frontiers in Applied Mathematics and Statistics

METHODS
published: 28 July 2017
doi: 10.3389/fams.2017.00015

Regression with Ordered Predictors via Ordinal Smoothing Splines

Nathaniel E. Helwig [1,2]*

[1] Department of Psychology, University of Minnesota, Minneapolis, MN, United States, [2] School of Statistics, University of Minnesota, Minneapolis, MN, United States

Many applied studies collect one or more ordered categorical predictors, which do not fit neatly within classic regression frameworks. In most cases, ordinal predictors are treated as either nominal (unordered) variables or metric (continuous) variables in regression models, which is theoretically and/or computationally undesirable. In this paper, we discuss the benefit of taking a smoothing spline approach to the modeling of ordinal predictors. The purpose of this paper is to provide theoretical insight into the ordinal smoothing spline, as well as examples revealing the potential of the ordinal smoothing spline for various types of applied research. Specifically, we (i) derive the analytical form of the ordinal smoothing spline reproducing kernel, (ii) propose an ordinal smoothing spline isotonic regression estimator, (iii) prove an asymptotic equivalence between the ordinal and linear smoothing spline reproducing kernel functions, (iv) develop large sample approximations for the ordinal smoothing spline, and (v) demonstrate the use of ordinal smoothing splines for isotonic regression and semiparametric regression with multiple predictors. Our results reveal that the ordinal smoothing spline offers a flexible approach for incorporating ordered predictors in regression models, and has the benefit of being invariant to any monotonic transformation of the predictor scores.

Keywords: isotonic regression, monotonic regression, nonparametric regression, ordinal data, smoothing spline, step function

OPEN ACCESS

Edited by:
Sou Cheng Choi,
Illinois Institute of Technology,
United States

Reviewed by:
Katerina Hlavackova-Schindler,
University of Vienna, Austria
Junhui Wang,
City University of Hong Kong,
Hong Kong

***Correspondence:**
Nathaniel E. Helwig
helwig@umn.edu

Specialty section:
This article was submitted to Mathematics of Computation and Data Science, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *02 May 2017*
Accepted: *12 July 2017*
Published: *28 July 2017*

Citation:
Helwig NE (2017) Regression with Ordered Predictors via Ordinal Smoothing Splines. Front. Appl. Math. Stat. 3:15. doi: 10.3389/fams.2017.00015

1. INTRODUCTION

1.1. Motivation

The General Linear Model (GLM) (see [1]) is one of the most widely applied statistical methods, with applications common in psychology [2], education [3], medicine [4], business [5], and several other disciplines. The GLM's popularity in applied research is likely due to a combination of the model's interpretability and flexibility, as well as easy availability through R [6] and commercial statistical softwares (e.g., SAS, SPSS, etc.). The GLM and its generalized extension (GzLM; see [7]) are well-equipped for modeling relationships between variables of mixed types, i.e., unordered categorical and continuous variables can be simultaneously included as predictors in a regression model. However, many studies collect one (or more) ordered categorical variables, which do not fit neatly within the GLM framework.

For example, in finance it is typical to rate the risk of investments on an ordinal scale (very low risk, low risk, …, high risk, very high risk), and a typical goal is to model expected returns given an investment's risk. In medical studies, severity of symptoms (very low, low, …, high, very high) and adherence to treatment (never, rarely, …, almost always, always) are often measured on ordinal scales, and a typical goal is to study patient outcomes in response to different treatments

after controlling for symptom severity and treatment adherence. Psychological attributes such as personality traits and intelligence are typically measured on an ordinal scale, e.g., using questionnaires consisting of Likert scale items (strongly disagree, disagree, ..., agree, strongly agree), and many psychological studies hope to understand how individual and group differences in psychological traits relate to differences in observed behavioral outcomes.

The examples mentioned in the previous paragraph represent just a glimpse of the many ways in which ordinal variables are relevant to our day-to-day financial, physical, and mental health. When it comes to modeling ordinal outcome (response) variables, there are a multitude of potential methods discussed in the literature (see [8–12]). However, when it comes to including ordinal variables as predictors in a GLM (or GzLM), the choices are slim. In nearly all cases, ordinal predictors are treated as either nominal (unordered) or continuous variables in regression models, which can lead to convoluted and possibly misleading results.

Suppose X is an ordered categorical variable with K categories (levels) $1 < \cdots < K$, and suppose we want to include X as a predictor in a regression model. The naive method would be to include $K - 1$ dummy variables in the model design matrix, and let the intercept absorb the K-th level's effect. We refer to this method as naive for two reasons: (i) this approach ignores the fact that the levels of X are *ordered* and, instead, parameterizes X as $K - 1$ *unrelated* variables, and (ii) this approach makes it cumbersome (and possibly numerically unstable) to examine interaction effects involving ordinal predictors, which could be of interest in a variety of studies. Although the dummy coding approach will suffice for certain applications, this method is far from ideal. Furthermore, if the number of levels K is large, the dummy coding approach could be infeasible.

Another possibility for including an ordinal predictor X in a regression model is to simply treat X as a continuous variable. In some cases, researchers make an effort to code the levels of an ordinal predictor X such that the relationship between X and Y is approximately linear. This approach is parsimonious because the ordinal predictor with K categories (requiring $K - 1$ coefficients) will be a reduced to a continuous predictor with a linear effect (requiring 1 coefficient). However, this approach is problematic for several reasons (i) the slope coefficient in such a model has no meaning, (ii) different researchers could concoct different coding schemes for the same data, which would hinder research comparability and reproducibility, and (iii) the ordinal nature of the predictor X is ignored, which is undesirable.

Penalized regression provides a promising framework for including ordinal predictors in regression models [13], given that an appropriate penalty can simultaneously induce order information on the solution and stabilize the estimation. Gertheiss and Tutz [13] discuss how a binary design matrix in combination with a squared difference penalty can be used to fit regression models with ordinal predictors. This approach is implemented in the R package `ordPens` [14], which fits models containing additive effects of ordinal and metric (continuous) predictors. Adding a penalty to impose order and stabilize the solution is a promising approach, but this method still parameterizes X as $K - 1$ *unrelated* variables in the model design matrix. As a result, this approach (and the `ordPens` R package) offer no method for examining interaction effects between multiple ordinal predictors or interaction effects between ordinal and metric predictors.

1.2. Purpose

In this paper, we discuss the benefits of taking a smoothing spline approach [15, 16] to the modeling of ordinal predictors. This approach has been briefly mentioned as a possibility [15, p. 34], but a thorough treatment of the ordinal smoothing spline is lacking from the literature. Expanding the work of Gertheiss and Tutz [13] and Gu [15], we (Section 3.1) derive the analytical form of the reproducing kernel function corresponding to the ordinal smoothing spline, which makes it possible to efficiently compute ordinal smoothing splines. Our results reveal that the reproducing kernel function only depends on rank information, so the ordinal smoothing spline estimator is invariant to any monotonic transformation of the predictor scores. We also (Section 3.2) propose an ordinal smoothing spline isotonic regression estimator via factoring the reproducing kernel function into monotonic (step) functions. Furthermore, we (Section 3.3) prove a correspondence between the ordinal and linear smoothing spline for large values of K, and (Section 3.4) develop computationally scalable approximations for fitting ordinal smoothing splines with large values of K. Finally, we demonstrate the potential of the ordinal smoothing spline for applied research via a simulation study and two real data examples. Our simulation study (Section 4) reveals that the ordinal smoothing spline can outperform the linear smoothing spline and classic isotonic regression algorithms when analyzing monotonic functions with various degrees of smoothness. Our real data results (Section 5) demonstrate that the ordinal smoothing spline—in combination with the powerful smoothing spline ANOVA framework [15]—provides an appealing approach for including ordinal predictors in regression models.

2. SMOOTHING SPLINE BACKGROUND

2.1. Reproducing Kernels

Let $\mathcal{H}$ denote a Hilbert space of functions on domain $\mathcal{X}$, and let $\langle \cdot, \cdot \rangle$ denote the inner-product of $\mathcal{H}$. According to the *Riesz representation theorem* [17, 18], if L is a continuous linear functional in a Hilbert space $\mathcal{H}$, then there exists a unique function $\zeta_L \in \mathcal{H}$ such that $L\eta = \langle \zeta_L, \eta \rangle$ for all $\eta \in \mathcal{H}$. The function ζ_L is referred to as the *representer* of the functional L. The *evaluation functional* $[x]\eta = \eta(x)$ evaluates a function $\eta \in \mathcal{H}$ at an input $x \in \mathcal{X}$. If the evaluation functional $[x]\eta = \eta(x)$ is continuous in $\mathcal{H}$ for all $x \in \mathcal{X}$, then the space $\mathcal{H}$ is referred to as a *reproducing kernel Hilbert space* (RKHS) (see [15, 16]). If $\mathcal{H}$ is a RKHS with inner-product $\langle \cdot, \cdot \rangle$, then there exists a unique function $\rho_x \in \mathcal{H}$ that is the representer of the evaluation functional $[x](\cdot)$, such that $[x]\eta = \langle \rho_x, \eta \rangle = \eta(x)$ for all $\eta \in \mathcal{H}$ and all $x \in \mathcal{X}$. The symmetric and non-negative definite function

$$\rho(x, y) = \rho_x(y) = \rho_y(x) = \langle \rho_x, \rho_y \rangle \tag{1}$$

is called the *reproducing kernel* of $\mathcal{H}$ because it satisfies the reproducing property $\langle \rho(x,y), \eta(y)\rangle = \eta(x)$ for all $\eta \in \mathcal{H}$ and all $x, y \in \mathcal{X}$, see Aronszajn [19]. If $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ with $\mathcal{H}_0 \cap \mathcal{H}_1 = \emptyset$, then

$$\rho(x,y) = \rho_0(x,y) + \rho_1(x,y) \tag{2}$$

where $\rho_0 \in \mathcal{H}_0$ and $\rho_1 \in \mathcal{H}_1$ are the reproducing kernels of $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively, such that

$$\begin{aligned}\langle \rho(x,y), \eta(y)\rangle &= \langle \rho_0(x,y), \eta_0(y)\rangle_0 + \langle \rho_1(x,y), \eta_1(y)\rangle_1 \\ &= \eta_0(x) + \eta_1(x) = \eta(x)\end{aligned} \tag{3}$$

with $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_0 + \langle \cdot, \cdot \rangle_1$, $\eta = (\eta_0 + \eta_1) \in \mathcal{H}$, $\eta_0 \in \mathcal{H}_0$, and $\eta_1 \in \mathcal{H}_1$.

2.2. Smoothing Spline Definition

Assume a nonparametric regression model (see [15, 16, 20–23])

$$y_i = \eta(x_i) + \epsilon_i \tag{4}$$

for $i = 1, \ldots, n$, where $y_i \in \mathbb{R}$ is the real-valued response for the i-th observation, $x_i \in \mathcal{X}$ is the predictor for the i-th observation where $\mathcal{X}$ is the predictor domain, $\eta \in \mathcal{H}$ is an unknown smooth function where $\mathcal{H}$ is a RKHS with inner-product $\langle \cdot, \cdot \rangle$, and $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$ is a Gaussian error term. A *smoothing spline* is the function η_λ that minimizes the penalized least squares functional

$$\frac{1}{n}\sum_{i=1}^{n}\big(y_i - \eta(x_i)\big)^2 + \lambda J(\eta) \tag{5}$$

where $J(\cdot)$ is a quadratic penalty functional such that larger values correspond to less smoothness, and $\lambda \geq 0$ is a smoothing parameter that balances the trade-off between fitting and smoothing the data. Let $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ denote a tensor sum decomposition of $\mathcal{H}$, where $\mathcal{H}_0 = \{\eta : J(\eta) = 0\}$ is the *null space* of J and $\mathcal{H}_1 = \{\eta : 0 < J(\eta) < \infty\}$ is the *contrast space*. Similarly, let $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_0 + \langle \cdot, \cdot \rangle_1$ denote the corresponding decomposition of the inner-product of $\mathcal{H}$. Note that, by definition, the quadratic penalty functional J is the inner-product of the contrast space $\mathcal{H}_1$, i.e., $J(\eta) = \langle \eta, \eta \rangle_1$. Given λ, the *Kimeldorf-Wahba representer theorem* [24] states that the η_λ minimizing Equation (5) has the form

$$\eta_\lambda(x) = \sum_{v=0}^{m-1} d_v \phi_v(x) + \sum_{i=1}^{n} c_i \rho_1(x, x_i) \tag{6}$$

where the functions $\{\phi_v\}_{v=0}^{m-1}$ span the penalty's null space $\mathcal{H}_0$, the function ρ_1 is the reproducing kernel of the contrast space $\mathcal{H}_1$, and $\mathbf{d} = \{d_v\}$ and $\mathbf{c} = \{c_i\}$ are the unknown coefficient vectors (see [15, 16, 25]). The reproducing property implies that the quadratic penalty functional $J(\eta_\lambda) = \langle \eta_\lambda, \eta_\lambda \rangle_1$ has the form

$$J(\eta_\lambda) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \rho_1(x_i, x_j) \tag{7}$$

which can be easily evaluated given ρ_1, $\mathbf{c}$, and $\mathbf{x}$.

TABLE 1 | Ingredients for forming different types of smoothing splines.

Type	Reproducing Kernel Hilbert space components
Nominal	$\mathcal{X}$: $x \in \{1, \ldots, K\}$
	$\mathcal{H}$: $\eta \in \mathbb{R}^K$
	$\mathcal{H}_0$: $\eta_0 \in \{\eta : \eta(1) = \cdots = \eta(K)\}$
	$\langle \eta, \xi \rangle_0$: $K\bar{\eta}\bar{\xi}$ where $\bar{\eta} = \frac{1}{K}\sum_{x=1}^{K}\eta(x)$ and $\bar{\xi} = \frac{1}{K}\sum_{x=1}^{K}\xi(x)$
	$\rho_0(x,y)$: $1/K$
	$\mathcal{H}_1$: $\eta_1 \in \{\eta : \sum_{x=1}^{K}\eta(x) = 0\}$
	$\langle \eta, \xi \rangle_1$: $\sum_{x=1}^{K}[\eta(x) - \bar{\eta}][\xi(x) - \bar{\xi}]$
	$\rho_1(x,y)$: $\delta_{xy} - 1/K$ where $\delta_{xy} = 1$ if $x = y$ and $\delta_{xy} = 0$ otherwise
Ordinal	$\mathcal{X}$: $x \in \{1, \ldots, K\}$ where $1 < \cdots < K$
	$\mathcal{H}$: $\eta \in \mathbb{R}^K$
	$\mathcal{H}_0$: $\eta_0 \in \{\eta : \eta(1) = \cdots = \eta(K)\}$
	$\langle \eta, \xi \rangle_0$: $K\bar{\eta}\bar{\xi}$ where $\bar{\eta} = \frac{1}{K}\sum_{x=1}^{K}\eta(x)$ and $\bar{\xi} = \frac{1}{K}\sum_{x=1}^{K}\xi(x)$
	$\rho_0(x,y)$: $1/K$
	$\mathcal{H}_1$: $\eta_1 \in \{\eta : \sum_{x=1}^{K}\eta(x) = 0\}$
	$\langle \eta, \xi \rangle_1$: $\sum_{x=2}^{K}[\eta(x) - \eta(x-1)][\xi(x) - \xi(x-1)]$
	$\rho_1(x,y)$: $1 - x \vee y + \frac{1}{2K}[x(x-1) + y(y-1)] + \frac{(K-1)(2K-1)}{6K}$
Polynomial	$\mathcal{X}$: $x \in [0, 1]$
	$\mathcal{H}$: $\eta \in \{\eta : \int_0^1 [\eta^{(m)}(x)]^2 dx < \infty\}$ where $\eta^{(m)} = \frac{d^m}{dx^m}\eta(x)$
	$\mathcal{H}_0$: $\eta_0 \in \{\eta : \eta^{(m)} = 0\}$
	$\langle \eta, \xi \rangle_0$: $\sum_{v=0}^{m-1}(\int_0^1 \eta^{(v)}dx)(\int_0^1 \xi^{(v)}dx)$
	$\rho_0(x,y)$: $\sum_{v=0}^{m-1}\kappa_v(x)\kappa_v(y)$
	$\mathcal{H}_1$: $\eta_1 \in \{\eta : \int_0^1 \eta^{(v)} = 0, v = 0, \ldots, m-1, \int_0^1 [\eta^{(m)}(x)]^2 dx < \infty\}$
	$\langle \eta, \xi \rangle_1$: $\int_0^1 \eta^{(m)}\xi^{(m)}dx$
	$\rho_1(x,y)$: $\kappa_m(x)\kappa_m(y) + (-1)^{m-1}\kappa_{2m}(\lvert x - y\rvert)$
Thin-Plate	$\mathcal{X}$: $x \in \mathbb{R}^d$
	$\mathcal{H}$: $\eta \in \{\eta : \langle \eta, \eta \rangle_1 < \infty\}$
	$\mathcal{H}_0$: $\eta_0 \in \{\eta : \langle \eta, \eta \rangle_1 = 0\}$
	$\langle \eta, \xi \rangle_0$: $\frac{1}{R}\sum_{j=1}^{R}\eta(\tilde{x}_j)\xi(\tilde{x}_j)$ where $\{\tilde{x}_j\}_{j=1}^{R} \subseteq \{x_i\}_{i=1}^{n}$ are the knots
	$\rho_0(x,y)$: $\sum_{v=1}^{M}\phi_v(x)\phi_v(y)$ where $\langle \phi_u, \phi_v \rangle_0 = \delta_{uv}$ and $M = \binom{d+m-1}{d}$
	$\mathcal{H}_1$: $\eta_1 \in \{\eta : 0 < \langle \eta, \eta \rangle_1 < \infty\}$
	$\langle \eta, \xi \rangle_1$: $\sum_{v_1+\cdots+v_d=m}\frac{m!}{v_1!\cdots v_d!}\int_{\mathbb{R}}\cdots\int_{\mathbb{R}}(\frac{\partial^m \eta}{\partial x_1^{v_1}\cdots\partial x_d^{v_d}})(\frac{\partial^m \xi}{\partial x_1^{v_1}\cdots\partial x_d^{v_d}})dx_1\cdots dx_d$
	$\rho_1(x,y)$: $(I - P_{(\mathbf{x})})(I - P_{(\mathbf{y})})\gamma(\lVert\mathbf{x} - \mathbf{y}\rVert)$

For ordinal splines, the notation $x \vee y$ denotes the maximum of x and y.
For polynomial splines, the κ_v functions denote scaled Bernoulli polynomials (see [35]).
For thin-plate splines, $P_{(\mathbf{x})} = (P\eta)(\mathbf{x}) = \sum_{v=1}^{M}\langle \eta, \phi_v \rangle_0 \phi_v(\mathbf{x})$ denotes the projection of η onto $\mathcal{H}_0$, and $\gamma(x) \propto x^{2m-d}\log(x)$ if d is even or $\gamma(x) \propto x^{2m-d}$ if d is odd.

2.3. Types of Splines

The type of spline will depend on the forms of the RKHS $\mathcal{H}$ and the inner-product $\langle \cdot, \cdot \rangle$, which will depend on the predictor domain $\mathcal{X}$. The essential components for the formation of a smoothing spline include: (i) form a tensor sum decomposition of the RKHS $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, (ii) partition the inner-product of the RKHS $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_0 + \langle \cdot, \cdot \rangle_1$ such that $\langle \cdot, \cdot \rangle_k$ defines an inner-product in $\mathcal{H}_k$ for $k \in \{0, 1\}$, (iii) define the smoothing spline penalty as $J(\eta) = \langle \eta, \eta \rangle_1$, and (iv) derive the reproducing kernels of $\mathcal{H}_0$ and $\mathcal{H}_1$. **Table 1** provides the information needed

to form three common smoothing splines (nominal, polynomial, and thin-plate), as well as the ordinal smoothing spline. See Gu [15] and Helwig and Ma [26] for more information about nominal and polynomial smoothing splines, and see Gu [15], Helwig and Ma [26], Duchon [27], Meinguet [28], and Wood [29] for more information about thin-plate splines. More information about the ordinal smoothing spline will be provided in Section 3.

2.4. Tensor Product Smoothers

Suppose we have a model with predictor vector $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathcal{X} = \prod_{j=1}^{d} \mathcal{X}_j$ where $X_j \in \mathcal{X}_j \; \forall j$. If the predictors are all continuous with similar scale, a thin-plate spline can be used. However, if the predictors differ in type and/or scale (e.g., some nominal and some continuous), another approach is needed. Let $\mathbf{x} = (x_1, \ldots, x_d)'$ and $\mathbf{y} = (y_1, \ldots, y_d)'$ denote two realizations of $\boldsymbol{X}$ and let

$$\rho_{\mathcal{X}_j}(x_j, y_j) = \rho_{0j}(x_j, y_j) + \rho_{1j}(x_j, y_j) \tag{8}$$

denote the reproducing kernel function corresponding to $\mathcal{H}_{\mathcal{X}_j} = \mathcal{H}_{0j} \oplus \mathcal{H}_{1j}$, which denotes a RKHS of functions on $\mathcal{X}_j$ for $j = 1, \ldots, d$. The symmetric and non-negative definite function

$$\rho(\mathbf{x}, \mathbf{y}) = \prod_{j=1}^{d} \rho_{\mathcal{X}_j}(x_j, y_j) = \rho_0(\mathbf{x}, \mathbf{y}) + \rho_1(\mathbf{x}, \mathbf{y}) \tag{9}$$

defines the unique tensor product reproducing kernel function corresponding to the tensor product RKHS

$$\mathcal{H} = \mathcal{H}_{\mathcal{X}_1} \otimes \cdots \otimes \mathcal{H}_{\mathcal{X}_d} = \mathcal{H}_0 \oplus \mathcal{H}_1 \tag{10}$$

where $\rho_0 \in \mathcal{H}_0$ with $\mathcal{H}_0 = \mathcal{H}_{01} \otimes \mathcal{H}_{02} \otimes \cdots \otimes \mathcal{H}_{0d}$ denoting the tensor product null space and $\rho_1 \in \mathcal{H}_1$ with $\mathcal{H}_1 = \mathcal{H} \ominus \mathcal{H}_0$ denoting the tensor product contrast space.

A *tensor product smoothing spline* solves the penalized least squares problem in Equation (5) in a tensor product RKHS $\mathcal{H} = \otimes_{j=1}^{d} \mathcal{H}_{\mathcal{X}_j} = \mathcal{H}_0 \oplus \mathcal{H}_1$. In this case, the contrast space can be decomposed such as $\mathcal{H}_1 = \mathcal{H}_1^* \oplus \cdots \oplus \mathcal{H}_s^*$ where the $\mathcal{H}_k^*$ are orthogonal subspaces with inner-products $\langle \cdot, \cdot \rangle_k^*$. The different $\mathcal{H}_k^*$ may have different modules, so it is helpful to reweight the relative influence of each $\mathcal{H}_k^*$. Specifically, when solving the penalized least squares functional, we can define

$$\begin{aligned} J(\eta) &= \sum_{k=1}^{s} \theta_k^{-1} \langle \eta, \eta \rangle_k^* \\ \rho_1(\mathbf{x}, \mathbf{y}) &= \sum_{k=1}^{s} \theta_k \rho_k^*(\mathbf{x}, \mathbf{y}) \end{aligned} \tag{11}$$

where $\rho_k^* \in \mathcal{H}_k^*$ is the reproducing kernel of $\mathcal{H}_k^*$ and $\theta_k > 0$ are additional (non-negative) smoothing parameters that control the influence of each $\mathcal{H}_k^*$ [see 15, 25, 26]. We can decompose any $\eta \in \mathcal{H}$ such as

$$\begin{aligned} \eta(\mathbf{x}) &= \eta_0(\mathbf{x}) + \sum_{k=1}^{s} \eta_k^*(\mathbf{x}) \\ &= \langle \rho_0(\mathbf{x}, \mathbf{y}), \eta_0(\mathbf{y}) \rangle_0 + \sum_{k=1}^{s} \theta_k^{-1} \langle \theta_k \rho_k^*(\mathbf{x}, \mathbf{y}), \eta_k^*(\mathbf{y}) \rangle_k^* \end{aligned} \tag{12}$$

where $\eta_0 \in \mathcal{H}_0$ and $\eta_k^* \in \mathcal{H}_k^*$ for $k = 1, \ldots, s$. Note that the different η_k^* functions correspond to different nonparametric main and interaction effects between predictors, so different statistical models can be fit by removing different $\mathcal{H}_k^*$ subspaces (and corresponding η_k^*) from the model.

2.5. Smoothing Spline Computation

Applying the Kimeldorf-Wahba representer theorem, we can approximate Equation (5) as

$$n^{-1} \|\mathbf{y} - \mathbf{K}\mathbf{d} - \mathbf{J}\mathbf{c}\|^2 + \lambda \mathbf{c}' \mathbf{Q} \mathbf{c} \tag{13}$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$ is the response vector, $\mathbf{K} = \{\phi_v(x_i)\}_{n \times m}$ is the null space basis function matrix, $\mathbf{J} = \{\rho_1(x_i, \tilde{x}_j)\}_{n \times R}$ denotes the contrast space basis function matrix with $\{\tilde{x}_j\}_{j=1}^{R} \subseteq \{x_i\}_{i=1}^{n}$ denoting the selected knots, and $\mathbf{Q} = \{\rho_1(\tilde{x}_i, \tilde{x}_j)\}_{R \times R}$ denotes the penalty matrix. The full solution uses all unique x_i as knots, but using a subset of $R \ll n$ knots has been shown to perform well in practice, as long as enough knots are included to reasonably span $\mathcal{X}$ [26, 30–32]. Given λ, the optimal coefficients are

$$\begin{pmatrix} \hat{\mathbf{d}} \\ \hat{\mathbf{c}} \end{pmatrix} = \begin{pmatrix} \mathbf{K}'\mathbf{K} & \mathbf{K}'\mathbf{J} \\ \mathbf{J}'\mathbf{K} & \mathbf{J}'\mathbf{J} + n\lambda\mathbf{Q} \end{pmatrix}^{\dagger} \begin{pmatrix} \mathbf{K}' \\ \mathbf{J}' \end{pmatrix} \mathbf{y} \tag{14}$$

where $(\cdot)^{\dagger}$ denotes the Moore-Pensore pseudoinverse [33, 34]. The fitted values can be written as

$$\hat{\boldsymbol{\eta}} = \mathbf{K}\hat{\mathbf{d}} + \mathbf{J}\hat{\mathbf{c}} = \mathbf{S}_\lambda \mathbf{y} \tag{15}$$

where

$$\mathbf{S}_\lambda = \begin{pmatrix} \mathbf{K} & \mathbf{J} \end{pmatrix} \begin{pmatrix} \mathbf{K}'\mathbf{K} & \mathbf{K}'\mathbf{J} \\ \mathbf{J}'\mathbf{K} & \mathbf{J}'\mathbf{J} + n\lambda\mathbf{Q} \end{pmatrix}^{\dagger} \begin{pmatrix} \mathbf{K}' \\ \mathbf{J}' \end{pmatrix} \tag{16}$$

is the smoothing matrix, which depends on λ and any θ_k parameters embedded in ρ_1. The smoothing parameters can be obtained by minimizing the Generalized Cross-Validation (GCV) criterion [35]

$$\text{GCV}(\lambda) = \frac{n^{-1} \|\mathbf{y} - \mathbf{S}_\lambda \mathbf{y}\|^2}{[1 - \text{tr}(\mathbf{S}_\lambda)/n]^2} \tag{17}$$

where $\text{tr}(\mathbf{S}_\lambda)$ is often considered the effective degrees of freedom of a smoothing spline.

2.6. Bayesian Interpretation

The smoothing spline solution in Equation (15) can be interpreted as a Bayesian estimate of a Gaussian process [36–40]. Suppose that $\eta = (\eta_0 + \eta_1) \in \mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $\eta_k \in \mathcal{H}_k$ for $k \in \{0, 1\}$. Furthermore, assume that (i) $\eta_0(x) = \boldsymbol{\phi}'\mathbf{d}$ where $\boldsymbol{\phi}' = (\phi_1(x), \ldots, \phi_m(x))$ and $\mathbf{d} \sim \mathrm{N}(\mathbf{0}_m, \tau^2\mathbf{I}_m)$, (ii) $\eta_1(x) = \boldsymbol{\rho}'\mathbf{c}$ where $\boldsymbol{\rho}' = (\rho_1(x, \tilde{x}_1), \ldots, \rho_1(x, \tilde{x}_R))$ and $\mathbf{c} \sim \mathrm{N}(\mathbf{0}_R, \frac{\sigma^2}{n\lambda}\mathbf{Q}^\dagger)$, and (iii) η_0 and η_1 are independent of one another and of the error term ϵ. Applying Bayes's Theorem, the posterior distribution of $\boldsymbol{\beta} = (\mathbf{d}', \mathbf{c}')'$ is multivariate normal with mean and covariance

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{\beta}|\mathbf{y}} &= \boldsymbol{\Sigma}_{\beta y}'\boldsymbol{\Sigma}_y^{-1}\mathbf{y} \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}|\mathbf{y}} &= \boldsymbol{\Sigma}_\beta - \boldsymbol{\Sigma}_{\beta y}'\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{\beta y} \end{aligned} \tag{18}$$

where $\boldsymbol{\Sigma}_\beta = \mathrm{bdiag}(\tau^2\mathbf{I}_m, \frac{\sigma^2}{n\lambda}\mathbf{J}\mathbf{Q}^\dagger)$ is a block-diagonal matrix, $\boldsymbol{\Sigma}_{y\beta} = (\tau^2\mathbf{K}\ \frac{\sigma^2}{n\lambda}\mathbf{J}\mathbf{Q}^\dagger)$, and $\boldsymbol{\Sigma}_y = \tau^2\mathbf{K}\mathbf{K}' + \frac{\sigma^2}{n\lambda}\mathbf{J}\mathbf{Q}^\dagger\mathbf{J}' + \sigma^2\mathbf{I}_n$. Letting $\tau^2 \to \infty$ corresponds to a diffuse prior on the null space coefficients, and

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\beta}|\mathbf{y}} = \lim_{\tau^2\to\infty} \boldsymbol{\mu}_{\boldsymbol{\beta}|\mathbf{y}} = \begin{pmatrix} \mathbf{K}'\mathbf{K} & \mathbf{K}'\mathbf{J} \\ \mathbf{J}'\mathbf{K} & \mathbf{J}'\mathbf{J} + n\lambda\mathbf{Q} \end{pmatrix}^\dagger \begin{pmatrix} \mathbf{K}' \\ \mathbf{J}' \end{pmatrix} \mathbf{y} \tag{19}$$

makes the posterior mean equivalent to the smoothing spline coefficient estimates from Equation (14). The corresponding Bayesian covariance matrix estimator is

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}|\mathbf{y}} = \lim_{\tau^2\to\infty} \boldsymbol{\Sigma}_{\boldsymbol{\beta}|\mathbf{y}} = \sigma^2 \begin{pmatrix} \mathbf{K}'\mathbf{K} & \mathbf{K}'\mathbf{J} \\ \mathbf{J}'\mathbf{K} & \mathbf{J}'\mathbf{J} + n\lambda\mathbf{Q} \end{pmatrix}^\dagger. \tag{20}$$

Using this Bayesian interpretation, one can form a $100(1-\alpha)\%$ Bayesian "confidence interval" around the smoothing spline estimate

$$\boldsymbol{\psi}'\hat{\boldsymbol{\mu}}_{\boldsymbol{\beta}|\mathbf{y}} \pm Z_{\alpha/2}\sqrt{\boldsymbol{\psi}'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}|\mathbf{y}}\boldsymbol{\psi}} \tag{21}$$

where $\boldsymbol{\psi}' = (\boldsymbol{\phi}', \boldsymbol{\rho}')$ and $Z_{\alpha/2}$ is a critical value from a standard normal distribution. These confidence intervals have a desirable across-the-function coverage property, such that the intervals are expected to contain about $100(1-\alpha)\%$ of the function realizations, assuming that the GCV has been used to select the smoothing parameters [37, 38, 40].

3. ORDINAL SMOOTHING SPLINE

3.1. Formulation

Real-valued functions on $\mathcal{X} = \{1, \ldots, K\}$ correspond to real-valued vectors of length K, where the function evaluation is equivalent to coordinate extraction, i.e., $\eta(x)$ for $x \in \{1, \ldots, K\}$ can be viewed as extracting the x-th element of the vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)'$. Thus, equipped with an inner product, the space of functions on $\mathcal{X} = \{1, \ldots, K\}$ is a RKHS, denoted by $\mathcal{H} = \mathbb{R}^K$, which can be decomposed into the tensor summation of two orthogonal subspaces: $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$. The null space $\mathcal{H}_0 = \{\eta : \eta(1) = \cdots = \eta(K)\}$ consists of all constant functions (vectors of length K), and the contrast space $\mathcal{H}_1 = \{\eta : \sum_{x=1}^K \eta(x) = 0\}$ consists of all functions (vectors of length K) that sum to zero across the K elements.

For ordinal predictors $x \in \mathcal{X} = \{1, \ldots, K\}$ with $1 < \cdots < K$, we can use the penalty functional

$$J(\eta) = \sum_{x=2}^K [\eta(x) - \eta(x-1)]^2 \tag{22}$$

which penalizes differences between adjacent values [13, 15]. Letting $\boldsymbol{\eta} = [\eta(1), \ldots, \eta(K)]'$, we can write the penalty functional as

$$J(\eta) = \|\mathbf{D}\boldsymbol{\eta}\|^2 \tag{23}$$

where $\mathbf{D} = \{d_{ij}\}_{K-1\times K}$ is a first order difference operator matrix with elements defined as

$$d_{ij} = \begin{cases} 1 & \text{if } j = i+1 \\ -1 & \text{if } j = i \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

As an example, with $K = 5$ the matrix $\mathbf{D}$ has the form

$$\mathbf{D} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}. \tag{25}$$

The penalty functional in Equation (22) corresponds to the inner-product

$$\begin{aligned} \langle \eta, \xi \rangle &= K\bar{\eta}\bar{\xi} + \sum_{x=2}^K [\eta(x) - \eta(x-1)][\xi(x) - \xi(x-1)] \\ &= \boldsymbol{\eta}' \left(\frac{1}{K}\mathbf{1}_K\mathbf{1}_K' \right) \boldsymbol{\xi} + \boldsymbol{\eta}'\mathbf{D}'\mathbf{D}\boldsymbol{\xi} \end{aligned} \tag{26}$$

where $\bar{\eta} = (1/K)\sum_{x=1}^K \eta(x)$, $\bar{\xi} = (1/K)\sum_{x=1}^K \xi(x)$, $\langle \eta, \xi \rangle_0 = \boldsymbol{\eta}\left(\frac{1}{K}\mathbf{1}_K\mathbf{1}_K'\right)\boldsymbol{\xi}$ is the inner product of the null space, and $\langle \eta, \xi \rangle_1 = \boldsymbol{\eta}'\mathbf{D}'\mathbf{D}\boldsymbol{\xi}$ is the inner product of the contrast space. Note that $\left(\frac{1}{K}\mathbf{1}_K\mathbf{1}_K'\right)\mathbf{D}'\mathbf{D} = \mathbf{0}_{K\times K}$ by definition. Furthermore, note that the ordinal smoothing spline estimates η within the same RKHS as the nominal smoothing spline (see **Table 1**), however the ordinal smoothing spline uses a different definition of the inner-product for the contrast space, which induces a different reproducing kernel.

Theorem 3.1. *Given the ordinal smoothing spline inner-product in Equation (26), the null space $\mathcal{H}_0 = \{\eta : \eta(1) = \cdots = \eta(K)\}$ has reproducing kernel $\rho_0(x, y) = 1/K$, and the contrast space $\mathcal{H}_1 = \{\eta : \sum_{x=1}^K \eta(x) = 0\}$ has reproducing kernel*

$$\begin{aligned} \rho_1(x, y) &= \sum_{k=1}^{K-1} \left(1_{\{x\le k\}} - \frac{k}{K}\right)\left(1_{\{y\le k\}} - \frac{k}{K}\right) \\ &= 1 - x \vee y + \frac{1}{2K}\left[x(x-1) + y(y-1)\right] + \tilde{K} \end{aligned}$$

for $x, y \in \{1, \ldots, K\}$ where $1_{\{\cdot\}}$ is an indicator function, $x \vee y = \max(x, y)$, and $\tilde{K} = \frac{(K-1)(2K-1)}{6K}$.

See the Supplementary Material for a proof. Note that Theorem 3.1 provides the reproducing kernel needed to numerically evaluate the optimal η_λ from Equation (6) for the ordinal smoothing spline. Considering the model in Equation (4) with $\mathcal{X} = \{1, \ldots, K\}$ where $1 < \cdots < K$ are elements of an ordered set, the penalized least squares problem has the form

$$n^{-1}\|\mathbf{y} - d\mathbf{1}_n - \mathbf{G}\mathbf{Q}\mathbf{c}\|^2 + \lambda \mathbf{c}'\mathbf{Q}\mathbf{c} \tag{27}$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$ is the response vector, d is the unknown intercept, $\mathbf{G} = \{g_{ij}\}_{n\times K}$ is a selection matrix such that $g_{ij} = 1$ if $x_i = j$ for $j \in \{1, \ldots, K\}$ and $g_{ij} = 0$ otherwise, $\mathbf{Q} = \{\rho_1(\tilde{x}_i, \tilde{x}_j)\}_{K\times K}$ with $\tilde{x}_i = i$ and $\tilde{x}_j = j$ for $i, j = 1, \ldots, K$, and $\mathbf{c} = (c_1, \ldots, c_K)'$ are the unknown function coefficients. Given λ, the optimal coefficients and smoothing matrix can be obtained using the result in Equations (14) and (16) with $\mathbf{K} = \mathbf{1}_n$ and $\mathbf{J} = \mathbf{G}\mathbf{Q}$. When applying the Bayesian interpretation in Section 2.6, note that the pseudoinverse of $\mathbf{Q}$ has Toeplitz form for the internal rows and columns with deviations from Toeplitz form in cells (1,1) and (K, K). As an example, with $K = 5$ the pseudoinverse of $\mathbf{Q}$ is given by

$$\mathbf{Q}^\dagger = \mathbf{D}'\mathbf{D} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix} \tag{28}$$

where $\mathbf{c} \sim \mathrm{N}(\mathbf{0}, [\sigma^2/(n\lambda)]\mathbf{Q}^\dagger)$ is the assumed prior distribution for the contrast space coefficients.

3.2. Relation to Isotonic Regression

Using the reproducing kernel definition in the first line of Theorem 3.1, $\mathbf{Q} = \{\rho_1(x, y)\}_{K\times K} = \mathbf{M}\mathbf{M}'$ where the elements of the matrix $\mathbf{M} = \{m_{ij}\}_{K\times K-1}$ have the form

$$m_{ij} = \begin{cases} (j - K)/K & \text{if } i \leq j \\ j/K & \text{if } i > j \end{cases} \tag{29}$$

As an example, with $K = 5$ the matrix $\mathbf{M}$ has the form

$$\mathbf{M} = \frac{1}{5}\begin{pmatrix} -4 & -3 & -2 & -1 \\ 1 & -3 & -2 & -1 \\ 1 & 2 & -2 & -1 \\ 1 & 2 & 3 & -1 \\ 1 & 2 & 3 & 4 \end{pmatrix}. \tag{30}$$

Note that the columns of $\mathbf{M}$ sum to zero, which implies that $\mathbf{Q}_0\mathbf{Q} = \mathbf{0}_{K\times K}$, where $\mathbf{Q}_0 = \left(\frac{1}{K}\mathbf{1}_K\mathbf{1}_K'\right)$ is the reproducing kernel matrix for the null space. To visualize the ordinal smoothing spline reproducing kernel function, **Figure 1** plots the columns of $\mathbf{M}$ and the columns of the corresponding reproducing kernel matrix $\mathbf{Q}$ for the situation with $K = 5$. Note that by definition (i) the $K - 1$ columns of $\mathbf{M}$ are (linearly independent) monotonic increasing functions of $x \in \{1, \ldots, K\}$ and (ii) the K columns of $\mathbf{Q}$ are linearly dependent given that $\mathbf{Q}$ has rank $K - 1$.

The factorization of the reproducing kernel matrix $\mathbf{Q} = \mathbf{M}\mathbf{M}'$ implies that the ordinal smoothing spline can be reformulated to form an isotonic regression estimator. Specifically, we could reparameterize the ordinal smoothing spline problem as

$$n^{-1}\|\mathbf{y} - d\mathbf{1}_n - \mathbf{G}\mathbf{M}\mathbf{c}_m\|^2 + \lambda\|\mathbf{c}_m\|^2 \tag{31}$$

where $\mathbf{c}_m = \mathbf{M}'\mathbf{c}$ and $\mathbf{Q} = \mathbf{M}\mathbf{M}'$ by definition. Note that columns of $\mathbf{M}$ are monotonic increasing such that the k-th column contains two unique values with the "jump" occurring between rows k and $k + 1$ for $k \in \{1, \ldots, K - 1\}$, see **Figure 1**. Thus, by constraining the elements of $\mathbf{c}_m$ to be non-negative, we can constrain the ordinal smoothing spline to be a monotonic increasing function across the values $1 < \cdots < K$. Specifically, for a fixed λ we seek the solution to the problem

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}'\left(\mathbf{X}'\mathbf{X} + n\lambda\mathbf{I}_K^*\right)\boldsymbol{\beta} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} \quad \text{subject to } \mathbf{A}\boldsymbol{\beta} \geq \mathbf{0}_K \tag{32}$$

where $\boldsymbol{\beta} = (d, \mathbf{c}_m')'$ is the coefficient vector, $\mathbf{X} = [\mathbf{1}_n, \mathbf{G}\mathbf{M}]$ is the design matrix, $\mathbf{I}_K^*$ is a $K \times K$ identity matrix with cell (1,1) equal to zero, and $\mathbf{A} = (\mathbf{0}_{K-1}, \mathbf{I}_{K-1})$ is the $(K - 1) \times K$ constraint matrix.

Theorem 3.2. *The coefficient vector $\boldsymbol{\beta}$ minimizing the inequality constrained optimization problem in Equation (32) has the form*

$$\hat{\boldsymbol{\beta}}^* = \left(\mathbf{I}_K + \mathbf{C}^{-1}\mathbf{A}_2'\mathbf{B}_2\mathbf{A}\right)\hat{\boldsymbol{\beta}}$$

where $\mathbf{C} = \mathbf{X}'\mathbf{X} + n\lambda\mathbf{I}_K^*$, $\hat{\boldsymbol{\beta}} = \mathbf{C}^{-1}\mathbf{X}'\mathbf{y}$ is the unconstrained ordinal smoothing spline solution, and the $\mathbf{A}_2$ and $\mathbf{B}_2$ matrices depend on the active constraints.

See the Supplementary Material for the proof. Note that Theorem 3.2 reveals that the ordinal smoothing spline isotonic regression estimator is a linear transformation of the unconstrained estimator. Furthermore, Theorem 3.2 implies that $\hat{\boldsymbol{\eta}} = \mathbf{S}_\lambda^*\mathbf{y}$ where the smoothing matrix has the form $\mathbf{S}_\lambda^* = \mathbf{X}\left(\mathbf{I}_K + \mathbf{C}^{-1}\mathbf{A}_2'\mathbf{B}_2\mathbf{A}\right)\mathbf{C}^{-1}\mathbf{X}'$. Thus, when the non-negativity constraints are active, $\mathrm{tr}(\mathbf{S}_\lambda^*)$ can be used as an estimate of the effective degrees of freedom for smoothing parameter selection.

3.3. Reproducing Kernel as $K \to \infty$

We now provide a theorem on the behavior of the ordinal smoothing spline reproducing kernel as the number of levels K of the ordered set $\mathcal{X} = \{1, \ldots, K\}$ approaches infinity.

Theorem 3.3. *As $K \to \infty$, we have $\rho_1(x, y)/(K - 1) \to \tilde{\rho}_1(\tilde{x}, \tilde{y})$ where*

$$\rho_1(x, y) = 1 - x \vee y + \frac{1}{2K}\left[x(x - 1) + y(y - 1)\right] + \frac{(K - 1)(2K - 1)}{6K}$$

is the ordinal smoothing spline reproducing kernel function for $x, y \in \mathcal{X} = \{1, \ldots, K\}$ with $x \vee y = \max(x, y)$ denoting the maximum of x and y, and

$$\tilde{\rho}_1(\tilde{x}, \tilde{y}) = (\tilde{x} - 1/2)(\tilde{y} - 1/2) + (1/2)\left\{\left(|\tilde{x} - \tilde{y}| - 1/2\right)^2 - 1/12\right\}$$

is the linear smoothing spline reproducing kernel function for $\tilde{x}, \tilde{y} \in \tilde{\mathcal{X}} = [0, 1]$ with $\tilde{x} = (x - 1)/(K - 1)$ and $\tilde{y} = (y - 1)/(K - 1)$ by definition.

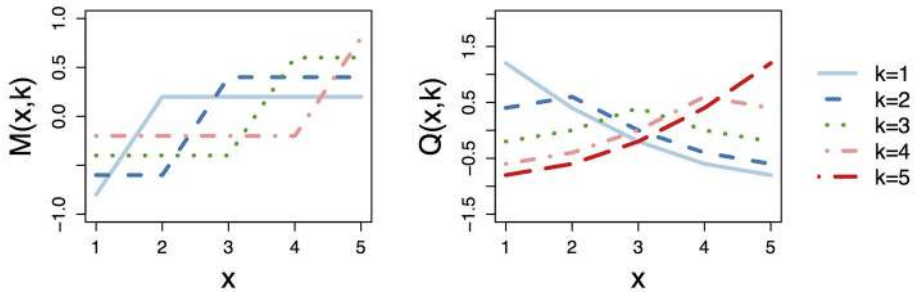

FIGURE 1 | Visualization of the ordinal smoothing spline reproducing kernel function with $K = 5$ levels of X. Left plot shows the columns of the factorized kernel matrix $\mathbf{M}$, whereas right plot shows the columns of reproducing kernel matrix $\mathbf{Q} = \mathbf{M}\mathbf{M}'$.

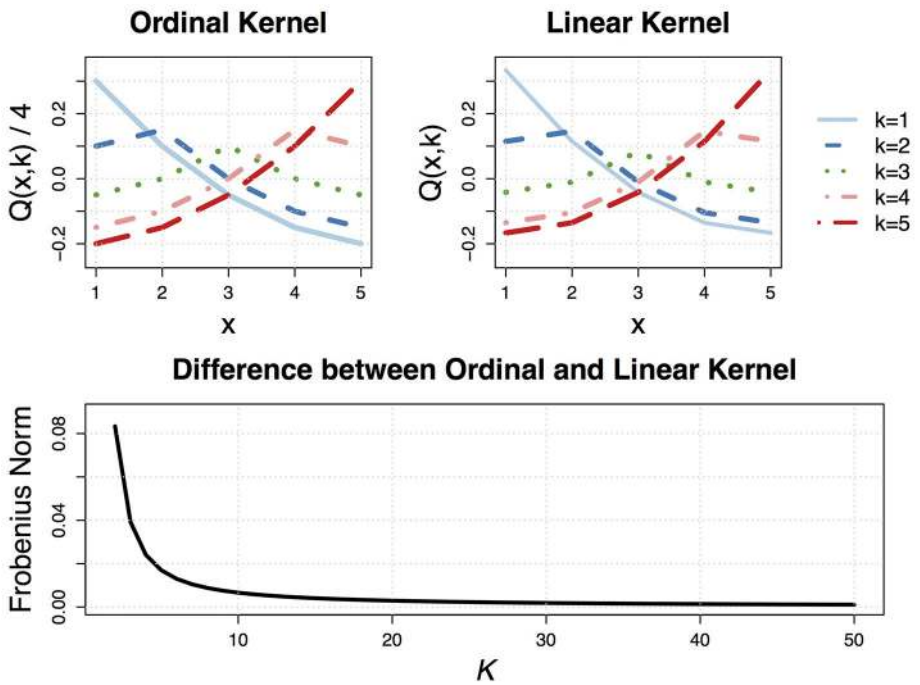

FIGURE 2 | (Top) Visualization of the (standardized) ordinal and linear smoothing spline reproducing kernel functions with $K = 5$ levels of X. **(Bottom)** Frobenius norm between the (standardized) ordinal and linear smoothing spline reproducing kernel matrices $\mathbf{Q}$ as a function of K.

See the Supplementary Material for a proof. Note that Theorem 3.3 implies that ordinal and linear smoothing splines will perform similarly as the number of levels of the ordered set K increases. In **Figure 2**, we plot the reproducing kernel function for the (standardized) ordinal and linear smoothing spline (with $K = 5$), along with the Frobenius norm between the (standardized) ordinal and linear smoothing spline reproducing kernel matrices for different values of K. This figure reveals that the two kernel functions are practically identical for $K \geq 20$, and can look quite similar even for small values of K (e.g., $K = 5$).

3.4. Approximation for Large *K*

If the number of elements of the ordered set K is quite large, then using all K levels as knots would be computationally costly. In such cases, the unconstrained ordinal smoothing spline solution can be fit via the formulas in Equations (14) and (16) with a set of knots $\{\tilde{x}_j\}_{j=1}^{R} \subset \{1, \ldots, K\}$. When monotonicity constraints are needed, the ordinal smoothing spline would still be computationally costly for large K. This is because the factorization of the reproducing kernel into the outer product of monotonic functions, i.e., $\mathbf{Q} = \mathbf{M}\mathbf{M}'$, depends on K. For a scalable approximation to the ordinal smoothing spline isotonic regression estimator, the penalty functional itself can be approximated such as

$$J(\eta) = \sum_{j=2}^{R} [\eta(\tilde{x}_j) - \eta(\tilde{x}_{j-1})]^2 \tag{33}$$

where the knots $1 = \tilde{x}_1 < \cdots < \tilde{x}_R = K$ are assumed to be ordered and unique. The modified RKHS is

$$\mathcal{H} = \{\eta : \eta(x) = \eta(\tilde{x}_j) \text{ if } x \in \{\tilde{x}_{j-1} + 1, \ldots, \tilde{x}_j\}\} \subset \mathbb{R}^K \quad (34)$$

for $x \in \{1, \ldots, K\}$ and $j \in \{2, \ldots, R\}$. The modified RKHS $\mathcal{H}$ has the modified inner product

$$\langle \eta, \xi \rangle = \boldsymbol{\eta}' \left(\frac{1}{R} \tilde{\mathbf{1}}_K \tilde{\mathbf{1}}_K' \right) \boldsymbol{\xi} + \boldsymbol{\eta}' \tilde{\mathbf{D}}' \tilde{\mathbf{D}} \boldsymbol{\xi} \quad (35)$$

where $\tilde{\mathbf{1}}_K$ is a $K \times 1$ vector that contains ones in the positions corresponding to the R knots and zeros elsewhere, and $\tilde{\mathbf{D}}$ is a sparse $R-1 \times K$ first order difference operator matrix with entries defined as

$$\tilde{d}_{ij} = \begin{cases} 1 & \text{if } j = \tilde{x}_{i+1} \\ -1 & \text{if } j = \tilde{x}_i \\ 0 & \text{otherwise.} \end{cases} \quad (36)$$

As an example, with $K = 5$ and $R = 3$ knots $(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = (1, 3, 5)$, the matrix $\tilde{\mathbf{D}}$ has the form

$$\tilde{\mathbf{D}} = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{pmatrix}. \quad (37)$$

Theorem 3.4. *Given the ordinal smoothing spline inner-product in Equation (35), the null space* $\mathcal{H}_0 = \{\eta : \eta(\tilde{x}_1) = \cdots = \eta(\tilde{x}_R)\}$ *has reproducing kernel* $\rho_0(x, y) = 1/R$*, and the contrast space* $\mathcal{H}_1 = \{\eta : \sum_{j=1}^R \eta(\tilde{x}_j) = 0\}$ *has reproducing kernel*

$$\rho_1(x, y) = \sum_{j=1}^{R-1} \left(1_{\{x \leq \tilde{x}_j\}} - \frac{j}{R} \right) \left(1_{\{y \leq \tilde{x}_j\}} - \frac{j}{R} \right)$$

for $x, y \in \{1, \ldots, K\}$ *where* $1_{\{\cdot\}}$ *is an indicator function and* $1 = \tilde{x}_1 < \cdots < \tilde{x}_R = K$ *are the knots.*

See the Supplementary Material for a proof. Note that Theorem 3.4 provides the reproducing kernel needed to compute the ordinal smoothing spline isotonic regression estimator using the knot-based approximation for large K. Specifically, Theorem 3.4 reveals that we can write $\tilde{\mathbf{J}} = \{\rho_1(x_i, \tilde{x}_j)\}_{n \times R} = \tilde{\mathbf{P}} \tilde{\mathbf{M}}'$ and $\tilde{\mathbf{Q}} = \{\rho_1(\tilde{x}_i, \tilde{x}_j)\}_{R \times R} = \tilde{\mathbf{M}} \tilde{\mathbf{M}}'$ where $\tilde{\mathbf{P}} = \{\tilde{p}_{ij}\}_{n \times R-1}$ with $\tilde{p}_{ij} = (j/R) - 1_{\{x_i \leq \tilde{x}_j\}}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, R-1$, and $\tilde{\mathbf{M}} = \{\tilde{m}_{ij}\}_{R \times R-1}$ with $\tilde{m}_{ij} = (j/R) - 1_{\{\tilde{x}_i \leq \tilde{x}_j\}}$ for $i = 1, \ldots, R$ and $j = 1, \ldots, R-1$. Note that $\tilde{\mathbf{P}} = \tilde{\mathbf{G}} \tilde{\mathbf{M}}$ where $\tilde{\mathbf{G}} = \{\tilde{g}_{ij}\}_{n \times R}$ is a selection matrix such that $\tilde{g}_{ij} = 1$ if $x_i = j = 1$ or $x_i \in (\tilde{x}_{j-1}, \tilde{x}_j]$ for $j = 2, \ldots, R$, and $\tilde{g}_{ij} = 0$ otherwise. Using the modified reproducing kernel, the reparameterized ordinal smoothing spline problem is

$$n^{-1} \|\mathbf{y} - d\mathbf{1}_n - \tilde{\mathbf{G}} \tilde{\mathbf{M}} \tilde{\mathbf{c}}_m\|^2 + \lambda \|\tilde{\mathbf{c}}_m\|^2 \quad (38)$$

where $\tilde{\mathbf{c}}_m = \tilde{\mathbf{M}}' \mathbf{c}$, so analogs of the results in Section 3.2 can be applied.

4. SIMULATION STUDY

To investigate the performance of the ordinal smoothing spline, we designed a simulation study that manipulated two conditions: (i) the sample size (4 levels: $n \in \{50, 100, 200, 500\}$), and (ii) the data generating function η (2 forms, see **Figure 3**). To simulate the data, we defined x_i to be an equidistant sequence of n points spanning the interval [0,1], and then defined $y_i = \eta(x_i) + \epsilon_i$ where the error terms were (independently) randomly sampled from a standard normal distribution. Note that each x_i is unique, so $K = n$ in this case. We compared four different methods as a part of the simulation: (a) linear smoothing spline, (b) ordinal smoothing spline, (c) monotonicity constrained ordinal smoothing spline, and (d) isotonic regression implemented through the `isoreg` function in `R` [6]. Methods (a)–(c) were fit using the `bigsplines` package [41] in `R`. For the smoothing spline methods, we used the same

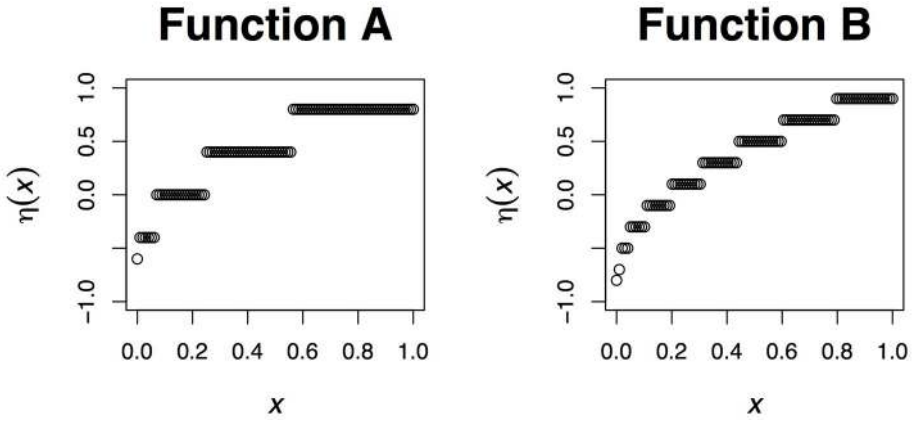

FIGURE 3 | The data generating functions from the simulation study. In each case, the function is defined as $\eta(x) = \frac{1}{K} \sum_{k=1}^{K-1} \text{sign}(\sqrt{x} - x_k^*)$ where the $\{x_k^*\}_{k=1}^{K-1}$ are a sequence of equidistant knots spanning the interval [0,1].

sequence of 20 points as knots to ensure that differences in the results are not confounded by differences in the knot locations. For each of the 8 ($4\ n \times 2\ \eta$) cells of the simulation design, we generated 100 independent samples of data and fit each of the four models to each generated sampled.

To evaluate the performance of each method, we calculated the root mean squared error (RMSE) between the truth η and the estimate $\hat{\eta}$. The RMSE for each method is displayed in **Table 2** and **Figure 4**, which clearly reveal the benefit of the ordinal smoothing spline. Specifically, **Figure 4** reveals that the unconstrained ordinal smoothing spline outperforms the linear smoothing spline for smaller values of $n = K$, and performs similarly to linear smoothing spline for large values of K. **Figure 4** also reveals that the monotonicity constrained estimator (using the knot-approximated reproducing kernel function described in Section 3.4) performed slightly better than the unconstrained ordinal smoothing spline, particularly in the $n = 50$ condition. Furthermore, the simulation results in **Figure 4** demonstrate that the ordinal smoothing spline systematically outperforms R's (default) isotonic regression routine. Thus, the ordinal smoothing spline offers an effective alternative to classic nonparametric and isotonic regression approaches, and—unlike the linear smoothing spline—the ordinal smoothing spline has the benefit of being invariant to any monotonic transformation of x.

TABLE 2 | Median root mean squared error (RMSE) across 100 simulation replications.

Method	Function A				Function B			
	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 50$	$n = 100$	$n = 200$	$n = 500$
lin	0.299	0.230	0.161	0.132	0.280	0.195	0.156	0.120
ord	0.253	0.211	0.166	0.130	0.239	0.193	0.162	0.123
mon	0.248	0.190	0.161	0.128	0.224	0.187	0.148	0.116
iso	0.321	0.269	0.201	0.148	0.325	0.263	0.204	0.154

5. EXAMPLES

5.1. Example 1: Income by Education and Sex

To demonstrate the power of the monotonic ordinal smoothing spline, we use open source data to examine the relationship between income and educational attainment. The data were collected as a part of the 2014 US Census and are freely available from the Integrated Public Use Microdata Series (IPUMS) USA website (https://usa.ipums.org/usa/), which is managed by the Minnesota Population Center at the University of Minnesota. Note that the original data file `usa_00001.dat` contains data from $n^* = 3,132,610$ subjects. However, we restrict our analyses to individuals who were 18+ years old and earned a non-zero income as reported on their 2014 census, resulting in a sample size of $n = 2,214,350$ subjects. We fit the monotonic ordinal smoothing spline separately to the males' data ($n = 1,093,949$) and females' data ($n = 1,120,401$) using all 11 unique education levels as knots. Due to the positive skewness of the income data, we fit the model to $y = \log(\text{income})$. Finally, to ensure our results were nationally representative, we fit the model via weighted penalized least squares using the person weights (`PERWT`) provided with the IPUMS data.

The fit models explain about 15% of the variation in the male and female income data, and reveal noteworthy differences in expected incomes as a function of education and sex. The predicted mean incomes for each educational attainment level and sex are plotted in **Figure 5**, which has some striking trends. First, note that expected income is low (about \$15,000 for males and \$10,000 for females) for individuals without a high school education. Completing high school results in an approximate \$7,000 expected increase in income for men, but only a \$4,500 expected increase in income for women. Similarly, completing 1 year of college results in an expected \$1,700 pay increase for men, but only an expected \$1,000 increase for women. This disturbing trend continues to magnify at each education level, such that women receive a smaller expected return on their education. At the highest level of educational attainment, the gender pay gap is

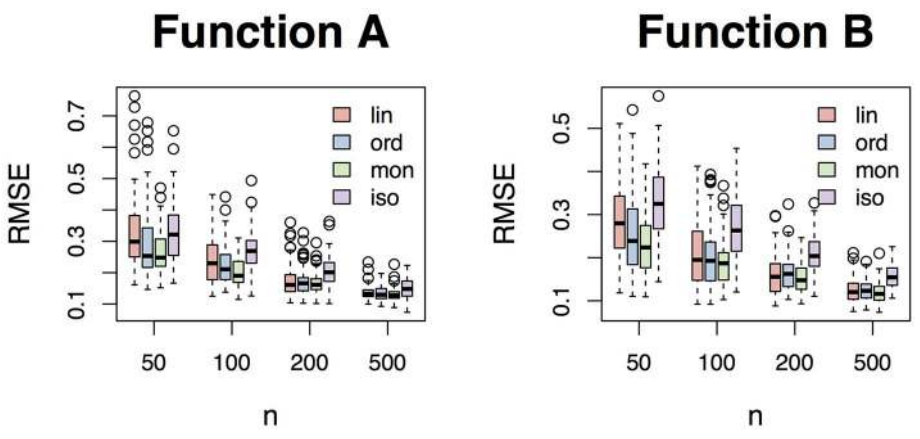

FIGURE 4 | Box plots of the root mean squared error (RMSE) between η and $\hat{\eta}$ for each method.

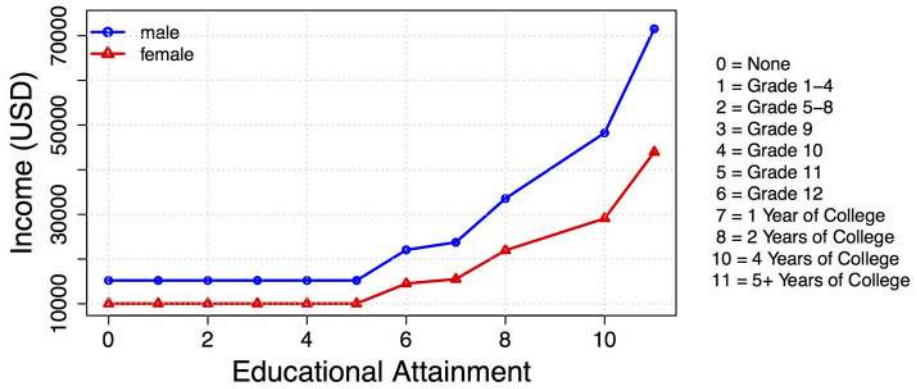

FIGURE 5 | Monotonic ordinal smoothing spline solution showing the relationship between education and income for males (blue circle) and females (red triangle).

most pronounced such that women tend to make about $28,000 less per year than men.

5.2. Example 2: Student Math Performance

As a second example, we use math performance data from Portuguese secondary students. The data were collected by Paulo Cortez at the University of Minho in Portugal [42], and are available from the UCI Machine Learning Repository [43]. In this example, we use the math performance data (student-mat.csv), which contains math exam scores from $n = 395$ Portuguese students. We focus on predicting the students' scores on the first exam during the period (G1). Unlike Cortez et al., we model the first exam (instead of the final grade) because we hope to identify factors that cause students to fall behind early in the semester. By discovering factors that relate to poor math performance on the first exam, it may be possible to create student-specific interventions (e.g., tutoring or more study time) with hopes of improving the final grade.

In **Table 3**, we describe the 15 predictor variables that we include in our model. To model the math exam scores, we fit a semiparametric regression model of the form

$$\begin{aligned} \text{math}_i = \beta_0 &+ \beta_1 \text{school}_i + \beta_2 \text{sex}_i + \beta_3 \text{famsup}_i + \beta_4 \text{paid}_i \\ &+ \beta_5 \text{activities}_i + \beta_6 \text{nursery}_i \quad + \eta_1(\text{age}_i) \\ &+ \eta_2(\text{failures}_i) + \eta_3(\text{absences}_i) + \eta_4(\text{Medu}_i) \\ &+ \eta_5(\text{traveltime}_i) + \eta_6(\text{studytime}_i) + \eta_7(\text{goout}_i) \\ &+ \eta_8(\text{Walc}_i) + \eta_9(\text{health}_i) + \epsilon_i \end{aligned} \tag{39}$$

where math_i is the i-th student's score on the first math exam, β_0 is an unknown regression intercept, $(\beta_1, \ldots, \beta_6)$ are unknown regression coefficients corresponding to the six binary predictors (school, sex, famsup, paid, activities, nursery), η_j is the j-th effect function for $j = 1, \ldots, 9$, and $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$ is the i-th student's model error term. We used a cubic smoothing spline marginal reproducing kernel for the integer valued variables (age, failures, absences) because these variables are measured on a ratio scale. We used the ordinal smoothing spline reproducing kernel (see Theorem 3.1) for the other six nonparametric effects because these variables were measured on an ordinal scale (see **Table 3**). The full smoothing spline solution was fit, i.e., all $n = 395$ data points were used as knots.

TABLE 3 | Predictor variables for math performance example.

Variable	Type	Range/levels
Student's School (school)	Binary	1 = Gabriel Pereira, 0 = Mousinho da Silveira
Student's Sex (sex)	Binary	1 = male, 0 = female
Educational Support (famsup)	Binary	1 = yes, 0 = no
Extra Paid Classes (paid)	Binary	1 = yes, 0 = no
Extra-Curricular Activities (activities)	Binary	1 = yes, 0 = no
Attended Nursery School (nursery)	Binary	1 = yes, 0 = no
Student's Age (age)	Integer	15, 16, …, 22
Number of Failures (failures)	Integer	0, 1, 2, 3
Number of Absences (absences)	Integer	0, 1, …, 75
Mother's Education (Medu)	Ordinal	None, <4th, 5th-9th, secondary, higher
Travel Time to School (traveltime)	Ordinal	<15 min, 15–30 min, 30–60 min, >60 min
Study Time per Week (studytime)	Ordinal	<2 h, 2–5 h, 5–10 h, >10 h
Goes Out with Friends (goout)	Ordinal	1 = very low, …, 5 = very high
Weekend Alcohol Consumption (Walc)	Ordinal	1 = very low, …, 5 = very high
Health Status (health)	Ordinal	1 = very bad, …, 5 = very good

The smoothing parameters were chosen by minimizing the GCV criterion [35]. To avoid the computational expense of simultaneously tuning the GCV criterion with respect to 9 smoothing parameters (one for each η_j), we used a version of Algorithm 3.2 described by Gu and Wahba [25] to obtain initial values of the smoothing parameters. Then we fit the model with the 9 local smoothing parameters fixed at the initialized values, and optimized the GCV criterion with respect to the global smoothing parameter λ. This approach has been shown to produce results that are essentially identical to the fully optimal solution (see [15, 26]).

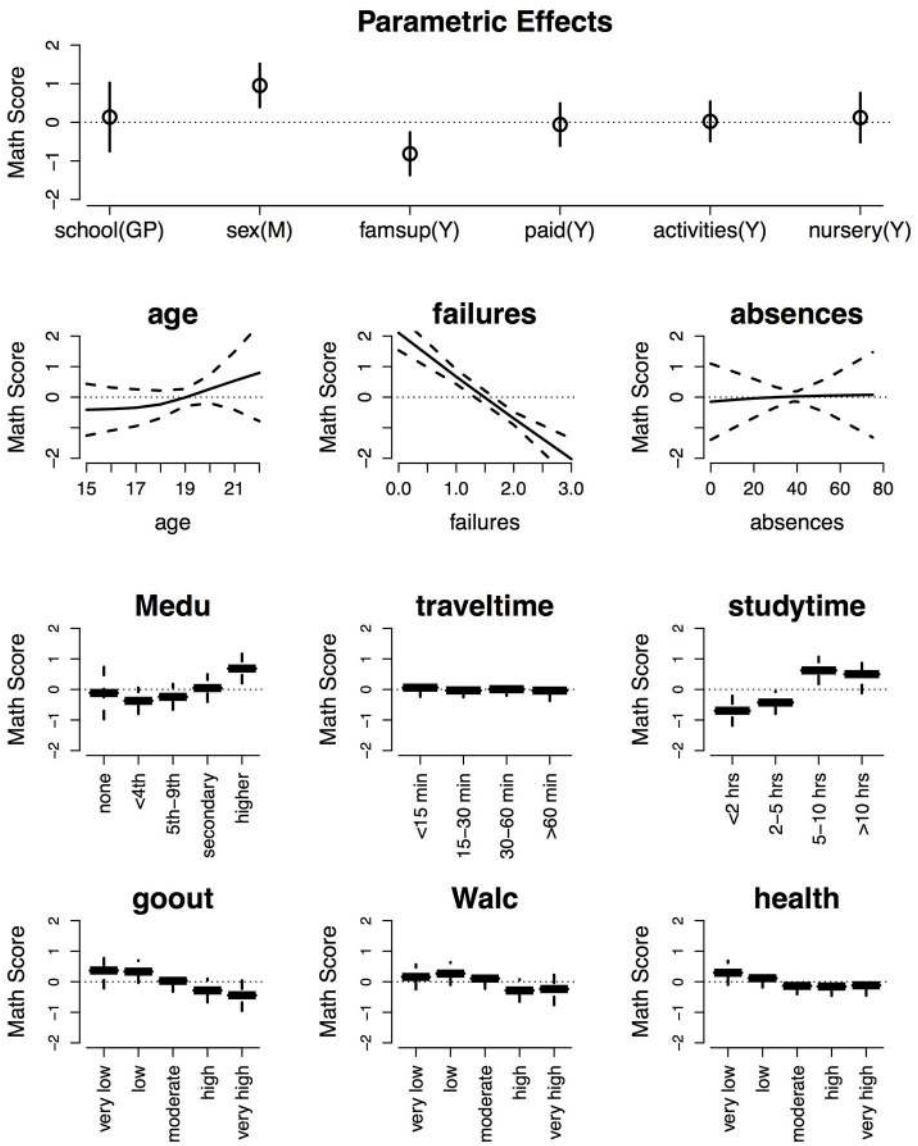

FIGURE 6 | Estimated effects of the different predictor variables on expected math scores with 90% Bayesian confidence intervals (dashed lines) around the predictions.

The fit model explains about 23% of the variation in the students' scores on the first math exam. The estimated regression coefficients $(\hat{\beta}_1, \ldots, \hat{\beta}_6)$ and effect functions $(\hat{\eta}_1, \ldots, \hat{\eta}_9)$ are plotted in **Figure 6**, along with 90% Bayesian confidence intervals [37, 38, 40] around the effect functions. Examining the top row of **Figure 6**, it is evident that only two of the six parametric effects has a significant effect: sex and famsup. The signs and magnitudes of these significant coefficients reveal that (i) males tend to get higher math exam scores than females, and (ii) students who receive extra educational support from their families tend to get lower math exam scores. This second point may seem counter-intuitive, because one may think that extra educational support should lead to higher grades. However, it is likely that the students who receive extra support are receiving this extra support for a reason.

Examination of the remaining subplots in **Figure 6** reveals that the number of prior course failures has the largest (negative) effect on the expected math scores, which is not surprising. There is a slight trend such that larger ages lead to better grades, but the trend is not significantly different from zero according to the 90% Bayesian confidence intervals. Given the other effects in the model, the number of absences has no effect on the expected math scores, which is surprising. Having a mother who completed higher education increases a student's expected math exam score, whereas there were no significant differences between the other four lesser levels of the mother's

education. Studying for 5 or more hours per week increases a student's expected scores, whereas studying less than 2 h per week decreases a student's expected scores. Going out with friends and/or drinking on the weekend with high or very high frequency significantly decreases a student's expected math exam scores. Given the other effects, travel time to school and a student's health did not have significant effects on the math exam scores.

6. DISCUSSION

6.1. Summary of Findings

Our simulation and real data examples reveal the flexibility and practical potential of the ordinal smoothing spline. The simulation study investigated the potential of the ordinal smoothing spline for isotonic regression, as well as the relationship between the ordinal and linear smoothing spline reproducing kernel functions. The simulation results demonstrated that (i) the ordinal smoothing spline can outperform the linear smoothing spline at small samples, (ii) the ordinal smoothing spline performs similar to the linear smoothing spline for large K, and (iii) monotonic ordinal smoothing splines can outperform standard isotonic regression approaches. Thus, the simulation study illustrates the results in Theorems 3.1–3.3, and also reveals that the knot-approximated reproducing kernel proposed in Theorem 3.4 offers an effective approximation to the monotonic ordinal smoothing spline solution.

The first example (income by education and sex) offers a practical example of the potential of the ordinal smoothing spline for discovering monotonic trends in data. Unlike classic approaches to isotonic regression, the ordinal smoothing spline uses a reproducing kernel (and knot-based) approach, making it easily scalable to large survey samples such as the US Census data. Using a large and nationally representative sample of US citizens ages 18+, this example reveals a clear gender pay gap, such that women receive less return on their educational investments, i.e., less pay for the same level of education. And note that the predictor could be monotonically transformed without changing the ordinal smoothing spline solution, which is one of the primary benefits of using the ordinal smoothing spline estimator for variables such as Educational Attainment—which has no clear unit of measurement.

The second example (math grades) demonstrates the effectiveness of using ordinal smoothing splines to include multiple ordinal predictors in a regression model along with other nominal and metric (continuous) predictors. Furthermore, the second example illustrates that many ordinal relationships do not have a clear linear in/decrease across the K levels of the ordinal variable. This reveals that it could be difficult and/or misleading to code ordinal predictors as if they were continuous (e.g., linear) effects. So, despite this common practice in the social sciences, our results make it clear that this sort of approach does not have an obvious solution for all ordinal variables, and could be particularly problematic when multiple ordinal variables are included in the same model.

6.2. Concluding Remarks

This paper discusses the ordinal smoothing spline, which can be used to model functional relationships between ordered categorical predictor variables and any exponential family response variable. This approach makes it straightforward to incorporate one (or more) ordinal predictors in a regression model, and also provides a theoretically sound method for examining interaction effects between ordinal and other predictors. In this paper, we present the ordinal smoothing spline reproducing kernel function (Theorem 3.1), which has the benefit of being invariant to any monotonic transformation of the predictor. We also discuss how the ordinal smoothing spline estimator can be adjusted to impose monotonicity constraints (Theorem 3.2). Furthermore, we reveal an interesting asymptotic relationship between ordinal and linear smoothing splines (Theorem 3.3), and develop large sample approximations for ordinal smoothing splines (Theorem 3.4). Finally, we have demonstrated the practical potential of ordinal smoothing splines using a simulation study (Section 4) and real data examples from two distinct disciplines (Section 5).

In nearly all applications of the GLM or GzLM, ordered categorical predictors are treated as either unordered categorical predictors or metric predictors for modeling purposes. In most cases, it is not obvious which approach one should choose. Neither approach is ideal for ordinal data, so one must ultimately decide which aspect of the data should be misrepresented in the model. Treating ordinal data as unordered (by definition) ignores the ordinal nature of the data, which is undesirable. Whereas, treating ordinal data as metric (continuous) ignores the discrete nature of the data, which is undesirable. In contrast, the ordinal smoothing spline is the obvious solution to this problem, because the ordinal smoothing spline (i) respects the ordinal nature of the predictor, and (ii) provides a theoretically sound framework for incorporating ordinal predictors in regression models.

AUTHOR CONTRIBUTIONS

The sole author (NH) contributed to all aspects of this work.

FUNDING

The sole author (NH) was funded by start-up funds from the University of Minnesota.

ACKNOWLEDGMENTS

The author acknowledges the two data sources: (i) the Minnesota Population Center at the University of Minnesota, and (ii) the Machine Learning Repository at University of California Irvine.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fams.2017.00015/full#supplementary-material

REFERENCES

1. Christensen R. *Plane Answers to Complex Questions: The Theory of Linear Models. 3rd Edn*. New York, NY: Springer-Verlag (2002).
2. Chartier S, Faulkner A. General linear models: an integrated approach to statistics. *Tutorial Quant Methods Psychol*. (2008) **4**:65–78. doi: 10.20982/tqmp.04.2.p065
3. Graham JM. The general linear model as structural equation modeling. *J Educ Behav Stat*. (2008) **33**:485–506. doi: 10.3102/1076998607306151
4. Casals M, Girabent-Farrés M, Carrasco JL. Methodological quality and reporting of generalized linear mixed models in clinical medicine (2000–2012): a systematic review. *PLoS ONE* (2014) **9**:e112653. doi: 10.1371/journal.pone.0112653
5. de Jong P, Heller GZ. *Generalized Linear Models for Insurance Data*. Cambridge: Cambridge University Press (2008).
6. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna (2017). Available online at: http://www.R-project.org/
7. McCullagh P, Nelder JA. *Generalized Linear Models. 2nd Edn*. London: Chapman and Hall (1989).
8. McCullagh P. Regression model for ordinal data (with discussion). *J R Stat Soc Ser B* (1980) **42**:109–42.
9. Armstrong B, Sloan M. Ordinal regression models for epidemiologic data. *Am J Epidemiol*. (1989) **129**:191–204. doi: 10.1093/oxfordjournals.aje.a115109
10. Peterson B, Harrell FE Jr. Partial proportional odds models for ordinal response variables. *J R Stat Soc Ser C* (1990) **39**:205–17. doi: 10.2307/2347760
11. Cox C. Location-scale cumulative odds models for ordinal data: a generalized non-linear model approach. *Stat Med*. (1995) **14**:1191–203. doi: 10.1002/sim.4780141105
12. Liu I, Agresti A. The analysis of ordinal categorical data: an overview and a survey of recent developments. *Test* (2005) **14**:1–73. doi: 10.1007/BF02595397
13. Gertheiss J, Tutz G. Penalized regression with ordinal predictors. *Int Stat Rev*. (2009) **77**:345–65. doi: 10.1111/j.1751-5823.2009.00088.x
14. Gertheiss J. *ordPens: Selection and/or Smoothing of Ordinal Predictors*. R package version 0.3-1 (2015). Available online at: https://CRAN.R-project.org/package=ordPens
15. Gu C. *Smoothing Spline ANOVA Models. 2nd Edn*. New York, NY: Springer-Verlag (2013). doi: 10.1007/978-1-4614-5369-7
16. Wahba G. *Spline Models for Observational Data*. Philadelphia, PA: Society for Industrial and Applied Mathematics (1990).
17. Riesz F. Sur une espèce de géométrie analytique des systèmes de fonctions sommables. *C R Acad Sci Paris* (1907) **144**:1409–11.
18. Riesz F. Sur les opérations fonctionnelles linéaires. *C R Acad Sci Paris* (1909) **149**:974–7.
19. Aronszajn N. Theory of reproducing kernels. *Trans Am Math Soc*. (1950) **68**:337–404. doi: 10.1090/S0002-9947-1950-0051437-7
20. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge: Cambridge University Press (2003).
21. Hastie T, Tibshirani R. *Generalized Additive Models*. New York, NY: Chapman and Hall/CRC (1990).
22. Ramsay JO, Silverman BW. *Functional Data Analysis. 2nd Edn*. New York, NY: Springer (2005).
23. Wood SN. *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman & Hall (2006).
24. Kimeldorf G, Wahba G. Some results on Tchebycheffian spline functions. *J Math Anal Appl*. (1971) **33**:82–95. doi: 10.1016/0022-247X(71)90184-3
25. Gu C, Wahba G. Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J Sci Stat Comput*. (1991) **12**:383–98. doi: 10.1137/0912021
26. Helwig NE, Ma P. Fast and stable multiple smoothing parameter selection in smoothing spline analysis of variance models with large samples. *J Comput Graph Stat*. (2015) **24**:715–32. doi: 10.1080/10618600.2014.926819
27. Duchon J. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In Schempp W, Zeller K editors. *Constructive Theory of Functions of Several Variables. Lecture Notes in Mathematics,* Vol. 571. Berlin; Heidelberg: Springer (1977). p. 85–100.
28. Meinguet J. Multivariate interpolation at arbitrary points Made simple. *J Appl Math Phys*. (1979) **30**:292–304. doi: 10.1007/BF01601941
29. Wood SN. Thin plate regression splines. *J R Stat Soc Ser B* (2003) **65**:95–114. doi: 10.1111/1467-9868.00374
30. Gu C, Kim YJ. Penalized likelihood regression: general formulation and efficient approximation. *Can J Stat*. (2002) **30**:619–28. doi: 10.2307/3316100
31. Kim YJ, Gu C. Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *J R Stat Soc Ser B* (2004) **66**:337–56. doi: 10.1046/j.1369-7412.2003.05316.x
32. Helwig NE, Ma P. Smoothing spline ANOVA for super-large samples: scalable computation via rounding parameters. *Stat Its Interface* (2016) **9**:433–44. doi: 10.4310/SII.2016.v9.n4.a3
33. Moore EH. On the reciprocal of the general algebraic matrix. *Bull Am Math Soc*. (1920) **26**:394–5.
34. Penrose R. A generalized inverse for matrices. *Math Proc Cambridge Philos Soc*. (1950) **51**:406–13. doi: 10.1017/S0305004100030401
35. Craven P, Wahba G. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* (1979) **31**:377–403. doi: 10.1007/BF01404567
36. Kimeldorf G, Wahba G. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann Math Stat*. (1970) **41**:495–502. doi: 10.1214/aoms/1177697089
37. Wahba G. Bayesian "confidence intervals" for the cross-validated smoothing spline. *J R Stat Soc Ser B* (1983) **45**:133–50.
38. Nychka D. Bayesian confidence intervals for smoothing splines. *J Am Stat Assoc*. (1988) **83**:1134–43. doi: 10.1080/01621459.1988.10478711
39. Gu C. Penalized likelihood regression: a Bayesian analysis. *Stat Sin*. (1992) **2**:255–64.
40. Gu C, Wahba G. Smoothing spline ANOVA with component-wise Bayesian "confidence intervals". *J Comput Graph Stat*. (1993) **2**:97–117.
41. Helwig NE. *Bigsplines: Smoothing Splines for Large Samples*. R package version 1.1-0 (2017). Available online at: http://CRAN.R-project.org/package=bigsplines
42. Cortez P, Silva A. Using data mining to predict secondary school student performance. In Brito A, Teixeira J editors. *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*. Porto: EUROSIS (2008). p. 5–12.
43. Lichman M. *UCI Machine Learning Repository*. (2013). Available online at: http://archive.ics.uci.edu/ml
44. Liew CK. Inequality constrained least-squares esitmation. *J Am Stat Assoc*. (1976) **71**:746–51.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.