Semiparametric Estimations under Shape Constraints with Applications to Production Functions

Ximing Wu* Robin Sickles†

January 8, 2013

Abstract

Economic theories often provide guidance on econometric modeling. We propose semiparametric estimators consistent with monotonicity and concavity constraints using the method of penalized splines. The shape constraints are maintained via integral transformations of spline basis expansions. We present an effective algorithm, the large sample properties, methods of inferences and of smoothing parameter selection of our estimators. We also consider multiple regressions under the framework of additive models. We conduct Monte Carlo simulations to illustrate the finite sample performance and usefulness of the proposed method. We applied the proposed methods to estimate: (a) the relationship between individuals' degree of optimism and risk tolerance; (b) a production function with multiple inputs.

1 Introduction

Economic theories can provide useful guidance on the modeling of real world data. For instance, utility functions associated with rational preference are monotone; under convex preference, we also obtain quasiconcavity. Demand functions of normal goods are downward sloping (Matzkin, 1991). According to the duality theorem, profit functions are concave in

*Department of Agricultural Economics, Texas A&M University; email: xwu@tamu.edu
†Department of Economics, Rice University; email: rsickles@rice.edu

output price, and cost functions are monotonically increasing and concave in input price. Researchers, when trying to model economic relationships, often face two challenges, fidelity to economic theories and flexibility in functional forms (cf. Diewert and Walers, 1987). In addition, these two goals are often at odds: conformity to theories usually dictates rigid functional forms, while flexible parameterizations sometimes lead to implausible predictions.

One fruitful approach of research to tackle this dilemma is to use nonparametric or seimparametric methods subject to restrictions suggested by economic theories. For general reviews of this body of research, see Matzkin (1994) and Chapter 6 of Yatchew (2003). Following this line of thought, in this paper we present a flexible semiparametric estimator that incorporates shape constraints. We focus on functional relationships with two shape constraints: monotonicity and concavity because this is the class of functions encountered most frequently in economic studies. Functional relationships with either one of these two constraints are special cases of our estimator. Convexity can be easily accommodated by a simple negation of one parameter in our model.

Following Ramsay's (1998) monotone smooth estimator, we use integral transformation defined by differential equations to impose shape restrictions. A key advantage of this transformation approach is that it transforms a constrained problem into an unconstrained one. We subsequently model the unconstrained problem using penalized spline methods, resulting in a nonlinear semiparametric estimator. We show that careful choice of transformation and model-based penalty can simplify the estimation considerably.

We propose an iterative algorithm to solve the proposed estimator. We establish the consistency of the estimator and present approximate methods of inferences and smoothing parameter selection. We then extend our estimators to multiple regressions under the framework of additive models. We illustrate the finite sample performance and usefulness of our methods with Monte Carlo simulations and two empirical applications.

The rest of the paper is organized as follows. Section 2 briefly reviews the relevant literature and then presents a transformation-based model to accommodate shape restrictions. Section 3 proposes a Gauss-Jordan algorithm to solve the estimator. Sections 4 and 5 discuss methods of inferences and model specification. Section 6 extends the model to multiple regressions. Sections 7 and 8 report Monte Carlo simulations and two empirical examples.

The last section concludes. A technical appendix gathers all proofs.

2 Model and Estimator

Several approaches have been used to impose restrictions in statistical estimations. A simple approach is the transformation of variables. For instance, logarithmic transformation is commonly used to ensure positiveness of the predicted outcomes, and Box-Cox transformation offers a more flexible alternative. In the estimation of production functions, the Cobb-Douglas, constant elasticity of substitution (CES), trans-log and generalized Leontief specifications are commonly employed. These functional forms are often chosen because they satisfy certain theoretical properties and also out of their simplicity. Simple parametric forms, however, can sometimes entail nontrivial restrictions. For example, a logarithm transformation of the dependent variable implies multiplicative errors rather than the usual additive ones; in addition, all inputs are restricted to have positive marginal risk (measured by the conditional variance of output). Interested readers are referred to Just and Pope (1978) for a general treatment on the economic implications of specification of production functions.

To avoid rigid functional forms, semiparametric and nonparametric methods have been used to accommodate shape restrictions. An early example is Brunk's (1955) isotonic estimator, which essentially produces a monotone step function. Mukerjee (1988) and Mammen (1991) developed kernel-based isotonic regression techniques which consist of a kernel smoothing step and an isotonization step to ensure monotonicity. Instead of isotonization, Hall and Huang (2001) suggested a penalized kernel method to obtain monotonicity. Their method is further generalized by Racine and Parmeter (2008) and Ma and Racine (2012) to allow for more general constraints. Another popular family of smoothers, the spline-based methods, has also been called upon. For example, Ramsay (1988), Kelly and Rice (1990), and Mammen and Thomas-Agnam (1999) proposed monotone estimators based on shape preserving spline basis functions. A third possibility is to use the technique of rearrangement or data sharpening (cf. Braun and Hall (2001) and Chernozhukov, et al. (2007)).

Our estimator is inspired by the smooth monotone estimator of Ramsay (1998). Suppose $y = f(x)$ is a smooth monotone function of x. For simplicity, we assume that $x \in [0, 1]$.

Ramsay (1998) proposed to model an unknown monotone function via the following integral transformation:

$$f(x) = \int_0^x \exp(r(s))ds, \tag{1}$$

where r is a square integrable function on $[0,1]$. Since $f'(x) = \exp(r(x)) > 0$ for all x, the monotone restriction is satisfied. Unlike some penalty-based monotone estimators that impose observation-specific monotonicity, (1) is globally monotone thanks to the positive exponential functional embedded in the integral transformation.

Since $f''(x) = f'(x)r'(x)$ and $f'(x) > 0$, $f(x)$ is concave if $r'(x) < 0$ for all x. Our strategy is to use the integration transformation (1) one more time to ensure $r'(x) < 0$. In particular, we consider the following parameterization

$$f(x) = \int_0^x \exp(-\int_0^s g(t)dt)ds. \tag{2}$$

It follows that $f'(x) = \exp(-\int_0^x g(t)dt) > 0$ and $f''(x) = -f'(x)g(x)$, implying that $f''(\cdot) < 0$ if $g(\cdot) > 0$. Thus under (2), the monotonicity and concavity constraints are reduced to a simple positiveness constraint that $g(x) > 0$ for all x. Natural candidates of g include $g(x) = x^2$ and $g(x) = \exp(x)$; other choices are certainly possible. Below we will show that $g(x) = x^2$ is particularly appealing for the proposed method on theoretical and practical grounds.

The parameterization (2) can be characterized by the following differential equation

$$g(x) = -\frac{f''(x)}{f'(x)}.^{1}$$

The solution is given by

$$f(x) = \beta_0 + \beta_1 \int_0^x \exp(-\int_0^s g(t)dt)ds,$$

where β_0 and β_1 are generic constants.

Given an iid random sample $\{Y_i, X_i\}_{i=1}^n$ with $X_i \in [0,1]$, we consider the following sta-

[1]The quantity g reflects the relative velocity of f. Interestingly, we note that this is also the parameterization used to derive Arrow-Pratt utility.

tistical model for a monotone and concave functional relationship

$$Y_i = f(X_i) + e_i = \beta_0 + \beta_1 \int_0^{X_i} \exp(-\int_0^s g(t)dt)ds + e_i, \tag{3}$$

where $g(\cdot) > 0$ and e_i are iid error terms with mean zero and a finite variance σ^2. Furthermore, we will model $g(\cdot)$ by $g \circ h(\cdot)$, where h is a square integrable function defined on $[0, 1]$ free of constraints.

One major advantage of the transformation-based approach to incorporate constraints is that we can transform a constrained problem into an unconstrained one. In our case, this boils down to the modeling of h. Lacking theoretical guidance or a priori information, we opt to model h using a flexible nonparametric estimator. Specifically, we use the spline method since it is relatively easy to embed smoothers in nonlinear functionals or to implement additive structures in multiple regressions using splines. A spline is defined as a piecewise polynomial that is smoothly connected at its joints (knots). Because of their local nature, do not suffer from the oscillations associated with global polynomials such as the power series.

There exist many types of splines, such as the truncated power series, B-splines, radial splines, periodic splines and thin-plate splines, just to name a few (cf. de Boor (2001) for a general treatment of splines). Let $0 < k_1 < \cdots < k_M < 1$ be a series of knots of the spline basis functions. The popular truncated power series splines are given by

$$\Phi(x) = (1, x, \ldots, x^p, (x - k_1)_+^p, \cdots, (x - k_M)_+^p)^T,$$

where $(x)_+ = \max(x, 0)$, and p is a positive integer. Define $h(x) = c^T\Phi(x)$ with c being a vector of coefficients with compatible dimension. This construction, a linear combination of spline basis functions, is a powerful tool of curve fitting. The degree of smoothness of spline approximation is controlled by p: a linear combination of spline basis functions of degree p is a p degree polynomial on each subinterval $[k_m, k_{m+1}]$ has $p - 1$ continuous derivatives on its entire domain. The global polynomials control the overall shape of a curve, while the spline basis functions pick up local features. For flexibility and numerical stability, a common practice in spline approximation is to employ a large number of low order spline basis functions (i.e., large M, small p).

In practice, truncated power series are often transformed to B-splines, which are the maximally differentiable interpolative basis functions. The B-splines are generalizations of Bézier curve and can be constructed recursively (cf. Eilers and Marx (1996)). B-splines sometimes facilitate theoretical analysis and usually produce better finite sample performance.

Let $P = 1 + p + M$. Φ is a P-dimensional basis functions. We consider the following model

$$Y_i = f(X_i; \beta, c) + e_i = \beta_0 + \beta_1 \int_0^{X_i} \exp\left(-\int_0^s g(c^T \Phi(t))dt\right) ds + e_i. \tag{4}$$

We need the intercept β_0 and 'slope' β_1 here for identification because the parameterization of f does not allow for free location and scale parameters. To see this, consider the simplest case $g(x) = a$, where a is a non-zero constant. It follows that $f(x) = (1 - \exp(-ax))/a$, whose location and scale can not independently vary.

Model (4) is a semiparametric model with two parametric coefficients and a nonparametric smoother g. To balance fidelity to the data and smoothness of the estimator, we adopt the approach of penalized spline estimation. This method uses a relatively generous spline basis and shrinks all coefficients towards zero to avoid overfitting. We choose this approach because the delicate balance between goodness-of-fit and smoothness is governed by a single smoothing parameter and therefore easier to implement.[2]

In particular, we estimate model (4) using the penalized least squares, minimizing the sum of squared residuals plus a penalty on the roughness of f. The objective function is given by

$$Q_\lambda(\beta, c) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(X_i; \beta, c)\right)^2 + \lambda D(f), \tag{5}$$

where $D(f) > 0$ reflects the roughness of f. For the pth degree splines, a popular choice of the penalty is the integrated squared qth derivative of f, $q \leq p$. For example, the integrated quadratic penalty with $q = 2$ is commonly used, which leads to the natural cubic splines in smoothing splines.

In penalized spline estimations, we can in principle specify the basis functions and the

[2]An alternative to the penalized spline method is the regression splines method, which balances the goodness-of-fit and smoothness trade-off through judicious selection of spline basis functions. The selection of basis functions for regression splines can be a daunting task, especially in multiple regressions. Consider a candidate set of P basis functions. A complete subset selection, which exhausts all possible combinations of the basis functions, entails 2^P evaluations of candidate models.

penalty separately. Nonetheless for nonlinear models, careful choice of penalty with respect to the form of f can sometimes improve the estimation considerably. For instance, Heckman and Ramsay (2000) showed that proper model-based penalties can reduce the number of spline basis functions and the approximation bias at the same time, resulting in smaller mean square errors. In our case a natural choice of the penalty is the integrated relative curvature; that is, $D(f) = -\int_0^1 f''(x)/f'(x)dx = \int_0^1 g(x)dx$. Since $g(x) > 0$ by construction, it consists a valid roughness penalty. This penalty on the relative curvature penalizes not only the curvature of f but also small values of f'. Consequently, it prevents the 'boundary' solutions where $f'(x) = 0$.

3 Algorithm

Denote the solution to the proposed nonlinear estimation (5) by $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^T$ and $\hat{c}$. Let $m(x; c) = \int_0^x \exp(-\int_0^s g(c^T\Phi(t))dt)ds$. It follows that $D(f) = D(m)$. Define $\hat{m}(X_i) = m(X_i; \hat{c})$ and $g'(x) = dg(x)/dx$. Replacing β with $\hat{\beta}$ and applying Taylor expansion to m in (5) with respect to c around $\hat{c}$ yields

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 m(X_i;\hat{c}) - \hat{\beta}_1\hat{Z}_i(c-\hat{c})\right)^2 + \lambda D, \tag{6}$$

where

$$\hat{Z}_i = \frac{\partial\hat{m}(X_i;\hat{c})}{\partial c} = -\int_0^{X_i}\left\{\int_0^s (\Phi(t)g'(\hat{c}^T\Phi(t))dt)\exp(\int_0^s -g(\hat{c}^T\Phi(t))dt)\right\}ds.$$

The first order condition of (6) with respect to c is given by

$$-\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1\hat{Z}_i^T(Y_i - \hat{\beta}_0 - \hat{\beta}_1\hat{m}(X_i) - \hat{\beta}_1\hat{Z}_i(c-\hat{c})) + \lambda D' = 0, \tag{7}$$

where

$$D' = \frac{\partial D}{\partial c} = \int_0^1 \Phi(x)g'(c^T\Phi(x))dx.$$

Next denote $\hat{D} = D(\hat{m})$ and $\hat{D}'$ and $\hat{D}''$ its first and second derivatives with respect to c

evaluated at $\hat{c}$. Taking Taylor expansion of D' with respect to c around $\hat{c}$ yield

$$-\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1\hat{Z}_i^T(Y_i-\hat{\beta}_0-\hat{\beta}_1\hat{m}(X_i)-\hat{\beta}_1\hat{Z}_i(c-\hat{c}))+\lambda\hat{D}'+\lambda\hat{D}''(c-\hat{c})\approx 0. \tag{8}$$

Define $\hat{e}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1\hat{m}(X_i)$. Plugging $\hat{e}_i$ into (7) and rearranging terms yield

$$(\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1^2\hat{Z}_i^T\hat{Z}_i+\lambda\hat{D}'')(c-\hat{c})\approx\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1\hat{Z}_1^T\hat{e}_i-\lambda\hat{D}'. \tag{9}$$

Expression (9) suggests a Gauss-Jordan algorithm to solve for the proposed estimator. Let $\hat{c}_-$ be the current estimate of c and $\hat{m}(X_i), \hat{Z}_i, \hat{D}', \hat{D}''$ and $\hat{e}_i$ be evaluated at $c = \hat{c}_-$. Denote $Y = (Y_1, \ldots, Y_n)^T$ and $\hat{m} = (\hat{m}(X_1), \ldots, \hat{m}(X_n))^T$. Taking $\hat{m}$ as given, we calculate $\hat{\beta}$ via the ordinary least squares by regressing Y on $\hat{m}$ and a constant one. Next holding $\hat{\beta}$ constant, we update c according to the following formula:

$$\hat{c} = \hat{c}_- + \left(\frac{1}{n}\hat{\beta}_1^2\hat{Z}^T\hat{Z}+\lambda\hat{D}''\right)^{-1}\left(\frac{1}{n}\hat{\beta}_1\hat{Z}^T\hat{e}-\lambda\hat{D}'\right), \tag{10}$$

where $\hat{e} = (\hat{e}_1, \ldots, \hat{e}_n)^T$ and $\hat{Z} = \left(\hat{Z}_1^T, \ldots, \hat{Z}_n^T\right)^T$. $\hat{\beta}$ and $\hat{c}$ are updated alternatively in this fashion until convergence.

Remark 1. *The penalty $D(\hat{m})$ and their derivatives $\hat{D}'$ and $\hat{D}''$ generally depend on the current estimate $\hat{c}_-$ and therefore need to be calculated anew at each stage of the updating. This updating process is simplified when $g(x) = \frac{1}{2}x^2$. Recall that $h(x) = c^T\Phi(x)$. Define $K = \int_0^1 \Phi(x)\Phi^T(x)dx$. It follows that $D(m) = \frac{1}{2}c^TKc$ and the updating formula* (8) *simplifies to*

$$\hat{c} = \hat{c}_- + \left(\frac{1}{n}\hat{\beta}_1^2\hat{Z}^T\hat{Z}+\lambda K\right)^{-1}\left(\frac{1}{n}\hat{\beta}_1\hat{Z}^T\hat{e}-\lambda K\hat{c}_-\right).$$

Thus with a quadratic g, the penalty weight matrix remains a constant that does not depend on unknown parameters. Moreover, the Taylor expansion given by (8) *is exact.*

Remark 2. *The convergence of the estimation is usually quite speedy. To assure that each step improves the penalized objective function, we also implement a step-halving procedure. Whenever an updating step in c fails to improve the objective function* (6)*, we divide it by*

two to mitigate overshooting. This adjustment further improves the numerical stability of the proposed algorithm.

4 Large Sample Properties and Inferences

Despite the popularity of penalized spline methods, their theoretical properties have not been well understood. Some earlier results were provided in Wand (1999), Aerts et al. (2002) and Yu and Ruppert (2002) under the framework that the dimension of the spline basis is sufficiently large and fixed. Hall and Opsomer (2005) investigated this problem using a white noise representation. Claeskens et al. (2008) showed that if the number of knots increases with sample size to infinity, the asymptotic properties of penalized splines are similar either to those of regression splines or of smoothing splines.[3] Kauerrmann et al. (2009) studied the asymptotic properties of penalized splines for generalized linear models under the regression splines scenario. Li and Ruppert (2008) used the device of equivalent kernel to study the smooth splines scenario.

Following Wand (1999), Aerts et al. (2002) and Yu and Ruppert (2002), we study the asymptotic behavior of the proposed methods under the premise that the fixed number of spline basis functions is sufficiently large such that approximation error is negligible. As in nonparametric modeling, the model is flexible enough to adapt to regression functions of unknown form. As in parametrci modeling, the number of parameter is fixed, and the parameters can be estimated at $\sqrt{n}$ rates. Fixed-knot asymptotics converges to a known normal distribution and thus facilitates inferences.

To facilitate the derivation, we first present an alternative representation of solution (10). Given current estates $\hat{\beta}$ and $\hat{c}_-$, define the 'pseudo regressand' $\tilde{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 \hat{m}(X_i) + \hat{\beta}_1 \hat{Z}_i \hat{c}_-$. Plugging $\tilde{Y}_i$ into (7) and rearranging terms yields

$$(\frac{1}{n}\hat{\beta}_1^2 \hat{Z}^T \hat{Z} + \lambda \hat{D}'')\hat{c} \approx \frac{1}{n}\hat{\beta}_1 \hat{Z}^T \tilde{Y} + \lambda(\hat{D}' - \hat{D}''\hat{c}_-),$$

where $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^T$. Holding $\hat{\beta}$ constant, we can update c using the following alternative

[3]Smoothing spline is a special case of the penalized spline estimation where the number of basis functions equals the number of unique observations, cf. Wahba (1990) for a general treatment of smoothing splines.

formula:

$$\hat{c} = (\frac{1}{n}\hat{\beta}_1^2 \hat{Z}^T \hat{Z} + \lambda \hat{D}'')^{-1} \quad \frac{1}{n}\hat{\beta}_1 \hat{Z}^T \tilde{Y} + \lambda(\hat{D}' - \hat{D}''\hat{c}_-) \quad . \tag{11}$$

Remark 3. *When $g = \frac{1}{2}x^2$, we have $D(m) = \frac{1}{2}c^T Kc$ and $D' - D''c = 0$, resulting in a simpler updating process*

$$\hat{c} = (\frac{1}{n}\hat{\beta}_1^2 \hat{Z}^T \hat{Z} + \lambda K)^{-1} \quad \frac{1}{n}\hat{\beta}_1 \hat{Z}^T \tilde{Y} \quad .$$

Since $\hat{\beta}$, $\hat{Z}$ and $\tilde{Y}$ all depend on the current estimate $\hat{c}_-$, iterations are still called for.

This representation (11) of c as a linear function of $\tilde{Y}$ allows us to use known results on linear smoothers for inferences.

Denote $\theta(\lambda) = (\beta(\lambda), c(\lambda))$. We emphasize the dependence of the estimators on the smoothing parameter in this section as the asymptotics depends on whether λ is fixed or goes to zero asymptotically. In particular, we shall denote by λ a fixed smoothing parameter and by λ_n one dependent on the sample size.

We need the following assumptions to obtain consistency.

Assumption 1. $\{X_i, Y_i\}$ are iid random samples such that

$$Y_i = f(X_i;\theta) + e_i = \beta_0 + \beta_1 \int_0^{X_i} \exp\left(-\int_0^s g(c^T\Phi(t))dt\right) ds + e_i, \tag{12}$$

where e_i's are iid random error with mean zero and finite variance $\sigma^2 > 0$.

Assumption 2. For all x, the conditional mean function $f(x;\theta)$ is continuous in $\theta \in \Theta$, which is compact.

Assumption 3. (a) $\frac{1}{n}\sum_{i=1}^n \{f(x_i;\theta^*) - f(x_i;\theta)\}^2$ converges to some limit function uniformly in $\theta^*, \theta \in \Theta$; (b)

$$\mathcal{Q}(\theta) = \lim_n \frac{1}{n}\sum_{i=1}^n (f(X_i;\theta) - f(X_i;\theta^0))^2.$$

has a unique minimum at $\theta = \theta^0$.

Theorem 1. *Under assumptions 1-3, if the smooth parameter $\lambda_n = o(1)$, then a sequence of penalized least estimators minimizing the objective function* (5) *exists and $\hat{\theta}(\lambda_n) \xrightarrow{p} \theta^0$ as*

$n \to \infty$.

Remark 4. *The variance of $\hat{\theta}(\lambda_n)$ goes to 0 as n tends to ∞ whether or not λ_n tends to 0. However, if $\lambda_n \to 0$ as $n \to \infty$, then the bias also tends to 0, and consistency can be established.*

Next we derive the asymptotic normality. We first derive the asymptotics with λ fixed. This is needed for finite sample inference. Denote $\theta(\lambda) = (\beta(\lambda), c(\lambda)$, coefficients dependent on λ. Let $W(\lambda)$ be a $n \times 2$ matrix with the ith row $W_i = (1, m(X_i; c(\lambda))), i = 1, \ldots, n$. Define

$$\begin{aligned} P_W(\lambda) &= W(\lambda)(W(\lambda)^T W(\lambda))^{-1} W(\lambda)^T, \\ P_Z(\lambda) &= (\beta_1(\lambda) Z(\lambda))(\beta_1^2(\lambda) Z(\lambda)^T Z(\lambda) + n\lambda D'')^{-1}(\beta_1(\lambda) Z^T(\lambda)), \end{aligned} \tag{13}$$

and $\hat{W}(\lambda)$, $\hat{P}_W(\lambda)$ and $\hat{P}_Z(\lambda)$ their analogs evaluated at $\hat{\theta}(\lambda)$, the penalized least squares estimators.

Under the assumption of iid errors, the variance σ^2 is estimated by the sum of squared residuals divided by proper degrees of freedom. Our semiparametric estimator has two parametric parameters β_0 and β_1, and a nonparametric smoother $m(X; c)$. The degrees of freedom of the smoother, which can be viewed as its equivalent number of coefficients to that of a power series approximation, is calculated as $\mathrm{tr}(\hat{P}_Z(\lambda))$. Therefore we estimate σ^2 with

$$s^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n - \mathrm{tr}(\hat{P}_Z(\lambda)) - 2}.$$

Remark 5. *Alternatively, we can use the degrees of freedom of the residuals in the calculation of variance. For linear smoothers, the residual degrees of freedom is given by $2tr(\hat{P}_Z(\lambda)) - tr(\hat{P}_Z^2(\lambda))$, cf. Ruppert et al. (2003) and references therein. In practice, these two specifications often give similar results.*

We make the following assumptions for asymptotic normality.

Assumption 4. Suppose the following penalized objective function

$$\mathcal{Q}_\lambda(\theta) = \mathcal{Q}(\theta) + \lambda D(f(\theta))$$

has a unique minimum at $\theta(\lambda)$ and $\theta(\lambda) \in \Theta$, where λ is positive and finite.

Assumption 5. The conditional mean function $f(\cdot;\theta)$ is twice continuously differentiable in a neighborhood of $\theta(\lambda)$, and $P_W(\lambda)$ and $P_Z(\lambda)$ converge uniformly in θ in a neighborhood of $\theta(\lambda)$.

Below we present an asymptotic normality of the estimator. We choose to report results for the predicted values because the coefficients of the models are not of direct interest. We can construct confidence intervals for quantity of interest, for instance the marginal value of productivity in the estimation of production functions, based on the asymptotic properties of the estimators.

Theorem 2. *Given a fixed smoothing parameter λ. Under assumptions 1, 2, 3(a), 4 and 5, a sequence of penalized spline estimators $\hat{\theta}(\lambda) \xrightarrow{p} \theta(\lambda)$ as $n \to \infty$. Denote $Y(\lambda) = f(X;\theta(\lambda))$ and $\hat{Y}(\lambda) = f(X;\hat{\theta}(\lambda))$. Then $\sqrt{n}(\hat{Y}(\lambda) - Y(\lambda) \xrightarrow{d} \mathcal{N}(0, V(\lambda))$ as $n \to \infty$, where*

$$V(\lambda) = \sigma^2(P_W(\lambda) + P_Z^2(\lambda)). \tag{14}$$

Define $\hat{V}(\lambda) = s^2(\hat{P}_W(\lambda) + \hat{P}_Z^2(\lambda))$. $\hat{V}(\lambda) \xrightarrow{p} V(\lambda)$ as $n \to \infty$.

Denote by $\hat{V}_i(\lambda)$ the ith diagonal element of $\hat{V}(\lambda)$. We construct the asymptotic $(1-\alpha)\%$ confidence interval of $\hat{Y}_i$ by

$$\hat{Y}_i \pm z_{1-\alpha/2}\sqrt{\hat{V}_i(\lambda)}, \tag{15}$$

where $z_{1-\alpha/2}$ is the critical value from the standard normal distribution at the confidence level α.

Remark 6. *The confidence interval* (15) *is about $Y(\lambda) = E[f(\cdot;\hat{\theta}(\lambda)]$, the best projection, rather than $f(\cdot;\theta^0)$. This is a well-known issue with series-based nonparametric estimations, of which the bias terms are generally not available. Although bias is inherent in nonparametric regression, approximate unbiasedness is often assumed and* (15) *can be interpreted as approximate confidence interval. Since this approximate confidence interval is oftentimes over optimistic, Hastie and Tibshirani (1990) suggested replacing $z_{1-\alpha/2}$ in* (15) *with $t_{1-\alpha/2,df}$, where df is the proper degrees of freedom for nonparametric regressions. Eubank (1999) suggested Bonferroni methods to calculate confidence bands. Ruppert et al. (2003) discussed bias-corrected confidence intervals.*

Remark 7. *Our estimator is semiparametric with two parametric coefficients. Taking $\hat{m}$ as nuisance parameters, the estimator can be viewed as a two-step estimator with nonparametric first step estimates. Newey (1994) and Ai and Chen (2007) discussed the estimation of asymptotic semiparametric variance of the second stage estimates. Recently Acerberg et al. (2011) showed that the asymptotic parametric variance that ignores the nonparametric nature of the first stage (for instance, the method of Newey (1984)) is numerically identical to the semiparametric variance. In particular, Acerberg et al. (2011) provided several examples that use sieve estimators in the first step. The penalized spline estimator investigated in this study fits into their framework naturally.*

Remark 8. *We present the alternative representation* (11) *to facilitate the asymptotic analysis. Our numerical experiments indicate that the Gauss-Jordan algorithm given in the previous section is usually more robust and converges faster, especially when a non-qudratic g is used. We recommend the Gauss-Jordan algorithm for the calculation of our estimator.*

Lastly, we derive the asymptotics with $\lambda_n \to 0$, corresponding to the limiting case where the shrinkage bias is asymptotically negligible. Define $P_W^0 = P_W(\theta^0)$ and $P_Z^0 = P_Z(\theta^0)$. We can then establish the following result.

Theorem 3. *Under the assumptions 1, 2, 3, and assumptions 4 and 5 with $\lambda = 0$, if the smoothing parameter $\lambda_n = o(n^{-1/2})$, then a sequence of penalized spline estimator $\hat{\theta}(\lambda_n) \xrightarrow{p} \theta^0$ as $n \to \infty$. Denote $\hat{Y}(\lambda_n) = f(X; \hat{\theta}(\lambda_n))$. Then $\sqrt{n}(\hat{Y}(\lambda_n) - Y) \xrightarrow{d} \mathcal{N}(0, V^0)$ as $n \to \infty$, where*

$$V^0 = \sigma^2(P_W^0 + P_Z^0). \tag{16}$$

Remark 9. *The limiting $P_Z(\lambda_n)$ is obtained by setting $\lambda_n = 0$, yielding*

$$P_Z^0 = Z(Z^T Z)^{-1} Z^T.$$

Since P_z^0 is now idempotent, we have P_Z^0 instead of $(P_Z^0)^2$ as in (14)*. For finite sample inference, one would expect V^0 overestimate the variance of $\hat{\theta}(\lambda_n)$ for a given $\lambda_n > 0$.*

5 Specification of Spline Basis and Smoothing Parameter

Implementation of the penalized spline estimators entails the specification of spline basis functions and smoothing parameters. The former includes the type of splines, number and location of knots. Commonly used splines include the truncated power series, B-splines and radial basis splines. The spline literature indicates that the practical differences among these splines are oftentimes quite small.

Because the penalized splines estimations normally use a relatively generous spline basis, the number and location of knots play a relatively minor role in the estimation. We follow the automatic knot selection rule in Ruppert (2002), where the number of knots is given by

$$M = \min(\frac{1}{4} \times \text{number of unique}\, X,\, 35), \tag{17}$$

and the knots are placed at the $m/(M+1)$th sample quantile of the unique X's for $m = 1, \ldots, M$.

It is well known that spline estimators depend crucially on the smoothing parameter (cf. Ruppert, 2002). A commonly used approach of smoothing parameter selection is the principle of cross validation (CV). Let $\hat{Y}_{(i)}$ be the prediction of Y_i by a given estimator that uses all but the ith observation. The 'leave-one-out' least squares cross validation criterion, in terms of sum of squared residuals, is given by

$$CV = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{(i)} \right)^2.$$

Direct implementation of the cross validation is straightforward but often costly, especially for nonparametric estimators without analytical solutions. For linear estimators, there exists an exact formula to evaluate the least squares cross validation criterion function, using only regression results based on the full sample. This exact solution usually does not exist for nonlinear estimations. Nonetheless, there exist approximate formulations that have been shown to give rather close results. Below we derive an approximate formula of the cross validation criterion for the proposed estimator. For $i = 1, \ldots, n$, denote by $\hat{c}_{(i)}$ the solution

to

$$\frac{1}{n}\sum_{k=1,k\neq i}^{n}\left(Y_k-\beta_0-\beta_1 m(X_k;c)\right)^2+\lambda D(m(x)),$$

and $\hat{Y}_{(i)}$ be the prediction of Y_i evaluated at $\hat{c}_{(i)}$. We establish the following result.

Theorem 4. *Let s_i be the ith diagonal element of P_Z given in (13) and $\hat{s}_i$ its corresponding sample analog, $i=1,\ldots,n$. The Cross Validation (CV) criterion satisfies*

$$CV=\sum_{i=1}^{n}(Y_i-\hat{Y}_{(i)})^2=\sum_{i=1}^{n}\left(\frac{Y_i-\hat{Y}_i}{1-\hat{s}_i}\right)^2+o_p(1). \tag{18}$$

The Generalized Cross Validation (GCV) is a popular and often more robust alternative to the CV criterion . It can be obtained by replacing $1-\hat{s}_i$ in (18) with $1-\frac{1}{n}\mathrm{tr}(\hat{P}_Z)$ (cf. Wahba, 1990). One can infer readily from Theorem 4 that in our case

$$\mathrm{GCV}\approx\sum_{i=1}^{n}\left(\frac{Y_i-\hat{Y}_i}{1-\frac{1}{n}\mathrm{tr}(\hat{P}_Z)}\right)^2.$$

Remark 10. *An alternative criterion of smoothing parameter selection is the estimated risk (cf. Eubank 1999). Although conceptually simple, this criterion requires a proper prior estimate of σ^2, which complicates the issue since optimal smoothing parameter for conditional mean estimations generally is not optimal for variance estimations.*

Remark 11. *Both the cross validation and estimated risk criteria fall into the category of model selection approach of smoothing parameter selection. Another possibility is likelihood based method that treats spline coefficients as random coefficients. In particular, spline coefficients are assumed to follow zero mean Gaussian distributions and estimated using mixed effect random coefficient models. Cf. Wand (2006) for an overview of this approach.*

6 Multiple regressions

In this section we consider the case where y is a function of $J(\geq 2)$ variables, being monotone and concave in each one. For multiple regressions, we adopt the convention that all notations,

whenever necessary, are indexed by a subscript to make explicit their dependence on the specific coordinate $j = 1, \ldots, J$. For simplicity, we focus on the case of additive models:

$$Y_i = \beta_0 + \sum_{j=1}^{J} \beta_j m_j(X_{j,i}) + e_i,\ m_j' > 0 \,\text{and}\, m_j'' < 0.$$

For a general treatment of additive models, see Hastie and Tibshirani (1990).

We estimate the additive model using the penalized spline estimator by minimizing the following objective function:

$$\frac{1}{n}\sum_{i=1}^{n} \quad Y_i - \beta_0 - \sum_{j=1}^{J} \beta_j m_j(X_{j,i}) \quad ^2 + \sum_{j=1}^{J} \lambda_j D_j,$$

where $D_j = D(m_j(x))$ and λ_j is the corresponding smoothing parameter for $j = 1, \ldots, J$. To ease notation, we suppress the dependence of various quantities on λ in this section. The Gauss-Jordan algorithm described above for the single covariate case can be extended to the multiple covariates case by updating the coefficients $c_j, j = 1, ..., J$, sequentially via back-fitting. Alternatively, we can update all coefficients simultaneously for possible efficiency gains. For $j, k \in (1, \ldots, J)$, let

$$\hat{S}_j = \hat{\beta}_j \hat{Z}_j^T \hat{e} - \lambda_j \hat{D}_j',$$

and

$$\hat{R}_{j,k} = \begin{cases} \frac{1}{n}\hat{\beta}_j^2 \hat{Z}_j^T \hat{Z}_j + \lambda_j \hat{D}_j'', & \text{if } j = k; \\ \frac{1}{n}\hat{\beta}_j \hat{\beta}_k \hat{Z}_j^T \hat{Z}_k, & \text{if } j \neq k, \end{cases}$$

where $\hat{Z}_j = (\hat{Z}_{j,1}^T, \ldots, \hat{Z}_{j,n}^T)^T$ with $\hat{Z}_{j,i} = \partial m_j(X_{j,i}; \hat{c}_j)/\partial c_j$. Further define $\hat{c} = (\hat{c}_1^T, \ldots, \hat{c}_J^T)^T$, $\hat{S} = (\hat{S}_1^T, \ldots, \hat{S}_J^T)^T$, and

$$\hat{R} = \begin{bmatrix} \hat{R}_{1,1} & \cdots & \hat{R}_{1,J} \\ \vdots & \ddots & \vdots \\ \hat{R}_{J,1} & \cdots & \hat{R}_{J,J} \end{bmatrix}.$$

The coefficients $\hat{c}$ are then updated simultaneously according to

$$\hat{c} = \hat{c}_- - \hat{R}^{-1}\hat{S}. \tag{19}$$

Given the current estimate $\hat{c}$, $\beta = (\beta_0, \ldots, \beta_J)^T$ is calculated using the ordinary least squares estimator. This process is iterated to update c and β alternatively until convergence.

Let W be a n by $J+1$ matrix with the ith row $W_i = (1, m_1(X_{1i}), \ldots, m_J(X_{Ji}))$ and $B = (\beta_1 Z_1^T, \ldots, \beta_J Z_J^T)^T$. Define

$$P_W = W(W^T W)^{-1} W^T,$$
$$P_Z = B^T R B,$$

where R is defined analogously to $\hat{R}$. The residual variance is then estimated by

$$s^2 = \frac{\sum_{i=1}^{n} \hat{e}^2}{n - 1 - J - \mathrm{tr}(\hat{P}_Z)}.$$

The variation bounds of the prediction of the additive model can then be calculated as

$$\hat{V} = s^2(\hat{P}_W + \hat{P}_Z^2).$$

A detailed investigation of the theoretical properties of the multiple regressions is beyond the scope of the current paper. We envision that under similar regularity conditions, the large sample properties derived in the previous section apply here (cf. Aerts, et al. (2002) for asymptotics of penalized spline estimators for additive models.)

7 Monte Carlo Simulations

In this section we use Monte Carlo simulations to assess the finite sample performance of our proposed estimator. We consider the following experiments:

- Experiment I:
$$Y_i = f_1(X_i) + e_i = 1 + \log(0.1 + X_i) + e_i$$

- Experiment II:
$$Y_i = f_2(X_i) + e_i = 5 - 5 \times \exp(1 - X_i) + e_i$$

- Experiment III:

$$\begin{aligned} Y_i &= f_{21}(X_{1i}) + f_{22}(X_{2i}) + e_i \\ &= 1 + 2 \times \log(0.01 + X_{1i}) + 3 \times \log(0.01 + X_{2i}) + e_i \end{aligned}$$

In all three experiments, we set the sample size $n = 100$, X be iid random variables from the standard uniform distribution, and e be iid random variables from the standard normal distribution. Each experiment is repeated 300 times. Experiments I and II study univariate monotone and concave functions, while Experiment III examines an additive function with two components, each being monotone and concave.

In each experiment, we estimate the underlying relationship using the proposed estimator. We use the cubic B-spline basis, and the number of knots is determined according to the automatic knot selection rule (17). We experiment with the CV, GCV and the likelihood based method of smoothing parameter selection. The results are quantitatively similar. To save space, we only report results based on the GCV.

For comparison, we consider two alternative estimators: the cubic smoothing spline estimator and the cubic polynomial estimator. The smoothing spline estimator is most flexible and does not impose any shape constraints. The cubic polynomial estimator represents the other extremum, which is the limiting case of cubic smoothing spline estimator when its smoothing parameter approaches infinity.

We use two criteria to assess the performance of these competing estimators. For goodness-of-fit, we report the mean and median of the mean squared errors across all repetitions. To check their compliances with shape restrictions, we report the percentage of observation-specific monotonicity and concavity of the fitted curves evaluated at sample values.

Denote by 'S-Spline' the shape-restricted semiparametric estimator, and by 'Polynomial' and 'Spline' the cubic polynomial and smoothing spline estimators respectively. Table 1 reports the simulation results. It is seen that S-Spline outperforms the other two estimators in all three experiments in terms of mean-MSE and median-MSE. By construction, monotonicity and concavity are satisfied globally under S-Spline. For the other two estimators,

Table 1: Simulation Results

	Estimator	Experiment I	Experiment II	Experiment III	
Mean-MSE	S-Spline	307	336	1811	
	Polynomial	360	406	2616	
	Spline	388	366	1907	
Median-MSE	S-Spline	254	250	1629	
	Polynomial	314	324	2549	
	Spline	331	277	1846	
Monotonicity (%)	S-Spline	100	100	100	100
	Polynomial	93	96	99	99
	Spline	95	98	92	94
Concavity (%)	S-Spline	100	100	100	100
	Polynomial	70	66	69	68
	Spline	51	51	66	65

we calculate their first and second order derivatives numerically on each sample points. In Experiment III, the monotonicity and concavity percentages are reported separately for the two additive components. Our results show that monotonicity is satisfied in most cases, while the percentages of estimates satisfying concavity range from 50 to 70 percent. This is not unexpected considering that higher order derivatives are generally more difficult to estimate.

Some illustrative results are presented in Figure 1. The left panel reports a typical picture of the regression results. The black curve is the constrained estimate, which is monotone and concave. The red line depicts the polynomial estimate, which appears to be concave on the observed range but is not monotone increasing near the right end. The smoothing splines estimate is the most flexible and exhibits multiple violations of monotonicity and concavity. The right panel plots one estimated curve by the constrained estimator, along with its approximate 95% variation bound by red lines. Also reported is the bootstrapped variation bound, based on 100 re-sampled estimates, in blue lines. One can see that the asymptotic bound closely tracks that produced by a bootstrap procedure, which is computationally more expensive.

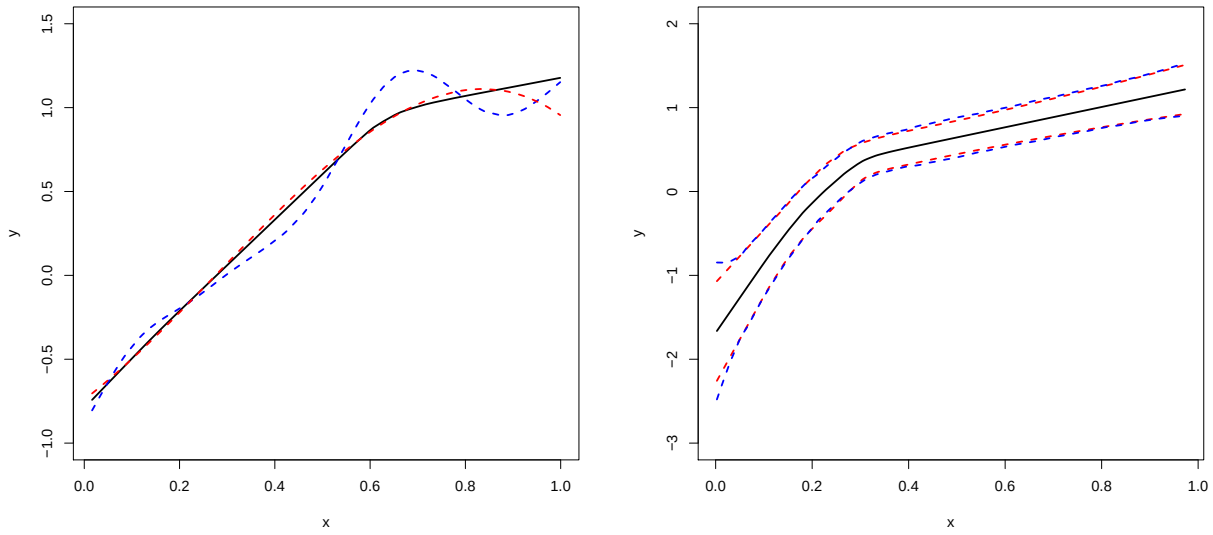

Figure 1: Light: estimated curves (black: S-Spline; red: Polynomial blue: S-Spline); Right: 95% variation bound (red: asymptotic; blue: bootstrap)

8 Empirical applications

In this section, we present two illustrative applications of the proposed method. The first application investigates the relationship between revealed risk attitude and optimism. The data come from a survey conducted by Mansour et al. (2008). In this survey, participants were offered the opportunity to enter a heads-and-tails game. A coin is flipped ten times; each time it comes up heads, the participant is supposed to get 10 euros. The participant is then asked for his own estimation of the number of times heads will occur. The participant is also asked to reveal the maximum amount he is willing to pay in order to take part in this game. The aim of this experiment is to obtain measures of individual levels of optimism and risk aversion. The sample has $n = 1,536$ observations. Summary statistics of the data are reported in the top panel of Table 2. On average, the participants are pessimistic (the average expectation is less than 5, the unbiased expectation) and risk averse (the average WTP 16.3 is below the fair expectation 50 and also below 39, which is the expected risk neutral WTP given the average expectation of 3.9).

For $i = 1, \ldots, n$, let Y_i be individual i's estimation of the number of heads, and X_i his

Table 2: Summary statistics

	Mean	S.D.	Min.	Max.
Risk and Optimism Data				
Optimism	3.9	1.8	0	10
WTP	12.0	13.6	0	100
Production Data				
Output	16.3	8.3	1.7	37.1
Capital	4.8	2.8	9.6	0.3
Labor	57.7	27.2	1.1	98.9

maximum willingness to pay. We are interested in estimating the relationship between these two measures. According to preference and utility function theories, there exists a monotone relationship between risk aversion and optimism (see Mansour et al. (2008) and references therein). Taking the WTP as a proxy for degree of risk aversion or risk loving, one expects a monotone increasing relationship between Y_i and X_i. Since measures of optimism is naturally bounded from above by 10, we expect the Y_i as a function of X_i to level off as X_i gets large (there is no upper bound for X_i; but as expected, no participants offered more than 100 euros). Therefore, it is plausible that $Y = f(X)$ is monotone increasing and concave.

The upper right plot of Figure 2 shows the participants' answers to the two questions, clearly implying a monotone and possibly concave relationship between these two measures. In our investigation, we apply the proposed method to the following model:

$$Y_i = f(X_i) + e_i, i = 1, \ldots, n,$$

where $f' > 0$, $f'' < 0$, and e_i are iid errors with mean zero and finite variance. For comparison, we also estimate the model using the cubic polynomial estimator and the cubic smoothing spline estimator. The estimation results are reported Figure 2. All three estimators capture the general patterns of the data. Also plotted are the 95% confidence intervals. Across all three estimates, the confidence intervals are tighter for small values of WTP and gradually increase with WTP, largely due to that the number of observations decreases rapidly with WTP.

The smoothing spline estimate clearly fails to be monotone increasing. A close exami-

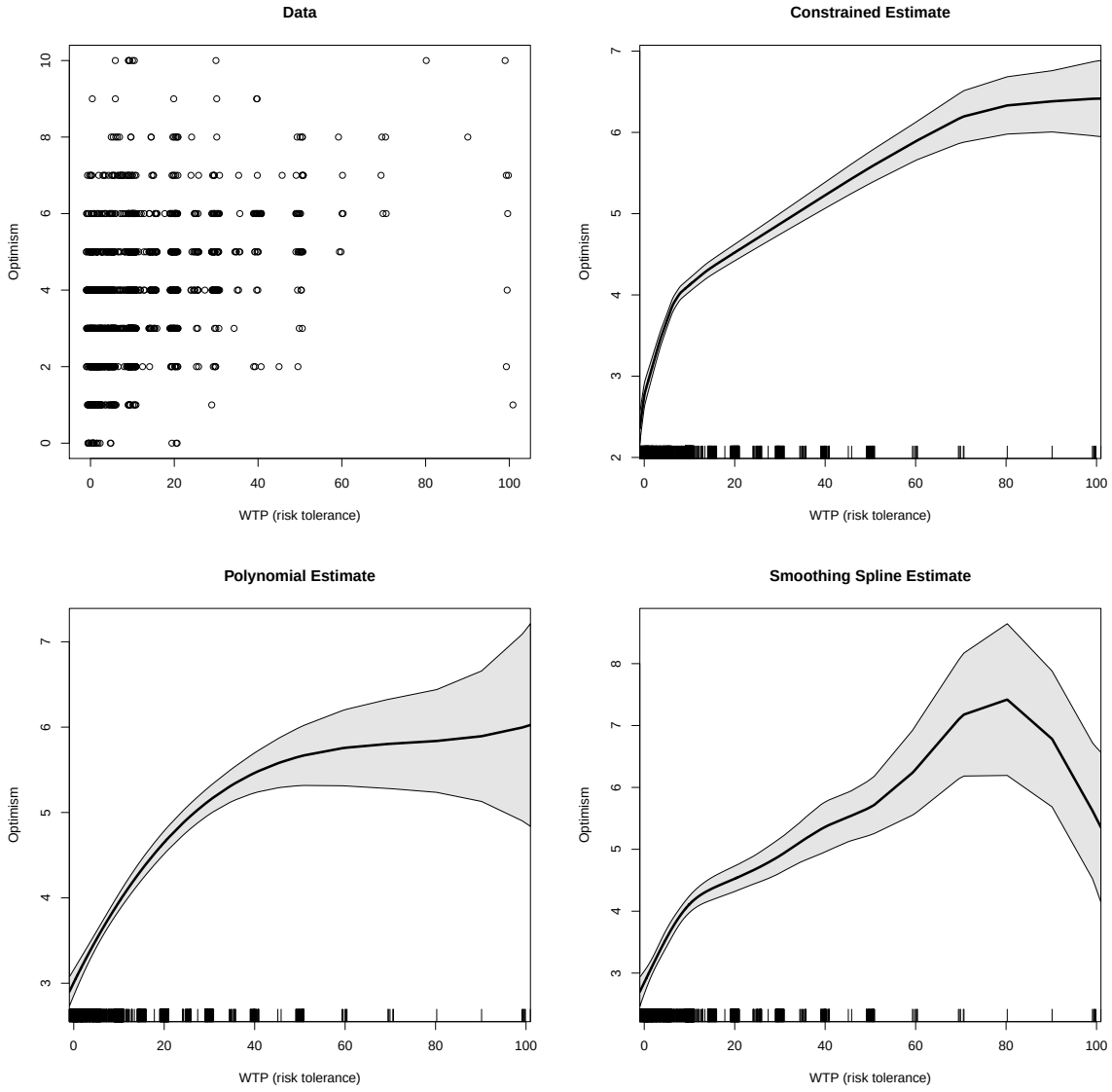

Figure 2: Risk tolerance vs optimism: data and estimates (shaded areas represent 95% variation bounds)

nation of the data indicates there are several possible outliers with low degree of optimism but high WTP, suggesting inconsistency in their preferences. These outliers appear to exert a disproportionately large influence on the smoothing spline estimate. The polynomial estimate is closer to the constrained estimates, but is not concave. The slight acceleration in optimism near the top end of the WTP range doesn't seem to be supported by the data. This spurious pattern can probably be attributed to oscillations typically associated with 'global'

projection estimators especially power series: fitting in one region of the curves might affect that in a different region. In contrast, local smoothers, such as kernel or spline estimators, do not suffer this kind of global oscillations.

The second example concerns the estimation of a production function. According to the standard economic theories, production functions are monotone increasing and concave with respect to inputs (cf. Diewert and Wales (1987)). We use a benchmark data in Coelli (1996). The data set contains information on the level of output and capital and labor inputs of 60 firms. The bottom panel of Table 2 reports summary statistics of the data set.

We assume that the production function in question takes the following additive form:

$$Q_i = f_1(C_i) + f_2(L_i) + e_i, i = 1, \ldots, n,$$

where Q, C and L denote output, capital and labor respectively, e_i are iid errors with mean zero and finite variance, and $n = 60$. We also assume that $f_j' > 0$ and $f_j'' < 0$ for $j = 1, 2$. As in previous example, we estimate the model using the constrained estimator, the polynomial estimator and the spline estimator. The estimation results are reported in Figure 3. The right panel plots the estimated surface and the left panel their corresponding contours. The general shape captured by the three estimators are similar. However, the monotonicity condition is clearly violated in the polynomial and smoothing spline estimates.

9 Concluding Remarks

We have proposed a semiparametric estimator that accommodates shape restrictions such as monotonicity and concavity. Unlike penalty-based estimators, our method employs an integral transformation to achieve desired shape constraints. The resulting estimates satisfy the constraints globally. We use penalized splines to achieve flexibility while maintaining shape constraints. We have proposed an iterative algorithm and an approximate cross validation criterion for smoothing parameter selection. We have derived an asymptotic variation bound of the proposed estimator. We have further extended the proposed method to multiple regressions within the framework of additive models. Our Monte Carlo simulations and two empirical examples illustrate the good finite sample performance and usefulness of the

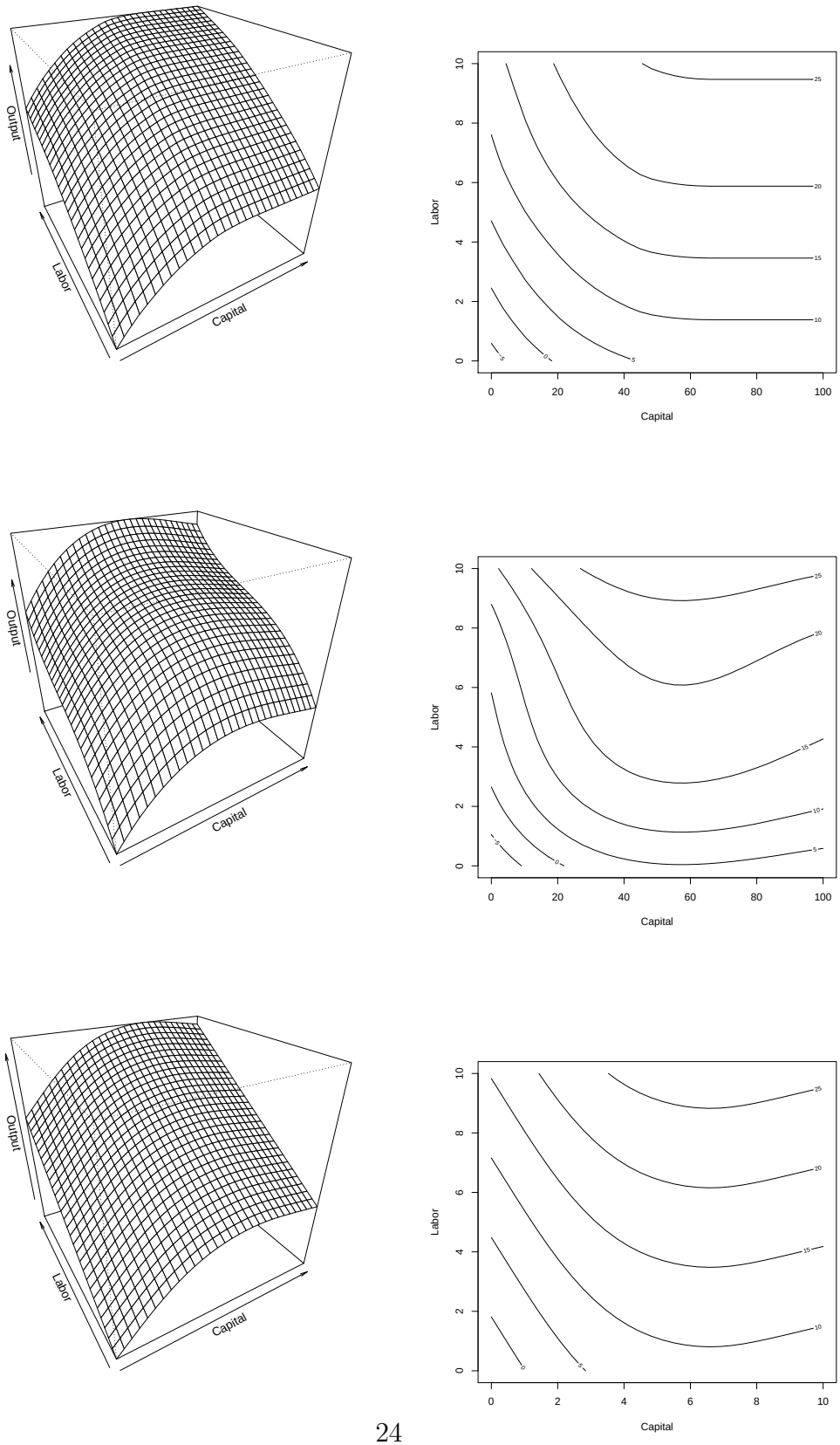

Figure 3: Estimated production function (Top: Sonstrained estimate; Middel: Polynomial estimate; Bottom: Smoothing spline estimate)

proposed method.

We conclude this work by suggesting some possible generalizations of the proposed method. First, our current model only considers continuous outcomes. Generalization to discrete or range-limited variables in the framework of the generalized linear models appears to be a natural extension of the current work. Second, we envision that our methods can be generalized to accommodate inter-temporally or spatially correlated errors, or composite errors as in the case of panel data analysis. Third, we restrict ourselves to additive models in this study. Relaxations of this restriction to accommodate interactions or more general non-separable structures while maintaining shape constraints may be of interest for future research. Lastly, we acknowledge that it is desirable to be able to test the validity of constraints implied by economic theories. Heckmam and Ramsay (2000) presented the L-spline estimators, whose model-based penalties are defined via linear differential functions. Their method provides a natural framework to test the validity of constraints implied by differential equations, as is in our estimator.

Regarding the estimation of production functions, we plan to generalize our estimators to accommodate heteroskedastic errors as in Just and Pope (1978) or stochastic production frontier analysis as in Aigner et al. (1977). Either scenario violates the iid assumption of the error terms; the former entails an extra generalized least squares (GLS) step, while the later can be tackled either in the GLS or maximum likelihood framework. Our preliminary results on these extensions have been encouraging.

References

[1] Ai C, Chen X (2007) Estimating of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics* 141: 5–43.

[2] Aigner, Lovell and Schmidt (1977) Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6: 21-37.

[3] Aerts M, Claeskens G, Wand M (2002) Some theory for penalized spline additive models. *Journal of Statistical Planning and Inferences* 103: 455–470.

[4] Ackerberg D, Chen X, Hahn J (2011) A practical asymptotic variance estimator for two-step semiparametric estimators. Working paper.

[5] Braun WJ, Hall P (2001) Data sharpening for nonparametric inference subject to constraints. *Journal of Computational and Graphical Statistics* 10: 786-806.

[6] Brunk HD (1955) Maximum likelihood estimates of monotone parameters. *Annals of Mathematical Statistics* 26: 607-616.

[7] Chambers R (1988) Applied Production Analysis: A Dual Approach. Cambridge University Press.

[8] Chernozhukov V, Fernandez-Val I, Galichon A (2007) Improving estimates of monotone functions by rearrangement. Mimeo.

[9] Coelli, TJ (1996) A Guide to FRONTIER Version 4.1: A Computer Program for Stochastic Frontier Production and Cost Function Estimation. CEPA Working Paper 96/7, Department of Econometrics, University of New England, Armidale NSW Australia.

[10] de Boor C (2001) A practical guide to splines. Springer.

[11] Diewert WE, Wales T J (1987) Flexible functional forms and global curvature conditions. *Econometrica* 55(1): 43-68.

[12] Eilers P, Marx B (1996) Flexible smoothing with B-splines and penalties. *Statistical Science* 11(2): 89–121.

[13] Eubank LE (1999) Nonparametric regression and simple smoothing. Marcel Dekker.

[14] Hall P, Huang H (2001) Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics* 29(3): 624-647.

[15] Hall, P, Opsomer JD (2005) Theory for penalized spline regression. *Biometrika*, 92: 105-118.

[16] Heckman NE, Ramsay JO (2000) Penalized regression with model-based penalties. *The Canadian Journal of Statistics* 28(2): 241–258.

[17] Härdle W, Sylvie H, Mammen E, Sperlich S (2004) Bootstrap inference in semiparametric generalized additive models. *Econometric Theory* 20(2): 265-300.

[18] Hastie, TJ, Tibshirani, R J (1990). Generalized additive models. London: Chapman & Hall.

[19] Just R, Pope RD (1978) Stochastic specification of production function and economic implications. *Journal of Econometrics* 7: 67-86.

[20] Kauermann G, Krivobokova T, Fahrmeir L (2009) Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statitical Society, Series B* 71, 487-503.

[21] Kelly C, Rice J (1990) Monotone smoothing with application to dose response curves and the assessment of synergism. *Biometrics* 46: 1071–1085.

[22] Li Q, Racine JS (2007) Nonparametric econometrics: Theory and practice. Princeton University Press.

[23] Li Y, Ruppert D (2008) On the asymptotics of penalized splines. *Biometrika* 95: 415–436.

[24] Ma S, Racine JS (2012) Additive regression splines with irrelevant categorical and continuous regressors. *Statistica Sinica*, forthcoming.

[25] Mammen E (1991) Estimating a smooth monotone regression function. *Annals of Statistics* 19(2): 724-740.

[26] Mammen E, Thomas-Agnam C (1999) Smoothing splines and shape restrictions. *Scandinavian Journal of Statistics* 26: 239-252.

[27] Matzkin, RL (1991) Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica* 59(5): 1315–1327.

[28] Matzkin, RL (1994) Restrictions of economic theory in nonparametric methods. in R. F. Engel and D. L. McFadden (eds.) Handbook of Econometrics, Vol. 4.

[29] Mukerjee H (1988) Monotone nonparametric regression. *Annals of Statistics* 16: 741-750.

[30] Newey WK (1984) A method of moment interpretation of sequential estimators. *Economic Letters* 14: 201–206.

[31] Newey WK (1994) The asymptotic variance of semiparametric estimators. *Econometrica* 62: 1349–1382.

[32] Racine JS, Parmeter, C (2010) Constrained Nonparametric Kernel Rgression: Estimation and Inference. Mimeo.

[33] Ramsay JO (1988) Monotone regression splines in action (with comments). *Statistical Science* 3: 425-461.

[34] Ramsay JO (1998) Estimating smooth monotone functions. *Journal of the Royal Satistical Society Series B* 60 (2): 365–375.

[35] Ruppert D (2002) Selecting the number of knots for penalized splines. *Journal of Computational Statistics* 11: 735–757.

[36] Ruppert D, Wand W, Carroll R (2003) Semiparametric regression. Cambirdge University Press.

[37] Wahba G (1990) Spline models for observational data. SIAM.

[38] Wand M (1999) On the optimal amount of smoothing in penalized spline regression. *Biometrika* 86: 936–940.

[39] Wand M (2006) Smoothing and Mixed Models. *Computational Statistics* 18: 223–249.

[40] Yachew, A (2003) Semiparametric regression for the applied econometrician. Cambridge University Press.

[41] Yu Y, Ruppert D (2002) Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* 97:469, 1042–1054.

Appendix

Proof of Theorem 1. We can rewrite the objective function as

$$\begin{aligned} Q_{\lambda_n}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\{Y_i - f(X_i;\theta^0) + f(X_i;\theta) - f(X_i;\theta^0)\}^2 + \lambda_n D(f(\theta)) \\ &= \frac{1}{n}\sum_{i=1}^{n} e_i^2 + \frac{2}{n}\sum_{i=1}^{n}\{f(X_i;\theta^0) - f(X_i;\theta)\}^2 e_i + \frac{1}{n}\sum_{i=1}^{n}\{f(X_i;\theta^0) - f(X_i;\theta)\}^2 + \lambda_n D(f(\theta)). \end{aligned}$$

Under assumptions 1, 2, and 3a, the first and third terms converge to σ^2 and $\mathcal{Q}(\theta)$ respectively, and the second term converges to zero. In addition, the last term vanishes if $\lambda_n \to 0$. It follows that

$$Q_\lambda(\theta_n) \xrightarrow{p} \mathcal{Q}(\theta) + \sigma^2$$

if $\lambda_n = o(1)$.

Next let $\hat{\theta}(\lambda_n)$ be the penalized least square estimators. It follows that

$$Q_{\lambda_n}(\hat{\theta}(\lambda_n)) \leq Q_{\lambda_n}(\theta^0).$$

Under assumption 3.a, the left hand side converges to, say, $\mathcal{Q}(\theta') + \sigma^2, \theta' \in \Theta$. It follows that

$$\mathcal{Q}(\theta') + \sigma^2 \leq \mathcal{Q}(\theta^0) + \sigma^2 = \sigma^2,$$

implying $Q(\theta') = 0$. Thus under assumption 3a and 3b, we have $\theta' = \theta^0$, which establishes the consistency of the penalized least square estimator. □

Proof of Theorem 2. Rewrite

$$\hat{Y}(\lambda) - Y(\lambda) = \{\hat{W}(\lambda) - W(\lambda)\}\hat{\beta}(\lambda) + W(\lambda)(\hat{\beta}(\lambda) - \beta(\lambda)).$$

It follows that

$$\begin{aligned}\mathrm{Var}(\hat{Y}(\lambda)) =& \mathrm{Var}((\hat{W}(\lambda) - W(\lambda))\hat{\beta}(\lambda)) + \mathrm{Var}(W(\lambda)(\hat{\beta}(\lambda) - \beta(\lambda))) \\ &+ 2\mathrm{cov}((\hat{W}(\lambda) - W(\lambda))\hat{\beta}(\lambda), W(\lambda)(\hat{\beta}(\lambda) - \beta(\lambda))). \end{aligned} \tag{A.1}$$

First note that the third term vanishes asymptotically. Since $\beta(\lambda) = (W(\lambda)^T W(\lambda))^{-1} W(\lambda) Y(\lambda)$, it follows readily that

$$\mathrm{Var}(W(\lambda)(\hat{\beta}(\lambda) - \beta(\lambda))) = \sigma^2 P_W(\lambda). \tag{A.2}$$

From (11), we have under assumption 5 that

$$\mathrm{Var}(\sqrt{n}(\hat{c}(\lambda) - c(\lambda)) = \Omega(\lambda),$$

with

$$\Omega(\lambda) = \sigma^2(\beta_1(\lambda) Z(\lambda)(\beta_1^2(\lambda) Z(\lambda)^T Z(\lambda) + n\lambda D'')^{-2}(\beta_1(\lambda) Z^T(\lambda)).$$

Next note that

$$\begin{aligned}(\hat{W}(\lambda) - W(\lambda))\hat{\beta}(\lambda) &= (\hat{W}(\lambda) - W(\lambda))\beta(\lambda) + o_p(1) \\ &= \beta_1(\lambda)(m(X; \hat{c}(\lambda)) - m(X; c(\lambda))) + o_p(1) \\ &= \beta_1(\lambda) Z(\lambda)(\hat{c}(\lambda) - c(\lambda)) + o_p(1).\end{aligned}$$

It follows that

$$\mathrm{Var}((\hat{W}(\lambda) - W(\lambda))\hat{\beta}(\lambda)) = (\beta_1 Z^T(\lambda))\Omega(\lambda)(\beta_1(\lambda) Z^T(\lambda)) = \sigma^2 P_Z^2(\lambda). \tag{A.3}$$

Combining (A.2) and (A.3) then yields (14). Under assumptions 1, 2, 3(a), 4 and 5, the asymptotic normality can be readily established under the central limite theorem.

Lastly the variance of the error terms is estimated by $(\sum_{i=1}^{n} \hat{e}_i^2)/(d.o.f.)$, where the degrees of freedom is given by n subtracted the effective number of parameters. The proposed semiparametric estimator has two parametric parameters, and the effective number of parameters (rank of the smoother) for the nonparametric part is calculated as $\mathrm{tr}(\hat{P}_Z)(\lambda)$ (Cf. Ruppert et al. (2003)). It follows that $s^2 \xrightarrow{p} \sigma^2$ as $n \to \infty$. In addition, it is straightforward

to show that $\hat{\beta}(\lambda), \hat{P}_W(\lambda)$ and $\hat{P}_Z(\lambda)$ converge in probability to $\beta(\lambda), P_W(\lambda)$ and $P_Z(\lambda)$ as $n \to \infty$. It follows that under assumption 5, $\hat{V}(\lambda) \xrightarrow{p} V(\lambda)$ as $n \to \infty$, which completes the proof of this theorem. □

Proof of Theorem 3. From (11) we have

$$\begin{aligned}
\hat{c}(\lambda_n) &= (\frac{1}{n}\hat{\beta}_1^2 \hat{Z}^T \hat{Z} + \lambda_n \hat{D}'')^{-1} \quad \frac{1}{n}\hat{\beta}_1 \hat{Z}^T \tilde{Y} + \lambda_n (\hat{D}' - \hat{D}'' \hat{c}_-) \\
&= (\frac{1}{n}\hat{\beta}_1^2 \hat{Z}^T \hat{Z} + \lambda_n \hat{D}'')^{-1} \quad \frac{1}{n}\hat{\beta}_1 \hat{Z}^T \tilde{Y} + o_p(1) \\
&\equiv (\frac{1}{n} B^T B + \lambda_n \hat{D}'')^{-1} (\frac{1}{n} B^T \tilde{Y} + o_p(1)).
\end{aligned}$$

A Taylor expansion of the above with respect to λ_n around zero, using that $(I + \lambda A)^{-1} = I - \lambda A + o(\lambda A)$ as $\lambda \to 0$ yields

$$\begin{aligned}
\hat{c}(\lambda_n) &= ((\frac{1}{n} B^T B)^{-1} \frac{1}{n} B^T B + \lambda_n \hat{D}'')^{-1} (\frac{1}{n} B^T B)^{-1} (\frac{1}{n} B^T \tilde{Y} + o_p(1)) \\
&= (I - \lambda_n \hat{D}'' + o(\lambda_n \hat{D}''))(\frac{1}{n} B^T B)^{-1} (\frac{1}{n} B^T \tilde{Y} + o_p(1)) \\
&= (B^T B)^{-1} B^T \tilde{Y} + o(\lambda_n \hat{D}'') + o_p(1) \\
&= c^0 + o(\lambda_n \hat{D}'') + o_p(1),
\end{aligned}$$

where the last equality is due to the consistency of $\hat{c}(\lambda_n)$ as $\lambda_n \to 0$ given in Theorem 1.

Next we can show that the variance of $\hat{c}(\lambda_n)$ is of order σ^2/n. It follows that $\text{MSE}(\hat{c}(\lambda_n)) = O_p(\sigma^2/n + \lambda_n^2)$ for bounded $\hat{D}''$ (which is implied by the compactness of Θ). Thus for the asymptotic bias to vanish, we need $\lambda_n = o(n^{-1/2})$. The asymptotic normality of the limiting case can then be established using essentially the same proof as for Theorem 2 and replacing the fixed λ with zero, the limiting value of λ_n. □

Proof of Theorem 4. Let $\hat{c}(i, w)$ be the solution to the following optimization

$$(w - \beta_0 - \beta_1 f(X_i))^2 + \sum_{k=1, k=i}^{n} (Y_k - \beta_0 - \beta_1 f(X_k))^2 + \lambda D(f(x)). \tag{A.4}$$

It follows that $\hat{c}(i, \hat{Y}_{(i)}) = \hat{c}_{(i)}$.

Let $\Delta_{(i)}$ be an $n \times 1$ vector of zeros except that the ith element equals $\hat{Y}_{(i)} - Y_i$. We can then write

$$\hat{c}_{(i)} = (\hat{\beta}_1^2 \hat{Z}^T \hat{Z} + \lambda \int_{\mathcal{X}} D''(x)dx)^{-1} \hat{\beta}_1 \hat{Z}^T (\tilde{Y} + \Delta_{(i)}).$$

It follows that

$$\begin{aligned}
\tilde{Y}_{(i)} =& \hat{\beta}_1 \hat{Z}_i^T \hat{c}_{(i)} \\
=& \hat{\beta}_1 \hat{Z}_i^T (\hat{\beta}_1^2 \hat{Z}^T \hat{Z} + \lambda \int_{\mathcal{X}} D''(x)dx)^{-1} \hat{\beta}_1 \hat{Z}^T \tilde{Y} \\
&+ \hat{\beta}_1 \hat{Z}_i^T (\hat{\beta}_1^2 \hat{Z}^T \hat{Z} + \lambda \int_{\mathcal{X}} D''(x)dx)^{-1} \hat{\beta}_1 \hat{Z}^T \Delta_{(i)} \\
=& \tilde{Y}_i + s_i (\hat{Y}_{(i)} - Y_i). \qquad \text{(A.5)}
\end{aligned}$$

Next we use the Taylor approximation on $\hat{Y}_{(i)}$ to obtain

$$\begin{aligned}
\tilde{Y}_{(i)} =& \hat{Y}_{(i)} - \hat{\beta}_0 - \hat{\beta}_1 f(X_i; \hat{c}) - \hat{\beta}_1 \hat{Z}_i^T (\hat{c}_{(i)} - \hat{c}) + \hat{\beta}_1 \hat{Z}_i^T \hat{c}_{(i)} + o_p(1) \\
=& \hat{Y}_{(i)} - \hat{\beta}_0 - \hat{\beta}_1 f(X_i; \hat{c}) + \hat{\beta}_1 \hat{Z}_i^T \hat{c} + o_p(1).
\end{aligned}$$

It follows that

$$\tilde{Y}_{(i)} - \tilde{Y}_i = \hat{Y}_{(i)} - \hat{Y}_i + o_p(1). \qquad \text{(A.6)}$$

Plugging (A.6) into (A.5) and rearranging terms yields

$$Y_i - \hat{Y}_{(i)} = \frac{Y_i - \hat{Y}_i}{1 - s_i} + o_p(1),$$

which gives (A.4) readily. □