Semi-nonparametric smooth isotonic regression

Xianzheng Huang

Department of Statistics, University of South Carolina, Columbia, South Carolina 29208, U.S.A.

Abstract

We propose a new method for smooth isotonic regression analysis. Unlike most existing methods for isotonic regression, the proposed method is akin to parametric regression without order restriction. To account for smoothness and isotonicity simultaneously, we exploit the flexible class of semi-nonparametric densities to model isotonic regression functions. Under this framework, the full range of inference techniques for parametric regression models become applicable for model estimation and model validation in isotonic regression.

Keywords: Cross validation; Linearity; Semi-nonparametric (SNP).

1 Introduction

Due to physical considerations, it is often assumed in regression models that the mean of a response is a monotone function of a predictor. Examples includes models for growth curves, those in dose-response analysis, and models arise in reliability theory. Imposing monotonicity in regression analysis when it is scientifically well justified can enhance the efficiency of statistical inference and the interpretability of inference results.

The majority of existing methods for isotonic regression are nonparametric in nature. Among early developments, the most well-known algorithm is the pool-adjacent-violator algorithm (PAVA) (Robertson, Wright, and Dykstra, 1988), which yields a piecewise linear isotonic function as the estimated regression function. The outcome from PAVA is unsatisfactory when a smooth functional relationship is desired. To achieve smooth regression functions, methods based on splines and kernels were proposed (Delecroix and Thomas-Agnan, 2000). Ramsay (1988, 1998), Kelly and Rice (1990), Mammen and Thomas-Agnan (1999), Wang and Feng (2008), among several others, studied smooth splines regression under order constraints. Mammen (1991) realized smooth isotonic regression analyses in two separate steps successively in two different orders, which are an isotonizing step through PAVA and a smoothing step via kernel estimation. Hall and Huang (2001) proposed a method to adjust the weights in a kernel estimator to satisfy the monotonicity constraint. Dette, Neumeyer, and Pilz (2006) avoided constrained optimization by combining unconstrained regression and density estimation, which requires users to specify two kernels and two bandwidths. In general, the strategies that exploit smoothing splines or kernels entail choice of knots and tuning parameter for smoothness penalty or choice of kernel and bandwidth. Recently, Wang (2011) modeled the regression function using Bernstein polynomials, with constraints imposed on the coefficients associated with the Bernstein polynomial basis vectors to achieve monotonicity of the regression function. Under the Bayesian framework, Bornkamp and Ickstadt (2009) proposed nonparametric isotonic regression, where they used a mixture of shifted and scaled para-

metric probability distribution functions to model the regression function. Meyer, Hackstadt, and Hoeting (2011) modeled smooth isotonic regression functions using quadratic *B*-spline basis functions, and developed a reversible-jump Markov chain Monte Carlo (MCMC) algorithm to allow for free knots. A battery of statistical inference techniques for nonparametric isotonic regression have been developed by Bartholomew (1959), Barlow, et al. (1972), Banerjee and Wellner (2001), Banerjee, Biswas, and Ghosh (2006), Banerjee (2007), and Pal and Banerjee (2008), among many others.

In this article we propose a new method that achieves smoothness and isotonicity simultaneously in one modeling step based on a class of semi-nonparametric (SNP) densities defined by Gallant and Nychka (1987). The proposed method does not involve choice of knots, smoothing penalty, or bandwidth, although, as discussed in Section 3, it requires one to choose a quantity that controls the flexibility of the regression function being modeled. The idea behind the new method is first motivated by the elementary concept that a valid cumulative distribution function (cdf) is nondecreasing. Hence one can construct an isotonic function based on a cdf. To attain smoothness and flexibility, we use the cdf of a flexible smooth distribution family and relax constraints required for the validity of a cdf, such as ranging from zero to one, to model the regression function. The use of SNP representation for the regression function greatly simplifies follow-up inference procedures such as parameter estimation, standard error estimation, and model validation. In effect, the proposed method sets isotonic regression, traditionally treated nonparametrically, back in the parametric framework without order restriction and makes the full range of parametric inferential techniques applicable.

Because the SNP representation is the backbone of the proposed method, we devote Section 2 to introducing the particular family of SNP used in this article and reviewing existing relevant works that employ SNP. We then present the new approach for isotonic regression in Section 3. A test for linearity of the regression function is also developed in this section. Section 4 presents simulation studies to illustrate the implementation and performance of the proposed methods. In

Section 5, these methods are applied to a real data example. In Section 6, we point out future research that refine the current proposal and make further use of SNP in regression analyses.

2 Semi-nonparametric representation

Gallant and Nychka (1987) defined a class of flexible probability density functions (pdf) termed as semi-nonparametric densities. The construction of SNP densities involves Sobolev norm, which is reviewed next. For a function $f(\mathbf{z})$ on the support of d-dimensional real space, $\mathbb{R}^d$, the Sobolev norm with respect to a weight function $w(z)$ is defined by

$$\|f\|_{m,p,w} = \begin{cases} \left\{ \sum_{|\lambda|\leq m} \int |D^\lambda f(\mathbf{z})|^p w(\mathbf{z}) d\mathbf{z} \right\}^{1/p} & \text{if } 1 \leq p < \infty \\ \max_{|\lambda|\leq m} \sup_{\mathbf{z}\in\mathbb{R}^d} |D^\lambda f(\mathbf{z})| w(\mathbf{z}) & \text{if } p = \infty, \end{cases} \tag{1}$$

where

$$D^\lambda f(\mathbf{z}) = \left(\frac{\partial^{\lambda_1}}{\partial z_1^{\lambda_1}} \right) \cdots \left(\frac{\partial^{\lambda_d}}{\partial z_d^{\lambda_d}} \right) f(\mathbf{z})$$

is the partial derivative of $f(\mathbf{z})$, $\lambda = (\lambda_1, \ldots, \lambda_d)$, and $|\lambda| = \sum_{k=1}^d \lambda_k$. The class of SNP densities is defined by

$$\mathcal{H} = \left\{ h(\mathbf{z}) : h(\mathbf{z}) = f^2(\mathbf{z}) + \epsilon_0 h_0(\mathbf{z}) \right\}, \tag{2}$$

where ϵ_0 is some small positive number, $h_0(\mathbf{z})$ is a strictly positive density function that satisfies $\|h_0\|_{m_0,2,w_0} < \mathcal{B}_0$ for some positive bound $\mathcal{B}_0$, $m_0 > d/2$, $w_0(\mathbf{z}) = (1 + \mathbf{z}^T\mathbf{z})^{\delta_0}$, $\delta_0 > d/2$, and $f(\mathbf{z})$ also satisfies $\|f\|_{m_0,2,w_0} < \mathcal{B}_0$. The effect of setting an upper bound $\mathcal{B}_0$ on $\|f\|_{m_0,2,w_0}$ and $\|h_0\|_{m_0,2,w_0}$ is to impose certain degree of smoothness restriction on $h(z)$. Adding $\epsilon_0 h_0(\mathbf{z})$ in $h(\mathbf{z})$ is to force a lower bound to avoid zero density. Assuming the true density, $h^*(\mathbf{z})$, belongs to $\mathcal{H}$, Gallant and Nychka (1987) showed that the number of derivatives of $h^*(\mathbf{z})$ that can be estimated consistently via maximum likelihood is $m_0 - d/2$. In practice, $\epsilon_0 h_0(\mathbf{z})$ is usually omitted in $h(\mathbf{z})$ without causing

noticeable numerical effect. In most existing SNP literature, $f^2(\mathbf{z})$ is approximated by the following truncated Hermite series,

$$\left\{\sum_{|\lambda|<K} a_\lambda(\mathbf{R}^{-1}\mathbf{z})^\lambda\right\}^2 \mathbf{R}^{-1}\exp\left(-\mathbf{z}^T\mathbf{R}^{-T}\mathbf{R}^{-1}\mathbf{z}/2\right), \tag{3}$$

which satisfies

$$\lim_{K\to\infty}\left\| f(\mathbf{z}) - \sum_{|\lambda|<K} a_\lambda(\mathbf{R}^{-1}\mathbf{z})^\lambda \sqrt{\mathbf{R}^{-1}\exp\left(-\mathbf{z}^T\mathbf{R}^{-T}\mathbf{R}^{-1}\mathbf{z}/2\right)}\right\|_{m_0-d/2,\infty,w_0} = 0,$$

where $\{a_\lambda, |\lambda| < K\}$ are a series of coefficients constrained so that (3) integrates to one over the support of $\mathbf{z}$, $\mathbf{R}^{-T}$ denotes the transpose of $\mathbf{R}^{-1}$, and for $\mathbf{t} \in \mathbb{R}^d$, $\mathbf{t}^\lambda = \prod_{l=1}^d t_l^{\lambda_l}$. Simply put, $\sum_{|\lambda|<K} a_\lambda(\mathbf{R}^{-1}\mathbf{z})^\lambda$ is a polynomial in $\mathbf{z}$ of order K. Noticing that $\mathbf{R}^{-1}\exp\left(-\mathbf{z}^T\mathbf{R}^{-T}\mathbf{R}^{-1}\mathbf{z}/2\right)$ is the kernel of a normal distribution with mean zero and variance-covariance matrix $\mathbf{R}^T\mathbf{R}$, a variant of (3) is given by

$$\left\{P_K(\mathbf{R}^{-1}\mathbf{z})\right\}^2 \mathbf{R}^{-1}\phi(\mathbf{R}^{-1}\mathbf{z}), \tag{4}$$

where $P_K(\mathbf{t})$ is a polynomial of $\mathbf{t} \in \mathbb{R}^d$ of order K, and $\phi(\cdot)$ denotes the d-dimensional standard normal pdf.

To recap, the class of SNP densities for practical use contains pdf members whose format is given by (4). The structure of a polynomial (quantity squared) multiplying a normal pdf leads to both mathematical and computational convenience, which makes SNP a popular tool for modeling distributions, especially when one wishes to avoid restrictive distributional assumptions. For example, Davidian and Gallant (1993) modeled random effects in nonlinear mixed models using SNP; Zhang and Davidian (2001) utilized SNP densities to model random effects in linear mixed models; Chen, Zhang, and Davidian (2002) applied SNP to random effects in generalized linear mixed models; Song, Davidian, and Tsiatis (2002a,b) employed SNP densities for random effects in joint models; Zhang and Davidian (2008) adopted SNP representation for time-to-event to allow arbitrarily censoring patterns; Irincheeva, Cantoni, and Genton (2012) used SNP to specify

the distribution of latent variables in generalized linear latent variable models. In contrast to these works, where SNP representation is exclusively used for distribution construction, our use of SNP as the basis of formulating smooth isotonic functions is a new contribution. The flexibility and smoothness of SNP pdf are inherited in the resulting isotonic regression functions.

Besides convenience in numerical implementation, using SNP to approximate density functions and regression functions also has solid theoretical justification. Gallant and Nychka (1987) showed that, if $h^*(\mathbf{z}) \in \mathcal{H}$, letting n be the sample size and $\hat{h}(\mathbf{z})$ be the estimated density obtained via maximum likelihood when (3) is used to approximate $h^*(\mathbf{z})$, then $\lim_{n\to\infty} \|\hat{h} - h^*\|_{m_0-d/2,\infty,w_0} = 0$ almost surely when $\lim_{n\to\infty} K_n = \infty$. In other words, $\hat{h}$ is consistent in the notion of Sobolev norm for appropriately chosen K_n. Here, K_n is K in (3), with a subscript to stress its dependence on n when studying asymptotics. Fenton and Gallant (1996a) further showed that, in order to achieve consistent $\hat{h}$, the rate at which K_n approaches infinity as $n \to \infty$ should depend on the highest order of derivative one assumes for $h^*(\mathbf{z})$. Moreover, a direct consequence of this consistency is that functionals of $h^*(\mathbf{z})$, such as $\int g(\mathbf{z})h^*(\mathbf{z})d\mathbf{z}$, can also be estimated consistently in the same notion, for some function $g(\mathbf{z})$. This property is especially important for our study because, assuming $d = 1$, we will base the construction of an isotonic function on the particular functional of $h^*(z)$ given by $H^*(z) = \int_{z_L}^{z} h^*(t)dt$, where z_L is the lower bound of the support of interest. Because an isotonic function may not range from zero to one, as a cdf does, the constraint on the coefficients, a_λ, in (3) is relaxed. More specifically, we define a rich class of flexible smooth isotonic functions as follows,

$$\mathcal{G} = \left\{ H(z) : H(z) = \int_{z_L}^{z} \left\{P_K(\mathbf{R}^{-1}t)\right\}^2 \mathbf{R}^{-1}\phi(\mathbf{R}^{-1}t)dt + s,\ s \in \mathbb{R},\ K = 1, 2, \ldots. \right\},$$

where the constraint on a_λ in $P_K(\cdot)$ required in (3) is removed, and $s(= H(z_L))$ is a location parameter (noting that $H^*(z_L) = 0$). If the true isotonic function falls in this rich SNP class, then the isotonic function can be estimated consistently via maximum likelihood as long as one chooses K appropriately. Relating to the I-splines used to model monotone regression functions, e.g., in

Ramsay (1988) and Meyer (2008), the construction of $H^*(z)(= \int_{Z_L}^{z} h^*(t)dt)$ is similar to that of the *I*-splines, which integrate over *M*-splines.

3 SNP isotonic regression

3.1 Model and estimation

Suppose that the observed data, $\{(Y_i, X_i),\ i = 1, \ldots, n\}$, are independent and identically distributed (i.i.d.) realizations of the following regression model,

$$Y = \theta(X) + \epsilon, \tag{5}$$

where $\theta(X)$ is a smooth isotonic function, and ϵ is the random error independent of X. The central interest is to estimate $\theta(\cdot)$. For ease in exposition, we consider a scalar X for the majority of this article. If one assumes that the true regression function, $\theta^*(x)$, belongs to $\mathcal{G}$, then $\theta^*(x)$ can be formulated as the integral of a truncated Hermite series for some K plus a shift,

$$\theta_K^*(x) = \int_{x_L}^{x} \{P_K(t)\}^2 e^{-t^2/2} dt + s, \tag{6}$$

where $s = \theta^*(x_L)$ and x_L denotes the lower bound of the support of X, $\mathcal{X}$. By the general SNP theory reviewed in Section 2, if $\theta^*(\cdot) \in \mathcal{G}$, then applying the maximum likelihood method to the resultant parametric model leads to an estimated regression function $\hat{\theta}(\cdot)$ that is consistent in the following sense,

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} |\hat{\theta}(x) - \theta^*(x)|(1 + x^2) = 0 \text{ as } K \to \infty. \tag{7}$$

Elaborating and translating the arguments in Section 2 in the current context, for $\theta^*(\cdot) \in \mathcal{G}$, its derivative $\theta^{*\prime}(x)$ behaves like $f^2(\cdot)$ in (2), except for the integrate-to-one constraint. And thus, parallel to the condition on $f(\cdot)$ following (2), $\theta^*(\cdot) \in \mathcal{G}$ implies that

$$\int_{\mathcal{X}} \theta^{*\prime}(x)(1 + x^2)dx + 0.25 \int_{\mathcal{X}} \frac{\left\{\theta^{*\prime\prime}(x)\right\}^2}{\theta^{*\prime}(x)}(1 + x^2)dx < C, \tag{8}$$

for some positive constant C, where $\theta^{*''}(x)$ is the second derivative of $\theta^*(x)$. In plain language, in order to achieve (7) for the estimated regression function resulting from maximum likelihood method, the underlying true regression function needs to be sufficiently smooth. Regression functions that violate (8) include those with abrupt jumps, sudden kinks, or oscillatory behaviors within $\mathcal{X}$, as these behaviors can cause the integrals in (8) blow up.

It is straightforward to program the regression function in (6) for any K because of the following easily derived recursive formulas. Denote by $\Phi(\cdot)$ the standard normal cdf. Define $\gamma_K(x) = \int_{x_L}^{x} \{P_K(t)\}^2 e^{-t^2/2} dt$ and $I_K(x) = \int_{x_L}^{x} t^K e^{-t^2/2} dt$, for $K = 0, 1, \ldots$. Then $\theta_K^*(x) = \gamma_K(x) + s$, where

$$
\begin{aligned}
\gamma_0(x) &= a_0^2 I_0(x), \\
\gamma_K(x) &= \gamma_{K-1}(x) + a_K^2 I_{2K}(x) + 2a_K \sum_{j=0}^{K-1} a_j I_{j+K}(x), \text{ for } K > 0, \text{ in which} \\
I_1(x) &= \exp\left(-x_L^2/2\right) - \exp\left(-x^2/2\right), \text{ and} \\
I_K(x) &= x_L^{K-1} \exp\left(-x_L^2/2\right) - x^{K-1} \exp\left(-x^2/2\right) + (K-1) I_{K-2}(x), \text{ for } K > 1.
\end{aligned}
$$

Substituting $\theta(x)$ in (5) with $\theta_K^*(x)$ yields an explicit parametric form. Define $\boldsymbol{\tau} = (a_0, a_1, \ldots, a_K, s)^T$, and let σ_ϵ be the unknown parameter(s) in the distribution of ϵ. Then $\boldsymbol{\Omega} = (\boldsymbol{\tau}^T, \sigma_\epsilon)^T$ is the collection of unknown parameters in a parametric model as an SNP representation of (5). Once $\boldsymbol{\tau}$ is estimated, an explicit estimated regression function becomes available for prediction and for use in other inference procedures, which is one advantage of the new method compared to spline-based and kernel-based methods. This appealing feature is underscored by Fenton and Gallant (1996b), who stated that SNP estimators "compress the information in the data to a set of coefficients whose number is a fractional power of the sample size" whereas estimators from methods using spline or kernel "must be recreated afresh from the data at every use." They also provided results to support that SNP estimators are "both quantitatively and asymptotically similar to the kernel estimator which is optimal". The advantages of the SNP method compared to kernel and spline methods in modeling regression functions are discussed in greater details in Eastwood and Gallant (1991).

The key message delivered in their article is that, with very similar statistical properties for the final inference results, the SNP method is more convenient in many aspects of implementation. Such advantages carry over to modeling isotonic regression functions.

If $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, then $Y_i|X_i \sim N(\theta(X_i;\boldsymbol{\tau}), \sigma_\epsilon^2)$, for $i = 1, \ldots, n$. The maximum likelihood estimator (MLE) for $\boldsymbol{\tau}$, denoted by $\hat{\boldsymbol{\tau}}$, is equivalent to the least squares estimator,

$$\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}\in\mathbb{R}^{p_\tau}} \sum_{i=1}^{n} \left\{Y_i - \theta_K^*(X_i;\boldsymbol{\tau})\right\}^2,$$

where p_τ is the dimension of $\boldsymbol{\tau}$. An estimator for σ_ϵ^2 is given by the residual sum of squares adjusted by the degrees of freedom,

$$\hat{\sigma}_\epsilon^2 = (n - p_\tau)^{-1} \sum_{i=1}^{n} \left\{Y_i - \theta_K^*(X_i;\hat{\boldsymbol{\tau}})\right\}^2. \tag{9}$$

Equivalently, estimating $\boldsymbol{\Omega}$ for a fixed K is to solve the following system of estimating equations for $\boldsymbol{\Omega}$,

$$\sum_{i=1}^{n} \boldsymbol{\psi}(Y_i, X_i; \boldsymbol{\Omega}) = \sum_{i=1}^{n} \begin{pmatrix} \boldsymbol{\psi}_{\boldsymbol{\tau}}(Y_i, X_i; \boldsymbol{\tau}) \\ \boldsymbol{\psi}_{\sigma_\epsilon^2}(Y_i, X_i; \boldsymbol{\Omega}) \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} -n^{-1}\left\{Y_i - \theta_K^*(X_i;\boldsymbol{\tau})\right\} \dfrac{\partial \theta_K^*(X_i;\boldsymbol{\tau})}{\partial \boldsymbol{\tau}} \\ n^{-1}\sigma_\epsilon^2 - (n-p_\tau)^{-1}\left\{Y_i - \theta_K^*(X_i;\boldsymbol{\tau})\right\}^2 \end{pmatrix} = \mathbf{0}. \tag{10}$$

The Jacobian corresponding to (10) is given by

$$\mathbf{A} = \begin{pmatrix} \displaystyle\sum_{i=1}^{n} n^{-1}\left[\frac{\partial \theta_K^*(X_i;\boldsymbol{\tau})}{\partial \boldsymbol{\tau}} \frac{\partial \theta_K^*(X_i;\boldsymbol{\tau})}{\partial \boldsymbol{\tau}^T} - \left\{Y_i - \theta_K^*(X_i;\boldsymbol{\tau})\right\} \frac{\partial^2 \theta_K^*(X_i;\boldsymbol{\tau})}{\partial \boldsymbol{\tau}\partial \boldsymbol{\tau}^T}\right] & \mathbf{0} \\ \displaystyle\sum_{i=1}^{n} 2(n-p_\tau)^{-1}\left\{Y_i - \theta_K^*(X_i;\boldsymbol{\tau})\right\} \frac{\partial \theta_K^*(X_i;\boldsymbol{\tau})}{\partial \boldsymbol{\tau}^T} & 1 \end{pmatrix}. \tag{11}$$

Based on (10) and (11), both of which can be derived explicitly, one can compute the sandwich variance estimator for $\hat{\boldsymbol{\Omega}}$ according to the M-estimation theory. An estimator for the variance of $\hat{\theta}(x_0) = \theta_K^*(x_0;\hat{\boldsymbol{\tau}})$ can also be obtained via the Delta method for any particular value of interest, x_0, within a plausible range. Also, confidence intervals for functionals of $\theta^*(x)$ are readily available. All these inference procedures are equally straightforward if one assumes a different error model for ϵ_i, such as a heteroscedastic error model with variance depending on X_i. Both weighted least squares method and maximum likelihood method can be carried out without extra complication.

3.2 Choosing K

The preceding inference procedure assumes a pre-specified (fixed) K. In practical implementation, one needs to choose a K first and the quality of the follow-up inference depends on this choice. Too small of a K can result in an estimated regression function not flexible enough to capture the shape of $\theta^*(x)$, and an unnecessarily large K causes inefficiency loss. In the sequel, $\hat{\theta}_K(x)$ denotes the estimated regression function when $\theta^*_K(x)$ in (6) is used to model $\theta(x)$.

Fenton and Gallant (1996a) derived the rate at which K_n tends to infinity as n increases in order to achieve a desired convergence notion of $\hat{\theta}_{K_n}(x)$ to $\theta^*(x)$. In particular, if it is assumed that the true density in $\mathcal{H}$ is second order differentiable, then $K_n \approx n^{1/5}$ is needed to achieve consistency. Fenton and Gallant (1996b) found that using a deterministic rule, such as $K_n \approx n^{1/5}$, is inferior to using an adaptive rule, according to which K is chosen adaptively based on information criteria such as the Akaike information criterion (AIC). Davidian and Gallant (1993) compared use of AIC, Schwarz information criterion (BIC), and Hannan-Quinn criterion (HQ) to choose K, and suggested use of HQ because it often chooses a value of K that lies between those chosen by AIC and BIC. Zhang and Davidian (2001), Chen, Zhang, and Davidian (2002), Zhang and Davidian (2008), and Irincheeva, Cantoni, and Genton (2012) also considered AIC, BIC, and HQ as three model selection criterion and reached very similar conclusions as those in Davidian and Gallant (1993). Coppejans and Gallant (2002) explored the method of cross-validation (CV) to choose K. Eastwood (1991) proposed an F test to compare goodness of fit resulting from two choices of K. His test is valid only if the candidate K's are large enough to begin with.

We investigated all the aforementioned methods for choosing K in our context and found that the K chosen by information criteria tends to be larger than what is chosen by CV. To strive for parsimony without sacrificing flexibility, we propose to combine CV and Eastwood's F test to choose K as described next. Firstly, one employs an r-fold CV. More specifically, one partitions the observed data randomly into r subsets, as equal sized as possible. Then, for each candidate K,

the model is fitted r times, each time the jth subset ($j = 1, \ldots, r$) is kept as the validation data set, and the remaining $r-1$ subsets form the testing data set used to fit the SNP (of order K) regression model. After each model fitting, the mean squared error (MSE) is calculated using the validation data and the average of these r MSE's for each candidate K is calculated. Denote by K_0 the K where the first abrupt drop in the average MSE occurs or where the average MSE is the smallest, whichever comes first. Now one uses $\theta^*_{K_0}(x)$ to fit the regression model based on the entire data set. Denote by $\hat{\sigma}^2_{\epsilon,0}$ the estimated σ^2_ϵ resulting from this fit. Secondly, one repeats the estimation based on $\theta^*_{K_1}(x)$, where $K_1 = K_0 + c$, for some positive integer c (usually set at one). Denote by $\hat{\sigma}^2_{\epsilon,1}$ the estimates for σ^2_ϵ from this round of model fitting. Lastly, one computes Eastwood's F statistic defined by

$$F = \frac{(n - p_{\tau,0})\hat{\sigma}^2_{\epsilon,0} - (n - p_{\tau,1})\hat{\sigma}^2_{\epsilon,1}}{c(n - p_{\tau,0})\hat{\sigma}^2_{\epsilon,0}}, \tag{12}$$

where $p_{\tau,0}$ and $p_{\tau,1}$ are the dimension of $\boldsymbol{\tau}$ when K is equal to K_0 and K_1, respectively. The F statistic is compared with the critical point of an $F(c, n - p_{\tau,0})$ distribution to decide whether or not SNP of order K_1 gives a significantly better fit than SNP of order K_0. In this procedure, one uses CV to choose K_0 first to reduce the chance of applying Eastwood's F test with insufficiently large K, which, as pointed out earlier, can invalidate the test.

3.3 Test for linearity

In nonlinear regression, it is often of interest to justify if a nonlinear regression function is indeed necessary as opposed to a linear regression function. In this subsection, we propose a simple test for linearity of $\theta(x)$ based on the SNP representation.

The proposed test is motivated by the simple fact that, assuming $\theta(x)$ is second-order differentiable, if $\theta(x)$ is a linear function of x, then $\theta''(x) = 0$ for all x, where $\theta''(x)$ is the second derivative of $\theta(x)$. Hence, an empirical indicator of (non)linearity can be constructed based on $\hat{\theta}''_K(x)$, which can be easily derived, thanks to the SNP presentation. Formally, we propose to test $H_0 : \theta''(x) = 0$

versus $H_1 : \theta''(x) \neq 0$ for testing linearity. A natural statistic for this test is given by

$$Q_K = n^{-1} \sum_{i=1}^{n} \left\{ \hat{\theta}''_K(X_i) \right\}^2 . \tag{13}$$

Note that, provided that K is chosen appropriately, Q_K is degenerate at zero under H_0. This prohibits one from obtaining a critical point with which Q_K compares to make a testing decision in the traditional ways. To obtain an empirical p-value for this statistic, we exploit the method proposed by Härdle, Mammen, and Müller (1998) and implement a parametric bootstrap procedure as follows.

Step 1: Fit the SNP regression model with an appropriately chosen K. Compute Q_K.

Step 2: Fit a linear regression model, $Y = \beta_0 + \beta_1 X + \epsilon$, under the constraint that $\beta_1 \geq 0$. Denote by $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$ the estimate for $\boldsymbol{\beta} = (\beta_0, \beta_1)$, and by $\hat{\sigma}^{2*}_{\epsilon,0}$ the estimate for σ^2_ϵ.

Step 3: For $b = 1, \ldots, B$,

(i) Generate the bth set of artificial data, $Y_{i,b} = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\sigma}^*_{\epsilon,0} \epsilon^*_{i,b}$, for $i = 1, \ldots, n$, where $\{\epsilon^*_{i,b}\}_{i=1}^n$ are i.i.d. random errors with mean zero and variance one (such as from $N(0, 1)$).

(ii) Fit the SNP regression model as in step 1 using the bth artificial data set, $\{(Y_{i,b}, X_i)\}_{i=1}^n$. Denote by $\hat{\theta}_{K,b}(x)$ the estimated regression function.

(iii) Compute $Q_{K,b} = n^{-1} \sum_{i=1}^n \left\{ \hat{\theta}''_{K,b}(X_i) \right\}^2$.

Step 4: The empirical p-value associated with Q_K is defined by $B^{-1} \sum_{b=1}^B I\{Q_K \leq Q_{K,b}\}$.

The choice of K adopted in step 1 is crucial especially for power consideration. Under H_0, it is likely that a small K is chosen by the procedure described in Section 3.2 and linearity is often well preserved by $\hat{\theta}_K(x)$. Even with a K slightly higher than necessary, the size of the test usually will not inflate noticeably because the redundant coefficients among a_λ's are expected to be estimated as close to zero. But under H_1, if one chooses a K not high enough to capture the curvature in

$\theta^*(x)$, Q_K can lack power to detect nonlinearity; and if one chooses an unnecessarily large K, the efficiency loss can also compromise the power of Q_K. Note that tests for concavity/convexity of $\theta(x)$ are readily available by changing the above test to a one-sided test. One can also test for constant regression functions by defining $H_0 : \theta'(x) = 0$, then revising Q_K accordingly. In fact, it is possible to take the advantage of the explicit parametric form of $\theta^*_K(x)$ to formulate H_0 and revise Q_K in different ways to test many functional features of $\theta(x)$.

4 Simulation studies

In this section we present five examples designed to provide empirical evidence for the quality of SNP estimation, the operating characteristics of the proposed method to choose K, and the test for linearity. The first three examples have the following common simulation settings. For each of 300 Monte Carlo (MC) replicates, a random sample of size $n = 200$ is generated according to a model specified in each example, and the model errors, $\{\epsilon_i\}_{i=1}^{200}$, are i.i.d. $N(0, \sigma^2_\epsilon)$, where $\sigma^2_\epsilon = 0.25$. To find the empirical p-value for Q_K, we set $B = 300$ in the procedure described in Section 3.3. The 5-fold CV is employed to find K_0. After a K is chosen from CV combing with an F test, we conduct the test for linearity of $\theta(x)$. For notational convenience, define F_{st} as the F statistic for testing $K = s$ versus $K = t$, where $1 \leq s < t$. The significance level for all tests is 0.05. To pictorially compare the estimated regression function and the true function, we compute the MC average of 300 sets of $\hat{\boldsymbol{\tau}}$ and use this average as the parameter values plugged into the SNP representation for $\theta(x)$, resulting in an estimated regression function, denoted by $\tilde{\theta}_K(x)$. In contrast, the estimated regression function from one MC replicate is denoted by $\hat{\theta}_K(x)$. The fourth example presents a comparative simulation study, where the fit for the regression function using the proposed method is compared with the fits when three existing methods developed for smooth isotonic regression are used. Lastly, the fifth example considers bivariate isotonic regression.

Example 1 [Linear $\theta^*(x)$]: The true regression model is, $Y_i = 2X_i + \epsilon_i$, for $i = 1, \ldots, n$, where $\{X_i\}_{i=1}^n$ are generated from uniform(0, 1). The CV process chooses $K = 1$, associated with the average MSE smaller than those associated with $K = 2$, 3, and 4. The F test comparing $K = 1$ versus $K = 2$ yields a rejection rate of 6% across 300 MC replicates. This suggests that, with one data set, one would most likely fail to reject $K = 1$. The test for linearity based on the test statistic Q_1 results in a rejection rate of 4% across 300 MC replicates. This provides empirical evidence that the proposed test for linearity confers the right size. In plot (a) of Figure 1 we compare $\theta^*(x) = 2x$ (solid line) with $\tilde{\theta}_1(x)$ (dashed line) and two $\hat{\theta}_1(x)$ from the two randomly chosen MC replicates (dotted lines). All three latter curves exhibit close agreement with the true linear function.

Example 2 [Strongly nonlinear $\theta^*(x)$]: The true regression model is, $Y_i = X_i^3 + \epsilon_i$, for $i = 1, \ldots, n$, where $\{X_i\}_{i=1}^n$ are generated from uniform$(-2, 2)$. The CV process shows that changing K from one to two leads to an abrupt drop in the average MSE and it levels off for $K \geq 2$. The statistic, F_{23}, used to test $K = 2$ versus $K = 3$, has an empirical rejection rate of 6% across 300 MC replicates. In contrast, the rejection rate associated with F_{12} is 100%. Hence, with one data set, it is most likely that F_{23} is insignificant, suggesting that $K = 3$ does not yield sufficient improvement in fitting the true regression function than when $K = 2$; whereas F_{12} is highly significant, providing strong evidence that $K = 2$ leads to a much better fit for $\theta(x)$ than when $K = 1$. Fixing K at 2, the test for linearity based on Q_2 also has a 100% rejection rate. This provides empirical evidence that Q_K has high power to detect nonlinearity of $\theta(x)$ in this case. Plot (b) in Figure 1 depicts $\theta^*(x) = x^3$ (solid line), $\tilde{\theta}_1(x)$ (dashed line), $\tilde{\theta}_2(x)$ (dash-dotted line), and $\tilde{\theta}_3(x)$ (long dashed line). The two higher-order curves, $\tilde{\theta}_2(x)$ and $\tilde{\theta}_3(x)$, nearly overlap with each other and show nearly perfect fit for the true function. In contrast, $\tilde{\theta}_1(x)$ fails to capture the curvature in the region around $x = -1$ and $x = 1$.

In this example we also summarize the parameter estimates when $K = 1$ and $K = 2$ in Table 1. The MC averages of the estimated standard errors based on sandwich construction are compared

with the MC standard deviation of the parameter estimates. Note that when $K = 1$, $\hat{\sigma}_\epsilon^2$ is substantially biased upwards. This is expected as a great amount of variability due to the underlying $\theta(x)$ is not capture by $\hat{\theta}_1(x)$ and is instead counted as model error. When $K = 2$, the accuracy of $\hat{\sigma}_\epsilon^2$ is greatly improved and it seems to only account for the real model error instead of the lack of fit for $\theta(x)$. This phenomenon also reflects the rationale behind Eastwood's F test. Moreover, the sandwich estimates for the standard errors are also much more reliable when $K = 2$ than when $K = 1$.

Furthermore, we explore the partially linear model, $Y_i = \beta Z_i + X_i^3 + \epsilon_i$, where Z_i is another covariate and β is a regression coefficient. Even with an additional linear part in the regression model, βZ_i, we do not observe noticeable change in the operating characteristics of the procedures used to choose K and the test for linearity compared to the phenomena described above. Additionally, when $K = 2$, the MLE for β, $\hat{\beta}$, is also satisfactory, so is the sandwich standard error for $\hat{\beta}$.

Example 3 [Mildly curved $\theta^*(x)$]: In this example we consider a less dramatically nonlinear $\theta(x)$ than that in Example 2. Consider the model $Y_i = e^{X_i} + \epsilon_i$, where we design two settings with the first has a milder nonlinearity than the second: (i) $X_i \sim$ uniform(0, 1); (ii) $X_i \sim$ uniform(1, 2). For $K = 1, 2, 3$, besides $\mathbf{\Omega}$, we also estimate the value of $\theta(x)$ at the 25th, 50th, and 75th percentile of the support of X under each setting, denoted by $\hat{\theta}_{25}$, $\hat{\theta}_{50}$, and $\hat{\theta}_{75}$, respectively. These results, along with the rejection rates of the F test and test for linearity are summarized in Table 2.

When $X \sim$ uniform(0, 1), the nonlinearity is very mild and, consequently, CV combining with the F test suggest $K = 2$ may not be needed to model e^x over this range of x. The statistic Q_1 has moderate power to detect nonlinearity of the true regression function. Note that if one sets $K = 2$ under the first setting, Q_2 has a very low power to detect the nonlinearity. This indicates that unnecessarily high K can compromise the power of Q_K. Similar phenomenon is observed under the second setting, even though F_{12} rejects $K = 1$ more often than under the first setting, reflecting the increased curvature of e^x over the range of [1, 2]. Under both settings, F tests provide little

evidence that K as high as 3 is needed.

Plot (c) in Figure 1 depicts $\hat{\theta}_1(x)$ from one randomly chosen MC replicate (dotted line), $\tilde{\theta}_1(x)$ (dashed line), and $\tilde{\theta}_2(x)$ (dash-dotted line), compared with $\theta^*(x) = e^x$ for $x \in [0, 1]$ (solid line). In this rage of x, it appears from the plot that $\tilde{\theta}_1(x)$ matches $\theta^*(x)$ better than $\tilde{\theta}_2(x)$ does. Plot (d) shows $\tilde{\theta}_1(x)$ and $\tilde{\theta}_2(x)$ in contrast to the truth over the range of [1, 2]. Here, $\tilde{\theta}_2(x)$ appears to have a slightly better fit. This pictorial impression is also reflected in the quality of $\hat{\theta}_{25}$, $\hat{\theta}_{50}$, and $\hat{\theta}_{75}$.

Example 4 [Compare with other methods]: In this example, we generate X from uniform(0, 1) and set $\sigma_\epsilon^2 = 0.25$, $n = 100$ or 200, for each of 2000 MC replicates. Three existing methods are compared with the proposed method, with two spline-based and one kernel-based. One spline-based method is implemented in the R function, `pcls`, which realizes a penalized cubic spline with monotonicity constraint (Woods, 1994) with 5 or 10 knots evenly spread over [0, 1]. Another spline-based method is based on nondecreasing piecewise quadratic splines (Meyer, 2013) with 2 or 4 knots (adopting the recommendation in Meyer (2008)) evenly spread over [0, 1], implemented in Meyer's R function, `mspl` (`http://www.stat.colostate.edu/~meyer/msplh.R`). The kernel-based method is the nonparametric method proposed by Dette, Neumeyer, and Pilz (2006) and implemented in the R function, `monreg`. Their method involves a local linear regression step and a density estimation step. For each step, one needs to specify a kernel and a bandwidth. In our simulation study, as done in Dette, Neumeyer, and Pilz (2006), we set both kernels to be the Epanechnikov kernel and the bandwidth in the local linear regression step to be $h_r = (\hat{\sigma}^2/n)^{1/5}$, where

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{[i+1]} - Y_{[i]})^2,$$

in which $\{Y_{[i]}\}_{i=1}^n$ denote the responses sorted by the covariate values. With the so-chosen h_r, we consider two settings for the bandwidth used in the density estimation step, denoted by h_d. In one setting, $h_d = h_r^3$, and in the other setting, $h_d = 0.5h_r$, both of which are considered in Dette, Neumeyer, and Pilz (2006). Lastly, instead of using the plug-in method to obtain h_r as

above, we use the R function `monreg.wrapper` to choose h_r via 5-fold CV, followed by which this function sets $h_d = h_r^2$ and implements `monreg` with these bandwidths. We compare these four methods in terms of the mean absolute error (MAE), defined as MAE $= n^{-1} \sum_{i=1}^{n} |\theta^*(X_i) - \hat{\theta}_K(X_i)|$. Four regression functions are considered: $\theta^*(x) = \log(2x + 1)$, $\theta^*(x) = x^3$, $\theta^*(x) = e^{2x}$, and $\theta^*(x) = I(x \leq 0.6)10x/6 + I(x > 0.6)$. Using the method described in Section 3.2 to choose K for the SNP regression, we settle with $K = 1$ for the first regression function and $K = 2$ for the remaining three functions. The summary of MAE resulting from four methods is presented in Table 3.

As evident in Table 3, the SNP method performs competitively with relatively stable performance across different regression functions. Both spline-based methods can be sensitive to the number of knots. The method implemented in `monreg` can be sensitive to the choice of bandwidths, and the added CV procedure in `monreg.wrapper` does not consistently improve the performance. Figure 2 depicts the estimated regression functions from the SNP method, `pcls` with 10 knots, and `monreg` with $h_d = h_r^3$ for each regression function from one randomly selected MC replicate. As shown in plot (a), the spline-based method implemented in `pcls` can oscillate, trying to capture local features, which is a typical phenomenon for spline-based methods. The SNP method seems to be able to capture the underlying curvature of $\theta^*(x)$ more accurately, except for the nonsmooth regression function in plot (d), where all three estimates fail to reflect the sharp turn at $x = 0.6$. This is expected since these methods are developed to fit smooth regression functions. Overall, the empirical evidence suggest that the SNP method, combined with the proposed procedure of choosing K, performs at least as satisfactorily as the three existing methods in comparison.

Example 5 [Bivariate isotonic regression]: Different from the first four examples, where univariate covariates are considered, here we consider bivariate covariates, motivated by applications where the regression function $\theta(\cdot)$ is constrained to be monotone in each of the two covariates of interest. As in the univariate case, the SNP presentation of a nondecreasing (in each of the two coordinates)

function, $\theta^*_K(\mathbf{x}) = \gamma_K(\mathbf{x}) + s$, can be straightforwardly derived and coded using recursive formulas similar to those in Section 3.1. More specifically, now with $\gamma_K(\mathbf{x}) = \int_{x_{2L}}^{x_2}\int_{x_{1L}}^{x_1}\{P_K(\mathbf{t})\}^2 \exp\{-(t_1^2 + t_2^2)/2\}dt_1dt_2$, where $\mathbf{t} = (t_1, t_2)^T$, $P_K(\mathbf{t}) = \sum_{0\le j+j'\le K} a_{jj'}t_1^j t_2^{j'}$, $\mathbf{x} = (x_1, x_2)^T$, and $x_{\ell L}$ is the lower bound of the range of x_ℓ, for $\ell = 1, 2$, the recursive formulas become

$$\begin{aligned}
\gamma_0(\mathbf{x}) &= a_{00}^2 I_0(x_1)I_0(x_2), \\
\gamma_K(\mathbf{x}) &= \gamma_{K-1}(\mathbf{x}) + a_{K0}^2 I_{2K}(x_1)I_0(x_2) + 2a_{K0}\sum_{j=0}^{K-1} I_{j+K}(x_1)\sum_{j'=0}^{K-j} a_{jj'}I_{j'}(x_2), \text{ for } K > 0,
\end{aligned}$$

where, for $\ell = 1, 2$,

$$\begin{aligned}
I_0(x_\ell) &= \sqrt{2\pi}\,\{\Phi(x_\ell) - \Phi(x_{\ell L})\}, \\
I_1(x_\ell) &= \exp\left(-x_{\ell L}^2/2\right) - \exp\left(-x_\ell^2/2\right), \text{ and} \\
I_K(x_\ell) &= x_{\ell L}^{K-1}\exp\left(-x_{\ell L}^2/2\right) - x_\ell^{K-1}\exp\left(-x_\ell^2/2\right) + (K-1)I_{K-2}(x_\ell), \text{ for } K > 1.
\end{aligned}$$

For illustration purpose, in this example, we assume the true regression function to be $\theta(\mathbf{x}) = x_1 + \exp(0.5x_2) + x_1x_2$, where $\mathbf{x} = (x_1, x_2)^T \in [0, 1]\times[0, 1]$. For each of 2000 MC replicates, we generate a random sample of size $n = 400$ according to the true regression model, with $\epsilon_i \sim N(0, 0.25)$, and $X_{1,i}$ independent of $X_{2,i}$, both generated from uniform(0, 1), for $i = 1, \ldots, 400$. Using the procedure described in Section 3.2, more than 75% of the time K is chosen to be 1 or 2. With $K = 1, 2$, the resultant estimated responses give MAE 0.404 (0.352) and 0.400 (0.352), respectively, with the corresponding $10^3\times$ standard error of the MC average in parentheses. For comparison, we implemented an existing bivariate isotonic regression algorithm described in Dykstra and Robertson (1982) using the R function, `biviso` (Bril et al., 1984), which uses successive one-dimensional smoothing subject to isotonic constraint. Across 2000 MC replicates, this algorithm produces estimated responses with MAE 0.539, and $10^3\times$ standard error being 0.417.

5 Application to indomethacin data

We now consider isotonic regression analyses using a data set from a study of the pharmacokinetics of indeomethacin following bolus intravenous injection (Davidian and Gallant, 1995, Section 2.1). The data include plasma concentrations of indomethacin measured at eleven time points for each of the six subjects participating the study. It is reasonable to assume that the plasma concentration (Y) is a nonincreasing function of time (X). For illustration purposes, suppose that the model in (5) with homoscedastic model error is a reasonable model for the association between plasma concentration and time for some nonincreasing $\theta(x)$. Next, we use the methods under comparison in Section 4 excluding `mspl` to fit the regression function. We exclude `mspl` in this example because it requires distinct predictor values, which is not the case for this data.

For the proposed method, we first choose K using the strategy described in Section 3.2. In the 5-fold CV, the first abrupt drop in MSE occurs when K is raised from 1 to 2. We then compute Eastwood's F statistic, F_{23}, and the associated p-value is 0.021. This implies that $K = 3$ is likely to provide a better fit for the current data. When we raise to $K = 4$, the F statistic, F_{34}, gives a p-value of 0.739. Hence, using SNP of order four may not provide a significant improvement in the fit compared to using SNP of order three. The fitted model using SNP of order 3 is plotted in Figure 3, along with the other two fitted curvs from `pcls` (with 10 knots) and `monreg` (with $h_d = h_r^3$). Pictorially it appears that the first two methods yield comparable results, and the fit from `monreg` is potentially problematic (with a negative fitted Y in the end). The MAE, now defined as $n^{-1}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$, associate with three methods are 0.121, 0.110, and 0.144 for SNP, `pcls`, and `monreg`, respectively. According to this numerical comparison along with the pictorial comparison, it appears that SNP method finds a nice balance between overfitting (likely observed for `pcls`) and underfitting (seemingly observed for `monreg`). Finally, the test for linearity results in an empirical p-value of 0.012, providing strong evidence of the nonlinearity of the regression function, as evident from the plot.

6 Discussion

We propose to use a class of SNP densities as the basis to model smooth isotonic regression functions flexibly. The mathematical representation of this class gives great advantages in achieving smoothness and isotonicity simultaneously in one simple step. We develop an adaptive procedure to choose the order of SNP to attain a parsimonious model that is flexible enough to yield a satisfactory fit for the underlying true regression function. We also construct a test for linearity of the regression function. The performance of the adaptive procedure and the test for linearity are satisfactory for both nonlinear models and partially linear models.

Considering the model formulation, the proposed method involves ingredients similar to existing semi-/non-parametric methods, with K playing the role parallel to a smoothing/tuning parameter in these literature, and the Hermite polynomials in the SNP formulation somewhat mimicking the base functions. From the numerical point of view, the R function, `optim`, is used to maximize the likelihood in the proposed method without any numerical difficulty in our simulation studies for K as high as eight when $d = 1$ and for K as high as four when $d = 2$, which is more than enough for the simulated examples and other real data applications we have looked into. It is possible that, for a very large K, more sophisticated algorithms are needed for more efficient optimization (especially when d is also not small). Among the four methods considered in Example 4 in Section 4, `mspl` is the least time-consuming and the other three methods are comparable in computing time.

The normal pdf in (4) is one choice of the base density when constructing an SNP density and other choices of pdf are also valid and will yield different classes of flexible pdf's. We have focused on the normal base density for the convenience in analytic derivation of the corresponding cdf. A minor drawback of this representation is lack of interpretation for the coefficients, a_λ's, in (3), which is a common pitfall among most flexible modeling methods. But because it is such a flexible class of functions, one can compare the SNP regression function with a particular regression function one may have in mind, of which the interpretation is scientifically more meaningful,

to see how close these two functions are. In other words, the SNP estimation can be used as a reference, which is supposed to represent the truth very well with an adequately chosen K. If a scientifically meaningful posited model provides similar fit as the SNP model, then one gains more confidence in the posited model.

We have not compared our method with the existing methods in the Bayesian framework, such as those proposed by Bornkamp and Ickstadt (2009) and Meyer, Hackstadt, and Hoeting (2011). One glitch of the proposed method is that inference procedures outlined in Section 3.1 assume a pre-specifed K and thus ignore the uncertainty in choosing K. Even though we did not observe sufficient numerical evidence of the tampering effect of ignoring this uncertainty on standard error estimation in the examples, it is desirable to account for this extra uncertainty. Solutions to this problem may lie within the Bayesian framework. For instance, the reversible-jump MCMC described in Meyer, Hackstadt, and Hoeting (2011) gives one a hint on how to incorporate the variability due to selecting K. There has been little work on SNP from a Bayesian perspective. Imposing a prior distribution on K was once brought up in Davidian and Gallant (1993) but has never been explored. It is of interest to look into this avenue of Bayesian SNP modeling and then compare the Bayesian SNP method with the other Bayesian isotonic regression methods.

Following the theoretical development in Fenton and Gallant (1996a,b), we conjecture that, with $K \to \infty$ at the right rate, the SNP estimator of the regression function has similar asymptotic properties as those from kernel-based or spline-based methods (Banerjee, 2007; Pal and Woodroffe, 2007). Noticing that Mammen, et al. (2001) also uses Sobolev norm as a measure of the distance between two functions and to impose smoothness penalty, we believe that the projection framework formulated in Mammen, et al. (2001) can be useful hints for us to study the asymptotic properties of our proposed methods more thoroughly. This theoretical consideration has been on the top of our follow-up research agenda.

Since SNP has been mostly used to model distributions of variables whose distribution may not be in a familiar family, it is natural to consider keeping this tradition while adding the new use

of SNP proposed in this article. We have started to apply both ideas to regression models with random effects and also contain isotonic component in the regression functions. We propose to use SNP densities to model the random effects to avoid stringent distributional assumptions on them, and at the same time, use SNP cdf to model the isotonic component in a regression function. With two flexible modeling combined in the same regression model, one faces the identifiability issue. Theories behind this double flexibility strategy and how to tackle the identifiability issue are topics worth further investigation.

References

Banerjee, M. (2007). Likelihood based inference for monotone response models. *The Annals of Statistics*, **35**, 931–956.

Banerjee, M., Biswas, P., and Ghosh, D. (2006). A semiparametric binary regression model involving monotonicity constraints. *Scandinavian Journal of Statistics*, **33**, 673–697.

Banerjee, M. and Wellner, J. A. (2001). Likelihood ratio tests for monotone functions. *The Annals of Statistics*, **29**, 1699–1731.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., Brunk, H. D. (1972). *Statistical Inference Under Order Restriction*. Wiley, New York.

Bartholomew, D. J. (1959). A test of homogeneity for ordered alternative. *Biometrika*, **46**, 36–48.

Bornkamp, B. and Ickstadt, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis. *Biometrics*, **65**, 198–205.

Bril, H., Dykstra, R., Pillers, C., and Robertson, T. (1984). Isotonic regression in two independent variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **33**, 352–357.

Chen, J., Zhang, D., and Davidian, M. (2002). Generalized linear mixed models with flexible distributions of random effects for longitudinal data. *Biostatistics*, **3**, 347–360.

Coppejans, M. and Gallant, A. R. (2002). Cross-validation SNP density estimates. *Journal of Econometrics*, **110**, 27–65.

Davidian, M. and Gallant, A. R. (1993). The nonlinear mixed effects models with a smooth random effects density. *Biometrika*, **80**, 475–488.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*. Chapman & Hall/CRC. Boca Raton, FL.

Delecroix, M. and Thomas-Agnan, C. (2000). Spline and kernel regression under shape restrictions. In M. G. Schimek (ed.). *Smoothing and Regression. Approaches, Computation and Application*. New York: Wiley.

Dette, H., Neumeyer, N., and Pilz, K. F. (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli*, **12**, 469–490.

Dykstra, R. L. and Robertson, T. (1982). An algorithm for isotonic regression for two or more independent variables. *The Annals of Statistics*, **10**, 708–716.

Eastwood, B. J. (1991). Asymptotic normality and consistency of semi-nonparametric regression estimators using an upwards F test truncation rule. *Journal of Econometrics*, **48**, 151–181.

Eastwood, B. J. and Gallant, A. R. (1991). Adaptive rules for seminonparametric estimators that achieve asymptotic normality. *Econometric Theory*, **7**, 307–340.

Fenton, V. M. and Gallant, A. R. (1996). Convergence rates of SNP density estimators. *Econometrica*, **64**, 719–727.

Fenton, V. M. and Gallant, A. R. (1996), Qualitative and asymptotic performance of SNP density estimators. *Journal of Econometrics*, **74**, 77-118.

Gallant, A. R and Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation.*Econometrica*, **55**, 363–390.

Hall, P. and Huang, L. S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics*, **29**, 624–647.

Härdle, W., Mammen, E., and Müller, M. (1998). Testing parametric versus semiparametric modeling in generalized linear models. *Journal of the American Statistical Association*, **93**, 1461–1474.

Irincheeva, I., Cantoni, E., and Genton, M. G. (2012). Generalized linear latent variable models with flexible distribution of latent variables. *Scandinavian Journal of Statistics*, **39**, 663–680.

Kelly, C. and Rice, J. (1990). Monotone smoothing with application to does-response curves and the assessment of synergism. *Biometrics*, **46**, 1071–1085.

Mammen, E. (1991). Estimating a smooth monotone regression function. *The Annals of Statistics*, **19**, 724–740.

Mammen, E., Marron, J. S., Turlach, B. A., Wang, M. P. (2001). A general projection framework for constrained smoothing. *Statistical Science*, **16**, 232–248.

Mammen, E. and Thomas-Agnan, C. (1999). Smoothing splines and shape restrictions. *Scandinavian Journal of Statistics*, **26**, 239–252.

Meyer, M. C. (2008). Inference using shape-restricted regression splines. *The Annals of Applied Statistics*, **2**, 1013–1033.

Meyer, M. C. (2013). A simple new algorithm for quadratic programming with applications in statistics. *Communications in Statistics–Theory and method*, to appear.

Meyer, M. C., Hackstadt, A. J., and Hoeting, J. A. (2011). Bayesian estimation and inference for generalised partial linear models using shape-restricted splines. *Journal of Nonparametric Statistics*, **23**,867–884.

Pal, J. K. and Banerjee, M. (2008). Estimation of smooth link functions in monotone response models. *Journal of Statistical Planning and Inference*, **138**, 3125–3143.

Pal, J. K. and Woodroffe, M. (2007). Large sample properties of shape restricted regression estimators with smoothness adjustments. *Statistica Sinica*, **17**, 1601-1616.

Ramsay, J. O. (1988). Monotone regression splines in action (with comments). *Statistical Science*, **3**, 425–461.

Ramsay, J. O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society, Series B*, **60**, 365–375.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

Song, X., Davidian, M., and Tsiatis, A. A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, **58**, 742–753.

Song, X., Davidian, M., and Tsiatis, A. A. (2002). An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics*, **3**, 511–528.

Wang, J. (2011). Shape restricted nonparametric regression with Bernstein polynomials. Ph.D. dissertation, North Carolina State University, Raleigh, North Carolina (http://www.lib.ncsu.edu/resolver/1840.16/7454).

Wang, X. and Feng, L. (2008). Isotonic smoothing spine regression. *Journal of Computational and Graphical Statistics*, **17**, 21–37.

Woods, S. (1994). Monotonic smoothing splines fitted by cross validation. *SIAM Journal on Scientific Computing*, **15**, 1126–1133.

Zhang, D. and Davidian, M. (2001). Linear mixed model with flexible distribution of random effects for longitudinal data. *Biometrics*, **57**, 795–802.

Zhang, M. and Davidian, M. (2008). "Smooth" semiparametric regression analysis for arbitrarily censored time-to-event data. *Biometrics*, **64**, 567–576.

Table 1: Averages of parameter estimates across 300 MC replicates in Example 2 in Section 4. Entries in parentheses next to the parameter estimates are corresponding MC standard deviations. Entries in parentheses next to the sandwich standard errors are (100 × MC standard error) of the averages of sandwich estimates.

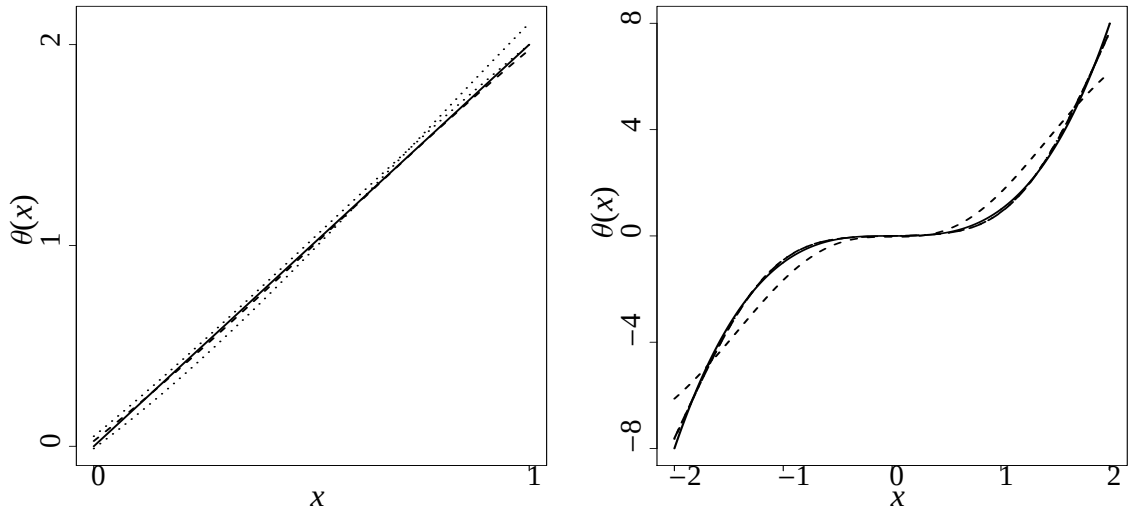
	$K = 1$		$K = 2$	
	Parameter estimates	Sandwich std. err.	Parameter estimates	Sandwich std. err.
a_0	0.012 (0.021)	0.034 (0.010)	0.295 (0.058)	0.059 (0.034)
a_1	2.576 (0.017)	0.030 (0.011)	0.004 (0.021)	0.022 (0.010)
a_2	(NA)	(NA)	1.957 (0.046)	0.048 (0.026)
s	−6.127 (0.102)	0.183 (0.094)	−7.643 (0.148)	0.149 (0.122)
σ_ϵ^2	0.574 (0.047)	0.060 (0.041)	0.256 (0.027)	0.026 (0.020)

Table 2: Simulation results from Example 3 in Section 4. The upper half of the table includes MC averages of the estimated quantities from the simulation, followed by (10 × MC standard error) in parentheses. True values of the five quantities being estimated, $(s, \sigma_\epsilon^2, \theta_{25}, \theta_{50}, \theta_{75})$, are (1, 0.25, 1.284, 1.649, 2.117) when $X \sim$ uniform(0, 1), and they are (2.718, 0.25, 3.490, 4.482, 5.755) when $X \sim$ uniform(1, 2). Entries following the row titles, F_{12}, F_{23}, and Q_K, are rejection rates of the test statistics across 300 MC replicates.

	$X \sim$ uniform(0, 1)		$X \sim$ uniform(1, 2)	
	$K = 1$	$K = 2$	$K = 1$	$K = 2$
$\hat{s}$	1.028 (0.062)	0.990 (0.075)	2.838 (0.064)	2.678 (0.087)
$\hat{\sigma}_\epsilon^2$	0.250 (0.015)	0.250 (0.015)	0.255 (0.015)	0.251 (0.015)
$\hat{\theta}_{25}$	1.275 (0.029)	1.286 (0.037)	3.431 (0.028)	3.499 (0.037)
$\hat{\theta}_{50}$	1.645 (0.031)	1.646 (0.032)	4.485 (0.029)	4.480 (0.028)
$\hat{\theta}_{75}$	2.131 (0.029)	2.121 (0.036)	5.829 (0.029)	5.763 (0.036)
F_{12}	0.050		0.340	
F_{23}	0.043		0	
Q_K	0.457	0.067	1	0.107

Table 3: MC averages of MAE across 2000 replicates from four methods compared in Example 4 in Section 4. Numbers in parentheses are ($10^3\times$ MC standard error) of the averages. "pcsl5" and "pcsl10" denote the method implemented by `pcsl` with 5 and 10 knots, respectively. "mspl2" and "mspl4" denote the method implemented by `mspl` with 2 and 4 knots, respectively. "monreg1" and 'monreg2" denote the method implemented by `monreg` with $h_d = h_r^3$ and $h_d = 0.5h_r$, respectively, where $h_r = (\hat{\sigma}^2/n)^{1/5}$. "monreg3" denotes the method implemented by `monreg.wrapper`, which implements 5-fold cross validation to choose h_r and then sets $h_d = h_r^2$.

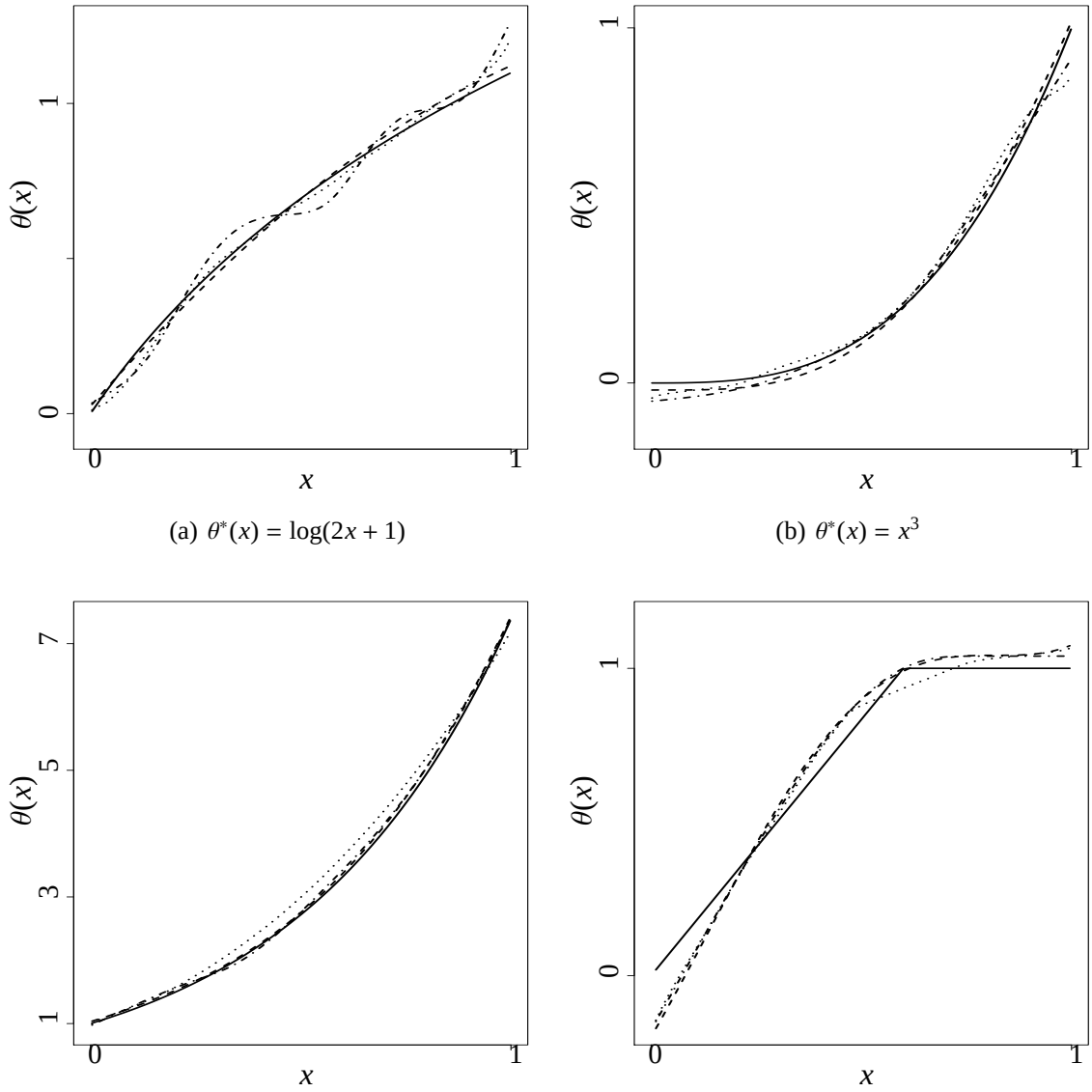
	$\theta^*(x) = \log(2x+1)$	$\theta^*(x) = x^3$	$\theta^*(x) = e^{2x}$	$\theta^*(x) = I(x \le 0.6)\frac{10x}{6} + I(x > 0.6)$
$n = 100$				
SNP	0.067 (0.647)	0.073 (0.654)	0.078 (0.662)	0.074 (0.616)
pcsl5	0.065 (0.584)	0.072 (0.656)	0.083 (0.621)	0.075 (0.621)
pcsl10	0.068 (0.601)	0.076 (0.659)	0.088 (0.657)	0.076 (0.622)
mspl2	0.074 (0.577)	0.072 (0.614)	0.084 (0.629)	0.074 (0.597)
mspl4	0.081 (0.576)	0.078 (0.585)	0.098 (0.617)	0.080 (0.616)
monreg1	0.068 (0.588)	0.073 (0.670)	0.186 (0.910)	0.073 (0.587)
monreg2	0.065 (0.596)	0.076 (0.639)	0.187 (0.902)	0.079 (0.537)
monreg3	0.076 (0.625)	0.077 (0.736)	0.108 (0.753)	0.083 (0.607)
$n = 200$				
SNP	0.048 (0.463)	0.051 (0.461)	0.055 (0.455)	0.054 (0.445)
pcls5	0.050 (0.418)	0.052 (0.469)	0.061 (0.424)	0.056 (0.455)
pcls10	0.051 (0.425)	0.054 (0.475)	0.064 (0.452)	0.056 (0.455)
mspl2	0.056 (0.419)	0.053 (0.443)	0.060 (0.448)	0.055 (0.423)
mspl4	0.061 (0.405)	0.058 (0.422)	0.070 (0.446)	0.059 (0.420)
monreg1	0.051 (0.425)	0.052 (0.462)	0.152 (0.616)	0.054 (0.429)
monreg2	0.049 (0.431)	0.055 (0.448)	0.153 (0.610)	0.060 (0.379)
monreg3	0.057 (0.461)	0.060 (0.454)	0.080 (0.543)	0.062 (0.433)



(a) $\theta^*(x) = 2x$ (b) $\theta^*(x) = x^3$

(c) $\theta^*(x) = e^x$ (d) $\theta^*(x) = e^x$

Figure 1: Estimated regression functions in the examples in Section 4. Solid lines are for the truth; dotted lines are for estimate from a single MC replicate; dashed lines are for $\tilde{\theta}_1(x)$; dash-dotted lines are for $\tilde{\theta}_2(x)$; long-dashed lines are for $\tilde{\theta}_3(x)$.



(a) $\theta^*(x) = \log(2x + 1)$

(b) $\theta^*(x) = x^3$

(c) $\theta^*(x) = e^{2x}$

(d) $\theta^*(x) = I(x \leq 0.6)10x/6 + I(x > 0.6)$

Figure 2: Estimated regression functions from one MC replicate resulting from three methods. Solid lines are for the truth; dashed lines are for the SNP estimation; dash-dotted lines are for pcls; dotted lines are for monreg.

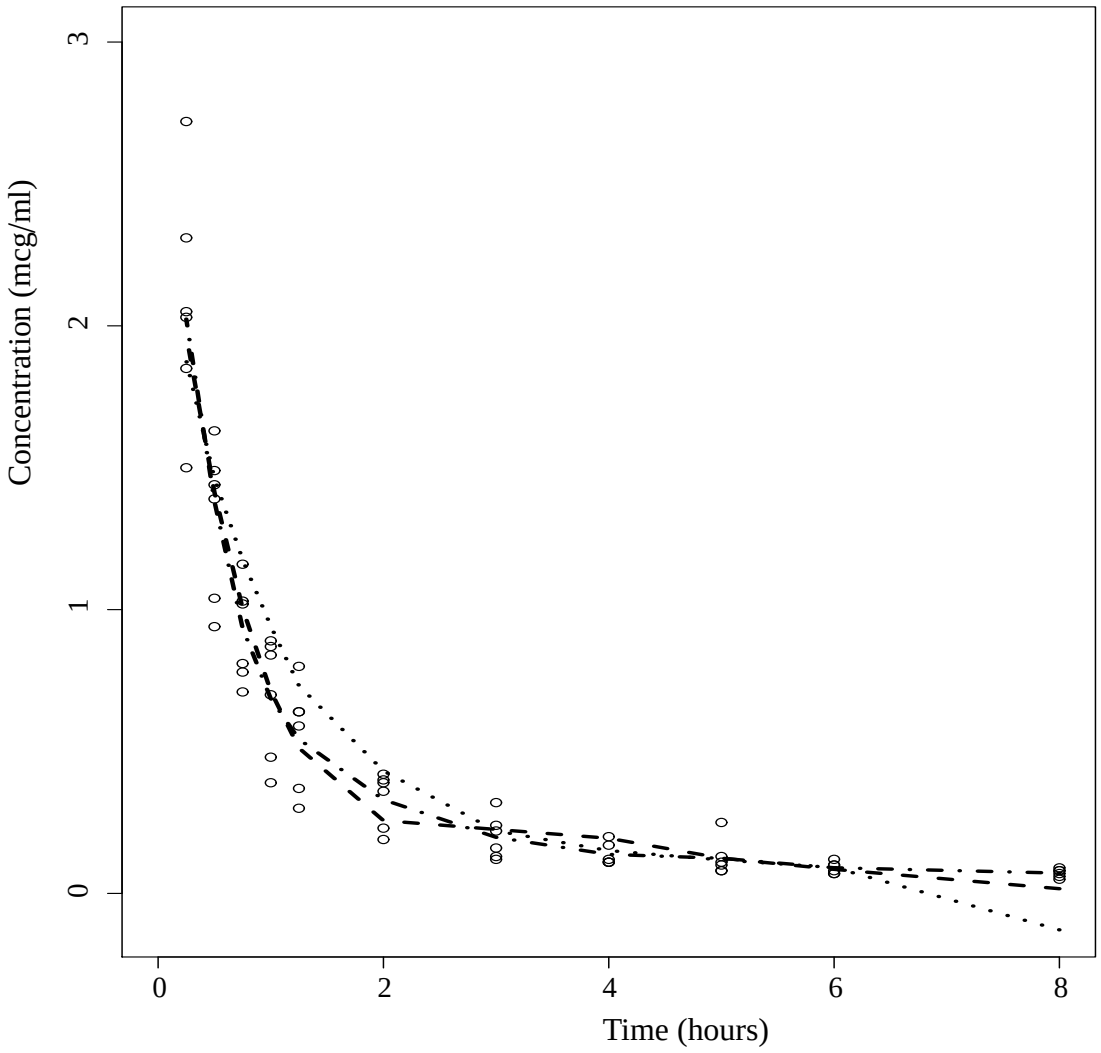


Figure 3: Fitted regression lines for the indomethacin data resulting from three methods: SNP with $K = 3$ (dashed line), pcls (dot-dashed line), and monreg (dotted line).