangle^\tau \beta_\ell$, where $\triangle \beta_\ell = \beta_\ell - \beta_{\ell-1}$, with $\tau = 2$ being the most common choice. It is worth mentioning that other sieve estimators might be used, see the survey in Chen (2007), but we found *B-splines* particularly useful for our purposes.

Since out ultimate goal is to develop a nonparametric statistical test for (9) using the consistent estimators $\widehat{s}_1, \widehat{s}_2 \ldots, \widehat{s}_J$, we want to be sure that functional properties in each class $\mathcal{M}_j$, $j = 1, \ldots, J+1$, can be captured by the properties of coefficients of *B-splines* approximating m on the respective interval $[s_{j-1}, s_j]$ in the partition of $[\underline{x}, \overline{x}]$, and that this representation by the properties of coefficients of approximating *B-splines* becomes both necessary and sufficient as the number of knots on $[s_{j-1}, s_j]$ goes to infinity.

Formally, this is stated in Condition C2 below. Before we formally introduce this condition, let us introduce some helpful notations. Let $\mathcal{B}_j(q_j, L_j)$ denote the set of all *B-splines* of degree q_j with knots that split $[s_{j-1}, s_j]$ into L_j' equally spaced intervals[4]. A generic element in this set is written as a linear combination in (13). Thus, any element in $\mathcal{B}_j(q_j, L_j)$ can be fully characterised by the vector $\beta_{all,j} \equiv (\beta_{1,j}, \ldots, \beta_{L_j,j})' \in \mathbb{R}^{L_j}$ and constraints on this vector can be mapped into constraints on the *B-spline*. We consider each vector $\beta_{all,j} \in \mathbb{R}^{L_j}$ to be embedded into the long vector $\beta_{all} = (\beta_{all,1}', \ldots, \beta_{all,J+1}')' \in \mathbb{R}^{\sum_{j=1}^{J+1} L_j}$.

Let $T_{\{(q_j, L_j)\}_{j=1}^{J+1}, s} \subset \mathbb{R}^{\sum_{j=1}^{J+1} L_j}$ denote a set that describes constraints on the vector of coefficients

4. The condition that these intervals are equally spaced is not important and is only imposed for the simplicity of the exposition. We only need that the system of knots has to become increasingly dense in $[s_{j-1}, s_j]$.

β_{all} for a given vector s pf ordered switch points. We can subsequently define

$$\mathcal{M}_{\{(q_j,L_j)\}_{j=1}^{J+1},s} = \left\{ m_{\mathcal{B}}(x;s) \text{ in the form of (14) } \mid \beta_{all} \in T_{\{(q_j,L_j)\}_{j=1}^{J+1},s} \right\}.$$

$\mathcal{M}_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ is, thus, a collection of functions that are joins of *B-splines* defined individually on the intervals $[s_{j-1}, s_j]$. $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ can contain restrictions that will guarantee that the whole $m_{\mathcal{B}}(\cdot;s)$ defined in such a piece-wise way is continuous, or, additionally, smooth or, more generally, r-th continuously differentiable (the choice of r would depend on the degrees q_j, $j = 1, \ldots, J+1$ of the *B-splines*). E.g., the continuity of the whole piece-wise approximation is ensured by the constraints

$$\beta_{L_j;j} = \beta_{1;j+1}, \quad j = 1, \ldots, J. \tag{17}$$

In order to guarantee the smoothness of the approximation $m_{\mathcal{B}}(\cdot;s)$, in addition to (17) we have to impose that[5]

$$\frac{q_j L_j' \left(\beta_{L_j;j} - \beta_{L_j-1;j}\right)}{s_j - s_{j-1}} = \frac{q_{j+1} L_{j+1}' \left(\beta_{2;j+1} - \beta_{1;j+1}\right)}{s_{j+1} - s_j}, \quad j = 1, \ldots, J, \tag{18}$$

which simplifies to

$$\beta_{L_j;j} - \beta_{L_j-1;j} = \beta_{2;j+1} - \beta_{1;j+1} = 0, \quad j = 1, \ldots, J \tag{19}$$

in the case when the switch point is a local minimum or a local maximum. Further restrictions can be derived to enforce the continuity of the second derivative, etc. shall a researcher want to impose higher order restrictions.

In the regularity conditions in Section 5.1 we require the regression function to be smooth and its first derivative to be Holder-continuous, therefore it is natural to narrow down $\mathcal{M}_0$ to include only those regression functions that satisfy those regulation conditions. We will denote this class as $\mathcal{M}_0^*$.

Condition C2 below formalizes our idea of approximating the properties in $\mathcal{M}_0^*$ by constraints on coefficients in the approximation $m_{\mathcal{B}}(\cdot;s)$ in a necessary and sufficient fashion.

Condition C2. *For each s there is a set $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s} \subset \mathbb{R}^{\sum_{j=1}^{J+1} L_j}$ that satisfies the following properties:*

5. this is in case the interior knots are equidistant within each $[s_{j-1}, s_j]$.

(i) For a given s, $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ does not depend on data $\{x_i\}_{i\in\mathbb{Z}}$ and, thus, is non-stochastic;

(ii) For any s, the boundary of $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ consists of a finite number of smooth surfaces;

(iii) Let $\mathcal{M}_{T_{\{(q_j,L_j)\}_{j=1}^{J+1}}}$ denote the union of $\mathcal{M}_{T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}}$ over all possible s and let $\mathcal{H}$ be the Hausdorff distance in the supremum norm in the space of continuous functions. Then

$$\mathcal{H}\left(\mathcal{M}_0^*, \mathcal{M}_{T_{\{(q_j,L_j)\}_{j=1}^{J+1}}}\right) = O\left(\frac{1}{\left(\min_{j=1,\ldots,J+1} L_j\right)^r}\right) \tag{20}$$

for $r > 2$ and $\min_{j=1,\ldots,J+1} L_j \to \infty$.

(iv) Let $\mathcal{M}_{0,s^0}^*$ denote the set of functions in class $\mathcal{M}_0^*$ with switch points s^0. For any pair (s^0, s)

$$\mathcal{H}\left(\mathcal{M}_{0,s^0}^*, \mathcal{M}_{T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}}\right) = \Omega\left(\|s^0 - s\|_\infty\right). \tag{21}$$

Condition C2(i) ensures that the constraints in the estimation can be constructed in a generic fashion and we can talk about a deterministic approximation of class $\mathcal{M}_0^*$ by $\mathcal{M}_{T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}}$. Condition C2$(ii)$ guarantees that for a given s the implementation of conditions $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ comes down to enforcing a finite number of constraints on coefficients in β_{all}. In practice, definitions of $T_{\{(q_j,L_j)\}_{j=1}^{J+1}}$ will often be sufficient to guarantee functional properties of $\mathcal{M}_0^*$. Condition C2(iii) ensures that these conditions become asymptotically necessary. Condition C2(iv) puts a lower bound on the quality of fit when we use a set of constraints which misspecify the switch point: we need the loss in fit to be large enough to allow us to estimate $\hat{s}$. Combined Condition C2(iii) and Condition C2(iv) ensure that, as long as $s \to s^0$, the constraints in $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ capture constraints in $\mathcal{M}_{0,s^0}^*$ in a necessary and sufficient way as the number of knots grows to infinity, and if the convergence of s to s^0 is sufficiently fast, the approximation rate in the constrained approximation with the enforced $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ is the same as the rate in the unconstrained *B-spline* approximation. We can interpret r as the number of continuous derivatives elements of $\mathcal{M}_0^*$.

Given Condition C2, our idea is to test the null hypothesis

$$H_0^B: \quad \beta_{all} \in T_{\{(q_j,L_j)\}_{j=1}^{J+1},s} \text{ for some } s \quad vs. \quad H_1^B: \quad \text{(negation of null)} \tag{22}$$

formulated in terms of the approximation for m.

Let us now illustrate Condition C2 and the approximation for the U-shape in Example 1.

Example 1 (continued). *In the case of the U-shape property the approximation consists of two B-spline joined at* s^0*:*

$$m_{\mathcal{B}}(x;s^0) = \underbrace{\sum_{\ell_1=1}^{L} \beta_{\ell_1;1} p_{\ell_1,L,[\underline{x},s^0],q}(x)}_{m_{\mathcal{B};1}(x;s^0)} \cdot 1[\underline{x},s^0) + \underbrace{\sum_{\ell_2=1}^{L} \beta_{\ell_2;2} p_{\ell_2,L,[s^0,\overline{x}],q}(x)}_{m_{\mathcal{B};2}(x;s^0)} \cdot 1[s^0,\overline{x}],$$

where for simplicity we took $q_1 = q_2 = q$ *(same degree of B-splines on both sides of* s^0*) and* $L_1 = L_2 = L$ *(same number of knots on both sides of* s^0*). To capture monotonicity patterns and also smoothness at* s^0 *described by (17)-(19), we take*

$$T_{\{(q,L)\}} = \Big\{ (\beta_{all,1}, \beta_{all,2}) \mid \beta_{\ell_1;1} \geq \beta_{\ell_1+1;1},\ \ell_1 = 1,\ldots,L-1, \quad \beta_{\ell_2;2} \leq \beta_{\ell_2+1;2},\ \ell_2 = 1,\ldots,L-1,$$
$$\beta_{L;1} = \beta_{L-1;1} = \beta_{1;2} = \beta_{2;2}, \quad \beta_{L_j-2;j} = \beta_{3;j+1}. \Big\}$$

Inequalities $\beta_{\ell_1;1} \geq \beta_{\ell_1+1;1}$, $\ell_1 = 1,\ldots,L-1$, *capture the fact that the function is decreasing on* $[\underline{x}, s^0]$, *while* $\beta_{\ell_2;2} \leq \beta_{\ell_2+1;2}$, $\ell_2 = 1,\ldots,L-1$, *capture the fact that it is increasing on* $[\underline{x}, s^0]$. *Equality* $\beta_{L;1} = \beta_{L-1;1}$ *for the continuity of the approximation at* s^0, *and the equalities* $\Delta\beta_{L;1} = \Delta\beta_{2;2} = 0$ *for smoothness of the approximation at* s^0 *as well as for the minimum of the approximation at* s^0 *together give us* $\beta_{L;1} = \beta_{L-1;1} = \beta_{1;2} = \beta_{2;2}$.

Now let's us show that C2 (iii) holds. From the B-spline theory we know (e.g. from De Boor (1978)) that the approximation of three-times differentiable $m|_{[\underline{x},s^0]}$ *and* $m|_{[s^0,\overline{x}]}$ *by unconstrained B-splines on the respective intervals* $[\underline{x}, s^0]$ *and* $[s^0, \overline{x}]$ *can be attained at the rate* $O\left(\frac{1}{L^3}\right)$. *Let us denote such approximations as* $\tilde{m}_{\mathcal{B};1}(\cdot)$ *and* $\tilde{m}_{\mathcal{B};2}(\cdot)$, *respectively:*

$$\tilde{m}_{\mathcal{B};1}(\cdot;s^0) = \sum_{\ell_1=1}^{L} \tilde{\beta}_{\ell_1;1} p_{\ell_1,L,[\underline{x},s^0],q}(x), \qquad \tilde{m}_{\mathcal{B};2}(\cdot;s^0) = \sum_{\ell_2=1}^{L} \tilde{\beta}_{\ell_2;2} p_{\ell_2,L,[s^0,\overline{x}],q}(x).$$

Let us show that because of $m|_{[\underline{x},s^0]}$ *strictly decreasing we can without a loss of generality take* $\tilde{\beta}_{\ell_1;1} \geq \tilde{\beta}_{\ell_1+1;1}$ *for all* $\ell_1 = 1,\ldots,L-1$, *in* $\tilde{m}_{\mathcal{B};1}(\cdot)$, *and analogously without a loss of generality take* $\tilde{\beta}_{\ell_2;2} \leq \tilde{\beta}_{\ell_2+1;2}$ *for all* $\ell_2 = 1,\ldots,L-1$, *in* $\tilde{m}_{\mathcal{B};2}(\cdot)$ *Indeed, from the approximation theory* *we know that*

$$\sup_{x\in[\underline{x},s^0]}\left|\sum_{\ell_1=1}^{L}\tilde{\beta}_{\ell_1;1}p'_{\ell_1,L,[\underline{x},s^0],q}(x)-m'|_{[\underline{x},s^0]}(x)\right|=O\left(\frac{1}{L^2}\right), \tag{23}$$

$$\sup_{x\in[s^0,\overline{x}]}\left|\sum_{\ell_2=1}^{L}\tilde{\beta}_{\ell_2;2}p'_{\ell_2,L,[s^0,\overline{x}],q}(x)-m'|_{[s^0,\overline{x}]}(x)\right|=O\left(\frac{1}{L^2}\right). \tag{24}$$

Using the formula for the derivative of B-spline, obtain

$$\sum_{\ell_j=1}^{L}\tilde{\beta}_{\ell_j;j}p'_{\ell_j,L,[s_{j-1},s_j],q}(x)=q\sum_{\ell_j=1}^{L-1}\frac{\triangle\tilde{\beta}_{\ell_j+1'j}}{t^{l_j+1+q;j}-t^{l_j+1;j}}p_{\ell_j+1,L,[s_{j-1},s_j],q-1}(x)\,,\quad j=1,2, \tag{25}$$

where $t^{l_j;j}$ denotes a knot on $[\underline{x},s^0]$ for $j=1$ and on $[s^0,\overline{x}]$ for $j=2$.

Taking into account (23)-(25), the fact that $\frac{\underline{K}_1}{L}t^{l_j+1+q;j}-t^{l_j+1;j}\leq\frac{\bar{K}_1}{L}$ for some constant $\underline{K}_1,\bar{K}_1>0$ as well as the facts that $m'|_{[\underline{x},s^0]}(x)\geq 0$ and $m'|_{[s^0,\overline{x}]}(x)\leq 0$ and

$$\sum_{\ell_j=1}^{L}p_{\ell_j,L,[s_{j-1},s_j],q}(x)=1\quad\textit{for all } x \textit{ in the respective interval,} \tag{26}$$

we conclude that

$$\triangle\tilde{\beta}_{\ell_1+1;1}\leq\frac{K_2}{L^3},\quad\triangle\tilde{\beta}_{\ell_2+1;2}\geq-\frac{K_2}{L^3},$$

for some constant $K_2>0$. Thus, to ensure that $\tilde{\beta}_{\ell_1+1;1}\leq 0$, $\ell_1=1,\ldots,L-1$, and $\tilde{\beta}_{\ell_2+1;2}\geq 0$, $\ell_2=1,\ldots,L-1$, which will guarantee the desired monotonicity patterns in the approximation, we have to change each coefficient $\tilde{\beta}_{\ell_1+1;1}$ by at most $\frac{K_2}{L^3}$. Because of the partitioning property (26), the B-splines with such potentially new coefficients that satisfy the desired inequalities will approximate functions $m|_{[\underline{x},s^0]}(\cdot)$ and $m|_{[s^0,\overline{x}]}(\cdot)$ at the same rate $O\left(\frac{1}{L^3}\right)$ as before.

Now let's show that imposing restrictions $\tilde{\beta}_{L-1;1}=\tilde{\beta}_{2;2}=\tilde{\beta}_{L;1}=\tilde{\beta}_{1;2}$, $\tilde{\beta}_{L_j-2;j}=\tilde{\beta}_{3;j+1}$ that ensure suitable smoothness of the approximation as well as the zero derivative at s^0, does not change the approximation rate.

Indeed, using the approximation properties of the B-splines as well as their derivatives, we have *the following sets of properties:*

$$\left|\sum_{\ell_j=1}^{L}\tilde{\beta}_{\ell_j;j}p_{\ell_j,L,[s_{j-1},s_j],q}(s^0)-m(s^0)\right|=O\left(\frac{1}{L^3}\right),\quad j=1,2,$$
$$\left|\sum_{\ell_j=1}^{L}\tilde{\beta}_{\ell_j;j}p'_{\ell_j,L,[s_{j-1},s_j],q}(s^0)\right|=O\left(\frac{1}{L^2}\right),\quad j=1,2,$$
$$\left|\sum_{\ell_j=1}^{L}\tilde{\beta}_{\ell_j,[s_{j-1},s_j],q}p''_{\ell_j,L;j}(s^0)-m''(s^0)\right|=O\left(\frac{1}{L}\right),\quad j=1,2,$$

where the second property also takes into account that $m'(s^0)=0$*.*

Note that

$$\sum_{\ell_1=1}^{L}\tilde{\beta}_{\ell_1;1}p_{\ell_1,L,[\underline{x},s^0],q}(s^0)=\tilde{\beta}_{L;1},\quad \sum_{\ell_2=1}^{L}\tilde{\beta}_{\ell_2;2}p_{\ell_2,L,[s^0,\overline{x}],q}(s^0)=\tilde{\beta}_{L;2},$$
$$\sum_{\ell_1=1}^{L}\tilde{\beta}_{\ell_1;1}p'_{\ell_1,L,[\underline{x},s^0],q}(s^0)=\frac{L\triangle\tilde{\beta}_{L;1}}{K_3},\quad \sum_{\ell_2=1}^{L}\tilde{\beta}_{\ell_2;2}p'_{\ell_2,L,[s^0,\overline{x}],q}(s^0)=\frac{L\triangle\tilde{\beta}_{2;2}}{K_4},$$
$$\sum_{\ell_1=1}^{L}\tilde{\beta}_{\ell_1;1}p''_{\ell_1,L,[\underline{x},s^0],q}(s^0)=\frac{L^2(2\triangle\tilde{\beta}_{L;1}-\triangle\tilde{\beta}_{L-1;1})}{K_5},\quad \sum_{\ell_2=1}^{L}\tilde{\beta}_{\ell_2;2}p''_{\ell_2,L,[s^0,\overline{x}],q}(s^0)=\frac{L^2(\triangle\tilde{\beta}_{3;2}-2\triangle\tilde{\beta}_{2;2})}{K_6},$$

for some constants $K_3>0$*,* $K_4>0$*,* $K_5>0$*,* $K_6>0$*.*

These imply that we may have to change the values of coefficients of $\tilde{\beta}_{\ell_1;1}$*,* $\ell_1=L-2,L-1,L$*, and* $\tilde{\beta}_{\ell_2;2}$*,* $\ell_2=1,2,3$*, by at most* $\frac{K_7}{L^3}$ *for some* $K_7>0$ *to ensure the desired equality constraints as well as to preserve the monotonicity patterns of the approximation. This means (taking into account Eq. (26) once again) that with coefficients possibly changed once again, the approximation rate of B-splines is still* $O\left(\frac{1}{L^3}\right)$*.* □

5 Testing methodology

5.1 Properties of the estimators

To formally prove that our testing procedure works, we need to establish the properties of our estimators. We start by listing regularity conditions.

Condition C3. *(i)* $\{(x_i, z_i', u_i)'\}_{i=1}^n$ *are i.i.d. random vectors. The support of* x *is normalized to* $[0,1]$ *and its density function* $f_X(x)$ *is bounded away from zero on the whole support.* $E(u_i|x_i, z_i) = 0$, $E(u_i^2|x_i, z_i) = \sigma_u^2 < \infty$, u_i *has finite 4th moments, there exists* $\nu > 0$ *such that* $E(|z_i|^{2+\nu}) < \infty$, *and* $E((z_i - E(z_i|x_i))(z_i - E(z_i|x_i))') \neq 0$.

(ii) $m(x)$ *is* $r \geq 3$ *times continuously differentiable.*

(iii) $\frac{\left(\min_{j=1,\ldots,J+1} L_j\right)^4}{n} \to 0$, $\frac{\left(\min_{j=1,\ldots,J+1} L_j\right)^{2r}}{n} \to \infty$ *as* $n \to \infty$.

Condition C3(i) ensures that no linear combination of z_i can be perfectly predicted by *B-splines* in x_i (we can think of it as no perfect multicollinearity condition: we cannot perfectly substitute between fitting $m_{\mathscr{B}}(x_i)$ and $\gamma' z_i$; adjusting γ cannot fully correct the overall fit if we chose an incorrect switch point). This assumption is needed for identification and root n consistency of the coefficients on z_i. One implication of this assumption is that z_i cannot include a constant. The homoskedasticity assumption could be weakened in a similar way as in Komarova and Hidalgo (2023). Condition C3(ii) on the smoothness of the estimated function determines the quality of *B-spline* approximation. Condition C3(iii) provides the rates at which the number of knots increases to infinity relative to n. This ensures that the bias term (due to the approximation using *B-splines*) is asymptotically negligible.

5.1.1 Consistency

We first show that under the null (9) the constrained estimator defined in (35)-(37) is consistent. To establish this, we consider the regression function $m(\cdot)$ to be a part of a certain compact set and we supplement (37) by additional constraints on coefficients $\beta_{\ell-j,j}$s (even though in practice such additional constraints most of the time will not be necessary). We rely on the consistency theorem in Newey and Powell (2003).

Since $m(\cdot)$ is smooth, it is bounded and has a finite Lipschitz constant. We take a very large pointwise bound $A_1 > 0$ and a very large Lipschitz constant A_2 on all the candidate regression functions under consideration (of course, these bounds should be large enough to be true for the underlying regression mean $E[y|x]$). In other words, we take the intersection

$$\Theta_0 = \mathcal{M}_0 \cap \left\{ m(\cdot) \,:\, \sup_{x \in [\underline{x}, \overline{x}]} |m(x)| \leq A_1, \, \sup_{[\underline{x}, \overline{x}]} |m'(x)| \leq A_2 \right\}. \tag{27}$$

Proposition 2. *Suppose that $m \in \Theta_0$ and $L_j \to \infty$, $j = 1, \ldots, J+1$, as $n \to \infty$. Then the estimator $\widehat{m}_{\mathcal{B}}(\cdot; \hat{s})$ obtained by solving (35)-(37) is consistent in the sense that*

$$\sup_{x \in [\underline{x}, \overline{x}]} |\widehat{m}_{\mathcal{B}}(x; \hat{s}) - m(x)| \xrightarrow{p} 0 \quad \textit{as } n \to \infty.$$

The consistency of $\widehat{m}_{\mathcal{B}}(\cdot; \hat{s})$ guarantees the consistency of the switch points, as established in the proposition below.

Corollary 1. *Under conditions of Proposition 2, the estimators $\hat{s}_j$ of switch points are consistent for s_j, $j = 1, \ldots, J$.*

We can also derive the rates at which the estiamtors converge to their limits:

Proposition 3. *Under conditions C1-C3:*

$$\hat{\beta} - \beta_0 = O_p\left(\sqrt{\frac{L}{n}}\right), \quad \hat{\gamma} - \gamma_0 = O_p\left(\sqrt{\frac{1}{n}}\right), \quad \hat{s} - s^0 = O_p\left(\frac{L}{\sqrt{n}}\right).$$

5.2 The test statistic

5.2.1 Testing procedure and the justification of the need for a transformation

We use a Lagrange multiplier type test[6]. From Assumption 2 we know that the true error terms are uncorrelated with any function of the regressors, i.e. $E(u_i f(x_i)) = 0$ for any function $f(\cdot)$. The idea of the test is to check if a similar property is satisfied by the regression residuals:

$$\hat{u}_i = y_i - \widehat{m}_{\mathcal{B}}(x_i; \hat{s}) - \hat{\gamma} z_i. \tag{28}$$

6. For more motivation behind this testing design see the discussion in Komarova and Hidalgo (2023).

A common choice of the function $f(\cdot)$ used in this type of tests, see e.g. Stute (1997b), is $f(x_i) = \mathbb{1}(x_i < x)$ for some $x \in [0, 1]$, which results in a test statistic of the form:

$$K(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x)\hat{u}_i. \tag{29}$$

Under the null hypothesis, $K(x)$ should be close to zero. However, finding the limiting distribution of this statistic turns out to be problematic. Consider the following expansion:

$$\begin{aligned} K(x) = &\underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x)u_i}_{T_0} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x)\left(m(x_i) - m_{\mathcal{B}}(x_i; \hat{s})\right)}_{T_1} \\ &+ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x)\left(m_{\mathcal{B}}(x_i; \hat{s}) - \widehat{m}_{\mathcal{B}}(x_i; \hat{s})\right)}_{T_2} + \underbrace{(\gamma - \hat{\gamma})' \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x)z_i}_{T_3}. \end{aligned}$$

It can be shown that

- $\sqrt{n}T_0 \xrightarrow{d} \sigma\mathcal{B}(F_X(x))$, where $\mathcal{B}(\cdot)$ denotes the standard Brownian motion and $F_X(x)$ is the cdf of x. This term has a well-defined limit which does not depend on the estimates. If this term dominated, we would be able to easily perform tests using standard critical values.

- $T_2 = O_p\left(\frac{1}{\sqrt{n}}\right)$ (follows from Lemma 5). This is the same rate of convergence as T_0, but unlike T_0 this term does not have a standard known distribution. Instead, the distribution depends on the estimated function in a non-trivial way. The presence of this term motivated the need for a transformation in Komarova and Hidalgo (2023).

- $T_1 = O_p(\frac{1}{\sqrt{n}})$ (follows from Lemma 6), which implies that this term is not negligible compared to T_0. This is a major difference between our case and that in Komarova and Hidalgo (2023), for whom the term of this form based on the known true s^0 was of a smaller order of magnitude than T_0. We need to modify the transformation to make sure it removes this term as well. An additional complication is that this term is non-linear in parameters, and the Khmaladze transformation relies on linear projections. Because of that, we do not remove this term entirely, but only up to a linear approximation. This is sufficient to ensure that the part which remains after the transformation is of a smaller order and does not affect the limiting distribution.

- $T_3 = O_p\left(\frac{1}{\sqrt{n}}\right)$ by the standard results on root n convergence of the linear part of a partially linear model, see e.g. Robinson (1988). Hence T_3 has the same convergence rate as the other terms, and just like T_2 it has a distribution which depends on the function we are estimating. We add another modification to the transformation from Komarova and Hidalgo (2023) to remove this term as well.

The last three terms are problematic because their asymptotic distributions depend on the estimated function. As a result, the limiting distribution of the test statistic is not standard. In order to achieve a limiting distribution which would allow us to perform testing using standard techniques, we would like to transform the test statistic in a way which removes the last three terms while leaving the asymptotic behavior of the first term unchanged. We describe a transformation which achieves this goal in the next section.

5.2.2 The Khmaladze's Transformation

The transformation which removes the problematic terms from $K(x)$ while keeping enough structure of the original statistic to allow for testing is a special case of a martingale transformation introduced by Khmaladze (1982). It can remove all terms linear in $\widetilde{\boldsymbol{P}}$, hence we define $\widetilde{\boldsymbol{P}}$ to include: *B-splines* basis functions (these are terms linear in βs, or in other words derivatives with respect to βs: $\frac{\partial \widehat{m}_{\mathscr{B}}(x_i;\hat{s})}{\partial \beta_l}$, these will remove T_2), zs (derivatives of the linear part with respect to γ_k, these will remove T_3) and linear approximation with respect to s: $\frac{\partial \widehat{m}_{\mathscr{B}}(x_i;\hat{s})}{\partial s}$ (this will remove the leading linear component in T_1). All of these are functions of the regressors (x, z), and we assume $E(u_i|x_i, z_i) = 0$, so the residual from regressing u_i on functions of z_i should be very close to u_i, hence the limiting behavior of the first term should be the same as without a transformation.

In Lemma 2 we show that:

$$\frac{\partial \widehat{m}_{\mathscr{B}}(x_i;\hat{s})}{\partial s_k} = \mathbb{1}\left[\hat{s}_{k-1}, \hat{s}_k\right) \frac{\hat{s}_{k-1} - x_i}{\hat{s}_k - \hat{s}_{k-1}} \frac{\partial \widehat{m}_{\mathscr{B}}(x_i;\hat{s})}{\partial x} + \mathbb{1}\left[\hat{s}_k, \hat{s}_{k+1}\right) \frac{x_i - \hat{s}_{k+1}}{\hat{s}_{k+1} - \hat{s}_k} \frac{\partial \widehat{m}_{\mathscr{B}}(x_i;\hat{s})}{\partial x}$$

where the derivative of B-spline is:

$$\frac{\partial \widehat{m}_{\mathscr{B}}(x_i;\hat{s})}{\partial x} = \sum_{\ell_j=1}^{L} \hat{\beta}_{\ell_j} p'_{\ell_j, L, [\hat{s}_{j-1}, \hat{s}_j], q}(x)$$

We let $\mathbf{P}_{L_j;j}(x)$ denote the L_j-dimensional vector of base *B-splines* on $[\hat{s}_{j-1}, \hat{s}_j]$, $j = 1, \ldots, J+1$, computed at x. The estimation uses the long system $\mathbf{P} \equiv (\mathbf{P}_{L_1;1}(x)', \ldots, \mathbf{P}_{L_{J+1};J+1}(x)')'$.

However, the constrained estimation under H_0^B results in some binding constraints. Once the binding constraints are enforced, we end up with a smaller system of relevant base *B-splines*[7]. We can refer to it as the system of "effective polynomials" and denote it as $\widetilde{\mathbf{P}}(x)$. Let

$$\widetilde{\boldsymbol{P}}_k \equiv \left(\widetilde{\mathbf{P}}(x_k)', z_k', \frac{\partial \widehat{m}_{\mathcal{B}}(x_k; \hat{s})}{\partial s}\right)'.$$

Note that the elements of $\widetilde{\boldsymbol{P}}$ are defined based on the estimates $\hat{\beta}, \hat{s}$ estimated using the entire sample, under the constraints of H_0^B.

In our setting a transformation $\mathcal{T}$ of a function $W(x)$ can be defined as:

$$(\mathcal{T}W)(x) = W(x) - \int_0^x \widetilde{\mathbf{P}}'(y) \left(\int_x^1 \widetilde{\mathbf{P}}(v)\widetilde{\mathbf{P}}'(v) f_X(v) dv\right)^+ \left(\int_y^1 \widetilde{\mathbf{P}}(w) W(dw)\right) f_X(y) dy \tag{30}$$

where A^+ denotes the Moore-Penrose generalized inverse of A. In practice, we cannot evaluate this transformation and instead we use its sample equivalent, $\mathcal{T}_n$. For technical reasons, we add a trimming which removes observations that fall just below knots. Let $\frac{1}{2} < \zeta < 1$ and

$$\mathcal{G} \equiv \left\{ i : x_i \in [0,1] \setminus \bigcup_{\ell}^{L} \left(t_\ell - n^{-\zeta}, t_\ell\right] \right\}. \tag{31}$$

where $\{t_\ell\}_{\ell=1}^L$ is the set of knots we use to define our constrained *B-spline* basis functions. We are now ready to define the transformation:

$$(\mathcal{T}_n W)(x) = W(x) - \frac{1}{n}\sum_{i \in \mathcal{G}} \widetilde{\boldsymbol{P}}_i' \left(\frac{1}{n}\sum_{k=1}^n \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}_k' \mathbb{1}(x_k \geq x_i)\right)^+ \int_{x_i}^1 \widetilde{\mathbf{P}}(w) W(dw) \mathbb{1}(x_i < x). \tag{32}$$

5.2.3 How the transformation removes the problematic terms

Suppose we apply the transformation to a step function $W(x)$ of the following form:

$$W(x) = \frac{1}{n}\sum_{i=1}^n g(x_i, z_i) \mathbb{1}(x_i < x)$$

7. e.g. we can enforce an equality constraint of the form $\beta_k = \beta_{k+1}$ by replacing the two *B-spline* basis functions p_k and p_{k+1} with a single term of the form $p_k + p_{k+1}$.

where $g(x_i, z_i)$ is some known function. By the properties of a Riemann-Stieltjes integrals with a step function as the integrator:

$$\int_{x_i}^{1} \widetilde{\mathbf{P}}(w)W(dw) = \sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k \left(\frac{1}{n}g(x_k, z_k)\right) \mathbb{1}(x_k \geq x_i) = \frac{1}{n}\sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k g(x_k, z_k)\mathbb{1}(x_k \geq x_i).$$

Then:

$$\begin{aligned}
&(\mathcal{T}_n W)(x) = \\
&= W(x) - \frac{1}{n}\sum_{i\in\mathcal{G}} \widetilde{\boldsymbol{P}}'_i \left(\frac{1}{n}\sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}'_k \mathbb{1}(x_k \geq x_i)\right)^{+} \left(\frac{1}{n}\sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k g(x_k, z_k)\mathbb{1}(x_k \geq x_i)\right) \mathbb{1}(x_i < x) \\
&= \frac{1}{n}\sum_{i\in\mathcal{G}} \left(g(x_i, z_i) - \widetilde{\boldsymbol{P}}'_i \left(\frac{1}{n}\sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}'_k \mathbb{1}(x_k \geq x_i)\right)^{+} \frac{1}{n}\sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k g(x_k, z_k)\mathbb{1}(x_k \geq x_i)\right) \mathbb{1}(x_i < x) \\
&\quad + \frac{1}{n}\sum_{i\notin\mathcal{G}} g(x_i, z_i)\mathbb{1}(x_i < x) \\
&= \frac{1}{n}\sum_{i\in\mathcal{G}:x_i<x} \left(g(x_i, z_i) - \widetilde{\boldsymbol{P}}'_i \left(\frac{1}{n}\sum_{k:x_k\geq x_i} \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}'_k\right)^{+} \frac{1}{n}\sum_{k:x_k\geq x_i} \widetilde{\boldsymbol{P}}_k g(x_k, z_k)\right) + \frac{1}{n}\sum_{i\notin\mathcal{G}:x_i<x} g(x_i, z_i)
\end{aligned}$$

The term in the first summation is a residual from regressing $g(x_i, z_i)$ on $\widetilde{\boldsymbol{P}}_i$, where the estimator is evaluated using only observations above x_i (i.e. x_k such that $x_k \geq x_i$). The transformed $\mathcal{T}_n W$ at a point x has a similar form to the original W, i.e. it is a weighted sum of functions of the observations x_i below x, but for the majority of indices which fall in $\mathcal{G}$ we use the part of $g(x_i, z_i)$ which cannot be explained by *B-splines* and zs for observations above x_i instead of the whole $g(x_i, z_i)$.

Consider the case where $g(x_i, z_i) = \widetilde{\boldsymbol{P}}'_i a$ for some constant vector a, i.e. where $g(x_i, z_i)$ is a linear combination the constrained *B-spline* functions evaluated at x_i, of z_i and of derivatives of the constrained *B-spline* with respect to the switch point. In this case

$$W(x) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\boldsymbol{P}}'_i a\mathbb{1}(x_i < x).$$

Then the transformed version of W is:

$$
\begin{aligned}
(\mathcal{T}_n W)(x) &= \frac{1}{n}\sum_{i\in\mathcal{G}:x_i<x}\left(\widetilde{\boldsymbol{P}}_i'a - \widetilde{\boldsymbol{P}}_i'\left(\frac{1}{n}\sum_{k:x_k\geq x_i}\widetilde{\boldsymbol{P}}_k\widetilde{\boldsymbol{P}}_k'\right)^{+}\frac{1}{n}\sum_{k:x_k\geq x_i}\widetilde{\boldsymbol{P}}_k\widetilde{\boldsymbol{P}}_k'a\right) + \frac{1}{n}\sum_{i\notin\mathcal{G}:x_i<x}\widetilde{\boldsymbol{P}}_i'a \\
&= \frac{1}{n}\sum_{i\in\mathcal{G}:x_i<x}\left(\widetilde{\boldsymbol{P}}_i' - \widetilde{\boldsymbol{P}}_i'\left(\frac{1}{n}\sum_{k:x_k\geq x_i}\widetilde{\boldsymbol{P}}_k\widetilde{\boldsymbol{P}}_k'\right)^{+}\frac{1}{n}\sum_{k:x_k\geq x_i}\widetilde{\boldsymbol{P}}_k\widetilde{\boldsymbol{P}}_k'\right)a + \frac{1}{n}\underbrace{\sum_{i\notin\mathcal{G}:x_i<x}\underbrace{\widetilde{\boldsymbol{P}}_i'a}_{\leq C}}_{=O_p\left(n^{1-\zeta}\right)} \\
&= 0 + O_p\left(n^{-\zeta}\right) = o_p\left(n^{-\frac{1}{2}}\right).
\end{aligned}
$$

The term inside the bracket in the second line is the residual from regressing $\widetilde{\boldsymbol{P}}_i$ on itself, and that residual is identically equal to zero for every i[8].

This shows that the transformation removes all terms that are linear combinations of constrained *B-splines*, zs and terms linearized in the switch point for $i \in \mathcal{G}$, and as the sample size increases, the number of $i \notin \mathcal{G}$ becomes insignificant. This proves the following results:

Proposition 4. *Let* $T_{2,3}(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(x_i < x)\left(m_{\mathcal{B}}(x_i;\hat{s}) - \widehat{m}_{\mathcal{B}}(x_i;\hat{s}) + (\gamma - \hat{\gamma})' z_i\right)$*. Then*

$$(\mathcal{T}_n T_{2,3})(x) = o_p\left(n^{-\frac{1}{2}}\right).$$

Proposition 5. *Let* $T_4(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(x_i < x)\left(\frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial s}(\hat{s} - s^0)\right)$*. Then*

$$(\mathcal{T}_n T_4)(x) = o_p\left(n^{-\frac{1}{2}}\right)$$

and

$$(\mathcal{T}_n T_1)(x) = o_p\left(n^{-\frac{1}{2}}\right).$$

5.2.4 The distribution of the test statistic

We have shown that the transformation removes the last two terms. The next two results show that the first term's distribution remains unchanged and the second term remains negligible.

8. For any generalized inverse we can define $P_X = X(X'X)^{-}+X'$, which is a matrix projecting onto the span of X. It has the property that $P_X X = X$, i.e. the projection of X onto X is X. For a given i we let X be the matrix containing columns $\widetilde{\boldsymbol{P}}_i, \widetilde{\boldsymbol{P}}_{i+1}, \ldots, \widetilde{\boldsymbol{P}}_n$ and think of the residual vector from a projection (regression) of X onto itself. The residuals from this regression are zero: $X - P_X X = 0$. The term in the bracket is just the first entry in the residual vector.

Proposition 6. *Under C1-C3, the transformation does not affect the limit of* $\sqrt{n}T_0(x)$*:*

$$\sqrt{n}(\mathcal{T}_n T_0)(x) \xrightarrow{weakly} \sigma\mathcal{B}(F_X(x))$$

for any $x \in [0,1]$.

Combining all of these results, we arrive at the pivotal asymptotic distribution of the transformed test statistic.

Theorem 1. *Under* H_0 *and conditions C1-C3:*

$$\sqrt{n}\,(\mathcal{T}K(x)) \xrightarrow{weakly} \sigma_u\mathcal{B}(F_X(x)),$$

$$\breve{\sigma}_u^2 \xrightarrow{p} \sigma_u^2.$$

In order to implement rests based on this asymptotic distribution, we rely on functionals such as Kolmogorov-Smirnov, Cramér-von-Mises and Anderson-Darling, as described in Section 5.3. The statistics achieve their respective distributions by Theorem 1 and continuous mapping theorem.

5.3 Algorithm outline

STEP 1 Order the sample $\{(x_i, z_i, y_i)\}_{i=1}^n$ in the ascending order of x. Without a loss of generality, we will assume that the original sample is already ordered in this way.

STEP 2 Find a constrained estimator $\widehat{m}_{\mathcal{B}}(\cdot, \hat{s})$ under H_0^B in (22) together with estimator $\widehat{\gamma}$ of γ_0 and compute the residuals $\widehat{u}_i = y_i - \widehat{m}_{\mathcal{B}}(x_i; \hat{s}) - z_i'\widehat{\gamma}$, $i = 1, ..., n$.

Let $\mathbf{P}_{L_j;j}(x)$ denote the L_j-dimensional vector of base *B-splines* on $[s_{j-1}, s_j]$, $j = 1, \ldots, J+1$, computed at x. The estimation uses the long system $\mathbf{P} \equiv (\mathbf{P}_{L_1;1}(x)', \ldots, \mathbf{P}_{L_{J+1};J+1}(x)')'$, However, the constrained estimation under H_0^B results in some binding constraints. Once the binding constraints are enforced, we end up with a much smaller system of relevant base *B-splines*. We can refer to it as the system of "effective polynomials"and denote it as $\widetilde{\mathbf{P}}(x)$. Let

$$\widetilde{\boldsymbol{P}}_k \equiv (\widetilde{\mathbf{P}}(x_k)', z_k')'.$$

STEP 3 For each $i = 1, \ldots, n$, compute the new residual

$$\widehat{v}_i = \widehat{u}_i - \widetilde{\boldsymbol{P}}_i' \Big(\sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}_k' \mathbb{1}(x_k \geq x_i) \Big)^{+} \sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k \mathbb{1}(x_k \geq x_i) \widehat{u}_k. \tag{33}$$

STEP 4 Compute the estimate of the variance of u_i, σ_u^2, as $\breve{\sigma}_u^2 = \frac{1}{n}\sum_{i=1}^{n} \breve{u}_i^2$, where $\breve{u}_i$ are unconstrained residuals $\breve{u}_i = y_i - \breve{m}_{\mathcal{B}}(x_i; \breve{s}) - z_i' \breve{\gamma}$ (details on how to obtain $\breve{m}_{\mathcal{B}}(x_i; \breve{s})$ are given later).

STEP 5 Compute $\widetilde{M}_{\tilde{n}}(x_i) = \frac{1}{\sqrt{\tilde{n}}} \sum_{k=1}^{\tilde{n}} \hat{v}_k \mathbb{1}(x_k \geq x_i)$ and calculate the values of standard functionals such as the Kolmogorov-Smirnov, Cramér-von-Mises and Anderson-Darling defined respectively as

$$\mathcal{KS}_{\tilde{n}} = \sup_{i=1,..,n} \left| \frac{\widetilde{M}_n(x_i)}{\breve{\sigma}_u} \right|, \quad \mathcal{CvM}_{\tilde{n}} = \sum_{i=1}^{\tilde{n}} \frac{\widetilde{M}_n(x_i)^2}{n \breve{\sigma}_u^2}, \quad \mathcal{AD}_{\tilde{n}} = \sum_{i=1}^{n} \frac{\widetilde{M}_n(x_i)^2/n}{\breve{\sigma}_u^2 \widehat{F}_X(x_i)}, \tag{34}$$

where $\widehat{F}_X$ denotes the empirical c.d.f. of X.[9] Compare them to the critical values $\mathcal{KS}_{\tilde{n}}^*(\alpha_0)$, $\mathcal{CvM}_{\tilde{n}}^*(\alpha_0)$, $\mathcal{AD}_{\tilde{n}}^*(\alpha_0)$, respectively, for a chosen significance level α_0. If. e.g., $\mathcal{KS}_{\tilde{n}} > \mathcal{KS}_{\tilde{n}}^*(\alpha_0)$, reject the null by Kolmogorov-Smirnov at the significance level α_0.

Conducting STEP 1 In the first step we estimate the regression function m under the null hypothesis (22). For that we approximate m on each subinterval $[s_{j-1}, s_j]$ by *B-spline* $m_{\mathcal{B};j}$ as defined in (13) and the approximation on the whole domain is described by a join in (14). The constraints in $T_{\{(q_j, L_j)\}_{j=1}^{J+1}, s}$ incorporate (17)-(18) which guarantee sufficient smoothness.

The fundamental difference between our approach and the approach in Komarova and Hidalgo (2023) is that here we do not take switch points s_j, $j = 1, \ldots, J+1$, as known but estimate them as well. As we will see, this leads not to just complications in the constrained estimation but also to significant theoretical complication in the testing methodology due to the need to control for the estimation error in these switch points.

The idea is to consider the objective function

$$\widehat{Q}^*(s, \beta_{all}, \gamma) = \frac{1}{n} \sum_{i=1}^{n} (y_i - m_{\mathcal{B}}(x_i; s) - z_i'\gamma)^2$$

9. One could, of course, center the process $\widetilde{M}_{\tilde{n}}(x)$ to ensure that it converges to a Brownian bridge indexed by the empirical c.d.f. of X. Then $\mathcal{AD}_{\tilde{n}}$ would be defined in a standard manner as follows: $\mathcal{AD}_{\tilde{n}} = \sum_{i=1}^{\tilde{n}} \frac{\widetilde{M}_{\tilde{n}}(x_i)^2/\tilde{n}}{\breve{\sigma}_u^2 \widehat{F}_X(x_i)(1-\widehat{F}_X(x_i))}$.

and then solve the problem

$$\min_{s,\beta_{all},\gamma} \widehat{Q}^*(s,\beta_{all},\gamma) \tag{35}$$

subject to the constraints

$$s_1 < s_2 < \ldots < s_J, \tag{36}$$

$$\beta_{all} \in T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}. \tag{37}$$

First of all, let us note that due to the requirement (7) on the functional properties in classes $\mathcal{M}_j$, $j = 1, \ldots, J+1$, in the overwhelming majority of applications, the properties in each $\mathcal{M}_j$ will be described by conditions on the derivatives of m (potentially on combinations of several derivatives). In cases when each $\mathcal{M}_j$ is described by inequalities on linear combinations of derivatives, all the constraints in $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ are linear inequalities[10]. This was illustrated earlier in the context of the U-shape property in Example 1. Thus, constraints (37) in such scenarios are especially easy to implement. However, the optimization is complicated by the fact that the switch points are unknown. The locations of switch points determine knots points on each subinterval and the values of the polynomials on the *B-spline* bases.

We can see two main approaches to such optimization. The first approach would be to use the closed-form expressions for *B-spline* base polynomials when programming the objective function in (35). These closed form expressions would explicitly account for the knots points which, in their turn, depend on the choice of switch points. Then non-linear optimization tools can be used.

Another approach, which may especially be convenient when dealing with a small number J of switch points, would be to conduct the grid search. Choose a grid on $[\underline{x}, \overline{x}]$, say of R points, and select all possible J-dimensional subsets from these R points. In these selected subsets J points are naturally ordered and can be treated as candidates for the set of switch points. Then the approximation (14) is constructed taking these points as candidate points for partitioning and then the problem (35) is solved subject to (37) only. In the end we select the sequence of switch point that delivers the smallest value of the objective function. Of course, such a grid search would result in a program conducting the estimation for $\binom{R}{J}$ subsets but, again, may be feasible for small values of J and especially in situations when there is only one switch point.

10. Equalities, of course, can be represented through inequalities.

Conducting STEP 2 In this step we need to define the system of "effective polynomials" by enforcing the biding constraints. Once again, this may be convenient to illustrate using U-shape as an main example. In this case if in the constrained estimate $\widehat{b}_{all}$ of β_{all} we have

$$\widehat{b}_{all,h_1} = \widehat{b}_{all,h_1+1} = \ldots = \widehat{b}_{all,h_2}$$

for some indices $h_1 < h_2$, and $\widehat{b}_{all,h_1-1} \neq \widehat{b}_{all,h_1}$, $\widehat{b}_{all,h_2} \neq \widehat{b}_{all,h_2+1}$, then instead of $h_2 - h_1 + 1$ different respective base *B-splines* we will include the sum of all these $h_2 - h_1 + 1$ base *B-splines* as one polynomial into $\widetilde{\mathbf{P}}(x)$.

Conducting STEP 3 is straightforward. It comes down to computing the projection of $\{v_k\}_{k=i}^n$ on $\{\widetilde{\mathbf{P}}(x_k)\}_{k=i}^n$ and then using the projection coefficient to compute the new residual for i. This can be conducted by recursive least squares.

Conducting STEP 4 involves finding an unconstrained estimator $\breve{m}_{\mathcal{B}}(x_i)$. This estimator can be found e.g. by either solving

$$\min_{\beta_{all},\gamma} \widehat{Q}^* \left(\hat{s}, \beta_{all}, \gamma\right)$$

subject to only suitable smoothness constraints in (37) and with $\hat{s}$ taken from the constrained estimation. Alternatively, one can use just one system of base *B-splines* on the whole interval $[\underline{x}, \overline{x}]$ and conduct unconstrained nonparametric estimation using that base.

Conducting STEP 5 is straightforward.

5.4 Bootstrap

Although our test statistic has a pivotal distribution and allows asymptotic testing, the performance may not be the best in small samples. As an alternative, we provide a valid bootstrap algorithm.

STEP 1 Let $\widetilde{m}_{\mathcal{B}}(x_i, \hat{s})$ and $\widetilde{\gamma}$ be the estimators analogous to $\widehat{m}_{\mathcal{B}}(x_i; \hat{s})$ and $\hat{\gamma}$ but evaluated without the constraints[11]. Compute the unconstrained residuals as:

$$\widetilde{\varepsilon}_i = y_i - \widetilde{m}_{\mathcal{B}}(x_i, \hat{s}) - \widetilde{\gamma}' z_i.$$

STEP 2 Draw a random sample from the empirical distribution of the unconstrained residuals centered at zero: $\left\{\widetilde{\varepsilon}_i - \frac{1}{n}\sum_{j=1}^{n} \widetilde{\varepsilon}_j\right\}_{i=1}^{n}$, denote it by $\{\varepsilon_i^*\}_{i=1}^{n}$. Construct the bootstrap outcomes y_i^* using the constrained estimators:

$$y_i^* = \widehat{m}_{\mathcal{B}}(x_i; \hat{s}) + \hat{\gamma}' z_i + \varepsilon_i^*.$$

STEP 3 Compute the bootstrap estimators $\widehat{m}_{\mathcal{B}}^*(x_i, \hat{s}^*)$ from (5.4). Use them to construct the bootstrap residuals

$$\hat{\varepsilon}_i^* = y_i^* - \widehat{m}_{\mathcal{B}}^*(x_i, \hat{s}^*) - \hat{\gamma}^{*'} z_i.$$

Use them to find the value of the bootstrap statistic:

$$\begin{aligned}&\sqrt{n}\,(\mathcal{T}K^*(x)) \\ &= \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{G}} \left(\hat{\varepsilon}_i^* - \widetilde{\boldsymbol{P}}'_i \left(\frac{1}{n} \sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}'_k \mathbb{1}(x_k \geq \widetilde{x}_i) \right)^{+} \frac{1}{n} \sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k \hat{\varepsilon}_k^* \mathbb{1}(x_k \geq \widetilde{x}_i) \right) \mathbb{1}(x_i < x).\end{aligned}$$

Theorem 2. *Under conditions C1-C3:*

$$\sqrt{n}\,(\mathcal{T}K^*(x)) \xrightarrow{weakly} \sigma_u \mathcal{B}(F_X(x))$$

in probability.

11. Note that we use the same set of knots as in the constrained case.

6 Extensions

6.1 Extended setting

In the extended setting, we allow the response variables y to depend on other covariates z in a parametric way. To be more specific,

$$y_i = m(x_i) + \psi(z_i, \gamma_0) + u_i, \tag{38}$$

$$E[u_i|x_i, z_i] = 0. \tag{39}$$

Function $m \in C^1[\underline{x}, \overline{x}]$ is unknown, whereas $\psi(\cdot,\cdot)$ is a *known* function, and $\gamma_0 \in \mathbb{R}^k$ is an unknown finite-dimensional parameter.

Now, of course, an additional issue will be identifying γ_0 in addition to identifying the ordered sequence of switch points. A popular choice of function ψ in the empirical work will be $\psi(z, \gamma_0) = z'\gamma_0$.

7 Monte-Carlo simulations

In Scenarios 1-3 we consider

$$y = m(x) + \gamma_0' z + u, \quad u \sim \mathcal{N}(0, \sigma^2)$$

Subscenarios labelled A. we will have no additional covariates – thus, we will take it as given that $\gamma_0 = 0$. We will take x to be uniformly distributed on $[0, 1]$.

Subscenarios labelled B. we will take $\gamma_0 = -2$ and treat γ_0 as unknown in our estimation. We will take x and z to be uniformly distributed on $[0, 1]$ and independent.

Subscenarios labelled C. we will take $\gamma_0 = -2$ and treat γ_0 as unknown in our estimation. We will take x and z to be

$$x = 0.8w_1 + 0.2v,$$

$$z = 0.25 - 0.25w_2 + 0.75v,$$

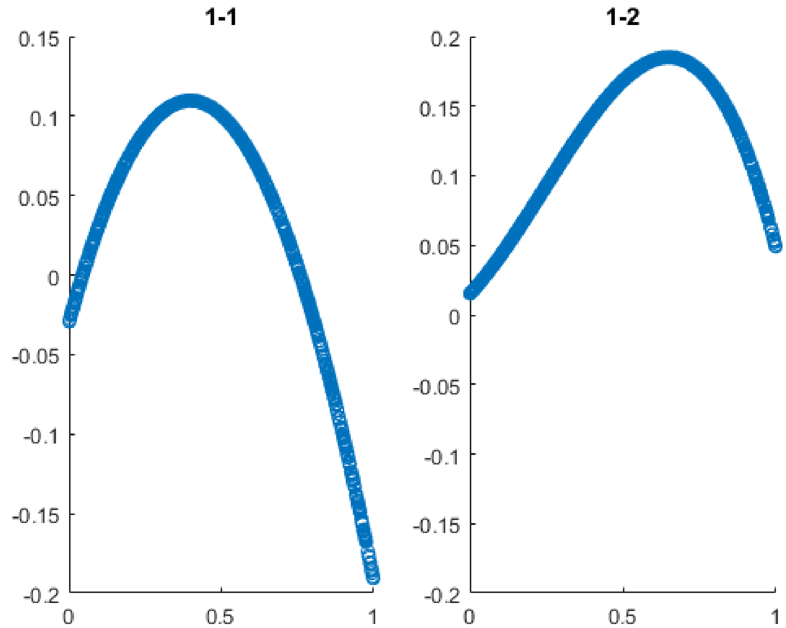


FIGURE 3: Graphs of functions $m(\cdot)$ in Scenarios 1-1 and 1-2.

where w_1, w_2 and v are uniformly distributed on $[0, 1]$ and mutually independent. Thus, on this subscenarios z will be correlated with the base B-splines.

Scenario 1. We consider several sub-scenarios within this scenario. Sub-scenarios 1-j, $j = 1, 2, 3$, can be described as

$$m(x) = -0.75(0.2 - x)^2 + 0.415 \log(1 + x) \quad \text{(1-1)},$$
$$m(x) = -0.75(x - 0.5)(0.2 - x)^2 + 0.415 \log(1 + x) \quad \text{(1-2)},$$
$$m(x) = 0.25(0.2 - x)^3 + 0.415 \exp(-80(x - 0.2)^2) \quad \text{(1-3)},$$

and σ is taken to be 0.05 (the findings under H_0 are quite robust with respect to the value of σ).

The graphs of the functions in Scenarios 1-1 and 1-2 are given in Figure 3. The graphs of the function in Scenario 1-3 as well as its derivative are given in Figure 4.

We start with sub-scenarios 1-jA, $j = 1, 2, 3$, we have $\gamma_0 = 0$ (in other words, there is no control for other covariates).

We apply our *B-spline* and *P-spline* methodology to test an inverse U-shape in m. Results are given in Tables 1-3.

		A				*B*				*C*			
Setting	Method	B-splines		P-splines		B-splines		P-splines		B-splines		P-splines	
		10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
$L_1' = L_2' = 4$ h	KS	0.076	0.048	0.088	0.042	0.072	0.028	0.08	0.058	0.0675	0.0375	0.0724	0.05
$N = 1000$	CvM	0.08	0.05	0.092	0.06	0.07	0.02	0.098	0.042	0.0875	0.045	0.09	0.045
$\sigma = 0.05$	AD	0.07	0.042	0.09	0.058	0.086	0.026	0.096	0.038	0.0775	0.04	0.1	0.05
$L_1' = L_2' = 6$	KS	0.074	0.03	0.085	0.0325	0.074	0.028	0.1	0.045	0.0575	0.035	0.0825	0.0325
$N = 1000$	CvM	0.074	0.03	0.085	0.035	0.078	0.028	0.0975	0.0525	0.095	0.0475	0.0975	0.05
$\sigma = 0.05$	AD	0.07	0.038	0.0775	0.035	0.07	0.032	0.1	0.0525	0.08	0.04	0.0725	0.045
$L_1' = L_2' = 9$	KS	0.088	0.052	0.085	0.045	0.068	0.03	0.098	0.058	0.0925	0.045	0.0925	0.045
$N = 1000$	CvM	0.102	0.048	0.095	0.05	0.072	0.038	0.098	0.06	0.095	0.065	0.095	0.065
$\sigma = 0.05$	AD	0.088	0.048	0.0875	0.045	0.076	0.034	0.09	0.056	0.0925	0.0575	0.0925	0.575

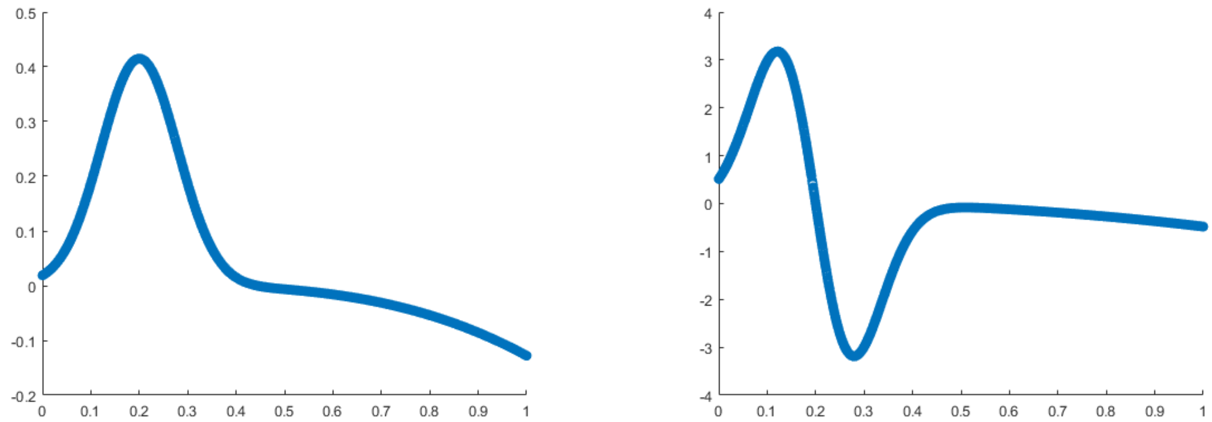
Table 1: Test for an inverse U-shape in Scenario 1-1.

		A				*B*				*C*			
Setting	Method	B-splines		P-splines		B-splines		P-splines		B-splines		P-splines	
		10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
$L_1' = L_2' = 4$	KS	0.052	0.018	0.064	0.024	0.074	0.026	0.095	0.0525	0.0625	0.025	0.08	0.0275
$N = 1000$	CvM	0.066	0.028	0.084	0.042	0.086	0.038	0.095	0.0425	0.0625	0.035	0.0875	0.0375
$\sigma = 0.05$	AD	0.068	0.032	0.094	0.032	0.084	0.03	0.08	0.0525	0.06	0.0225	0.095	0.0425
$L_1' = L_2' = 6$	KS	0.074	0.028	0.0925	0.0425	0.066	0.024	0.09	0.048	0.0825	0.05	0.1075	0.05
$N = 1000$	CvM	0.078	0.03	0.0925	0.0525	0.056	0.026	0.072	0.034	0.1175	0.0475	0.1	0.0475
$\sigma = 0.05$	AD	0.072	0.028	0.085	0.0425	0.054	0.02	0.074	0.046	0.105	0.0525	0.075	0.0375
$L_1' = L_2' = 9$	KS	0.09	0.046	0.09	0.046	0.09	0.05	0.085	0.0475	0.115	0.07	0.1	0.0475
$N = 1000$	CvM	0.088	0.052	0.088	0.052	0.106	0.048	0.0875	0.045	0.1225	0.0725	0.0925	0.0475
$\sigma = 0.05$	AD	0.09	0.05	0.09	0.05	0.086	0.062	0.0925	0.05	0.1325	0.08	0.0975	0.0425

Table 2: Test for an inverse U-shape in Scenario 1-2.

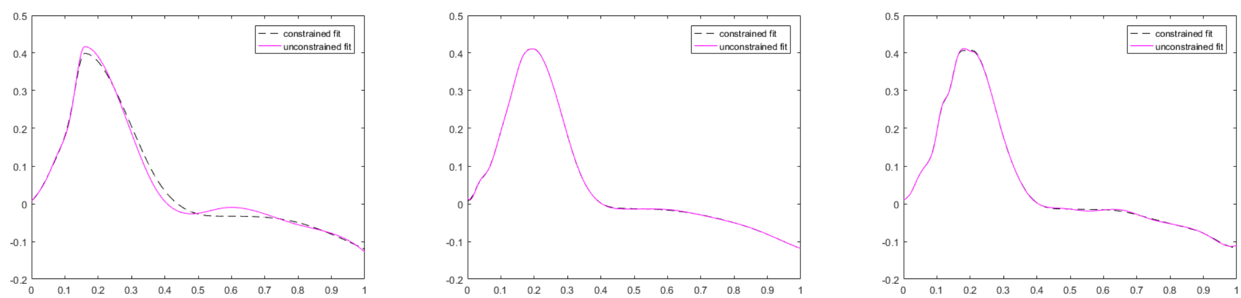
		A				*B*				*C*			
Setting	Method	B-splines		P-splines		B-splines		P-splines		B-splines		P-splines	
		10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
$L_1' = L_2' = 4$	KS	1	1	0.745	0.4675	1	0.998	0.72	0.5025	1	1	0.33	0.0675
$N = 1000$	CvM	1	1	0.6075	0.4275	1	0.998	0.5425	0.365	1	1	0.295	0.0575
$\sigma = 0.05$	AD	1	0.996	0.62	0.4375	1	1	0.5	0.2925	1	0.995	0.1025	0.0375
$L_1' = L_2' = 6$	KS	0.082	0.028	0.1	0.044	0.086	0.042	0.095	0.06	0.1025	0.05	0.0975	0.045
$N = 1000$	CvM	0.084	0.04	0.098	0.048	0.078	0.038	0.0925	0.045	0.1	0.045	0.1	0.045
$\sigma = 0.05$	AD	0.084	0.044	0.096	0.044	0.076	0.024	0.0975	0.04	0.0775	0.05	0.0825	0.0475
$L_1' = L_2' = 9$	KS	0.124	0.052	0.08	0.0375	0.084	0.036	0.0775	0.035	0.085	0.0275	0.0875	0.0425
$N = 1000$	CvM	0.132	0.052	0.095	0.0425	0.092	0.054	0.09	0.0525	0.0625	0.04	0.0875	0.055
$\sigma = 0.05$	AD	0.118	0.058	0.09	0.05	0.092	0.04	0.09	0.0475	0.065	0.025	0.0925	0.04

Table 3: Test for an inverse U-shape in Scenario 1-3.



(a) Function in 1-3. (b) Derivative of function in 1-3.

FIGURE 4: Graphs of function $m(\cdot)$ and it derivative in Scenario 1-3.

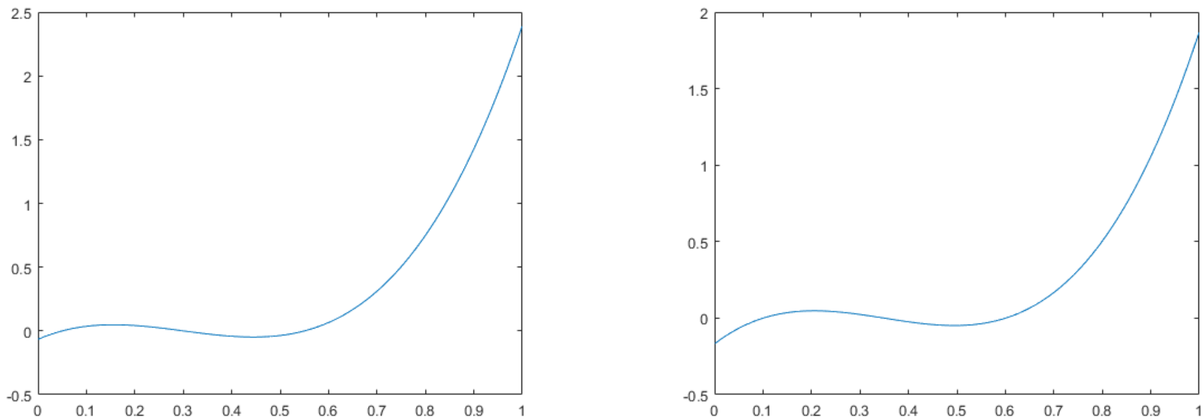


(a) $L_1' = L_2' = 4$ (b) $L_1' = L_2' = 6$ (c) $L_1' = L_2' = 9$

FIGURE 5: Typical *B-spline* fits of function $m(\cdot)$ in Scenario 1-3.

A particularly interesting case in this setting is the testing result in Table 3 where we see drastically different results for $L_1' = L_2' = 4$ compared to other cases of $L_1' = L_2' = 6$ and $L_1' = L_2' = 9$. The intuition for this can be obtained from Figure 4, where we see that the derivative of function m is close to constant on a subinterval. Since our bootstap draws residuals from the *unconstrained B-splines* fit, the drastic differences between unconstrained and constrained fits in that subinterval can create the high rejection rate. The typical *B-splines* fits with an adaptive choice of the turning point for our three cases of (L_1', L_2') are given in Figure 5. What we see is that for the case $L_1 = L_2' = 4$ the unconstrained *B-splines* typically estimates the function as being increasing on a part of the subinterval with the derivative close to zero, which explains rejection rates for that case in Table 3. This situation is no longer the case when $L_1' = L_2' = 6$ or $L_1' = L_2' = 9$, as can be seen from typical fits in 5 as well.

Scenario 2. In Scenario 2 we use $m(x) = x - a - 6(x-a)^2 + 8(x-a)^3$. The graphs of this function



(a) $a = 0.05$.

(b) $a = 0.1$

FIGURE 6: Graphs of functions in Scenario 2.

for $a = 0.05$ and $a = 0.1$ are given in Figure 6. As we can see, the functions are not U-shaped but it is a difficult case to reject U-shape as its violations only happen in a small domain near one of the support ends. It is harder to reject U-shape for $a = 0.05$ tan for $a = 0.1$.

8 Applications

8.1 "The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development", by Q.Ashraf and O. Galor, American Economic Review, 2013

The paper argues that in the course of the prehistoric exodus of Homo sapiens out of Africa, genetic diversity has had a persistent hump-shaped effect on the the logarithm of population density and on comparative economic development. The paper contains many various findings related to the presence or absence of hump-shaped effects. The authors use quadratics in all their specifications to establish the presence or absence of hump shapes.

We apply our method and compare our results to Ashraf and Galor (2013) Table 4, which contains robustness checks to using alternative distances. It is given in Figure 7. The authors' conclusion is that "the results presented in Table 4 indicate that migratory distance from East Africa is the only concept of distance that confers a significant nonmonotonic effect on log population

		$a = 0.05$				$a = 0.1$			
Setting	Method	B-splines		P-splines		B-splines		P-splines	
		10%	5%	10%	5%	10%	5%	10%	5%
$L_1' = 4, L_2' = 4$	KS	1	1	1	1	1	1	1	1
$N = 1000$	CvM	0.998	0.992	1	1	1	1	1	1
$\sigma = 0.05$	AD	1	1	1	1	1	1	1	1
$L_1' = 6,\ L_1' = 6$	KS	1	1	1	1	1	1	1	1
$N = 1000$	CvM	1	1	1	1	1	1	1	1
$\sigma = 0.05$	AD	1	1	1	1	1	1	1	1
$L_1' = 9,\ L_2' = 9$	KS	1	1	1	1	1	1	1	1
$N = 1000$	CvM	1	1	1	1	1	1	1	1
$\sigma = 0.05$	AD	1	1	1	1	1	1	1	1
$L_1' = 4, L_2' = 4$	KS	0.792	0.654	0.8475	0.795	1	1	1	1
$N = 1000$	CvM	0.636	0.448	0.7425	0.64	1	1	1	1
$\sigma = 0.1$	AD	0.908	0.812	0.92	0.88	1	1	1	1
$L_1' = 6,\ L_1' = 6$	KS	0.858	0.756	0.89	0.785	1	1	1	1
$N = 1000$	CvM	0.786	0.664	0.83	0.755	1	1	1	1
$\sigma = 0.1$	AD	0.95	0.904	0.95	0.925	1	1	1	1
$L_1' = 9,\ L_2' = 9$	KS	0.874	0.78	0.865	0.8025	1	1	1	1
$N = 1000$	CvM	0.8	0.678	0.8325	0.77	1	0.992	1	1
$\sigma = 0.1$	AD	0.964	0.934	0.9525	0.94	1	1	1	1
$L_1' = 4, L_2' = 4$	KS	0.132	0.066	0.164	0.096	0.748	0.536	0.732	0.55
$N = 1000$	CvM	0.144	0.084	0.186	0.1	0.688	0.486	0.668	0.564
$\sigma = 0.25$	AD	0.25	0.152	0.262	0.158	0.876	0.714	0.872	0.764
$L_1' = 6,\ L_1' = 6$	KS	0.188	0.086	0.198	0.136	0.794	0.62	0.8075	0.6725
$N = 1000$	CvM	0.196	0.11	0.246	0.16	0.73	0.544	0.73	0.5675
$\sigma = 0.25$	AD	0.286	0.17	0.32	0.228	0.876	0.738	0.8625	0.7675
$L_1' = 9,\ L_2' = 9$	KS	0.17	0.106	0.208	0.136	0.65	0.526	0.672	0.532
$N = 1000$	CvM	0.2	0.102	0.24	0.166	0.576	0.39	0.548	0.438
$\sigma = 0.25$	AD	0.278	0.202	0.324	0.222	0.728	0.58	0.728	0.596

Table 4: Test for U-shape in Scenario 2.

		$a = 0.05$				$a = 0.1$			
Setting	Method	B-splines		P-splines		B-splines		P-splines	
		10%	5%	10%	5%	10%	5%	10%	5%
$L'_1 = 4, L'_2 = 4$	KS	1	1	1	1	1	1	1	1
$N = 1000$	CvM	0.996	0.992	1	1	1	1	1	1
$\sigma = 0.05$	AD	1	1	1	1	1	1	1	1
$L'_1 = 6,\ L'_1 = 6$	KS	1	1	1	1	1	1	1	1
$N = 1000$	CvM	0.998	0.998	1	1	1	1	1	1
$\sigma = 0.05$	AD	1	1	1	1	1	1	1	1
$L'_1 = 9,\ L'_2 = 9$	KS	1	1	1	1	1	1	1	1
$N = 1000$	CvM	0.99	0.9475	1	1	1	1	1	1
$\sigma = 0.05$	AD	1	1	1	1	1	1	1	1
$L'_1 = 4, L'_2 = 4$	KS	0.796	0.65	0.904	0.83	1	1	1	1
$N = 1000$	CvM	0.672	0.486	0.864	0.752	1	1	1	1
$\sigma = 0.1$	AD	0.91	0.814	0.954	0.902	1	1	1	1
$L'_1 = 6,\ L'_1 = 6$	KS	0.78	0.662	0.8725	0.8075 1		1	1	
$N = 1000$	CvM	0.66	0.522	0.8175	0.7625	1	0.998	1	1
$\sigma = 0.1$	AD	0.888	0.822	0.9475	0.9175	1	1	1	1
$L'_1 = 9,\ L'_2 = 9$	KS	0.648	0.448	0.71	0.535	1	1	1	1
$N = 1000$	CvM	0.544	0.366	0.6775	0.4825	1	1	1	1
$\sigma = 0.1$	AD	0.816	0.676	0.8325	0.7057	1	1	1	1
$L'_1 = 4, L'_2 = 4$	KS	0.162	0.09	0.195	0.1125	0.662	0.484	0.7725	0.65
$N = 1000$	CvM	0.184	0.118	0.2075	0.1375	0.662	0.468	0.6825	0.5575
$\sigma = 0.25$	AD	0.276	0.16	0.2575	0.19	0.818	0.684	0.8225	0.75
$L'_1 = 6,\ L'_1 = 6$	KS	0.138	0.068	0.195	0.1	0.622	0.406	0.6775	0.4475
$N = 1000$	CvM	0.15	0.096	0.1875	0.105	0.594	0.352	0.6075	0.4675
$\sigma = 0.25$	AD	0.214	0.122	0.245	0.14	0.722	0.47	0.7725	0.6125
$L'_1 = 9,\ L'_2 = 9$	KS	0.184	0.086	0.245	0.1725	0.6	0.436	0.6475	0.5175
$N = 1000$	CvM	0.172	0.1	0.2625	0.1725	0.422	0.314	0.5725	0.4775
$\sigma = 0.25$	AD	0.278	0.162	0.31	0.185	0.62	0.47	0.6775	0.5725

TABLE 5: Test for U-shape in Scenario 2-B.

FIGURE 7: Table 4 from Ashraf and Galor (2013).

TABLE 4—ROBUSTNESS TO ALTERNATIVE DISTANCES

	log population density in 1500 CE				
Distance from:	Addis Ababa (1)	Addis Ababa (2)	London (3)	Tokyo (4)	Mexico City (5)
Migratory distance	0.138**		−0.040	0.052	−0.063
	(0.061)		(0.063)	(0.145)	(0.099)
Migratory distance square	−0.008***		−0.002	−0.006	0.005
	(0.002)		(0.002)	(0.007)	(0.004)
Aerial distance		−0.008			
		(0.106)			
Aerial distance square		−0.005			
		(0.006)			
log Neolithic transition timing	1.160***	1.158***	1.003***	1.047***	1.619***
	(0.144)	(0.138)	(0.164)	(0.225)	(0.277)
log percentage of arable land	0.401***	0.488***	0.357***	0.532***	0.493***
	(0.091)	(0.102)	(0.092)	(0.089)	(0.094)
log absolute latitude	−0.342***	−0.263***	−0.358***	−0.334***	−0.239***
	(0.091)	(0.097)	(0.112)	(0.099)	(0.083)
log land suitability for agriculture	0.305***	0.254**	0.344***	0.178**	0.261***
	(0.091)	(0.102)	(0.092)	(0.080)	(0.092)
Observations	145	145	145	145	145
R^2	0.67	0.59	0.67	0.59	0.63

Notes: This table establishes that, unlike migratory distance from East Africa, alternative concepts of distance, including aerial distance from East Africa and migratory distances from placebo points of origin in other continents across the globe, do not possess any systematic relationship, hump-shaped or otherwise, with log population density in 1500 CE while controlling for the timing of the Neolithic Revolution and land productivity. Heteroskedasticity-robust standard errors are reported in parentheses.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

density." We want to analyze and assess these findings using our methodology.

Our first series of tests is about the specification

$$\ln pd1500 = \alpha + \beta dist + \gamma dist^2 + z'\delta + u, \tag{40}$$

where $dist$ is a distance notion from Figure 7, z is the set of 4 controls used there in every column, and u is the error term.

Finding 1, The quadratic specification in (40) is rejected at the 5% significance level for Columns 2-5 in Table 4 in Ashraf and Galor (2013).

For this finding we use an approach based on Khmaladze's transformation but within the con-

	Kolmogorov-Smirnov			Cramer-von-Mises			Anderson-Darling		
	90%	95% cv	99% cv	90%	95% cv	99% cv	90%	95% cv	99% cv
Column (1)	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$
Column (2)	$<$	$<$	$<$	$>$	$>$	$>$	$>$	$>$	$<$
Column (3)	$>$	$>$	$>$	$>$	$>$	$>$	$>$	$>$	$>$
Column (4)	$<$	$<$	$<$	$>$	$>$	$>$	$>$	$<$	$<$
Column (5)	$>$	$>$	$<$	$>$	$>$	$<$	$<$	$<$	$<$

Table 6: Ashraf and Galor (2013). Test for quadratic specifications in Table 4 in Ashraf and Galor (2013). "cv" stands for the critical value. All critical values are based on 1000 bootstrap draws. $>$ ($<$) means that the test statistic for the functional indicated in the first row is greater (is less) than the respective critical value for that functional.

text of a semiparametric regression (rather than a nonparametric one) as discussed in Stute, Thies, and Zhu (1998). Based on that approach, quadratic specifications in distance (plus other covariates) in Table 4 in Ashraf and Galor (2013) are rejected for Columns (2)-(5) at the 5% level by at least one of our testing functionals. More detailed results are given in Table 6, where we can see that for Columns (2), (3) and (5) the quadratic specifications are rejected by at least two functionals we employ (for Column (3) it is rejected by all three functionals). Results for Column (1), thus, can be taken as supportive of Ashraf and Galor (2013) findings for that particular specification, which cannot be said for specifications in other columns used to justify the use of one particular migratory distance in Column (1).

An immediate conclusion here is that robustness to alternative distances needs to be analyzes through more general hump-shapes that go beyond quadratics. This naturally brings us to using our method.

For a distance of interest in a respective column we choose cubic *B-splines* on both sides of a candidate switch point with intervals on both sides being uniformly divided into 4 subintervals. This results in 12 base splines overall but the constraints of smoothness of the function at the switch point effectively reduce this number of unknown parameters with respect to the distance variable to 9 (for comparison, in the quadratic specification it is 3 unknown parameters). Table 7 shows the results of performing the test using our method with *B-splines*. As we can see, for models analogous to those in Ashraf and Galor (2013) Table 4 which differ from them only in a more general specification with respect to a distance variables, a hump-shape relation with respect to distance is not rejected for distances and in all of the columns.

These conclusions are very different from those reached by quadratic specifications used in Ashraf

	Kolmogorov-Smirnov			Cramer-von-Mises			Anderson-Darling		
	90% cv	95% cv	99% cv	90% cv	95% cv	99% cv	90% cv	95% cv	99% cv
Column (1)	<	<	<	<	<	<	<	<	<
Column (2)	>	<	<	<	<	<	<	<	<
Column (3)	<	<	<	<	<	<	<	<	<
Column (4)	<	<	<	<	<	<	<	<	<
Column (5)	<	<	<	<	<	<	<	<	<

TABLE 7: Ashraf and Galor (2013) data. *B-splines* based test for hump-shaped specifications in Table 4 in Ashraf and Galor (2013). All critical values are based on 1000 bootstrap draws. > (<) means that the test statistic for the functional indicated in the first row is greater (is less) than the respective critical value for that functional.

and Galor (2013). Namely, the aerial distance from East Africa and migratory distance from Tokyo have systematic hump-shaped effect on the logarithm of population density in 1500 CE.

A reader may say that our approach to testing hump-shaped relationship potentially allows only weak monotonicity on both sides of the turning points and, thus, potentially hump-shaped relations we find could exhibit a constant effect before or after the estimated turning point. To address this, we look at our *B-splines* hump-shaped fit, compute a) the difference between the fitted value at the lowest value of the distance and the fitted value at the switch point; b) the difference between the fitted value at the switch point and the fitted value at the largest value of the distance, and then we construct a 95% bootstrap confidence intervals for both these differences. The results are given in Table 8 and allow us to conclude that for Columns (1), (2) and (4) both parts of the fitted curve are strictly monotone at the 5% significance level. For column (3) the first part (increasing) is not rejected to be constant and for Column (5) the second part (decreasing) i snot rejected to be constant. Since our constrained estimation imposes difference 1 to be non-negative and difference 2 to be non-positive, what what be more informative for Columns (3) and (5) is the percentage of analogous bootstrap differences that are close to 0. In the case of Column (5), difference 2 is within 10^{-6} distance from 0 in 5.9% samples (so, 90% CI would have 0 on the boundary as well).

for at the 5% significance level the fitted function has a strict increase over the domain before the estimated switch point and a strict decreases after it.

Finally, taking into account the small size of the sample (just 145 observations) and the presence of additional controls in some specifications in Table 4, we use *P-splines* that is an effective tool for dealing with potential overfitting and avoiding fitted lines that are “too wiggly.” For

	difference 1	95% CI	difference 2	95% CI
Column (1)	1.3061	(0.3611,2.8020)	-3.2675	(-4.1464, -2.2730)
Column (2)	0.9621	(0.1208, 2.4071)	-1.2410	(-2.2153,-0.8581)
Column (3)	0.2132	$(1.7 \cdot 10^{-13}, 1.0135)$	-2.9977	(-3.8233,-2.0648)
Column (4)	0.4352	(0.0044,1.8457)	-3.3684	(-4.7242,-2.2614)
Column (5)	1.3980	(1.0919,2.7541)	-0.4366	$(-2.2800, -1.5 \cdot 10^{-13})$

TABLE 8: Ashraf and Galor (2013) data. Analysis whether there are statistically significant changes in the hump-shaped *B-splines* fit before the estimated switch point and also after it. All critical values are based on 1000 bootstrap draws.

	Kolmogorov-Smirnov			Cramer-von-Mises			Anderson-Darling		
	90% cv	95% cv	99% cv	90% cv	95% cv	99% cv	90% cv	95% cv	99% cv
Column (1)	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$
Column (2)	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$
Column (3)	$>$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$
Column (4)	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$
Column (5)	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$

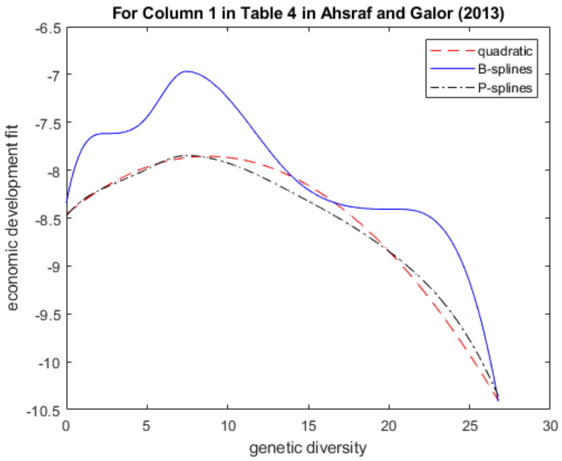
TABLE 9: Ashraf and Galor (2013) data. *P-splines* based test for hump-shaped specifications in Table 4 in Ashraf and Galor (2013). All critical values are based on 1000 bootstrap draws. $>$ ($<$) means that the test statistic for the functional indicated in the first row is greater (is less) than the respective critical value for that functional.

P-splines, we penalize second differences of coefficients choosing the same penalty on different sides of the switch point. The penalty is chosen by the cross validation approach in Eilers and Marx (1996). If fort a model the penalty is rather large then the fitted regression mean would have a shape closer to a quadratic one.

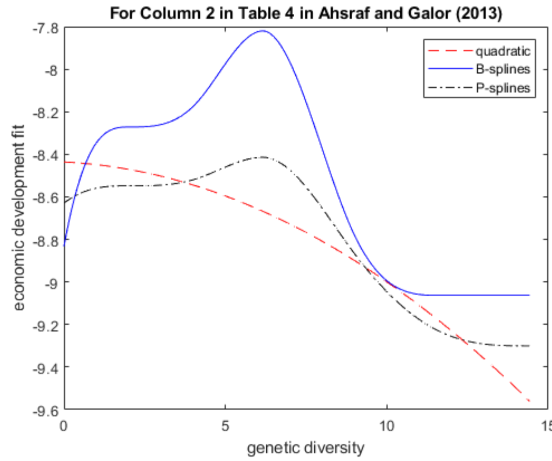
Test results using *P-splines* are given in Table 9. The substantive conclusions are largely similar to those in Table 7.

Finally, we present the following fitted curves for all the columns: first, obtained by quadratic specification in Ashraf and Galor (2013); second, obtained by our *B-spline* methodology under the hump-shape constraint; third, obtained by our *P-splines* methodology with cross-validated penalties enforcing the hump-shape constraint, these are contained in Figure 8.

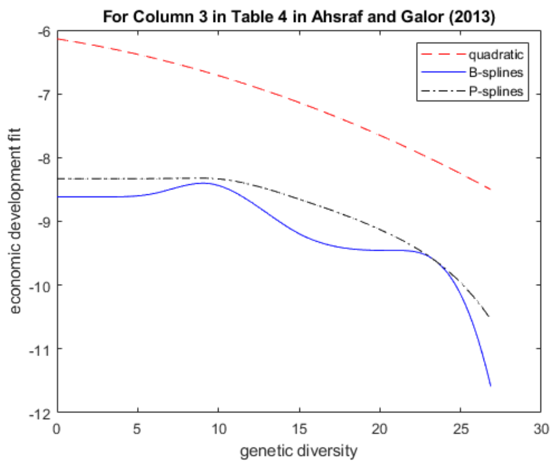
As we can see for the model in Column (1), the fit by *P-splines* is similar to the one provided by the quadratic function. However, for other columns the results are very different. For the model in Column (2), the quadratic specification gives us a monotonically decreasing fit on the domain of the distance, whereas both nonparametric fits indicate a hump-shaped pattern (recall



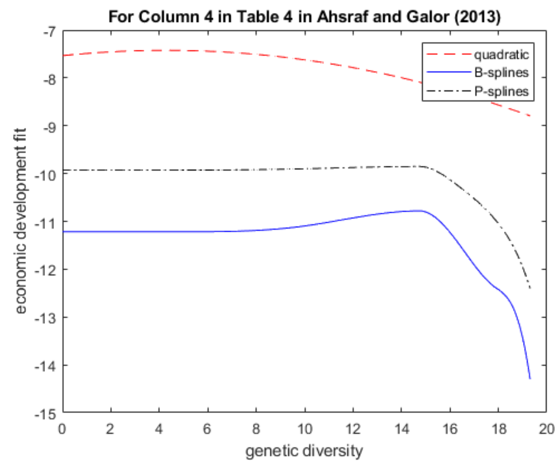
(a) Column (1)



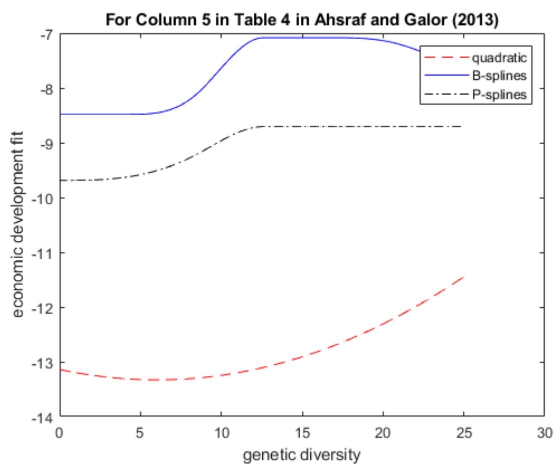
(b) Column (2)



(c) Column (3)



(d) Column (4)

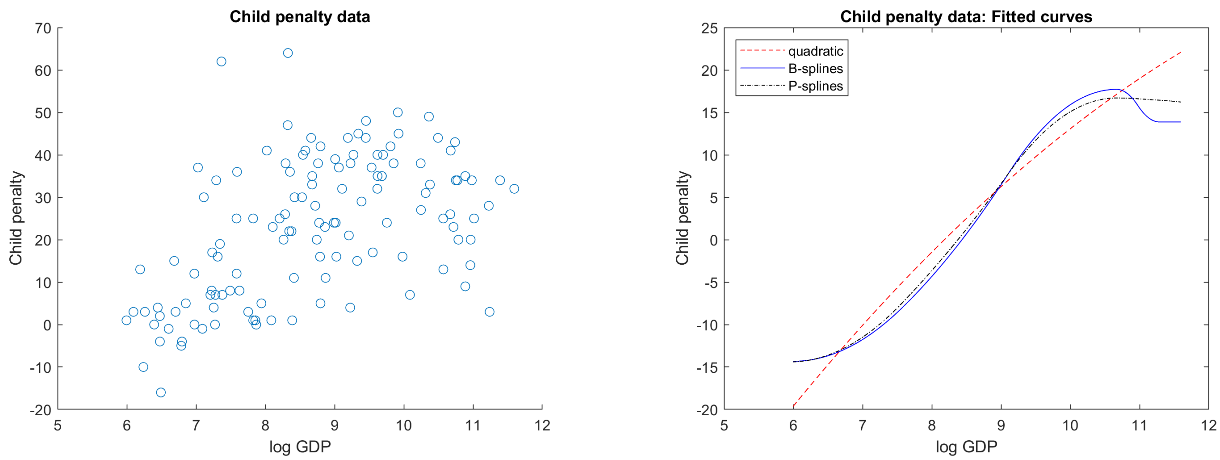


(e) Column (5)

Figure 8: Fitted curves for models in AG Table 4.

that they are not rejected by either *B-splines* or *P-splines*) with visible asymmetries around the turning point. For the model in Column (4) both non-parametric fits indicate a turning point much further to the right than that given by the quadratic fit. Also, in either non parametric fit the decrease after the turning point is much sharper compared to the increase before that (for *P-splines* the curve before the turning point looks almost flat even though statistically it is not). For the model in Column (5), the quadratic specification fit is U-shaped rather than hump-shaped (recall that in Table 4 in Ashraf and Galor (2013) that it is statistically insignificant at the 5% level) which is drastically different from the hump-shaped nonparametric fits exhibiting visible asymmetry around the turning point.

In summary, our methodology finds relationship between migratory distance and the log population density in 1500 CE in to be monotonic for specifications in Columns (3) and (5) (at the 5% level). Our findings for Column (1), including the estimation results by *P-splines*, are largely consistent with Ashraf and Galor (2013). Our findings for models and distances in Columns (2)-(5) are different from those in Ashraf and Galor (2013). Namely, in columns (2) and (4) we find hump-shaped relationship between migratory distances and the log population density in 1500 CE and they are different from quadratic ones. In Column (3) we do not reject at 5% level that find a monotonic weakly decreasing relationship, which is consistent with Ashraf and Galor (2013). However, we do have a statistically significant change in the monotonic relationship if we compare the values of our fitted function at the lower and upper support points (this is different from lack of statistical significance conclusions in Ashraf and Galor (2013)). In Column (5) we do not reject at 5% level that find a monotonic weakly increasing relationship and we also find the change in this monotone function over the domain to be statistically significant, with both of these features being different from findings in Ashraf and Galor (2013)). These differences are best explained by the fact that in Columns (2)-(5) the best fitted curves under the null of a hump-shape exhibit striking asymmetries around the turning points which is not allowed by quadratic specifications.



(a) Data

(b) Fitted curves

FIGURE 9: Left panel: data on log of GDP per capita and child penalty. Right panel: Fitted curves for the model (41).

9 Application 2: Child penalty

We consider the country-level model[12]

$$Child\_Penalty_i = m(\log(GDP\_per\_capita_i)) + \beta Employment\_Gap_i + u_i \qquad (41)$$

$$E[u_i | \log(GDP\_per\_capita_i), Employment\_Gap_i] = 0,$$

and test

$$H_0 : m \quad \text{is hump-shaped.}$$

The employment gap between women and men can reflect societal norms, policies, and labor market dynamics that influence the child penalty. Larger employment gaps e.g. might indicate less support for working mothers, which could exacerbate the child penalty.

The left panel of Figure 9 plots the data ($\log(GDP\_per\_capita), Child\_Penalty$) and the right hand plots the fitted curves $m(\log(GDP\_per\_capita))$ obtained by a) a quadratic specification $m(\log(GDP\_per\_capita)) = \gamma_0 + \gamma_1 Child\_Penalty + \gamma_2 Child\_Penalty^2$, b) *B-spline* specification for $m(\cdot)$, and c) *B-spline* specification for $m(\cdot)$ estimated with the use of penalty on the second-differences of coefficients as explained earlier (so *P-splines*).

As we can see, the quadratic specification finds a strictly increasing curve within the domain

12. We are grateful to Camille Landais and Gabriel Leite-Mariante for providing us with the data.

Kolmogorov-Smirnov			Cramer-von-Mises			Anderson-Darling		
90% cv	95% cv	99% cv	90% cv	95% cv	99% cv	90% cv	95% cv	99% cv
$>$	$>$	$>$	$>$	$>$	$>$	$>$	$>$	$<$

Table 10: Child penalty data. Test for a quadratic form of $m(\cdot)$ in (41). "cv" stands for the critical value. All critical values are based on 1000 bootstrap draws. $>$ ($<$) means that the test statistic for the functional indicated in the first row is greater (is less) than the respective critical value for that functional.

	Kolmogorov-Smirnov			Cramer-von-Mises			Anderson-Darling		
	90% cv	95% cv	99% cv	90% cv	95% cv	99% cv	90% cv	95% cv	99% cv
B-splines	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$
P-splines	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$

Table 11: Child penalty data data. *B-splines* and *P-splines* based tests for hump-shaped $m(\cdot)$ in (41). All critical values are based on 1000 bootstrap draws. $>$ ($<$) means that the test statistic for the functional indicated in the first row is greater (is less) than the respective critical value for that functional.

of log of GDP per capita. We start by applying our test for testing the quadratic specification of $m(\cdot)$ in (41) analogously to how it was conducted in the previous application. The results are in Table 10. As we can see, all three type of tests reject a quadratic form of $m(\cdot)$ at the 5% significance level. Therefore, quadratics do not look like a suitable approach in capturing a nonlinear relationship between log of GDP per capita and child penalty.

Our next step is to test the null hypothesis of hump-shape using *B-splines* and *P-splines* approach. We choose quadratic *B-splines* on both sides of a candidate switch point with intervals on both sides being uniformly divided into 4 subintervals. This results in 10 base splines overall but the constraints of smoothness of the function at the switch point effectively reduce this number of unknown parameters with respect to the distance variable to 7 (compared to three unknown parameters in a quadratic specification). The results are given in 11.

As we can see, the null of a hump-shaped relationship is not rejected at the 10% significance level. The switch point is found to be 10.67 (on the grid of equidistant 1001 grid points in the domain of log of GDP per capita).

10 Conclusion

This paper develops a robust nonparametric methodology for testing shape constraints in regression analysis, accommodating multiple shape changes across the domain of the regressor. Our approach extends beyond conventional U-shaped or hump-shaped patterns to a broad class of nonlinear shapes, including S-shapes, W-shapes. etc. Unlike previous methods that rely on parametric assumptions or require predetermined switch points, our approach identifies turning points adaptively within the data. This allows for greater flexibility and more accurate representation of complex nonlinear relationships, which are often misrepresented by simplistic parametric polynomial (in particular, quadratic) models.

The theoretical contributions of this paper include ensuring that the adaptive estimation of turning points does not compromise the statistical properties of the test statistics, both in finite samples and asymptotically. Practically, the methodology improves the power and interpretability of shape testing by reducing reliance on restrictive parametric forms. As our applications demonstrate, standard parametric approximations can miss or distort true underlying relationships, while our method captures these dynamics more precisely.

In summary, this paper provides a valuable tool for researchers across disciplines who require a flexible, rigorous approach to testing complex shape constraints. The methodology broadens the scope of nonparametric analysis in regression contexts, offering a unified framework that can be applied to partially linear models (or partially parametric models more generally) and expanded to incorporate multiple turning points. Future research may build on this work by further refining the estimation of turning points and exploring additional applications in diverse empirical settings.

References

Aghion, Philippe, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt. 2005. "Competition and innovation: An inverted-U relationship." *The quarterly journal of economics* 120 (2): 701–728.

Aghion, Philippe, John Van Reenen, and Luigi Zingales. 2013. "Innovation and institutional ownership." *American economic review* 103 (1): 277–304.

Andrews, Donald. 1997. “A Conditional Kolmogorov Test.” *Econometrica* 65 (5): 1097–1128.

Ashraf, Quamrul, and Oded Galor. 2013. “The “Out of Africa” hypothesis, human genetic diversity, and comparative economic development.” *American Economic Review* 103 (1): 1–46.

Bessec, Marie, and Julien Fouquau. 2008. “The non-linear link between electricity consumption and temperature in Europe: A threshold panel approach.” *Energy Economics* 30 (5): 2705–2721.

Blanchflower, David G. 2020. *Is Happiness U-shaped Everywhere? Age and Subjective Well-being in 132 Countries.* Working Paper, Working Paper Series 26641. National Bureau of Economic Research, January.

Brown, Robert L, James Durbin, and James M Evans. 1975. “Techniques for testing the constancy of regression relationships over time.” *Journal of the Royal Statistical Society Series B: Statistical Methodology* 37 (2): 149–163.

Calabrese, E.J., and L.A. Baldwin. 2001. “Hormesis: U-shaped dose responses and their centrality in toxicology.”

Chen, Xiaohong. 2007. “Large sample sieve estimation of semi-nonparametric models.” *Handbook of econometrics* 6:5549–5632.

Clark, Andrew. 2007. “Born to Be Mild? Cohort Effects Don’t (Fully) Explain Why Well-Being is U-Shaped in Age.”

Das, Somnath, and Baruch Lev. 1994. “Nonlinearity in the Returns-Earnings Relation: Tests of Alternative Specifications and Explanations.” *Contemporary Accounting Research* 11 (1): 353–379.

De Boor, Carl. 1978. *A practical guide to splines.* Vol. 27. springer-verlag New York.

Delgado, Miguel A., and Javier Hidalgo. 2000. “Nonparametric inference on structural breaks.” *Journal of Econometrics* 96 (1): 113–144.

Eddy, William F. 1980. “Optimum Kernel Estimators of the Mode.” *Ann. Statist.* 8 (4): 870–882.

Eilers, Paul H. C., and Brian D. Marx. 1996. “Flexible smoothing with B-splines and penalties.” *Statistical Science* 11 (2): 89–121.

Engle, Robert F., C. W. J. Granger, John Rice, and Andrew Weiss. 1986. "Semiparametric Estimates of the Relation Between Weather and Electricity Sales." *Journal of the American Statistical Association* 81 (394): 310–320.

Feder, Paul I. 1975. "On Asymptotic Distribution Theory in Segmented Regression Problems–Identified Case." *Ann. Statist.* 3 (1): 49–83.

Freeman, Robert N., and Senyo Y. Tse. 1992. "A Nonlinear Model of Security Price Responses to Unexpected Earnings." *Journal of Accounting Research* 30 (2): 185–209.

Gallant, A. Ronald. 1987. "Advances in Econometrics: Identification and consistency in semi-nonparametric regression." In *Advances in Econometrics: Volume 1: Fifth World Congress,* 145–169.

Ganz, Scott C. 2024. *A Goldilocks Approach to Testing Nonmonotonic Hypotheses in Management and Strategy.* Working Paper, Working Paper Series. Georgetown McDonough School of Business and AEI.

Goldin, Claudia. 1995. "The U-Shaped Female Labor Force Function in Economic Development and Economic History." In *Investment in Women's Human Capital and Economic Development,* edited by T. P. Schultz, 61–90. University of Chicago Press.

Groes, Fane, Philipp Kircher, and Iourii Manovskii. 2014. "The U-Shapes of Occupational Mobility." *The Review of Economic Studies* 82 (2): 659–692.

Heckman, Nancy E. 1986. "Spline Smoothing in a Partly Linear Model." *Journal of the Royal Statistical Society. Series B (Methodological)* 48 (2): 244–248.

Hidalgo, Javier, Heejun Lee, Jungyoon Lee, and Myung Hwan Seo. 2023. "Minimax Risk in Estimating Kink Threshold and Testing Continuity." In *Essays in Honor of Joon Y. Park: Econometric Theory,* 45:233–259. Advances in Econometrics. Emerald Group Publishing Limited.

Hidalgo, Javier, Jungyoon Lee, and Myung Hwan Seo. 2019. "Robust inference for threshold regression models." *Journal of Econometrics* 210 (2): 291–309.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47 (2): 263–291.

Khmaladze, E. V. 1982. "Martingale Approach in the Theory of Goodness-of-Fit Tests." *Theory of Probability & Its Applications* 26 (2): 240–257.

Komarova, Tatiana, and Javier Hidalgo. 2023. "Testing nonparametric shape restrictions." *The Annals of Statistics* 51 (6): 2299–2317.

Kostyshak, Scott. 2017. "Non-parametric Testing of U-shaped Relationships." *Available at SSRN 2905833.*

Lind, Jo Thori, and Halvor Mehlum. 2010. "With or Without U? The Appropriate Test for a U-Shaped Relationship*." *Oxford Bulletin of Economics and Statistics* 72 (1): 109–118.

Lyche, Tom, Carla Manni, and Hendrik Speleers. 2017. "B-Splines and Spline Approximation." Lecture notes.

Muller, Hans-Georg. 1989. "Adaptive Nonparametric Peak Estimation." *Ann. Statist.* 17 (3): 1053–1069.

———. 1992. "Change-Points in Nonparametric Regression Analysis." *Ann. Statist.* 20 (2): 737–761.

Newell, Jay, Ulrike Genschel, and Ni Zhang. 2014. "Media Discontinuance: Modeling the Diffusion "S" Curve to Declines in Media Use." *Journal of Media Business Studies* 11 (4): 27–50.

Newey, Whitney K. 1997. "Convergence rates and asymptotic normality for series estimators." *Journal of Econometrics* 79 (1): 147–168.

Newey, Whitney K., and James L. Powell. 2003. "Instrumental Variable Estimation of Nonparametric Models." *Econometrica* 71 (5): 1565–1578.

Pardo, Angel, Vicente Meneu, and Enric Valor. 2002. "Temperature and seasonality influences on Spanish electricity load." *Energy Economics* 24 (1): 55–70.

Parzen, Emanuel. 1962. "On Estimation of a Probability Density Function and Mode." *Ann. Math. Statist.* 33 (3): 1065–1076.

Procházková, Jana. 2005. "Derivative of B-spline function." *25. konference o geometrii a počítačové grafice.*

Robinson, P. M. 1988. "Root-N-Consistent Semiparametric Regression." *Econometrica* 56 (4): 931–954.

Rogers, Everett M. 2003. *Diffusion of innovations* [in eng]. 5th. New York, NY [u.a.]: Free Press, August.

Simonsohn, Uri. 2018. "Two Lines: A Valid Alternative to the Invalid Testing of U-Shaped Relationships With Quadratic Regressions." *Advances in Methods and Practices in Psychological Science* 1 (4): 538–555.

Speckman, Paul. 1988. "Kernel Smoothing in Partial Linear Models." *Journal of the Royal Statistical Society. Series B (Methodological)* 50 (3): 413–436.

Stute, Winfried. 1997a. "Nonparametric model checks for regression." *The Annals of Statistics,* 613–641.

———. 1997b. "Nonparametric model checks for regression." *Ann. Statist.* 25 (2): 613–641.

Stute, Winfried, Silke Thies, and Li-Xing Zhu. 1998. "Model checks for regression: an innovation process approach." *The Annals of Statistics* 26 (5): 1916–1934.

Sutton, John, and Daniel Trefler. 2016. "Capabilities, Wealth, and Trade." *Journal of Political Economy* 124 (3): 826–878.

Utterback, James M. 1996. *Mastering the dynamics of innovation.* USA: Harvard Business School Press.

Weiman, C.G. 1977. "A study of occupational stressor and the incidence of disease/risk."

Weiss, Alexander, James E. King, Miho Inoue-Murayama, Tetsuro Matsuzawa, and Andrew J. Oswald. 2012. "Evidence for a midlife crisis in great apes consistent with the U-shape in human well-being." *Proceedings of the National Academy of Sciences* 109 (49): 19949–19952.

Wunder, Christoph, Andrea Wiencierz, Johannes Schwarze, and Helmut Küchenhoff. 2013. "Well-being over the life span: Semiparametric evidence from british and german longitudinal data" [in English]. *Review of economics and statistics* 95 (1): 154–167.

A Appendix A: Nonparametric vs quadratic fits,

The purpose of this Appendix is to illustrate that the choice of quadratic specifications can be very misleading when one tries to estimate (inverse) U-shaped relations. Here we outline several scenarios.

We use the following setting: $y = m(x) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $x \sim \mathcal{U}[0,1]$, and ε is independent of x.

Scenario 1. $m(x) = (x^{1/4} - 0.5)^2$, $\sigma = 0.01$. The turning for this regression function is 1/16 but it is not symmetric around this point. This can be seen in Figure 10 which shows one set of generated data (1,000 points) from this model and a fitted line using a quadratic specification.

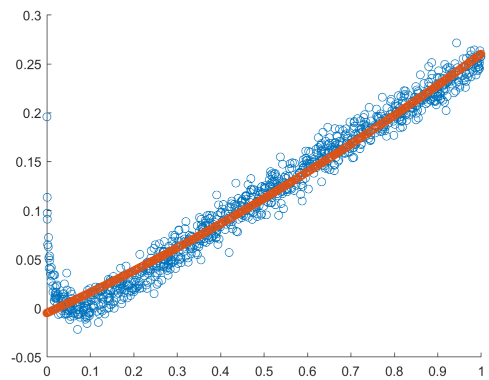
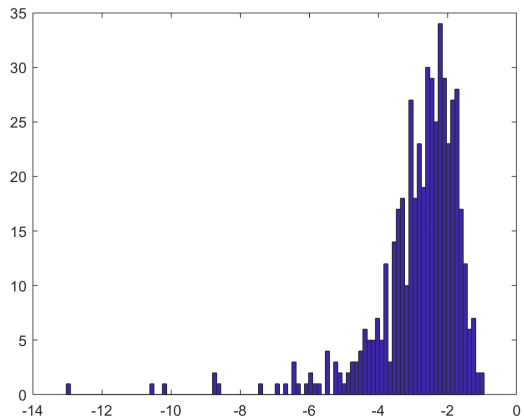


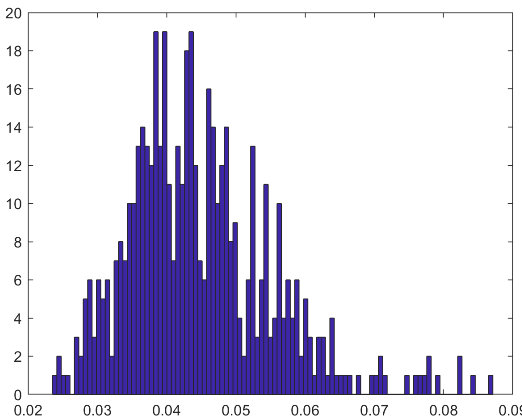
FIGURE 10: Scenario 1.

As we can be seen in 10, the fitted line is monotonic on the whole domain. Indeed, it turns out that the use of quadratic specification results in the estimated turning point being negative with probability almost 1. Figure 11 gives histograms for the estimated turning point in 500 simulations as well the basic summary statistics for those turning points in every sub-case (see the captions), including the quadratic specification subcase in Panel (a). *B-spline* specifications have a vastly superior performance to quadratic specifications, even though they seem to exhibit a negative finite sample bias when estimating the switch point. This is not surprising given the closeness of the switch point to the boundary. Also, with the sample size increase and the suitable increase in the number of knots. the estimated switch points will converge in probability to the true switch point 1/16.

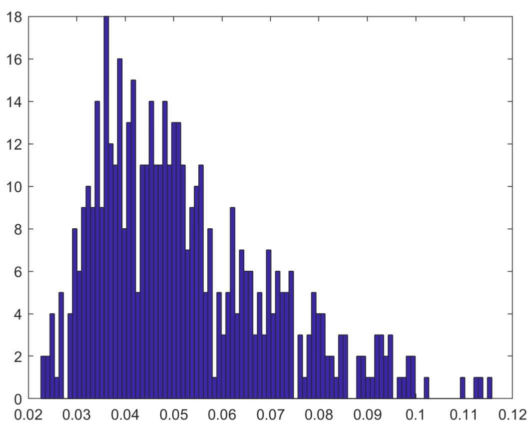
$m(x) = \Phi(\frac{x-0.5}{5}) \cdot 1\!\!1(x \leq 0.5) \left(1 - \Phi(\frac{x-0.1}{0.1})\right) \cdot 1\!\!1(x > 0.5)$, $\sigma = 0.01$. The turning for this regression function is 0.5 but it is not symmetric around this point and it is continuous at that point but not differentiable (the left and the right derivatives exist and are finite, but they take different values). This can be seen in Figure 12 which shows one set of generated data (500 points) from this model and a fitted line using a quadratic specification and a *B-splines* specification with an adaptive choice of a switch point.



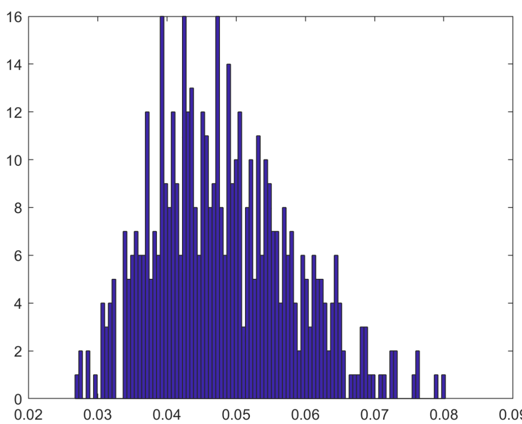
(a) quadratic specification, $n = 500$.
$mean = -2.8370$, $std = 1.3143$
$50th$, $95th$, $99th$ percentiles:
-2.5402, -1.5200, -1.2384



(b) cubic splines, 8 base splines on each side of the turning point, $n = 500$.
$mean = 0.0448$, $std = 0.0106$
$50th$, $95th$, $99th$ percentiles:
0.0433, 0.0636, 0.0784



(c) $5th$ degree splines, 13 base splines on each side of the turning point, $n = 2000$
$mean = 0.0522$, $std = 0.0180$
$50th$, $95th$, $99th$ percentiles:
0.0484, 0.0789, 0.0891



(d) cubic splines, 11 splines on each side of the turning point, $n = 2000$
$mean = 0.0482$, $std = 0.01$
$50th$, $95th$, $99th$ percentiles:
0.0473, 0.0651, 0.0744

FIGURE 11: Histograms and summary statistics of estimated switching points in Scenario 1 using various specifications,. Results are obtained in 500 simulations.

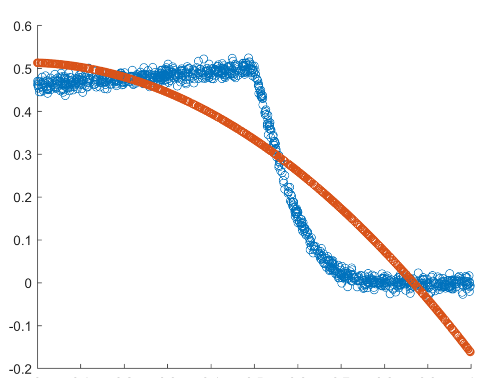
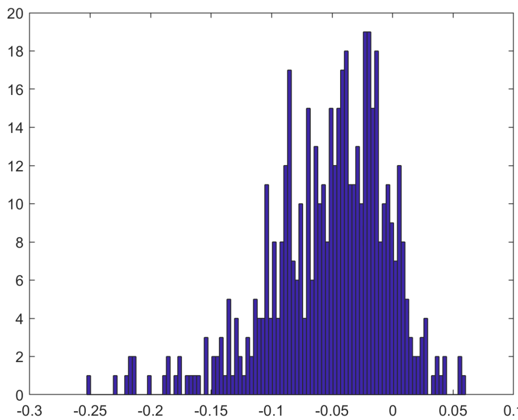
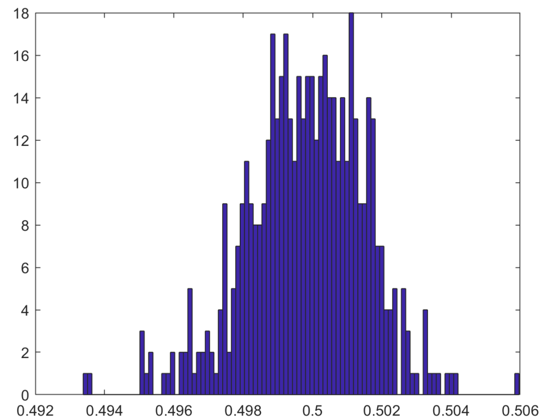


FIGURE 12: Scenario 2.



(a) quadratic specification, $n = 500$.
$mean = 0.0524$, $std = 0.0497$
$50th$, $95th$, $99th$ percentiles:
-0.0449, 0.0116, 0.0413



(b) cubic splines, 8 base splines on each side of the turning point, $n = 500$.
$mean = 0.4998$, $std = 0.0017$
$50th$, $95th$, $99th$ percentiles:
0.4999, 0.5023, 0.5035

FIGURE 13: Histograms and summary statistics of estimated switching points in Scenario 2 using quadratic and then *B-splines* specifications. Results are obtained in 500 simulations.

Figure 13 gives histograms for the estimated turning point in 500 simulations as well the basic summary statistics for those turning points in every sub-case (see the captions), including the quadratic specification subcase in Panel (a) and the *B-spline* specification in Panel (b). When fitting *B-splines*, we connect two pieces on each side of a turning point as to, first, ensure continuity only (consistent with the property of the original function) and, second, to ensure continuity and differentiability at the switch point.B-spline fit also ensures a hump-shaped relation.

B Appendix

Proof of Proposition 1. First of all, note that function $m(\cdot)$ is identified as the regression mean: $m(x) = E[y|x]$. Suppose, contrary to the statement of the proposition that there are two different ordered sequences $s_1 < s_2 < \ldots s_J$ and $\tilde{s}_1 < \tilde{s}_2 < \ldots \tilde{s}_J$ of switch points such that in addition to (6) it holds that

$$m|_{[\tilde{s}_j, \tilde{s}_{j+1}]} \in \mathcal{M}_{j+1}\left([\tilde{s}_j, \tilde{s}_{j+1}]\right), j = 0, \ldots, J, . \tag{42}$$

let j_0 be the minimum index such that $s_{j_0} \neq \tilde{s}_{j_0}$. Without a loss of generality, suppose that $s_{j_0} < \tilde{s}_{j_0}$. Using condition (7), we then have that on $[s_{j_0}, \tilde{s}_{j_0}]$ the regression function $m(\cdot)$ belongs to

$m_{j_0}([s_{j_0}, \tilde{s}_{j_0}])$ (as implied by (42)) and also to $m_{j_0+1}([s_{j_0}, \tilde{s}_{j_0}])$ (as implied by (6)). But according to (8) the intersection $m_{j_0}([s_{j_0}, \tilde{s}_{j_0}]) \cap m_{j_0+1}([s_{j_0}, \tilde{s}_{j_0}])$ is empty, which gives us a contradiction. Thus, we can conclude that the ordered sequence of switch points with the properties given in (6) is unique. ■

Proof of Proposition 2. By Arzela-Ascoli theorem, Θ_0 is relatively compact in the uniform metric. Therefore, its closure $\overline{\Theta}_0$ in the uniform metric is compact. We take $\overline{\Theta}_0$ as our parameter set, and, clearly, $m(x) = E[y|x] \in \overline{\Theta}_0$.

To ensure the compactness of the sample parameter space, as required in the Newey and Powell (2003) (see Appendix B.2), we use the Arzela-Ascoli theorem once again and obtain the relatively compact set by imposing conditions on the parameters in the *B-spline* approximation captured in the following definition of $\widehat{\Theta}$:[13]

$$\widehat{\Theta} = \left\{ m_{\mathcal{B}} \in m_{T_{\{(q_j, L_j)\}_{j=1}^{J+1}}} \;:\; |\beta_{\ell_j,j}| \leq A_1 + \Delta_1,\; \frac{L_j|\beta_{\ell_j+1,j} - \beta_{\ell_j,j}|}{s_j - s_{j-1}} \leq A_2 + \Delta_2,\; \forall \ell_j\; \forall j \right\}$$

for some positive constants $\Delta_1 > 0$ and $\Delta_2 > 0$.

As the sample parameter space, we consider the closure $\overline{\widehat{\Theta}}$ of $\widehat{\Theta}$ in the uniform norm. The proof of this proposition establishes, among other things, that every function from $\overline{\Theta}_0$ can be well approximated asymptotically in the uniform metric by functions from $\overline{\widehat{\Theta}}$.

We prove this consistency result by applying Lemma A.1 from Newey and Powell (2003) (see Appendix B.2). Let us verify all of its conditions. Our population and sample objective functions for the purpose of this proof are, respectively,[14]

$$Q(m(\cdot)) = E[(y - m(x))^2], \quad \widehat{Q}\left(m_{\mathcal{B}}(\cdot; s)\right) = \frac{1}{n} \sum_{i=1}^{N} \left(y - m_{\mathcal{B}}(x_i; s)\right)^2.$$

Condition (i) in Newey and Powell (2003) (Appendix B.2) about $m(\cdot)$ being the unique argmin of Q (up to almost everywhere) in $\overline{\Theta}_0$ follows from the property of the conditional mean as an optimiser and the fact that $m(x) = E[y|x]$ a.e..

For condition (ii) in Newey and Powell (2003) (Appendix B.2), note that both Q and $\widehat{Q}$ are obviously continuous in m and $m_{\mathcal{B}}$, respectively. Let us show that $\sup_{m \in \overline{\Theta}_0} |Q(m) - \widehat{Q}(m)| =$

13. The second condition in this definition is specific to having uniform knots inside each $[s_{j-1}, s_j]$ but could, of course, be easily extended to allow for a different choice of knots.

14. $\widehat{Q}(\cdot)$ is, of course, $\widehat{Q}^*(\cdot)$ rewritten as a function of the approximation itself.

$o_p(1)$. For that, we can use Lemma A.2 in Newey and Powell (2003) (Appendix B.2) and note that for any $\widetilde{m}, \widetilde{\widetilde{m}} \in \overline{\Theta}_0$

$$\left|\widehat{Q}(\widetilde{m}) - \widehat{Q}\left(\widetilde{\widetilde{m}}\right)\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{m}(x_i) - \widetilde{\widetilde{m}}(x_i)\right)\left(2m(x_i) + 2u_i - \widetilde{m}(x_i) - \widetilde{\widetilde{m}}(x_i)\right)\right|$$
$$\leq \sup_{[\underline{x},\overline{x}]}\|\widetilde{m}(x) - \widetilde{\widetilde{m}}(x)\| \cdot \left(4A_1 + \frac{1}{n}\sum_{i=1}^{n}|u_i|\right),$$

and of course, $\frac{1}{n}\sum_{i=1}^{n}|u_i| = O_p(1)$ implied by the assumption that u_i has finite fourth moment. Thus, by Lemma A.2 in Newey and Powell (2003) (see Appendix B.2) we can conclude that

$$\sup_{m\in\overline{\Theta}_0}\left|Q(m) - \widehat{Q}\,(m)\right| = o_p(1). \tag{43}$$

Finally, for condition (iii), we want to show that for every $m \in \overline{\Theta}_0$ there is a sequence of $m_{\mathcal{B}} \in \overline{\widehat{\Theta}}$ such that $\sup_x|m_{\mathcal{B}}(x;s) - m(x)| = o(1)$. Note that Condition C2 automatically implies that for every $m \in \Theta_0$ we can find an approximation $m_{\mathcal{B}} \in m_{T_{\{(q_j,L_j)\}_{j=1}^{J+1}}}$ such that $\sup_x|m_{\mathcal{B}}(x;s) - m(x)| = O\left(\frac{1}{\left(\min_{j=1,\dots,J+1}L_j\right)^r}\right)$ for some $r > 1$, which implies $\sup_x|m_{\mathcal{B}}(x;s) - m(x)| = o(1)$. Let us show that we can take such an approximation $m_{\mathcal{B}}$ to satisfy constraints in the definition of $\widehat{\Theta}$.

First, by *B-spline* properties, $\left|\beta_{\ell_j,j} - \frac{\underline{d}_j+\overline{d}_j}{2}\right| \leq D_{q_j,\infty}\frac{\overline{d}_j-\underline{d}_j}{2}$, where $[\underline{d}_j, \overline{d}_j]$ is the range of values of $m_{\mathcal{B}}(\cdot;s)$ on $[t_{\ell_j+1,j}, t_{\ell_j+q_j-1,j}]$ (see De Boor (1978), p. 133), where $t_{\ell_j,j}$ denotes the ℓ_j's knot on the interval $[s_j, s_{j+1}]$ and $D_{q_j,\infty}$ is a universal constant that does not depend on the system of knots and only depends on the degree of *B-splines* on $[s_j, s_{j+1}]$. Since

$$\left|\overline{d}_j - \underline{d}_j\right| \leq O\left(\frac{1}{L_j^r}\right) + A_2 O\left(\frac{1}{L_j}\right),$$
$$\left|\overline{d}_j + \underline{d}_j\right| \leq 2A_1 + O\left(\frac{1}{L_j^r}\right),$$

then

$$\left|\beta_{\ell_j,j}\right| \leq A_1 + O\left(\frac{1}{L_j}\right) \leq A_1 + \Delta_1$$

for large enough L_j.

Analogously, we can use the same property for the derivative of the *B-spline*. We now have

$$\left|\frac{q_j(\beta_{\ell_j+1,j}-\beta_{\ell_j,j})}{t_{\ell_j+1+q_j,j}-t_{\ell_j+1,j}}-\frac{\underline{e}_j+\overline{e}_j}{2}\right|\le E_{q_j-1,\infty}\frac{\overline{e}_j-\underline{e}_j}{2},$$

where $[\underline{e}_j,\overline{e}_j]$ is the range of values of $m'_{\mathscr{B}}(\cdot)$ on $[t_{\ell_j+1,j},t_{\ell_j+q_j-2}]$ and $E_{q_j,\infty}$ is a universal constant that does not depend on the system of knots. Once can show that

$$|\overline{c}_j-\underline{c}_j|=O\left(\frac{1}{L_j^{r-1}}\right),\quad |\overline{c}_j+\underline{c}_j|\le 2A_2+O\left(\frac{1}{L_j^{r-1}}\right).$$

Since $t_{\ell_j+1+q_j,j}-t_{\ell_j+1,j}$ is proportional to $\frac{1}{L_j}$ ($t_{\ell_j+1+q_j,j}-t_{\ell_j+1,j}$ takes possible values of $\frac{s_j-s_{j-1}}{L_j},2\frac{s_j-s_{j-1}}{L_j},\dots,q_j\frac{s_j-s_{j-1}}{L_j}$), then

$$\frac{L_j|\beta_{\ell_j+1,j}-\beta_{\ell_j,j}|}{s_j-s_{j-1}}\le\left|\frac{q_j(\beta_{\ell_j+1,j}-\beta_{\ell_j,j})}{t_{\ell_j+1+q_j,j}-t_{\ell_j+1,j}}\right|\le A_2+\Delta_2$$

for large enough L_j.

Now it is only left to consider $m\in\overline{\Theta}_0\backslash Int(\Theta_0)$, where $Int(\Theta_0)$ denotes the interior of the set Θ_0. For such m we can always find $\tilde{m}\in Int(\Theta_0)$ such that

$$\sup_x|m(x)-\tilde{m}(x)|\le\frac{K_0}{(\min_{j=1,\dots,J+1}L_j)^r}$$

for some $K_0>0$ (and even a faster rate by the definition of the boundary). Then, according to the discussion above, we can find $m_{\mathscr{B}}\in\overline{\overline{\Theta}}$ such that

$$\sup_x|\tilde{m}(x)-m_{\mathscr{B}}(x;s)|=O\left(\frac{1}{(\min_{j=1,\dots,J+1}L_j)^r}\right),$$

implying thus that $\sup_x|m(x)-m_{\mathscr{B}}(x;s)|=O\left(\frac{\tilde{K}_0}{(\min_{j=1,\dots,J+1}L_j)^r}\right)=o(1)$ as $\min\limits_{j=1,\dots,J+1}L_j\to\infty$. ■

Proof of Corollary 1. Suppose at least one $\hat{s}_j$ is not consistent for s_j. Let j_0 be the smallest index such that $\hat{s}_{j_0}-s_{j_0}\overset{p}{\nrightarrow}0$. This means that there is $\varepsilon_1>0$ and $\varepsilon_2>0$ we have $P(|\hat{s}_j-s_j|>\varepsilon_1)\ge\varepsilon_2$ on a subsequence of $\hat{s}_j$. Without a loss of generality, we can take that $P(\hat{s}_{j_0}<s_{j_0}-\varepsilon_1)\ge\varepsilon_2$. But we then conclude that on the interval $[s_{j_0}-\varepsilon_1,s_{j_0}]$ the subsequence of $\widehat{m}_{\mathscr{B}}(\cdot)$ with a probability bounded away from zero uniformly approximates the property of the

class $\mathcal{M}_{j_0+1}$, which contradicts the fact that the whole sequence $\widehat{m}_{\mathcal{B}}(\cdot)$ on $[s_{j_0} - \varepsilon_1, s_{j_0}]$ converges uniformly in probability to $m(\cdot)$ and on that interval $m(\cdot)$ has property $\mathcal{M}_{j_0}$. Since classes $\mathcal{M}_{j_0}$ and $\mathcal{M}_{j_0+1}$ don't intersect, we obtain a contradiction. Hence, all $\hat{s}_j$ are consistent. ■

Proof of Proposition 3.

The rates of convergence of *B-spline* coefficients and the coefficients of the partial linear model are standard. We focus on the rate of convergence of the switch point estiamtor.

In order to derive the rates we split the estimation procedure into two steps: in the first step we fix s and find the parameters $\hat{\beta}(s), \hat{\gamma}(s)$ which minimize the constrained optimization problem treating the given value of s as a parameter, and in the second step we minimize the redefined objective function $\widehat{Q}(s) = \widehat{Q}^*(s, \hat{\beta}(s), \hat{\gamma}(s))$ with respect to s only. Since $\hat{s}$ minimizes $\widehat{Q}(s)$:

$$\left.\frac{\partial \widehat{Q}(s)}{\partial s'}\right|_{s=\hat{s}} = 0 \tag{44}$$

and by Taylor expansion around the true s^0:

$$0 = \frac{\partial \widehat{Q}(\hat{s})}{\partial s'} = \frac{\partial \widehat{Q}(s^0)}{\partial s'} + \frac{\partial^2 \widehat{Q}(s^0)}{\partial s \partial s'}(\hat{s} - s^0) + o_p\left(\left|\hat{s} - s^0\right|\right). \tag{45}$$

Then:

$$\hat{s} - s^0 = \left(\frac{\partial^2 \widehat{Q}(s^0)}{\partial s \partial s'} + o_p(1)\right)^{-1} \frac{\partial \widehat{Q}(s^0)}{\partial s'}. \tag{46}$$

We show that $\frac{\partial \widehat{Q}(s^0)}{\partial s'}$ is $O_p\left(\frac{1}{\sqrt{n}}\right)$ and $\frac{\partial^2 \widehat{Q}(s^0)}{\partial s \partial s'} \xrightarrow{p} R$ for some invertible matrix R, hence the rate of convergence of $\hat{s}$ to s^0 is $O_p\left(\frac{1}{\sqrt{n}}\right)$[15].

Let

$$\mathcal{L}(s, \beta, \gamma, \lambda) = \widehat{Q}^*(s, \beta, \gamma) + \lambda g(\beta)$$

be the Lagrangian of the constrained minimisation problem, where the inequality constraints are listed as $g(\beta) \geq 0$ and λ are the corresponding Lagrange multipliers. By the envelope theorem:

$$\frac{\partial \widehat{Q}(s)}{\partial s'} = \frac{\partial \mathcal{L}(\hat{\beta}(s), \hat{\gamma}(s), s, \hat{\lambda}(s), \hat{\mu}(s))}{\partial s'} = \frac{\partial \widehat{Q}^*(\hat{\beta}(s), \hat{\gamma}(s), s)}{\partial s'},$$

15. A faster rate of convergence can be achieved in the case where the estimated function has a discontinuity (or discontinuity in a derivative) at the switch point, see e.g.Muller (1992). In applications where the researcher has knowledge of these kinds of changes in behavior at $\hat{s}$ alternative methods of estimation can be used to find the switch point before estimating the remaining parameters.

so to find the first derivative we only need to differentiate $\widehat{Q}^*(s,\beta,\gamma)$ directly with respect to s and evaluate at $\hat{\beta}(s), \hat{\gamma}(s)$. For any $j \in \{1, 2, \ldots, J\}$:

$$\begin{aligned}
\frac{\partial \widehat{Q}(s^0)}{\partial s_j} &= \frac{1}{n}\sum_{i=1}^{n} -2\left(y_i - \widehat{m}_{\mathcal{B}}(x_i; s^0) - \hat{\gamma}'\left(s^0\right) z_i\right) \frac{\partial \widehat{m}_{\mathcal{B}}(x_i; s^0)}{\partial s_j} \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n} -2\left(m(x_i) - \widehat{m}_{\mathcal{B}}(x_i; s^0) + \left(\gamma - \hat{\gamma}\left(s^0\right)\right)' z_i + u_i\right) \underbrace{\frac{\partial \widehat{m}_{\mathcal{B}}(x_i; s^0)}{\partial s_j}}_{=O_p(1)}}_{O_p\left(\frac{1}{\sqrt{n}}\right)} \\
&= O_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

We now justify the rates listed above.

Using derivation in Lemma 2 and Lemma 3, and remembering that due to the envelope theorem we only take the derivative with respect to the *B-spline* basis functions and not with respect to $\hat{\beta}(s)$:

$$\begin{aligned}
&\frac{\partial \widehat{m}_{\mathcal{B}}(x; s^0)}{\partial s_j} \\
&\quad = \frac{\partial \widehat{m}_{\mathcal{B}}(x; s^0)}{\partial x}\left(\left(\frac{s^0_{j-1} - x}{s^0_j - s^0_{j-1}}\right) \mathbb{1}\left(x \in [s^0_{j-1}, s^0_j)\right) + \left(\frac{x - s^0_{j+1}}{s^0_{j+1} - s^0_j}\right) \mathbb{1}\left(x \in [s^0_j, s^0_{j+1})\right)\right) \\
&\quad = O_p(1).
\end{aligned}$$

This term is stochastically bounded because the derivative of the spline function with respect to x is bounded (we allow coefficients $\hat{\beta}(s)$ from a space $\widehat{\Theta}$ which imposes a common bound of $A_2 + \Delta_2 < \infty$ on the derivative of $m_{\mathcal{B}}(\cdot, s^0)$ across all n and all possible values of x and the ratios $\left(\frac{s^0_{j-1}-x}{s^0_j - s^0_{j-1}}\right) \mathbb{1}\left(x \in [s^0_{j-1}, s^0_j)\right)$ and $\left(\frac{x - s^0_{j+1}}{s^0_{j+1} - s^0_j}\right) \mathbb{1}\left(x \in [s^0_j, s^0_{j+1})\right)$ are in $[0, 1]$.

The fact that $\frac{1}{n}\sum_{i=1}^{n} m(x_i) - \widehat{m}_{\mathcal{B}}(x_i; s^0) = O_p\left(\frac{1}{\sqrt{n}}\right)$ follows from Lemma 5, in the special case where we take $s = s^0$ (this removes the approximation error due to using incorrect switch point).

Finally, $\frac{1}{n}\sum_{i=1}^{n} \left(\gamma - \hat{\gamma}\left(s^0\right)\right)' z_i + u_i = O_p\left(\frac{1}{\sqrt{n}}\right)$ by standard results for rates of convergence of the linear part of a partly linear model (e.g. Robinson (1988), this could also be shown directly by the same arguments as in Lemma 5) and of i.i.d. random variables with bounded second moments (e.g. Lindeberg-Levy CLT).

To find the expression for the second derivative of the objective function we introduce the shorthand notation for the residual:

$$\hat{\varepsilon}_i \equiv y_i - \widehat{m}_{\mathcal{B}}(x_i; s) - \hat{\gamma}' z_i. \tag{47}$$

We have $\widehat{Q}(s) = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2$ and $\frac{\partial \widehat{Q}(s)}{\partial s'} = \frac{2}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i \frac{\partial \hat{\varepsilon}_i}{\partial s'}$, hence

$$\frac{\partial^2 \widehat{Q}(s)}{\partial s \partial s'} = \frac{2}{n}\sum_{i=1}^{n} \frac{\partial \hat{\varepsilon}_i}{\partial s'}\frac{\partial \hat{\varepsilon}_i}{\partial s} + \hat{\varepsilon}_i \frac{\partial^2 \hat{\varepsilon}_i}{\partial s \partial s'}$$

Let $m(x_i; s)$ denote the function providing the closest approximation to $m(x_i)$ which has a switch point at s. Assuming that $m(x_i; s)$ is smooth and has bounded second derivatives, $\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i \frac{\partial^2 \hat{\varepsilon}_i}{\partial s \partial s'} \xrightarrow{p} E\left(\varepsilon_i \frac{\partial^2 m(x_i;s) + \gamma' z_i}{\partial s \partial s'}\right) = E\left(E\left(\varepsilon_i | \, x_i, z_i\right) \frac{\partial^2 m(x_i;s) + \gamma' z_i}{\partial s \partial s'}\right) = 0$, i.e. the second term is negligible in the limit. The first term converges in probability to $R \equiv E\left(\frac{\partial m(x_i;s) + \gamma' z_i}{\partial s'} \frac{\partial m(x_i;s) + \gamma' z_i}{\partial s}\right)$. As argued in Lemma 4, a sufficient condition for R to be invertible is that the elements of the $\frac{\partial m(x_i;s) + \gamma' z_i}{\partial s'}$ are not linearly dependent. Linear dependence between the elements of the $\frac{\partial m(x_i;s) + \gamma' z_i}{\partial s'}$ would mean that worsening of fit due to changing some of the switch point locations could be perfectly corrected by adjusting other switch points or terms linear in z_i. This cannot be the case because different switch points affect different parts of the domain of $m(x)$ and under the assumption of no perfect multicollinearity we cannot perfectly substitute between fitting $m_{\mathcal{B}}(x_i)$ and $\gamma' z_i$ so adjusting γ cannot fully correct the overall fit.

It is also worth observing that the elements of the vector are non-zero: if they were, it would mean that the choice of s_j doesn't affect the quality of the best fitted function of x_i and z_i explaining y_i. This would contradict the assumption that there is no intersection between the neighboring classes $\mathcal{M}_j$ and no perfect multicollinearity between a function of x_i and the linear term in z_i. Forcing the fitted function to follow a different property to the true function on some interval must result in a worse fit. Some of the worsening in the fit could be accounted for by adjusting γ, but given the assumption of no perfect multicollinearity we cannot perfectly substitute between fitting $m_{\mathcal{B}}(x_i)$ and $\gamma' z_i$. This proves that R is almost surely invertible.

R is non-zero over a region in which we impose an incorrect constraint. As s approaches s^0 it becomes smaller than $\frac{1}{L}$, yet we are still imposing the constraints based on *B-splines*, and any incorrectly imposed constraint will worsen the fit over a region proportional to $\frac{1}{L}$, hence $R = O\left(\frac{1}{L}\right)$.

Finally, by applying the above results to equation (46):

$$|\hat{s}-s^0| = (R+o_p(1))^{-1}\, O_p\left(\frac{1}{\sqrt{n}}\right)$$

hence

$$|\hat{s}-s^0| = O_p\left(\frac{L}{\sqrt{n}}\right).$$

■

Proof of Proposition 4

This follows from the discussion in the main text, in Section 5.2.3. ■

Proof of Proposition 5

We wish to show that, after the transformation, the following term is small (i.e. $o_p\left(\frac{1}{\sqrt{n}}\right)$):

$$T_1 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x)\left(m(x_i) - m_{\mathcal{B}}(x_i;\hat{s})\right).$$

The transformation removes the terms linear in $\frac{\partial m_{\mathcal{B}}(x_i;\hat{s})}{\partial s_k}$.

By Taylor expansion, we have:

$$m_{\mathcal{B}}(x_i;\hat{s}) = m_{\mathcal{B}}(x_i;s^0) + \frac{\partial m_{\mathcal{B}}(x_i;\tilde{s})}{\partial s}(s^0-\hat{s})$$

for $\tilde{s}$ between $\hat{s}$ and s^0 (element-wise). The term inside the average in T_1 can be written as:

$$\begin{aligned}
m(x_i) - m_{\mathcal{B}}(x_i;\hat{s}) &= \underbrace{m(x_i) - m_{\mathcal{B}}(x_i;s^0)}_{=O(L^{-r})} + m_{\mathcal{B}}(x_i;s^0) - m_{\mathcal{B}}(x_i;\hat{s}) \\
&= O\left(L^{-r}\right) + \frac{\partial m_{\mathcal{B}}(x_i;\tilde{s})}{\partial s}(\hat{s}-s^0) \\
&= O\left(L^{-r}\right) + \frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial s}(\hat{s}-s^0) \\
&\quad + \Bigg(\underbrace{\frac{\partial m_{\mathcal{B}}(x_i;\tilde{s})}{\partial s} - \frac{\partial m_{\mathcal{B}}(x_i;\hat{s})}{\partial s}}_{=(\hat{s}-\tilde{s})'\frac{\partial^2 m_{\mathcal{B}}(x_i;\check{s})}{\partial s \partial s'} = O(\|\hat{s}-s^0\|_\infty)} + \underbrace{\frac{\partial m_{\mathcal{B}}(x_i;\hat{s})}{\partial s} - \frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial s}}_{=O\left(L^{-(r-1)}\right)}\Bigg)(\hat{s}-s^0) \\
&= O\left(L^{-r} + L^{-(r-1)}\left\|\hat{s}-s^0\right\|_\infty + \left\|\hat{s}-s^0\right\|_\infty^2\right) + \frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial s}(\hat{s}-s^0)
\end{aligned}$$

We have used the fact that the best *B-spline* approximation of kth derivative of r-times differentiable function is within $O\left(L^{k-r}\right)$ of the approximated function. We also rely on Taylor expansion (of the *B-spline* itself, as shown above, and of its derivative), where $\check{s}$ is another vector between $\hat{s}$ and s^0 (element-wise).

The final term is linear in $\frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial s}$ and gets removed by the Khmaladze transformation.

The first part is a common upper bound over all x_i: the second derivative of the *B-spline* with respect to the switch points is bounded over the whole domain of x_i, the bound on the fit is also taken uniformly over the whole domain of x_i. Given Proposition 3 and Condition C3:

$$\begin{aligned} L^{-r} + L^{-(r-1)} \left\|\hat{s} - s^0\right\|_\infty + \left\|\hat{s} - s^0\right\|_\infty^2 &= O_p\left(L^{-r} + \frac{1}{L^{r-2}\sqrt{n}} + \frac{L^2}{n}\right) \\ &= o_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned}$$

These terms are small before the transformation, and the transformation takes the form of a projection which can only make the terms smaller. Hence, the transformed term goes to zero faster than $\frac{1}{\sqrt{n}}$. ■

Proof of Proposition 6.

The proof follows the same steps as the proof of Theorem 1 in Komarova and Hidalgo (2023) and is omitted. The main differences are

1. we add regressors of the form z_k and $\frac{\partial \widehat{m}_{\mathcal{B}}(x_k;\hat{s})}{\partial s_l}$. The first type takes non-zero values over the whole domain, the second over $(\hat{s_{l-1}}, \hat{s_{l+1}})$. Both of these regions remain bounded away from zero as sample size increases (unlike the basis functions which have support proportional to $\frac{1}{L}$ that goes to zero, causing issues with eigenvalues of the matrix we use in the transformation). This addition doesn't cause any complications.

2. we use regressors based on the estimates $\hat{\beta}, \hat{s}$ derived from the whole sample. This is again not an issue because they are consistent for the true values β_0, s^0: we can show that the limiting behavior is the same when we use estimates as if we use d the true values.

■

Proof of Theorem 1.

The first statement follows straight from Proposition 4-6 and continuous mapping theorem. The second statement follows by the same arguments as in Proposition 1 in Komarova and Hidalgo (2023). ■

B.1 Proofs of supporting results

Lemma 1. B-splines *are continuous in the switch point almost everywhere.*

Proof. We want to analyse continuity of P_i in s. Each element of the vector P_i is of the form $p_{\ell,L_j,[s_{j-1},s_j],q}(x_i)$. The value of $p_{\ell,L_j,[s_{j-1},s_j],q}(x)$ does not depend on s_k for $k \notin \{j-1, j\}$, hence $p_{\ell,L_j,[s_{j-1},s_j],q}(x)$ is continuous in s_k for $k \notin \{j-1, j\}$.

To show continuity of $p_{\ell,L_j,[s_{j-1},s_j],q}(x)$ in s_{j-1} at $s_{j-1} = s$ we want to show that for almost all $x \in [0,1]$: $\lim_{\tilde{s} \longrightarrow s} p_{\ell,L_j,[\tilde{s},s_j],q}(x) = p_{\ell,L_j,[s,s_j],q}(x)$. We use the fact that *B-splines* are invariant under a translation and scaling of the knot sequence (see the result from e.g. Lyche, Manni, and Speleers (2017) restated in Lemma 7). $p_{\ell,L_j,[s_{j-1},s_j],q}(x_i)$ is defined on the knot sequence

$$t^{[s_{j-1},s_j],L'_j,q} = \left(\underbrace{s_{j-1}, \ldots, s_{j-1}}_{q+1 \text{ times}}, s_{j-1} + \frac{s_j - s_{j-1}}{L'_j}, s_{j-1} + 2\frac{s_j - s_{j-1}}{L'_j}, \ldots, \underbrace{s_j, \ldots, s_j}_{q+1 \text{ times}} \right).$$

and moving from $s_{j-1} = \tilde{s}$ to $s_{j-1} = s$ is equivalent to scaling by $\frac{s_j - s}{s_j - \tilde{s}}$ and shifting by $-\frac{(\tilde{s}-s)s_j}{s_j - \tilde{s}}$: $t^{[s,s_j],L'_j,q} = \left(\frac{s_j - s}{s_j - \tilde{s}}\right) t^{[\tilde{s},s_j],L'_j,q} - \frac{(\tilde{s}-s)s_j}{s_j - \tilde{s}}$. Hence for $x \in [0,1] \setminus \{s, s_j\}$:

$$\lim_{\tilde{s} \longrightarrow s} p_{\ell,L_j,[\tilde{s},s_j],q}(x) = \lim_{\tilde{s} \longrightarrow s} p_{\ell,L_j,[s,s_j],q} \left(\left(\frac{s_j - s}{s_j - \tilde{s}} \right) x - \frac{(\tilde{s} - s)\, s_j}{s_j - \tilde{s}} \right) = p_{\ell,L_j,[s,s_j],q}\,(x)$$

by continuity of $p_{\ell,L_j,[s,s_j],q}\,(x)$ in x on $x \in [0,1] \setminus \{s, s_j\}$ and the fact that regardless of the sequence of $\tilde{s}$ the points $\left(\frac{s_j - s}{s_j - \tilde{s}}\right) x - \frac{(\tilde{s}-s)s_j}{s_j - \tilde{s}}$ will eventually fall in (s, s_j) if $x \in (s, s_j)$ or in $[0, s) \cup (s_j, 1]$ if $x \in [0, s) \cup (s_j, 1]$ (where the *B-spline* is identically equal to zero).

Similarly, for continuity in s_j at $s_j = s$, we have $t^{[s_{j-1},s],L'_j,q} = \left(\frac{s - s_{j-1}}{\tilde{s} - s_{j-1}}\right) t^{[s_{j-1},\tilde{s}],L'_j,q} - \frac{(\tilde{s}-s)s_{j-1}}{\tilde{s} - s_{j-1}}$. Hence for $x \in [0,1] \setminus \{s_{j-1}, s\}$:

$$\lim_{\tilde{s} \longrightarrow s} p_{\ell,L_j,[s_{j-1},\tilde{s}],q}(x) = \lim_{\tilde{s} \longrightarrow s} p_{\ell,L_j,[s_{j-1},s],q} \left(\left(\frac{s - s_{j-1}}{\tilde{s} - s_{j-1}} \right) x - \frac{(\tilde{s} - s)\, s_{j-1}}{\tilde{s} - s_{j-1}} \right) = p_{\ell,L_j,[s_{j-1},s],q}\,(x)$$

by continuity of $p_{\ell,L_j,[s_{j-1},s],q}(x)$ in x on $x \in [0,1] \setminus \{s_{j-1}, s\}$ and the fact that regardless of the sequence of $\tilde{s}$ the points $\left(\frac{s_j - s}{s_j - \tilde{s}}\right) x - \frac{(\tilde{s}-s)s_j}{s_j - \tilde{s}}$ will eventually fall in (s_{j-1}, s) if $x \in (s_{j-1}, s)$ or in $[0, s_{j-1}) \cup (s, 1]$ if $x \in [0, s_{j-1}) \cup (s, 1]$.

We have shown that the individual elements of P_i are continuous in s on almost all x. The only potential points of discontinuity are $x = s_j$, but this is not a problem given that we are interested in $\beta' P_i$ and the β coefficients are constrained to give continuity at $x = s_j$ (only the last *B-spline* on $[s_{j-1}, s_j]$ and the first on $[s_j, s_{j-1}]$ have a discontinuity at $x = s_j$, they both take the value of 1 at that point, and we constrain their corresponding coefficients to be equal: $\beta_{L_j,j} = \beta_{1,j+1}$). □

Lemma 2. *The first derivative of a* B-spline *basis function $p_{\ell,L_j,[s_{j-1},s_j],q}(x)$ with respect to s_k is:*

$$\begin{aligned}
&\frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q}(x)}{\partial s_k} = \\
&= \frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q}(x)}{\partial x}\left(\left(\frac{s_{k-1}-x}{s_k - s_{k-1}}\right)\mathbb{1}\left(x \in [s_{k-1}, s_k)\right) + \left(\frac{x - s_{k+1}}{s_{k+1}-s_k}\right)\mathbb{1}\left(x \in [s_k, s_{k+1})\right)\right) \qquad (48)\\
&= q\left(\frac{p_{\ell,L_j,[s_{j-1},s_j],q-1}(x)}{t_{\ell+q}^{[s_{j-1},s_j],L_j',q} - t_{\ell}^{[s_{j-1},s_j],L_j',q}} - \frac{p_{\ell+1,L_j,[s_{j-1},s_j],q-1}(x)}{t_{\ell+q+1}^{[s_{j-1},s_j],L_j',q} - t_{\ell+1}^{[s_{j-1},s_j],L_j',q}}\right)\times \\
&\quad \times\left(\left(\frac{s_{k-1}-x}{s_k - s_{k-1}}\right)\mathbb{1}\left(x \in [s_{k-1}, s_k)\right) + \left(\frac{x - s_{k+1}}{s_{k+1}-s_k}\right)\mathbb{1}\left(x \in [s_k, s_{k+1})\right)\right).
\end{aligned}$$

Proof. Let

$$t^{[0,1],K,q} = \left(\underbrace{0,\ldots,0}_{q+1 \text{ times}}, \frac{1}{K}, \frac{2}{K}, \ldots, \underbrace{1,\ldots,1}_{q+1 \text{ times}}\right) \qquad (49)$$

be the set of knots on $[0,1]$ with K equally spaced intervals and endpoints repeated $q+1$ times. The degree q *B-splines* defined on this set of knots are $\left\{p_{\ell,K+q,[0,1],q}(x)\right\}_{\ell=1}^{K+q}$. Let us consider a set of knots t which can be written as

$$t^{[0,1],K,q} = \alpha(s)t - \beta(s)$$

with the corresponding set of degree q *B-splines* $\{p_{\ell,t,q}(x)\}_{\ell=1}^{K+q}$. By the invariance of *B-splines* to translation/scaling (see Lemma 7), for any x in the support of t:

$$p_{\ell,t,q}(x) = p_{\ell,K,[0,1],q}(\alpha(s)x + \beta(s))$$

where by construction $\alpha(s)x + \beta(s)$ is in $[0,1]$, the support of $t^{[0,1],K,q}$. Then for any s_k and any

x in the support of t:

$$\begin{aligned}
\frac{\partial p_{\ell,t,q}(x)}{\partial s_k} &= \frac{\partial p_{\ell,K,[0,1],q}\left(\alpha(s)x+\beta(s)\right)}{\partial s_k} \\
&= \frac{\partial p_{\ell,K,[0,1],q}\left(y\right)}{\partial y}\Big|_{y=\alpha(s)x+\beta(s)} \frac{\partial(\alpha(s)x+\beta(s))}{\partial s_k} \\
&= q\left(\frac{p_{\ell,K,[0,1],q-1}\left(\alpha(s)x+\beta(s)\right)}{t^{[0,1],K,q}_{\ell+q}-t^{[0,1],K,q}_{\ell}} - \frac{p_{\ell+1,K,[0,1],q-1}\left(\alpha(s)x+\beta(s)\right)}{t^{[0,1],K,q}_{\ell+q+1}-t^{[0,1],K,q}_{\ell+1}}\right)\left(\frac{\partial\alpha(s)}{\partial s_k}x+\frac{\partial\beta(s)}{\partial s_k}\right) \\
&= q\left(\frac{p_{\ell,K,[0,1],q-1}\left(\alpha(s)x+\beta(s)\right)}{\alpha(s)t_{\ell+q}+\beta(s)-\alpha(s)t_\ell-\beta(s)} - \frac{p_{\ell+1,K,[0,1],q-1}\left(\alpha(s)x+\beta(s)\right)}{\alpha(s)t_{\ell+q+1}+\beta(s)-\alpha(s)t_{\ell+1}-\beta(s)}\right)\times \\
&\quad\times\left(\frac{\partial\alpha(s)}{\partial s_k}x+\frac{\partial\beta(s)}{\partial s_k}\right) \\
&= q\left(\frac{p_{\ell,t,q-1}\left(x\right)}{t_{\ell+q}-t_\ell} - \frac{p_{\ell+1,t,q-1}\left(x\right)}{t_{\ell+q+1}-t_{\ell+1}}\right)\frac{1}{\alpha(s)}\left(\frac{\partial\alpha(s)}{\partial s_k}x+\frac{\partial\beta(s)}{\partial s_k}\right) \\
&= \frac{\partial p_{\ell,t,q}(x)}{\partial x}\frac{1}{\alpha(s)}\left(\frac{\partial\alpha(s)}{\partial s_k}x+\frac{\partial\beta(s)}{\partial s_k}\right).
\end{aligned}$$

For the set of knots defined on $[s_{j-1}, s_j]$:

$$t^{[s_{j-1},s_j],L'_j,q} = \left(\underbrace{s_{j-1},\ldots,s_{j-1}}_{q+1 \text{ times}}, s_{j-1}+\frac{s_j-s_{j-1}}{L'_j}, s_{j-1}+2\frac{s_j-s_{j-1}}{L'_j},\ldots,\underbrace{s_j,\ldots,s_j}_{q+1 \text{ times}}\right).$$

we can write

$$t^{[0,1],L'_j,q} = \frac{1}{s_j-s_{j-1}}t^{[s_{j-1},s_j],L'_j,q} - \frac{s_{j-1}}{s_j-s_{j-1}},$$

i.e. $\alpha(s) = \frac{1}{s_j-s_{j-1}}$ and $\beta(s) = -\frac{s_{j-1}}{s_j-s_{j-1}}$. Then:

$$\frac{1}{\alpha(s)}\left(\frac{\partial\alpha(s)}{\partial s_k}x+\frac{\partial\beta(s)}{\partial s_k}\right) = \begin{cases} \frac{x-s_j}{s_j-s_{j-1}} & \text{if } s_k = s_{j-1} \\ \frac{s_{j-1}-x}{s_j-s_{j-1}} & \text{if } s_k = s_j \\ 0 & \text{if } s_k \notin \{s_{j-1}, s_j\}. \end{cases}$$

Since for $x \in [s_{k-1}, s_k]$ only the $p_{\ell,L_j,[s_{k-1},s_k],q}(x)$ take non-zero values, for $x \in [0,1] \setminus$

$\{s_1, s_2, \ldots, s_J\}$:

$$\frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q}(x)}{\partial s_k} = \frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q}(x)}{\partial x}\left(\left(\frac{s_{k-1}-x}{s_k-s_{k-1}}\right)\mathbb{1}\left(x\in[s_{k-1},s_k)\right)+\left(\frac{x-s_{k+1}}{s_{k+1}-s_k}\right)\mathbb{1}\left(x\in[s_k,s_{k+1})\right)\right).$$

Note that $\left(\frac{s_{k-1}-x}{s_k-s_{k-1}}\right)\mathbb{1}\left(x\in[s_{k-1},s_k)\right)+\left(\frac{x-s_{k+1}}{s_{k+1}-s_k}\right)\mathbb{1}\left(x\in[s_k,s_{k+1})\right)$ is continuous for all $x\in[0,1]$, but due to potential discontinuity in $\frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q}(x)}{\partial x}$ at the switch points we need to rule out $x\in\{s_1, s_2, \ldots, s_J\}$.

While the derivatives of specific *B-spline* basis functions may be discontinuous at switch points, this is not a problem in our setting because of the continuity and smoothness constraints which ensure that the derivative with respect to x of the constrained *B-spline* $m_{\mathcal{B}}(x;s) \equiv \sum_{j=1}^{J}\sum_{\ell=1}^{L_j}\beta_{\ell,j}p_{\ell,L_j,[s_{j-1},s_j],q_j}(x)$ at $x=s_k$ is well-defined and continuous. □

Lemma 3. *$\widehat{Q}(\theta)$ is continuously differentiable.*

Proof. The derivatives with respect to β_ℓ and γ_k are clearly continuous:

$$\frac{\partial\widehat{Q}(\theta)}{\partial\beta_{\ell,j}} = \frac{1}{n}\sum_{i=1}^{n} -2\left(y_i - m_{\mathcal{B}}(x_i) - \gamma' z_i\right) p_{\ell,L_j,[s_{j-1},s_j],q_j}(x_i)$$

$$\frac{\partial\widehat{Q}(\theta)}{\partial\gamma_k} = \frac{1}{n}\sum_{i=1}^{n} -2\left(y_i - m_{\mathcal{B}}(x_i) - \gamma' z_i\right) z_{ik}$$

as the first bracket is continuous in β, s (see Lemma 1) and γ, and $p_{\ell,L_j,[s_{j-1},s_j],q_j}(x_i)$ and z_{ik} are constant. The derivative with respect to s is a bit more involved, but using Lemma 2 we can show that it is:

$$
\begin{aligned}
\frac{\partial \widehat{Q}(\theta)}{\partial s_k} &= \frac{1}{n}\sum_{i=1}^{n} -2\left(y_i - m_{\mathcal{B}}(x_i) - \gamma' z_i\right)\frac{\partial m_{\mathcal{B}}(x_i)}{\partial s_k} \\
&= \frac{1}{n}\sum_{i=1}^{n} -2\left(y_i - m_{\mathcal{B}}(x_i) - \gamma' z_i\right)\left(\sum_{j=1}^{J}\sum_{\ell=1}^{L_j}\beta_{\ell,j}\frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q_j}(x_i)}{\partial s_k}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} -2\left(y_i - m_{\mathcal{B}}(x_i) - \gamma' z_i\right)\left(\sum_{j=1}^{J}\sum_{\ell=1}^{L_j}\beta_{\ell,j}\frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q_j}(x_i)}{\partial x}\times\right. \\
&\qquad \left.\times\underbrace{\left(\left(\frac{s_{k-1}-x}{s_k - s_{k-1}}\right)\mathbb{1}\left(x\in[s_{k-1},s_k)\right) + \left(\frac{x - s_{k+1}}{s_{k+1}-s_k}\right)\mathbb{1}\left(x\in[s_k,s_{k+1})\right)\right)}_{\equiv A_{s_k}(x_i)}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} -2\left(y_i - m_{\mathcal{B}}(x_i) - \gamma' z_i\right)\sum_{j=1}^{J}\sum_{\ell=1}^{L_j}\beta_{\ell,j}\frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q_j}(x_i)}{\partial x}A_{s_k}(x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} -2\left(y_i - m_{\mathcal{B}}(x_i) - \gamma' z_i\right)\frac{\partial m_{\mathcal{B}}(x_i)}{\partial x}A_{s_k}(x_i).
\end{aligned}
$$

$m_{\mathcal{B}}$ is continuously differentiable in x (by properties of spline functions and by the assumption of smoothness at the minimum), hence $\frac{\partial m_{\mathcal{B}}(x_i)}{\partial x}$ is well-defined for all x_i, and $A_{s_k}(x_i)$ is continuous in both s and $x \in [0,1]$. Hence the derivatives with respect to all inputs are continuous. □

Lemma 4. *Let x be a k-dimensional vector of random variables. The matrix $E\left(xx'\right)$ is invertible if and only if the elements of x are not linearly dependent, i.e. there does not exist a constant vector $v \in \mathbb{R}^k \setminus \{0\}$ such that $x'v = 0$ a.s..*

Proof. For necessity, suppose $\exists v \in \mathbb{R}^k$ such that $v \neq 0$ and $x'v = 0$ *a.s.*. Then with probability one

$$0 = E\left(xx'v\right) = E\left(xx'\right)v$$

for $v \neq 0$, i.e. $rank\left(E\left(xx'\right)\right) < k$ and $E\left(xx'\right)$ is not invertible.

For sufficiency, suppose $E\left(xx'\right)$ is not invertible. Then there must exist a constant vector $v \in \mathbb{R}^k \setminus \{0\}$ such that $E\left(xx'\right)v = 0$. Then we also have

$$0 = v'0 = v'E\left(xx'\right)v = E\left(v'xx'v\right) = E\left(\left(v'x\right)^2\right)$$

which implies that $v'x = 0$ $a.s.$. $\square$

Lemma 5. $\frac{1}{n}\sum_{i=1}^{n}\widehat{m}_{\mathcal{B}}(x_i;s^0) - m(x_i) = O_p\left(\frac{1}{\sqrt{n}}\right)$.

Proof. This is stated for the case without additional covariates and when we use the true switch point s^0. The argument is identical if we look at the version where instead of $\widehat{m}_{\mathcal{B}}(x_i;s^0)$ we use $\widehat{m}_{\mathcal{B}}(x_i;s^0) + \widehat{\gamma}'z_i$ and instead of $m(x_i)$ we use $m(x_i) + \gamma' z_i$.

Let P denote the matrix of effective *B-splines* (i.e. after imposing all binding constraints) based on switch points s^0 and evaluated at all points $\{x_i\}_{i=1}^n$. Let m, $m_{\mathcal{B}}\left(s^0\right)$ and $\widehat{m}_{\mathcal{B}}\left(s^0\right)$ denote the vectors of the three functions evaluated at all points $\{x_i\}_{i=1}^n$, and let $\hat{\beta}$ and β_0 be vectors of coefficients such that $\widehat{m}_{\mathcal{B}}\left(s^0\right) = P\hat{\beta}$ and $m_{\mathcal{B}}\left(s^0\right) = P\beta_0$.

The term of interest is:

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{m}_{\mathcal{B}}(x_i;s^0) - m(x_i) = \frac{1}{n}\iota'(\widehat{m}_{\mathcal{B}} - m)$$

where ι is a vector of n 1s.

We use the property[16] that for a scalar random variable X_n:

$$X_n - E(X_n) = O_p\left(\sqrt{V\left(X_n\right)}\right).$$

For $X_n = \frac{1}{n}\sum_{i=1}^{n}\widehat{m}_{\mathcal{B}}(x_i;s^0) - m(x_i)$ we start by looking at the expectation. We firstly find an expression for $\widehat{m}_{\mathcal{B}}\left(s^0\right) - m_{\mathcal{B}}\left(s^0\right)$:

$$\begin{aligned}
\widehat{m}_{\mathcal{B}}\left(s^0\right) = P\hat{\beta} &= P(P'P)^{+}P'(m+u) \\
&= P(P'P)^{+}P'(P\beta_0 + m - P\beta_0 + u) \\
&= P\beta_0 + P(P'P)^{+}P'(m - P\beta_0 + u) \\
&= m_{\mathcal{B}}\left(s^0\right) + P(P'P)^{+}P'(m - m_{\mathcal{B}}\left(s^0\right) + u).
\end{aligned}$$

Each element of the $m - m_{\mathcal{B}}\left(s^0\right)$ vector is bounded above by $\|m(x_i) - m_{\mathcal{B}}(x_i;s^0)\|_\infty = O\left(L^{-r}\right)$,

16. This follows from Markov's inequality: $\forall\varepsilon > 0$ there exists $C = C(\varepsilon) > 0$ such that

$$P\left(|X_n| > C\sqrt{E\left(X_n^2\right)}\right) \leq \frac{E\left(X_n^2\right)}{C^2 E\left(X_n^2\right)} = C^{-2} < \varepsilon.$$

hence we have:

$$\begin{aligned}
\frac{1}{n}\iota'(m_{\mathscr{B}}\left(s^0\right) - m) &= \frac{1}{n}\sum_{i=1}^{n} m_{\mathscr{B}}(x_i; s^0) - m(x_i) \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \|m(x_i) - m_{\mathscr{B}}(x_i; s^0)\|_\infty \\
&= \|m(x_i) - m_{\mathscr{B}}(x_i; s^0)\|_\infty = O\left(L^{-r}\right).
\end{aligned}$$

We can find an upper bound on the length of the vector $m - m_{\mathscr{B}}\left(s^0\right)$ as:

$$\begin{aligned}
\|m - m_{\mathscr{B}}\left(s^0\right)\| &= \sqrt{\sum_{i=1}^{n}(m(x_i) - m_{\mathscr{B}}(x_i; s^0))^2} \\
&\leq \sqrt{n\|m(x_i) - m_{\mathscr{B}}(x_i; s^0)\|_\infty^2} \\
&= \sqrt{n}\|m(x_i) - m_{\mathscr{B}}(x_i; s^0)\|_\infty
\end{aligned}$$

It follows that:

$$\begin{aligned}
\frac{1}{n}\iota' P(P'P)^{-1}P'(m - m_{\mathscr{B}}\left(s^0\right)) &\leq \frac{1}{n}\|\iota\|\|P(P'P)^{-1}P'(m - m_{\mathscr{B}}\left(s^0\right))\| \\
&\leq \frac{1}{n}\sqrt{n}\|m - m_{\mathscr{B}}\left(s^0\right)\| \\
&\leq \frac{1}{n}\sqrt{n}\sqrt{n}\|m(x_i) - m_{\mathscr{B}}(x_i; s^0)\|_\infty \\
&= \|m(x_i) - m_{\mathscr{B}}(x_i; s^0)\|_\infty = O\left(L^{-r}\right).
\end{aligned}$$

The first inequality is by the Cauchy-Schwarz inequality and the second comes from the fact that projecting a matrix can only make it shorter.

Combining all of these facts, we have:

$$\begin{aligned}
E\left(\frac{1}{n}\iota'(\widehat{m}_{\mathcal{B}}\left(s^0\right)-m)\middle| X\right) &= \frac{1}{n}\iota' E(\widehat{m}_{\mathcal{B}}\left(s^0\right)-m_{\mathcal{B}}\left(s^0\right)|X) + \underbrace{\frac{1}{n}\iota'\left(m_{\mathcal{B}}\left(s^0\right)-m\right)}_{=O(L^{-r})} \\
&= \frac{1}{n}\iota' P(P'P)^{-1}P'(m-m_{\mathcal{B}}\left(s^0\right) + \frac{1}{n}\iota' \underbrace{E(u|X)}_{=0}) + O\left(L^{-r}\right) \\
&= \underbrace{\frac{1}{n}\iota' P(P'P)^{-1}P'(m-m_{\mathcal{B}}\left(s^0\right))}_{=O(L^{-r})} + O\left(L^{-r}\right) \\
&= O\left(L^{-r}\right).
\end{aligned}$$

Note that the bound of $2\|m(x_i)-m_{\mathcal{B}}(x_i;s^0)\|_\infty$ does not depend on X, it is the same for all X, hence by the law of iterated expectations we also have:

$$E\left(\frac{1}{n}\iota'(m_{\mathcal{B}}\left(s^0\right)-m)\right) = E\left(E\left(\frac{1}{n}\iota'(m_{\mathcal{B}}\left(s^0\right)-m)\middle| X\right)\right) = O\left(L^{-r}\right).$$

For variance:

$$\begin{aligned}
V\left(\frac{1}{n}\iota'(\widehat{m}_{\mathcal{B}}\left(s^0\right)-m)\middle| X\right) &= \frac{1}{n^2}\iota' V\left(P(P'P)^{-1}P'(m-m_{\mathcal{B}}\left(s^0\right)+u)+m_{\mathcal{B}}\left(s^0\right)-m\middle| X\right)\iota \\
&= \frac{1}{n^2}\iota' P(P'P)^{-1}P' V\left(u| X\right) P(P'P)^{-1}P'\iota \\
&= \frac{1}{n^2}\iota' P(P'P)^{-1}P'\sigma^2 I P(P'P)^{-1}P'\iota \\
&= \frac{\sigma^2}{n^2}\underbrace{\iota' P(P'P)^{-1}P'P(P'P)^{-1}P'\iota}_{=\|P(P'P)^{-1}P'\iota\|^2\leq\|\iota\|^2=n} \\
&\leq \frac{\sigma^2}{n}.
\end{aligned}$$

For the second equality we use the fact that $m_{\mathcal{B}}\left(s^0\right)$, m and P are deterministic functions of X. The final inequality comes from the fact that $P(P'P)^{-1}P'$ is a projection matrix, and projecting a vector can only make it shorter[17]. This is again a common bound for any choice of X.

17. In fact, B-splines sum to 1, so the vector if 1s is in the span of P and the projection should leave ι unchanged.

By the law of total variance:

$$
\begin{aligned}
V\left(\frac{1}{n}\iota'(\widehat{m}_{\mathscr{B}}\left(s^0\right)-m)\right) &= E\left(V\left(\frac{1}{n}\iota'(\widehat{m}_{\mathscr{B}}\left(s^0\right)-m)\middle| X\right)\right)+V\left(E\left(\frac{1}{n}\iota'(\widehat{m}_{\mathscr{B}}\left(s^0\right)-m)\middle| X\right)\right)\\
&= E\left(\frac{\sigma^2}{n}\right)+V\left(O\left(L^{-r}\right)\right)\\
&= O\left(\frac{1}{n}+L^{-2r}\right).
\end{aligned}
$$

Finally, using $L^{-r} \prec \frac{1}{\sqrt{n}}$:

$$
\frac{1}{n}\sum_{i=1}^{n}\widehat{m}_{\mathscr{B}}(x_i;s^0)-m(x_i) = O_p\left(\frac{1}{\sqrt{n}}+L^{-r}\right) = O_p\left(\frac{1}{\sqrt{n}}\right).
$$

□

Lemma 6.

$$
\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(x_i<x)\left(m(x_i)-m_{\mathscr{B}}(x_i;\hat{s})\right) = O_p\left(\frac{1}{\sqrt{n}}\right).
$$

Proof. Let $f(x_i,s) = m(x_i) - m_{\mathscr{B}}(x_i,s)$ where $m_{\mathscr{B}}(x_i,s)$ is the best possible *B-spline* approximation to $m(x_i)$ which satisfies the constraints under H_0^B. $f(x_i,s)$ is of order $O\left(L^{-r}\right)$ if $s = s^0$ or x_i is sufficiently far from a misspecified switch point. If $s \neq s^0$ and x_i is within a neighborhood of the misspecified switch point, the $f(x_i,s)$ is separated away from zero and does not go to zero as $n \to \infty$, at least for x_i between the true switch point and the switch point used to impose constraints[18].

In the proof of Proposition 3 we rely on the Taylor-expansion of the objective function around the true switch point:

$$
\begin{aligned}
\|\hat{s}-s^0\| &= \left(\frac{\partial^2\widehat{Q}(s^0)}{\partial s\partial s'}+o_p(1)\right)^{-1}\frac{\partial\widehat{Q}(s^0)}{\partial s'}\\
&= \left(\frac{\partial^2\widehat{Q}(s^0)}{\partial s\partial s'}\right)^{-1}O_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}
$$

18. e.g. when we use an incorrect switch point and impose a constraint of increasing function when the true one is decreasing, the best we can do is choose a constant function at some level between $m(s)$ and $m(s^0)$.

It can be shown that

$$\frac{\partial^2 \widehat{Q}(s^0)}{\partial s \partial s'} \simeq \frac{1}{n}\sum_{i=1}^{n} \frac{\partial f(x_i,s^0)}{\partial s}\frac{\partial f(x_i,s^0)}{\partial s'} \simeq \int \frac{\partial f(x,s^0)}{\partial s}\frac{\partial f(x,s^0)}{\partial s'}dx \sim \max_k \int \left(\frac{\partial f(x,s^0)}{\partial s_k}\right)^2 dx.$$

At the same time, the term of interest is:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x) f(x_i,\hat{s}) &\simeq \int_0^x f(x_i,\hat{s})dx \\
&\simeq \int_0^x \underbrace{f(x_i,s^0)}_{\sim L^{-r}} + \frac{\partial f(x_i,\hat{s})}{\partial s}(\hat{s}-s^0)dx \\
&\simeq O_p\left(L^{-r}\right) + \int_0^x \frac{\partial f(x_i,\hat{s})}{\partial s}dx \frac{O_p\left(\frac{1}{\sqrt{n}}\right)}{max_k \int \left(\frac{\partial f(x,s^0)}{\partial s_k}\right)^2 dx} \\
&= O_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

The last equality is because $\int \left(\frac{\partial f(x,s^0)}{\partial s_k}\right)^2 dx$ and $\int \frac{\partial f(x,s^0)}{\partial s_k} dx$ are proportional to each other (both are $O(1)$ over the same region). And the whole term is close to 0 is x is below the misspecified switch point. □

B.2 Useful results

Lemma 7 (B-splines are invariant under a translation and/or scaling of the knot sequence (see e.g. Lyche, Manni, and Speleers (2017)).)**.** *Let* $p_{\ell,t,q}(x)$ *be the* l*th* B-spline *function of order* q *based on the knot vector* t *evaluated at* x*, and let* $\alpha,\beta \in \mathbb{R}$ *with* $\alpha \neq 0$*. Then*

$$p_{\ell,\alpha t+\beta,q}(\alpha x+\beta) = p_{\ell,t,q}(x). \tag{50}$$

Lemma NP.A1. (Newey and Powell 2003) based on Gallant (1987): Consistency of an extremum estimator. *Let*

$$\hat{\theta}_n = \arg\min_{\theta \in \widehat{\Theta}} \widehat{Q}(\theta)$$

be an extremum estimator based on a sample of size n *and assume that there exists a function* $Q(\theta)$ *and a set* Θ *such that:*

(i) $Q(\theta)$ *has a unique minimum on* Θ *at* θ_0*;*

(ii) $\widehat{Q}(\theta)$ *and* $Q(\theta)$ *are continuous,* Θ *is compact, and* $\max_{\theta\in\Theta}\left|\widehat{Q}(\theta)-Q(\theta)\right| \xrightarrow{p} 0$*;*

(iii) $\widehat{\Theta}$ *are compact subsets of* Θ *such that for any* $\theta \in \Theta$ *there exists* $\widehat{\theta} \in \widehat{\Theta}$ *such that* $\widehat{\theta} \xrightarrow{p} \theta$*.*

Then

$$\hat{\theta}_n \xrightarrow{p} \theta_0.$$

Lemma NP.A2. (Newey and Powell 2003): Uniform convergence. *If*

(i) Θ *is a compact subset of a space with norm* $\|\theta\|$*;*

(ii) $\widehat{Q}(\theta) \xrightarrow{p} Q(\theta)$ *for all* $\theta \in \Theta$*;*

(iii) there is a $\delta > 0$ *and* $B_n = O_p(1)$ *such that for all* $\theta, \tilde{\theta} \in \Theta$, $|\widehat{Q}(\theta) - \widehat{Q}(\tilde{\theta})| \leq B_n\|\theta - \tilde{\theta}\|^{\delta}$,

then $Q(\theta)$ *is continuous and*

$$\sup_{\theta\in\Theta} |\widehat{Q}(\theta) - Q(\theta)| \xrightarrow{p} 0. \tag{51}$$