DISCUSSION PAPER SERIES

IZA DP No. 4103

Imposing Economic Constraints in Nonparametric Regression: Survey, Implementation and Extension

Daniel J. Henderson
Christopher F. Parmeter

March 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Imposing Economic Constraints in Nonparametric Regression: Survey, Implementation and Extension

Daniel J. Henderson
State University of New York at Binghamton
and IZA

Christopher F. Parmeter
Virginia Tech

Discussion Paper No. 4103
March 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 4103
March 2009

ABSTRACT

Imposing Economic Constraints in Nonparametric Regression: Survey, Implementation and Extension*

Economic conditions such as convexity, homogeneity, homotheticity, and monotonicity are all important assumptions or consequences of assumptions of economic functionals to be estimated. Recent research has seen a renewed interest in imposing constraints in nonparametric regression. We survey the available methods in the literature, discuss the challenges that present themselves when empirically implementing these methods and extend an existing method to handle general nonlinear constraints. A heuristic discussion on the empirical implementation for methods that use sequential quadratic programming is provided for the reader and simulated and empirical evidence on the distinction between constrained and unconstrained nonparametric regression surfaces is covered.

JEL Classification: J20, J30, C14

Keywords: constraint weighted bootstrapping, Hessian, concavity, identification, earnings function

Corresponding author:

Daniel J. Henderson
Department of Economics
State University of New York
Binghamton, NY 13902
USA
E-mail: djhender@binghamton.edu

* The research on this project has benefitted from the comments of participants in seminars at Cornell University, the University of California, Merced, the University of California, Riverside, the University of Nevada, Las Vegas, and the State University of New York at Albany as well as participants at the 5th annual Advances in Econometrics Conference held at Louisiana State University and the 3rd Annual New York Camp Econometrics. All GAUSS 8.0 code used in this paper is available from the authors upon request.

1. Introduction

Nonparametric estimation methods are a desirable tool for applied researchers since economic theory rarely yields insights into a model's appropriate functional form. However, when paired with the specific smoothness constraints imposed by an economic theory, such as monotonicity of a cost function in all input prices, this often increases the complexity of the estimator in practice. Access to a constrained nonparametric estimator that can handle general, multiple smoothness conditions is desirable.[1] Fortunately, a rich literature on constrained estimation has taken shape and a multitude of potential suitors have been proposed for various constrained problems. Given the potential need for constrained nonparametric estimators in applied economic research and the availability of a wide range of potential estimators, coupled with the dearth of detailed, simultaneous descriptions of these methods, a survey on the current state of the art is warranted.

In empirical studies on games, such as auctions, monotonicity of players strategies is a key assumption used to derive the equilibrium solution. This monotonicity assumption thus carries over to the estimated equilibrium strategy. And while parametric models of auctions have monotonicity 'built-in', their nonparametric counterparts impose no such condition. Thus, using a nonparametric estimator of auctions which allows monotonicity to be imposed is expected to be more competitive against parametric alternatives than an unconstrained estimator. Recently, Henderson, List, Millimet, Parmeter & Price (2008) have shown that random samples from equilibrium bid distributions can produce nonmonotonic nonparametric estimates for small samples. This suggests that being able to construct an estimator that is monotonic from the onset is important for analyzing auction data.

Analogously, convexity is theoretically required for either a production or cost function and the ability to impose this constraint in a nonparametric setting is thus desirable given that very few models of production yield reduced form parametric solutions. Cost functions are concave in input prices and outputs, nondecreasing and homogeneous of degree one in input prices. Thus, estimating a cost function requires the imposition of three distinct economic conditions. To our knowledge applied studies that nonparametrically estimate cost functions (Wheelock & Wilson 2001) do not impose these conditions directly. Thus, at the very least there is a loss of efficiency since these constraints are not imposed on the estimator. Moreover, since the constraints are not imposed, it is impossible to test whether these conditions are valid or not.

Before highlighting the potential methods available we aim to gauge the necessity of imposing smoothness constraints via a primitive example. Consider the univariate data generating process:

$$y_i = \ln(x_i) + \varepsilon_i, \ i = 1, 2, \ldots, n,$$

[1] An additional benefit of imposing constraints in a nonparametric framework is that it may provide nonparametric identification, see Matzkin (1994). Also, Mammen, Marron, Turlach & Wand (2001) show that when one imposes smoothness constraints on derivatives higher than first order the rate of convergence is faster than had the constraints not been imposed.

which is monotonic and concave. If we generate random samples under a variety of sample sizes and distributional assumptions for the pair (x_i, ε_i), we can gain insight into the *need* for a constrained estimator. Tables 1 and 2 provide the proportion of times, out of $9,999$ simulations, a local-constant kernel estimator, in unconstrained form, provides an estimate that is either monotonic or concave uniformly over a grid of points on the interior of the range of x. We use three different bandwidths for our simulations. Generically, we use bandwidths of the form $h = c\sigma_x n^{-1/5}$ where c is a user defined constant, σ_x is the standard deviation of the regressand, and n is the sample size being used. A traditional rule-of-thumb bandwidth is obtained by setting $c = 1.06$. We also use $c = 0.53$ (lesser-smoothed) and $c = 2.12$ (greater-smoothed) to gauge the impact the bandwidth has on the ability of the unconstrained estimator to satisfy the constraints without further manipulation.

We see that as the sample size is increased from 100 to 200 to 500 the proportion of trials where monotonicity is uniformly found over the grid of points approaches unity. However, concavity is violated much more often. There are many instances, especially when the bandwidth is relatively small, where there are no cases where concavity is found uniformly over the grid of points. This result may be unexpected to some given that we have nearly ten-thousand replications. Further, we see that as we increase the error variance, this leads to large decreases in the number of cases of both monotonicity and concavity.

Even with these alarming results we note that larger scale factors (c) increase the incidence of concavity. Somewhat surprising is that we do not always see that increasing the sample size leads to higher incidences of concavity. While increasing n increases the number of cases of concavity when we have large bandwidths, we often find the opposite result when $c = 1.06$. This conflicting result likely occurs because of two competing forces. First, the increase in the number of observations leads to more points in the neighborhood of x. This should lead to more cases of concavity. The second effect counteracts the first because increasing the number of observations decreases the bandwidth as $h \propto n^{-1/5}$. Finally, we note that the design of the experiment also has a noticeable effect on the likelihood of observing monotonicity or concavity without resorting to a constrained estimator. For instance, generating the regressor from the Gaussian distribution as opposed to the Uniform brings about much larger proportions of concave estimates when the bandwidth is relatively large (likely due to more data in the interior of x).[2]

The results from these tables suggest that constrained estimators are necessary tools for nonparametric analysis as in even very simple settings direct observation of an unrestricted estimator that satisfies the constraints is by no means the norm. One can imagine that with multiple covariates, multiple bandwidths and a variety of constraints to be imposed simultaneously that the likelihood the constraints are satisfied *de facto* are low.

In general, a wide variety of constrained nonparametric estimation strategies have been proposed to incorporate economic theory within the estimation procedure. While many of these estimators

[2]We also looked at the proportion of times a single point on the interior of the grid produced a monotonic or concave result. For example, when setting this value of x equal to the expected mean of each series, the incidence of both monotonicity and concavity increased. This percentage increase proved to be much larger for concavity. These results are available from the authors upon request.

TABLE 1. Likelihood of an Estimated Monotonic Regression (9,999 trials)

	$\varepsilon \sim N(0, 0.1)$			$\varepsilon \sim N(0, 0.2)$		
	100	200	500	100	200	500
$x \sim U[0.5, 1.5]$						
$c = 0.53$	0.996	0.999	1.000	0.731	0.825	0.933
$c = 1.06$	1.000	1.000	1.000	0.999	1.000	1.000
$c = 2.12$	1.000	1.000	1.000	1.000	1.000	1.000
$x \sim N(1, 0.25)$						
$c = 0.53$	0.978	0.993	0.999	0.584	0.699	0.841
$c = 1.06$	1.000	1.000	1.000	0.997	1.000	1.000
$c = 2.12$	1.000	1.000	1.000	1.000	1.000	1.000

TABLE 2. Likelihood of an Estimated Concave Regression (9,999 trials)

	$\varepsilon \sim N(0, 0.1)$			$\varepsilon \sim N(0, 0.2)$		
	100	200	500	100	200	500
$x \sim U[0.5, 1.5]$						
$c = 0.53$	0.000	0.000	0.000	0.000	0.000	0.000
$c = 1.06$	0.021	0.033	0.040	0.016	0.016	0.014
$c = 2.12$	0.016	0.004	0.008	0.022	0.007	0.003
$x \sim N(1, 0.25)$						
$c = 0.53$	0.000	0.000	0.000	0.000	0.000	0.000
$c = 1.06$	0.027	0.019	0.014	0.019	0.010	0.003
$c = 2.12$	0.445	0.527	0.683	0.397	0.427	0.498

are designed myopically for a specific smoothness constraint, a small but burgeoning literature has focused on estimators which can handle many arbitrary economic constraints simultaneously. Of note are the recent contributions of Racine, Parmeter & Du (2009) who develop a constrained kernel regression estimator and Beresteanu (2004) who developed a similar type of estimator but for use with spline based estimators.[3] In addition to providing a survey of the current menu of available constrained nonparametric estimators, we also shed light on the quantitative aspects for empirical implementation regarding the constrained kernel estimator of Racine et al. (2009). While they mention the ability of their method to handle general constraints, their existence results and simulated and real examples all focus on linear (defined in the appropriate sense) restrictions. We augment their discussion by providing existence results as well as heuristic arguments on the implementation of the method. Simulated and empirical evidence targeting imposing concavity on a regression surface is provided to showcase the full generality of the method.

The rest of this paper proceeds as follows. Section 2 reviews the literature on constrained nonparametric regression. Section 3 discusses imposing general, nonlinear constraints, specifically

[3]We should also recognize Yatchew & Bos (1997) who also developed a general framework for constrained nonparametric estimation in a series based setting. See also the recent application of their method in Yatchew & Härdle (2006).

concavity, using constraint weighted bootstrapping and shows how it can be implemented computationally. Section 4 presents a small scale simulation and an empirical discussion of estimation of an age-earnings profile. Section 5 presents several concluding remarks and directions for future research.

2. Available Constrained Estimators

Consider the standard nonparametric regression model

$$y_i = m(x_i) + \varepsilon_i, \quad \text{for } i = 1, \ldots, n, \tag{1}$$

where y_i is the dependent variable, $m(\cdot)$ is the conditional mean function with argument x_i, x_i is a $k \times 1$ vector of covariates and ε_i is a random variable with zero mean and unit variance. Our goal is to estimate the unknown conditional mean subject to economic constraints (e.g., concavity) in a smooth framework.

Imposing arbitrary constraints on nonparametric regression surfaces, while not new to econometrics, has not received as much attention as other aspects of nonparametric estimation, for instance bandwidth selection, at least not in the kernel regression framework. Indeed, one can divide the literature on imposing constraints in nonparametric estimation frameworks into two broad classes:

(1) Developing a nonparametric estimator to satisfy a particular constraint. Here the class of monotonically restricted estimators is a prime example.
(2) Developing a nonparametric estimator (either smooth or interpolated) that satisfies a class of constraints.

Our goal is to highlight the variety of existing methods and document the differences across the available techniques to guide the reader to an appropriate estimator for the problem at hand.

2.1. Isotonic Regression.

The first constrained nonparametric estimators were nonsmooth and fell under the heading of 'isotonic regression', initially proposed by Brunk (1955). Brunk's (1955) estimator was a minmax estimator that was designed to impose monotonicity on a regression function with a single covariate, while Hansen, Pledger & Wright (1973) extended the estimator to two dimensions and provided results on consistency of the estimator. To explain the estimator of Brunk, let $\mathcal{C}_B$ be the discrete cone of restrictions in R^n:

$$\{(z_1, z_2, \ldots, z_n) : z_1 \leq z_2 \leq \cdots \leq z_n\}.$$

We let y_i^* be a solution to the minimization problem

$$\min_{(y_1^*, \ldots, y_n^*) \in \mathcal{C}_B} \sum_{i=1}^{n} (y_i - y_i^*)^2.$$

This minimization problem has a unique solution that is expressed succinctly by a min-max formula.

Use $X_{(1)}, \ldots, X_{(n)}$ to denote the order statistics of X and $y_{[i]}$ the corresponding observation of $X_{(i)}$. Then our 'isotonized' fitted values can be represented as

$$y_i^* = \min_{s \geq i} \max_{t \leq i} \sum_{j=s}^{t} y_{[j]}/(t-s+1), \tag{2}$$

or

$$y_i^* = \max_{s \leq i} \min_{t \geq i} \sum_{j=s}^{t} y_{[j]}/(t-s+1). \tag{3}$$

In Brunk's (1955) approach there is no attempt to smooth the estimation results to values of x between the observation points. A simple approach would be to extend flatly between the values of x_i but this has been criticized for the presence of too many flat spots and a slow rate of convergence.[4]

Interestingly, Hildreth (1954) introduced a related method to that in Brunk (1955), but geared towards estimating a regression function that is restricted to be concave. His procedure amounts to conducting least-squares subject to discretized concavity restrictions. Similar to Brunk (1955), let $\mathcal{C}_H$ be the discrete cone of restrictions in R^n:

$$\left\{(z_1, z_2, \ldots, z_n) : \frac{z_{i+1} - z_i}{x_{i+1} - x_i} \geq \frac{z_{i+2} - z_{i+1}}{x_{i+2} - x_{i+1}}, i = 1, \ldots, n-2\right\}$$

then y_i^* is a solution of.

$$\min_{(y_1^*, \ldots, y_n^*) \in \mathcal{C}_H} \sum_{i=1}^{n} (y_i - y_i^*)^2. \tag{4}$$

An iterative procedure is required to solve the minimization as no closed form solution exists. However, unlike the monotonically constrained estimator of Brunk (1955), the concave restricted estimator of Hildreth (1954) extends between observation points linearly, thus falling into the classification of a least-squares spline estimator.

While both of these estimators construct restricted regression estimates predicated on simple concepts, they are not 'smooth' in the traditional sense. The classic isotonic regression estimator of Brunk (1955) was smoothed by Mukerjee (1988) and Mammen (1991*a*). An alternative way to characterize their estimators is to say that they forced the traditional Nadaraya-Watson regression smoother to satisfy a monotonicity constraint. The key insight was to use a two-step estimator that consisted of a smoothing step and an isotonizing step. Mukerjee (1988) proved that one could preserve the isotonization constructed in the first step by using a log-concave kernel to smooth in the second step. Thus, after one uses either (2) or (3) to isotonize the regressand, a smooth,

[4]Slower than conventional nonparametric rates.

nonparametric estimate of the unknown conditional mean is constructed as

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K((x - X_{(i)})/h_n) y_i^*}{\sum_{i=1}^{n} K((x - X_{(i)})/h_n)}, \tag{5}$$

where h_n is the bandwidth.[5] One does not need to use a special kernel, however, as a second order Gaussian kernel is log concave, thus making this method easy to implement. Mammen (1991*a*) proved that asymptotically the order of the steps is irrelevant. No equivalent estimator exists for the concave variant introduced by Hildreth (1954) and as such the generalizability of smoothing isotonic type estimators is unknown. Moreover, multivariate extensions to the traditional isotonic regression estimator are difficult to implement and often not available in closed form solutions.

2.2. **Constrained Spline/Series Estimation.** Both spline and series based functions provide the researcher with a flexible set of basis functions with which to construct a regression model that is linear in parameters, which is intuitively appealing. Early methods using splines or series based methods, designed to impose general economic constraints, include Gallant (1981, 1982) and Gallant & Golub (1984) who introduced the Fourier Flexible Form estimator (FFF), whose coefficients could be restricted to impose concavity, homotheticity and heterogeneity in a nonparametric setting.[6] Constrained spline smoothers were proposed by Dierckx (1980), Holm & Frisen (1985), Ramsay (1988), and Mammen (1991*b*), to name a few early approaches.

In what follows we describe the basic setup for constrained least-squares spline estimation.[7] We define our spline space to be $\mathcal{S}$.[8] Our least-squares spline estimate is a function m which represents a linear combination of spline functions from $\mathcal{S}$ that solves:

$$\min_{s \in \mathcal{S}} \sum_{i=1}^{n} (y_i - m(x_i))^2. \tag{6}$$

To impose constraints we note that positivity of either the first or second derivative at a given point $\tilde{x}$ of the function $m(\cdot)$ can be written equivalently as positivity of a linear combination of the associated parameters with respect to the chosen basis. Thus, monotonicity or concavity can be readily imposed on a discretized grid of points where each point adds additional *linear* constraints on the spline coordinates with the associated basis. It is a natural step to include these linear constraints directly into the least-squares spline problem.

Similar to isotonic regression, the literature appears to have focused on concavity first (Dierckx 1980) and then monotonicity (Ramsay 1988). In what will be a common theme in constrained nonparametric regression, Dierckx (1980) used a quadratic program to enforce *local* concavity or

[5]This has connections with both data sharpening (Section 2.5) and constraint weighted bootstrapping (Section 2.6).
[6]Monotonicity is not easily imposed in this setting.
[7]For a more detailed treatment of either series or spline based estimation we refer the reader to Eubank (1988) and Li & Racine (2007, chapt. 15).
[8]Unlike kernel smoothing where smoothing is dictated by a bandwidth, in series and spline based estimation, the smoothing is controlled by the dimension of the series or spline space.

convexity of a spline function. His function estimate, using normalized B-splines (see Schumaker 1981) with basis N_j, is

$$\hat{m}(x) = \sum_{j=-3}^{k} c_j^* N_j(x).$$

Here k denotes the total number of knots. The values c_j^* solve the quadratic program

$$\min_{\sum_{j=-3}^{k} d_{j,l} c_j e_j \leq 0} \sum_{i=1}^{n} \left(y_i - \sum_{j=-3}^{k} c_j N_j(x_i) \right)^2. \tag{7}$$

The e_j in equation (7) determine the type of constraint being imposed on the function locally. That is, $e_j = 1$ if the function is locally convex at knot ℓ, $e_j = 0$ if the function is unrestricted at the ℓ^{th} knot and $e_j = -1$ if the function is locally concave at knot ℓ. The numbers $d_{j,l}$ are derived from the second derivatives of the basis splines at each of the knots and have simple formula. We have

$$\begin{aligned}
d_{j,l} =& 0 && \text{if} \quad j \leq l-4 \quad \text{or} \quad j \geq 4 \\
d_{l-3,l} =& \frac{6}{(t_{l+1} - t_{l-2})(t_{l+1} - t_{l-1})} \\
d_{l-1,l} =& \frac{6}{(t_{l+2} - t_{l-1})(t_{l+1} - t_{l-1})} \\
d_{l-2,l} =& -(d_{l-3,l} + d_{l-1,l}),
\end{aligned}$$

where t_l refers to the l^{th} point under consideration. Ramsay (1988) developed a similar monotonically constrained spline estimator using I-splines. I-splines have a direct link to the B-splines used by Dierckx (1980). An I-spline of order M is an indefinite integral of a corresponding B-spline of the same order. Ramsay (1988) used I-splines because he was able to establish that they had the property that each individual I-spline is monotonic and that any linear combination of I-splines with positive coefficients is also monotonic. This made it easy to construct the associated monotonic spline estimator. Both of the aforementioned estimators can also be placed in the smoothing spline domain as well.

Yatchew & Bos (1997) develop a series based estimator that can handle general constraints. This estimator is constructed by minimizing the sum of squared errors of a nonparametric function relative to an appropriate Sobolev norm. The basis functions that make up the series estimation are determined from a set of differential equations that provide 'representors'. Representors of function evaluation consist of two functions spliced together, where each of these functions is a linear combination of trigonometric functions. In essence, one can 'represent' any function in Sobolev space through this process (see Yatchew & Bos 1997, Appendix 2). Let R be an $n \times n$ 'representor' matrix whose columns (equivalently rows) equal the representors of the function, evaluated at the

observations $x_1, \ldots, x_n$.[9] Then, arbitrary constrained estimation of a nonparametric function

$$\min_{f \in \mathcal{F}} \quad n^{-1} \sum_{i=1}^{n} (y_i - m(x_i))^2 \quad \text{s.t. } \|m\|_{Sob}^2 \leq L, \tag{8}$$

can be recast as

$$\min_{c} \quad n^{-1} \sum_{i=1}^{n} (y_i - Rc)^2 \quad \text{s.t. } c'Rc \leq L, c'R^{(1)}c \leq L^{(1)}, c'R^{(2)}c \leq L^{(2)}, \ldots, c'R^{(k)}c \leq L^{(k)}. \tag{9}$$

Here L denotes the upper bound on the squared Sobolev norm of our constrained function, c is an $n \times 1$ vector of coefficients and $\mathcal{F}$ is our constrained function space which we are searching over. Since we are interested in constraints that relate directly to the derivatives of the nonparametric function we are estimating, $R^{(1)}, \ldots, R^{(k)}$ represent the appropriate derivatives of the original representor matrix and $L^{(1)}, \ldots, L^{(k)}$ are the corresponding bounds. For example, if one wished to impose monotonicity, $L^{(1)} = 0$ and $R^{(1)}$ represents the representor matrix with each of the representors first order differentiated with respect to the corresponding column's variable (i.e., the fifth column of $R^{(1)}$ corresponds to the fifth covariate so the representors are first order differentiated with respect to that variable). Again, this is a quadratic programming problem with a quadratic constraint.[10]

Beresteanu (2004) introduced a spline based procedure that can handle multivariate data and impose multiple, general, derivative constraints. His estimator is solved via quadratic programming over an equidistant grid created on the covariate space. These points are then interpolated to create a globally constrained estimator. He employed his method to impose monotonicity and supermodularity of a cost function for the telephone industry. His estimation setup is similar to the approaches described above and involves putting together a set of appropriately defined constraint matrices for the shape constraint(s) desired and solving for a set of coefficients, then interpolating these points to construct the nonparametric function which satisfies the constraints over the appropriate interval. In essence, since Beresteanu (2004) is constructing his estimator first based on a grid of points and then interpolating, this estimation procedure can be viewed as a two-step series based equivalent of the isotonic regression discussed earlier (Mukerjee 1988).

2.3. **The Matzkin Approach.** The seminal work of Matzkin (1991, 1992, 1993, 1994, 1999) considered identification and estimation of general nonparametric problems with arbitrary economic constraints. One of her pioneering insights was that when nonparametric identification was not possible, imposing shape constraints tied to economic theory could provide nonparametric identification in certain estimation settings. Her work laid the foundations for a general operating theory of constrained nonparametric estimation. Her methods focused on standard economic constraints

[9]For more on the construction of representor matrices see Wahba (1990) or Yatchew & Bos (1997, Appendix 2).

[10]See the work of Yatchew & Härdle (2006) for an empirical application of constrained nonparametric regression using the series based method of Yatchew & Bos (1997). Yatchew & Härdle (2006) focus on nonparametric estimation of an option pricing model where the unknown function must satisfy monotonicity and convexity as well as the density of state prices being a true density (positivity and integrates to 1).

(monotonicity, concavity, homogeneity, etc.) but facilitated in more general settings than regression. Primarily, her work focused on binary-threshold crossing models and polychotomous choice models, although her definition of sub-gradients equally carried over to a regression context. One can suitably recast her estimation method in the regression context as nonparametric constrained least-squares.

For example, to impose concavity on a regression function she created 'subgradients', T^j, which were defined for any convex function $m : X \to \mathbb{R}^k$ where $X \subset \mathbb{R}$ is a convex set and $x \in X$ any vector $T \in \mathbb{R}^k$ such that $\forall y \in X \; m(y) \geq m(x) + T(y - x)$.[11] We use the notation T^j to denote that the subgradients are calculated for the observations. Matzkin (1994) showed how to use the subgradients to impose concavity and monotonicity simultaneously. Using the Hildreth (1954) constraints for concavity of a regression surface, Matzkin (1994) rewrites them as

$$m(x_i) \leq m(x_j) + T^j(x_i - x_j), \quad i, j = 1, \ldots, n.$$

She solves the minimization problem in (4) but the minimization is over $m(x_i) \; \forall i$ and $T^j \; \forall j$. To impose monotonicity one would add the additional constraint that $T^j > 0 \; \forall j$. Algorithms to solve the constrained optimization problem were first developed for the regression setup by Dykstra (1983), Goldman & Ruud (1992) and Ruud (1995) and for general functions by Matzkin (1999), who used a random search routine regardless of the function being minimized.

Implementation of these constrained methods is of the two-step variety (see Matzkin 1999). First, for the specified constraints, a feasible solution consisting of a finite number of points is determined through optimization of some criterion function (in Matzkin's choice framework set-ups this is a pseudo-likelihood function). Second, the feasible points are interpolated or smoothed to construct the nonparametric surface that satisfies the constraints. These methods can be viewed in the same spirit as that of Mukerjee (1988), but for a more general class of problems.

2.4. **Rearrangement.** Recent work on imposing monotonicity on a nonparametric regression function, known as rearrangement, is detailed in Dette, Neumeyer & Pilz (2006) and Chernozhukov, Fernandez-Val & Galichon (2007). The estimator of Dette et al. (2006) combines density and regression techniques to construct a monotonic estimator. The appeal of 'rearrangement' is that no constrained optimization is required to obtain a monotonically constrained estimator, making it computationally efficient compared to the previously described methods. Their estimator actually estimates the inverse of a monotonic function, which can then be inverted to obtain an estimate of the function of interest.

To derive this estimator let M denote a natural number that dictates the number of equi-spaced grid points to evaluate the function. Then, their estimator is defined as

$$\widehat{m}^{-1}(x) = \int_{-\infty}^{x} \frac{1}{Mh} \sum_{j=1}^{M} K\left(\frac{\hat{m}(j/M) - u}{h}\right) du, \tag{10}$$

[11] When $m(x)$ is differentiable at x the gradient of x is the *unique* subgradient of m at x.

where $\widehat{m}(x)$ is any unconstrained nonparametric regression function estimate (kernel smoothed, local polynomial, series, splines, neural network, etc.). The intuition behind this estimator is simple; the connection rests on the properties of transformed random variables.

Note that $m(x_i)$ is a transformation of the random variable x_i. The estimator

$$\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{m(x_i)-u}{h}\right),$$

represents the classical kernel density of the random variable $u = m(x_1)$ which has density

$$g(u) = f(x_1)|(m^{-1})'(x_1)|.$$

The integration in (10) is that of a probability density function and as such a CDF is constructed, which is always *monotonically increasing*. The equi-space grid is used for estimation since the evaluation points are then treated as though they came from a uniform density, making $f(j/M) = I[a, b]$, where a and b denote the lower and upper bounds of the support of X, respectively. Thus, the integration in this case amounts to integrating $|(m^{-1})'(x_1)|$ over its domain, which gives us $m^{-1}(x_1)$. Once this has been obtained, it is a simple matter to reflect this estimate across the $y = x$ line in Cartesian 2-space to obtain our monotonically restricted regression estimator. Chernozhukov et al. (2007) discuss implementation of this estimator in a multivariate setting and show that the constrained estimator *always* improves (reduces the estimation error) over an original estimate whenever the original estimate is not monotonic.

The name rearrangement comes from the fact that the point estimates are rearranged so that they are in increasing order (monotonic). This happens because the kernel density estimate of the first stage regression estimates sorts the data from low to high to construct the density, which is then integrated. This sorting, or rearranging, is how the monotonic estimate is produced. It works because monotonicity as a property is nothing more than a special ordering and the kernel density estimator is 'unaware' that the points it is smoothing over to construct a density are from an estimate of a regression function as opposed to raw data.

One issue with this estimator is that while it is intuitive, computationally simple and easy to implement with existing software, it requires the selection of two 'bandwidths'.[12] Additionally, the intuition underlying the ease of implementation does not readily extend itself to general constraints on nonparametric regression surfaces. No such transformation is obtainable to impose concavity using the same insights, for example.

2.5. **Data Sharpening.** Data sharpening derives from the work of Friedman, Tukey & Tukey (1980) and Choi & Hall (1999). These methods are designed to admit a wide range of constraints and are closely linked to biased-bootstrap methods (Hall & Presnell 1999). Data sharpening is inherently different than biased-bootstrapping and constraint weighted bootstrapping (to be discussed later)

[12]We use the word bandwidth loosely here as the first stage does not have to involve kernel regression. One could use series estimators in which case the selection would be over the number of terms. Or, if one uses splines then the number of knots would have to be selected in the first stage.

as it alters the data, but keeps the weights associated with each point fixed, whereas biased-bootstrapping and constraint weighted bootstrapping change the weights associated with each point, but keep the points fixed. Both of these methods, however, can be thought of as data tuning methods which in some sense alter the underlying empirical distribution to achieve the desired outcome. We discuss the method of Braun & Hall (2001) in what follows.

Let our original data be $\{x_1, \ldots, x_n\}$ and our sharpened data be $\{z_1, \ldots, z_n\}$. Define the distance between original and sharpened points as $D(x_i, z_i) \geq 0$. We choose $\mathcal{Z} = \{z_1, \ldots, z_n\}$, our set of sharpened data, to minimize

$$D(\mathcal{X}, \mathcal{Z}) = \sum_{i=1}^{n} D(x_i, z_i),$$

subject to our constraints of interest. Once the sharpened data have been obtained we apply our method of interest, in this setting nonparametric regression, to the sharpened data.

More formally, our kernel regression (local-constant, say) estimator is

$$\widehat{m}(x|\mathcal{X}, \mathcal{Y}) = \frac{\sum_{i=1}^{n} K\left((x_i - x)/h\right) y_i}{\sum_{i=1}^{n} K\left((x_i - x)/h\right)} = \sum_{i=1}^{n} A_i(x) y_i.$$

We want to impose an arbitrary constraint on the function, monotonicity for example, by 'sharpening' the y's. Thus, we minimize

$$D(\mathcal{Y}, \mathcal{Q}) = \sum_{i=1}^{n} D(y_i, q_i), \tag{11}$$

for a preselected distance function, subject to the constraints

$$\widehat{m}'(x|\mathcal{X}, \mathcal{Q}) = \sum_{i=1}^{n} A_i'(x) q_i > 0. \tag{12}$$

Notice the conditioning set for which the estimator is defined over has changed from $\mathcal{Y}$ to $\mathcal{Q}$. Thus, we *construct* our restricted estimator while simultaneously minimizing our criterion function. If one chose $D(r, t) = (r - t)^2$, we would have a standard quadratic programming problem provided the constraints were linear (which they are in our monotonicity example). Compared to rearrangement, given the fact that the data is smoothed, even though the response variables are moved around, the corresponding constrained curve is as smooth as the unconstrained curve. The rearranged curve will have ambiguous low order kinks where the non-monotonic portion of the curve is 'forced' to be monotonic resulting in a curve that is less smooth than its unconstrained counterpart.

2.6. **Constraint weighted bootstrapping.** Hall & Huang (2001) suggests an alternative smooth, monotonic nonparametric estimator that admits any number of covariates. Racine et al. (2009) have generalized the method to accommodate a variety of 'linear' constraints simultaneously. Start

again with the standard local-constant least-squares estimator

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} K\left((x_i - x)/h\right) y_i}{\sum_{i=1}^{n} K\left((x_i - x)/h\right)} = \frac{1}{n}\sum_{i=1}^{n} A_i(x) y_i, \tag{13}$$

where $A_i(x) = nK\left((x_i - x)/h\right) / \sum_{i=1}^{n} K\left((x_i - x)/h\right)$. Even though we are choosing to use the local-constant least-squares framework, this setup can be immediately extended to other types of kernel and local polynomial estimation routines. As it stands the regression estimator in (13) is not guaranteed to produce a monotonic estimator. Hall & Huang's (2001) insight was to introduce observation specific weights p_i instead of the $1/n$ that appears in (13). These weights can then be manipulated so that the estimator satisfies monotonicity. To be clear,

$$\widehat{m}(x|p) = \sum_{j=1}^{n} p_i A_i(x) y_i,$$

is the constraint weighted bootstrapping estimator. It is not necessarily monotonic unless we properly restrict the weights.

In the unconstrained setting we have $p = (p_1, \ldots, p_n) = (1/n, \ldots, 1/n)$ which represents weights drawn from a uniform distribution. If the bandwidth chosen produces an estimate that *is already* monotonic, the weights should be set equal to the uniform weights. However, if the function by itself is not monotonic then the weights are diverted away from the uniform case to create a monotonic estimate. In order to decide how to manipulate the weights a distance metric is introduced based on power divergence (Cressie & Read 1984):

$$D_\rho(p) = \frac{1}{\rho(1-\rho)}[n - \sum_{i=1}^{n}(np_i)^\rho], \quad -\infty < \rho < \infty. \tag{14}$$

where $\rho \neq 0, 1$. One needs to take limits for $\rho = 0$ or 1. They are given as

$$D_0(p) = -\sum_{i=1}^{n} \log(np_i); \qquad D_1(p) = \sum_{i=1}^{n} p_i \log(np_i).$$

This distance metric is quite general. If one uses $\rho = 1/2$, then this corresponds to Hellinger distance, whereas $nD_0(p) + n^2 \log(n)$ is equivalent to Kullback-Leibler divergence $\left(-\sum_{i=1}^{n} n \log(p_i/n)\right)$. This metric is minimized for a selected ρ subject to the constraint that

$$\widehat{m}'(\cdot \mid p) = \sum_{j=1}^{n} p_i A_i'(\cdot) y_i \geq \varepsilon,$$

on a grid of selected points. Here $\varepsilon \geq 0$ can be used to guarantee either weak or strict monotonicity. A nice feature of this estimator is that the kernel and bandwidth are chosen before the weights are selected. This means that the user can choose their desired kernel estimator and bandwidths selector to construct their nonparametric estimator and then constrain it to be monotonic. This

leaves the door open to straightforward modification of the estimator. In fact, there is nothing special about monotonicity for the method of Hall & Huang (2001) to work. Any constraint that is desired, could, in principle, be imposed on the regression surface.

Note that the monotonic constraint imposed in Hall & Huang (2001) can be written in the more general form:

$$\sum_{i=1}^{n} p_i \left[\sum_{\mathbf{s} \in \mathbf{S}} \alpha_{\mathbf{s}} A_i^{(\mathbf{s})}(x) \right] y_i - c(x) \geq 0, \tag{15}$$

where the inner sum is taken over all vectors $\mathbf{S}$ that correspond to our constraints of interest (monotonicity, say), $\alpha_{\mathbf{s}}$ are a set of constants used to generate various constraints and $c(x)$ is a known function. $\mathbf{S}$ indexes the order of the derivative associated with the kernel portion of the regression estimator. In our example of monotonicity, $\mathbf{s} = e_j$ is a k-vector (since we have $x \in \mathbb{R}^k$) with 1 in the j^{th} position and zeros everywhere else, $\alpha_{\mathbf{s}} = 1 \ \forall \mathbf{s} \in \mathbf{S}$ and $c(x) = 0$.[13] Racine et al. (2009) provide existence and uniqueness for a set of weights for constraints of the form (15). They call these constraints linear since they are linear with respect to the weights, $p_i \ \forall i$. Additionally, to make the constrained optimization computationally simple, they use the L_2 norm with respect to the uniform weights $(1/n)$, as opposed to the power divergence metric. This condenses the problem into a standard quadratic programming problem which can be solved using existing packages in almost all standard econometric software.

Note the subtle difference between the data sharpening methods discussed previously and the constraint weighted bootstrapping methods here. When one chooses to sharpen the data the actual data values are being transformed while the weighting is held constant. Here, the exact opposite occurs, the data is held fixed while the weights are changed. At the end of the day however, the two estimators can be viewed as 'visually' equivalent. That is, both estimators can be looked at as

$$\widehat{m}(x) = \sum_{j=1}^{n} A_i(x) y_i^*, \tag{16}$$

where y_i^* corresponds to either the sharpened values or $p_i y_i$ obtained from the constraint weighted bootstrapping approach. The difference between the methods is how to obtainy_i^*.[14] Also, note that both constraint weighted bootstrapping and data sharpening are *vertically* moving the data whereas rearrangement methods *horizontally* move the data.

2.7. **Summary of Methods.** While our discussion of existing methods has indicated a number of choices for the user, there does not exist one clear cut method for imposing arbitrary constraints on a regression surface for every given situation. Each of the methods discussed has computational or theoretical drawbacks when considered against the set of all available methods. Additionally, several of the key differences across the methods focuses on the choice of operating in a kernel, spline, or

[13]The notation $A^{(\mathbf{s})}$ refers to the order of the derivative of our weight function with respect to its argument.

[14]An interesting topic for future research would be to compare the performance of these methods across a variety of constraints.

series based framework, the selection of smoothing parameters, the smoothness of the estimator, the adaptability/generalizability of the method, whether to impose global or discrete constraints, and the ability to use the method to conduct inference on the constraints being imposed.

2.7.1. *Spline, Series and Kernels.* Given that the above constrained estimation methods discussed above use vastly differing nonparametric methods, this choice cannot be overlooked. Kelly & Rice (1990) mention that if the coefficients in the B-spline bases are nondecreasing, then so is the function (if one was imposing monotonicity) and Delecroix & Thomas-Agnan (2000) discuss that splines are defined as the solution to a minimization problem in general lend support for their use in constrained settings. However, given the prevalence of discrete data in applied settings, the seminal work of Racine & Li (2004) highlighting the fact that smoothing categorical data can lead to substantial finite sample efficiency gains, lends support for adopting a kernel based method. Alternatively, given the ease with which one may construct and employ series based methods, it is easy to advocate that these constrained methods are computationally easy to employ.

Given the adaptability of the methods of Yatchew & Bos (1997) (which is series based), Beresteanu (2004) (which is spline based) and Racine et al. (2009) (which is kernel based), we cannot advocate for a particular type of nonparametric method based on imposing general smoothness constraints. Nor do we advocate on behalf of the particular type of nonparametric smoothing one should engage in. However, given the ease with which one can implement a constrained estimator, we remark that the easiest method for which a researcher can incorporate the constraints should be used. Additionally, if a researcher traditionally uses a type of nonparametric method (spline say), then they may have more familiarity with employing one set of constrained methods over another which is an obvious benefit.

2.7.2. *Choice of Smoothing Parameter.* As with all nonparametric estimation methods, the choice of smoothing parameter plays a crucial role to the performance of the estimator both in practice and theory. No mention was given to the appropriate level of smoothing in the aforementioned constrained methods. Few results exist suggesting how the optimal level of smoothing should be imposed. For many of the methods described previously one could engage in cross-validation simultaneously with the constraint imposition. This may actually help in determination of the optimal smoothing parameter. The simulations of Delecroix & Thomas-Agnan (2000) show that the mean integrated square error (typically used in cross-validation) as a function of the smoothing parameter typically had a wider zone of stability around the optimal level of the smoothing parameter, suggesting it may be easer to determine the optimal level. It is well known that various forms of the cross-validation function are noisy, making determination of the optimal level difficult in certain settings.

However, engaging in cross-validation and constraint imposition simultaneously is unnecessary in particular methods. For example, the constraint weighted bootstrapping methods of Hall & Huang (2001) and Racine et al. (2009) show that the constrained kernel estimator should use a bandwidth of the standard, unconstrained optimal order. In this setting both the restricted and

unrestricted smooths will have the same level of smoothing. Further, tuning could be performed by cross-validation after the constraint weights have been found and simple checks to determine if the constraints were still satisfied (similar to that described above).

2.7.3. *Method Complexity.* The methods discussed above range from simple computation (rearrangement and univariate isotonic regression) to involving quadratic or nonlinear program solvers. These numerical methods may dissuade the user from adopting a specific approach, but we note that with the drastic reductions in computation time and the availability of solvers in most econometric software packages, these constraints will continue to lessen over time. Indeed, part of this survey discusses in detail the implementation of a sequential quadratic program to showcase its implementation in practice. Also, given the ease with which a quadratic program can be solved with linear constraints, the method of Racine et al. (2009) addresses the critique of Dette & Pilz (2006, Page 56) who note "[rearrangement offers] substantial computational advantages, because it does not rely on constrained optimization methods." We mention here that rearrangement requires slightly more sophistication when one migrates from a univariate to multivariate setting and so this concern is lessened in applied work.

2.7.4. *Numerical Comparisons.* Very little theoretical work exists to showcase the performance of one method against a set of competitors. Indeed, even numerical comparisons are scant. The most comprehensive study between methods is that of Dette & Pilz (2006) who conduct a Monte Carlo comparison of smooth isotonic regression, rearrangement, and the method of Hall & Huang (2001) for the constraint of monotonicity, in the univariate setting for a bevy of DGPs. Their findings suggest that rearrangement has desirable/equivalent finite sample performance compared to the other methods across all of the DGPs considered.

3. Imposing nonlinear constraints

We discuss a further generalization of Racine et al. (2009) that can handle general nonlinear constraints and discuss in detail the computational method of sequential quadratic programming required to implement nonparametric regression in this setting. Our choice for a deeper, prolonged discussion of this methods hinges on the necessity of sequential quadratic programming methods in several of the methods mentioned prior. Very rarely are the methods to obtain a solution discussed at length and given the use of these methods in both data sharpening and constraint weighted bootstrapping, we feel it requisite to highlight the implementation of this technique.

While we discuss general constrained estimation in the face of arbitrary nonlinear constraints, to cement our ideas we focus on the specific example of concavity. Concavity is a common assumption used in the characterization of production functions. Concavity of the production function implies diminishing marginal productivity of each input.[15] This assumption is widely agreed upon by economists and failure to impose it may lead to conclusions which are economically infeasible.

[15]Quasi-concavity does not imply diminishing marginal productivity to factor inputs. However, under constant returns to scale, quasi-concavity does guarantee diminishing marginal products. This is because quasi-concavity combined with constant returns to scale yields concavity. That being said, a major issue with constant returns to

In the case of a single factor, a twice continuously differentiable function $m(x)$ is said to be concave if $m''(x) \leq 0 \ \forall x \in \mathcal{S}(x)$. Extending this result to the case of multiple x's is relatively straight forward. Concavity implies that the Hessian matrix

$$H(m(x)) = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1k} \\ m_{21} & m_{22} & \cdots & m_{2k} \\ \vdots & & \ddots & \vdots \\ m_{k1} & m_{k2} & \cdots & m_{kk} \end{bmatrix},$$

where $m_{lk} \equiv \frac{\partial^2 m(x)}{\partial x_l \partial x_k}$, must be negative semi-definite. In other words, all the lth ($l = 1, 2, \ldots, k$) order principal minors of H are less than or equal to zero if l is odd and greater than or equal to zero if l is even (alternatively, all the eigenvalues of this matrix are negative). We could, instead, choose to impose concavity via the constraints given in Hildreth (1954), however, many formal definitions of concavity are linked to the Hessian and as such we enforce concavity using this.

Following Hall & Huang (2001), we have the following constrained nonlinear programming problem:

$$\min D_\rho(p) \ s.t. \ H(m(x|p)) \text{ is negative semi-definite } \forall x \in \mathcal{S}(x), \ p_i \geq 0 \ \forall i, \text{ and } \sum_{i=1}^{n} p_i = 1. \tag{17}$$

To solve this or any other constrained optimization problem in the spirit of Hall & Huang (2001) we need to use sequential quadratic programming.

3.1. **Sequential quadratic programming.** Although the steps to constructing a constrained nonparametric estimator seem straight forward, implementing these types of programs are often not discussed in detail in econometrics papers. In this sub-section we outline sequential quadratic programming (SQP).

Consider the inequality constrained problem

$$\min D(z) \text{ subject to } r_i(z) = 0, \ i \in \mathcal{E}, \text{ and } c_j(z) \geq 0, \ j \in \mathcal{I}. \tag{18}$$

Where $D : \mathbb{R}^{qo} \to \mathbb{R}$, $r_i : \mathbb{R}^{qo} \to \mathbb{R}^{q_1}$ and $c_j : \mathbb{R}^{qo} \to \mathbb{R}^{q_2}$ can all be nonlinear, but we require that all the functions are smooth in the argument z. The idea behind SQP is to convert the nonlinear programming problem in (18) into a conventional quadratic programming (QP) problem. To do this we need to 'linearize' our constraints and 'quadracize' our objective function. Before doing this we introduce some additional concepts.

The Lagrangian of our problem is defined as

$$\mathcal{L}(z, \lambda_r, \lambda_c) = D(z) - \lambda_r' r_i(z) - \lambda_c' c_j(z). \tag{19}$$

Also, define $B_r(z)' = [\nabla r_1(z), \nabla r_2(z), \ldots, \nabla r_n(z)]$ and $B_c(z)' = [\nabla c_1(z), \nabla c_2(z), \ldots, \nabla c_n(z)]$. Now pick an initial z, z_0, and an initial set of vectors of Lagrange multipliers, $\lambda_{r,0}$ and $\lambda_{c,0}$. Lastly,

scale is that it implies that both the average and marginal productivities of inputs are independent of the scale of production. In other words, they depend only on the relative proportion of inputs.

define $\nabla^2 \mathcal{L}_{zz}(z, \lambda_r, \lambda_c) = \nabla^2 D(z) - \nabla B_r(z)' \lambda_r - \nabla B_c(z)' \lambda_c$. We are now ready to describe how to solve our SQP problem.

Our QP at step 0 is

$$\min D(z_0) + \nabla D(z_0)' q + \frac{1}{2} q' \nabla^2_{zz} \mathcal{L}(z_0, \lambda_{r,0}, \lambda_{c,0}) q, \tag{20}$$

subject to

$$B_r(z_0) q + r(z_0) = 0 \text{ and } B_c(z) q + c(z_0) \geq 0. \tag{21}$$

The solution of this standard quadratic program, q_0, $\ell_{r,0}$, and $\ell_{c,0}$ can be used to update z_0, $\lambda_{r,0}$, and $\lambda_{c,0}$ as follows: $z_1 = z_0 + q_0$, $\lambda_{r,1} = \ell_{r,0}$, and $\lambda_{c,1} = \ell_{c,0}$. These updated values can then be plugged back into the SQP to repeat the whole process until convergence. SQP requires nothing more than repeated evaluation of the levels, first and second order derivatives of the objective and constraint functions. It is a simple matter to determine these derivatives, thus this simplification process requires nothing more than taking derivatives of a set of functions.

3.2. **Existence and uniqueness of a solution.** When the following assumptions hold:

(1) The constraint Jacobians, $B_r(z)$ and $B_c(z)$, have full row rank,
(2) The matrix $\nabla^2_{zz} \mathcal{L}(z, \lambda_r, \lambda_c)$ is positive definite on the tangent space of constraints,

our SQP has a unique solution that satisfies the constraints. Essentially, this result comes from the fact that one could have used Newton's method to solve the constrained optimization and the result here is obtained from the associated iterate from running Newton's method instead. These two assumptions are enough to guarantee that a unique solution holds if one were to use Newton's method instead of the one we outlined. However, Nocedal & Wright (2000, pgs. 531-532) show that these two procedures, in this setting, are equivalent. For more on existence of a *local* solution we direct the interested reader to Robinson (1974).

Additionally, since we have converted our general nonlinear programming problem into a QP problem, the conditions required for existence of a solution in QP problems are exactly the conditions we need to hold, at each iteration, to guarantee a solution exists in this setting. Thus, the results established in Racine et al. (2009) carry over to our setting, provided our nonlinear constraints are first order differentiable in the p's and satisfy our assumptions listed above, which are easily checked. Moreover, if the forcing matrix $(\nabla^2_{zz} \mathcal{L}(q, \lambda_r, \lambda_c))$ in the quadratic portion of our 'quadricized' objective function is positive semidefinite and if our solution satisfies the set of linearized equality/inequality constraints then our soluation is the unique, global solution to the problem (Nocedal & Wright 2000, Theorem 16.4). Positive semi-definiteness guarantees that our objective function is convex which is what yields a global solution. We note that this only shows uniqueness but does not guarantee a solution will even exist.

However, it should be noted that because the constraint weights are restricted to be nonnegative and sum to one, this implies that it may be difficult to impose a constraint that is 'far away' from being satisfied. In essence, the constraints imposed on the problem may be inconsistent if

a nonnegative weight or a weight greater than one is *needed* to satisfy the constraints of interest. However, the conditions needed to determine how far away is 'far away' are not investigated here. Our conjecture is that the distance from an observation and the underlying function is dependent on the error process that perturbs the data generating process.

In essence the weights act as vertical scaling factors and if the amount of scaling is restricted then it can be difficult to find a solution. Hall & Presnell (1999) note the difficulty in finding the appropriately sharpened points using essentially the same technique described here in roughly 10% of their simulations. They advocate for an approach similar to simulated annealing that always was able to arrive at a solution although that procedure was computational more intensive than SQP. An alternative, not followed here, would be to dispense with the power divergence metric and all constraints on the weights if no solution is found in the SQP format. In this setting one could use the L_2 norm of Racine et al. (2009) and linearize (provided the nonlinear constraints are differentiable) the nonlinear constraints, again engaging in an iterative procedure to determine the optimal set of weights which can be shown to always exist in this setting.

3.3. **SQP imposing concavity.** If we use the power divergence measure of Cressie & Read (1984):

$$D_\rho(p) = \frac{1}{\rho(1-\rho)}\left\{n - \sum_{i=1}^{n}(np_i)^\rho\right\},$$

for $-\infty < \rho < \infty$ and $\rho \neq 0$, 1, as our objective function to minimize, then we have the following set of functions that need to be estimated prior to solving our QP at any iteration (ℓ^{th}):

(i) $D_\rho(p_\ell) \equiv \frac{1}{\rho(1-\rho)}\left\{n - \sum_{i=1}^{n}(np_{i,\ell})^\rho\right\}$.
(ii) $\nabla D_\rho(p_\ell) = vec\left[\frac{-n}{1-\rho}(np_{i,\ell})^{\rho-1}\right]$.
(iii) $\nabla^2 D_\rho(p_\ell) = diag\left[n^2(np_{i,\ell})^{\rho-2}\right]$.
(iv) $r(z) \equiv \sum_{i=1}^{n} p_{i,\ell} - 1$.
(v) $B_r(p_\ell) = [1, 1, \ldots, 1]$, an n-vector of ones.
(vi) $\nabla B_r(p_\ell)$ which is an $n \times n$ matrix of zeros.

Our objective function is defined in (i) whereas (ii) and (iii) are the first and second partial derivatives of our objective function, respectively. Our equality constrained function (ensuring the weights sum to 1) is defined in (iv) and the first and second partial derivatives of this function are given in (v) and (vi).

Additionally, we have to calculate our inequality constrained functions as well as their first and second partial derivatives, which can be broken into two pieces. First, we focus directly on the linear inequality constraints, $p_i \geq 0\ \forall i$. For this we have

(i) $B_{c,1}(p_\ell) = [e_1, e_2, \ldots, e_n]$, where e_j is an $n-$vector of zeros with a 1 in the j^{th} spot.
(ii) $\nabla B_{c,1}(p_\ell)$ which is an $n \times n$ matrix of zeros.

We also have to calculate the first and second derivatives of the determinants of the principal minors of our Hessian matrix *for each* point we wish to impose concavity. In a local-constant

setting, the Hessian matrix is calculated as follows. Assume that we have q continuous covariates and we are smoothing with a standard product kernel with second order, individual Gaussian kernels. Then, defining

$$K_i(x) \equiv (2\pi)^{-q/2} \prod_{j=1}^{q} h_j^{-1} e^{\frac{-(x_j - x_{ji})^2}{2h_j^2}}$$

we can derive

$$\frac{\partial K_i(x)}{\partial x_s} = -\left(\frac{x_s - x_{si}}{h_s^2}\right) K_i(x), \tag{22}$$

and we can easily determine that

$$\frac{\partial^2 K_i(x)}{\partial x_s \partial x_r} = \left[\left(\frac{x_s - x_{si}}{h_s^2}\right)\left(\frac{x_r - x_{ri}}{h_r^2}\right) + \delta_{sr}\frac{1}{h_s^2}\right] K_i(x), \tag{23}$$

where $\delta_{sr} = 1$ when $s = r$ and zero otherwise.

Recalling that $A_i(x) = nK_i(x)/\sum_{i=1}^{n} K_i(x)$, we have

$$\begin{aligned}
\frac{\partial A_i(x)}{\partial x_s} &= \frac{n\frac{\partial K_i(x)}{\partial x_s}\sum_{i=1}^{n} K_i(x) - nK_i(x)\sum_{i=1}^{n}\frac{\partial K_i(x)}{\partial x_s}}{\left[\sum_{i=1}^{n} K_i(x)\right]^2} \\
&= A_i(x)\left[n^{-1}\sum_{i=1}^{n} D_i(x_s)A_i(x) - D_i(x_s)\right] = A_i(x)M_s(x), \qquad (24)
\end{aligned}$$

where $D_i(x_s) = \frac{x_s - x_{si}}{h_s^2}$. Similar arguments show that

$$\begin{aligned}
\frac{\partial^2 A_i(x)}{\partial x_s \partial x_r} &= \frac{\partial A_i(x)}{\partial x_r}M_s(x) + A_i(x)\frac{\partial M_s(x)}{\partial x_r} \\
&= A_i(x)M_s(x)M_r(x) + A_i(x)\left[M_r(x)n^{-1}\sum_{i=1}^{n} D_i(x_s)A_i(x)\right] \\
&= A_i(x)M_r(x)\left[2M_s(x) + D_i(x_s)\right]. \qquad (25)
\end{aligned}$$

Our first order partial derivatives of our local-constant smoother are

$$\frac{\partial \hat{m}(x|p)}{\partial x_s} = \sum_{i=1}^{n} p_i y_i \frac{\partial A_i(x)}{\partial x_s} = \sum_{i=1}^{n} p_i y_i A_i(x) M_s(x). \tag{26}$$

Note that we cannot pull $M_s(x)$ through the summation since it has a $D_i(x_s)$ inside of it so that it depends on the counter. To determine the second order partial derivatives of our smooth regression

function we use our results from equation (25) to obtain

$$\frac{\partial^2 \hat{m}(x|p)}{\partial x_s \partial x_r} = \sum_{i=1}^{n} p_i y_i \frac{\partial^2 A_i(x)}{\partial x_s \partial x_r} = \sum_{i=1}^{n} p_i y_i \left[A_i(x) M_r(x) \left(2M_s(x) + D_i(x_s) \right) \right]$$

$$= 2 \sum_{i=1}^{n} p_i y_i A_i(x) M_r(x) M_s(x) + \sum_{i=1}^{n} p_i y_i A_i(x) M_r(x) D_i(x_s). \tag{27}$$

One can save computation time by noting that terms required for calculation of $M_s(x)$, $M_r(x)$ and $D_i(x_s)$ are all calculated when $A_i(x)$ is calculated. We suggest using numerical techniques in the user's preferred software to calculate the first and second derivatives of the Hessian matrix to then pass to the SQP.[16] For k covariates, if one imposes concavity for each of the n points then this requires construction of n $k \times k$ Hessian matrices. There are k determinants of principal minors (or k eigenvalues) to be calculated for each Hessian representation, resulting in nk constraints to go with the $n+1$ constraints placed on the weights. This results in a total of $n(k+1)+1$ total constraints.[17] As noted in the introduction, imposing concavity over the entire support of the data may be burdensome since near the boundaries it will be harder to enforce the constraints. However, using an interior hypercube of the data will lessen the burden on the SQP since concavity will be less likely to be violated (assuming concavity holds in the limit) on the interior of the support.

4. Demonstration

4.1. **Simulated Examples.** This section uses Monte Carlo simulations to examine the finite sample performance of the nonlinearly constrained estimator described above. Following the focus on concavity, we choose to perform our simulations imposing concavity in models which should be concave. We consider the following data generating process used to motivate our problem in the introduction:

$$y = \ln(x) + u \tag{28}$$

where x is generated as uniform distribution from 0.5 to 1.5 and u is generated as normal with mean zero and variance equal to 0.1. Note that this data generating process produces a theoretically consistent concave function. However, both the unknown error and finite sample biases of the estimator itself may cause the kernel estimate to exhibit ranges of non-concavities.

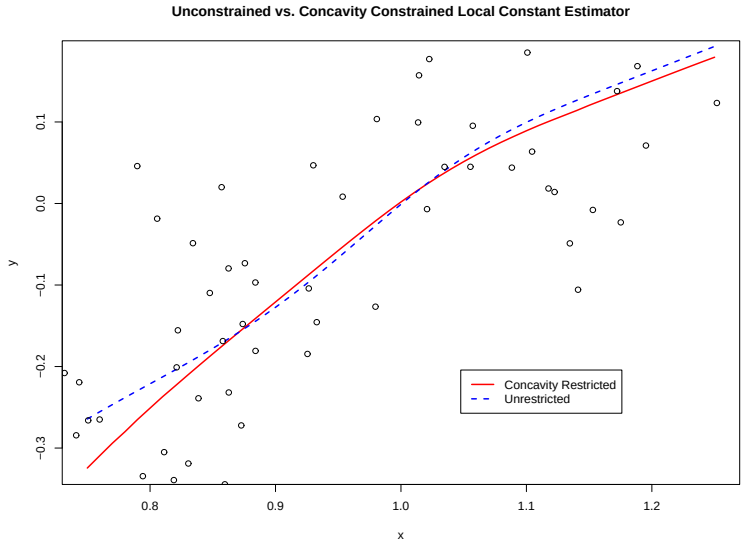
We consider samples of $n = 100$ and 500 for each of our 999 Monte Carlo replications. We present results using $\rho = 0.5$, but note that other choices for ρ do not significantly change the results. We use local-constant least-squares and a Gaussian kernel with $h = 1.06\sigma_x n^{-1/5}$. The weights ($p$) are found using the sequential quadratic programming routine SQPSolve in the programming language

[16] An alternative would be to solve analytically for all of these derivatives, perhaps with the assistance of a numerical software such as Maxima, Maple, or Mathematica.

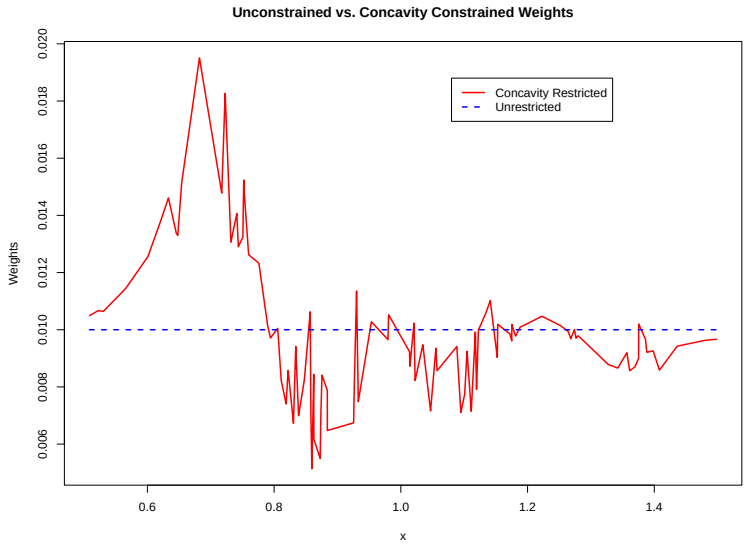
[17] If one can also assume monotonicity, then to impose concavity all one requires is that the second order derivatives are negative thus only $2n$ constraints need to be imposed which is always fewer constraints than imposing concavity without monotonicity.

GAUSS 8.0. While our problem is not a quadratic programming problem, this type of solver uses a modified quadratic program to find the step length for moving in the direction of a minimum.

FIGURE 1. Simulation for $n = 100$ corresponding to 95^{th} percentile of $D_{1/2}(p)$ for 999 simulations.



(a)



(b)

The simulation results for (28) are given in Figures 1 and 2 for $n = 100$ and 500, respectively. Each of the curves correspond to the 95th percentile of the distance metric for each sample size.[18] The dashed line in panel (a) of each figure is the corresponding unconstrained local-constant least-squares estimator and the solid line is the constrained local-constant least-squares estimator. We note that in each case the constrained estimator deviates from the unconstrained estimator where the second derivative is positive. This difference is shown by positive values for the distance metric. Specifically, in Figures 1 and 2 the values of the distance metric are 0.111 and 0.069, respectively. Note that the distance metric decreases with the sample size. It is easy to see that as the sample size increases that the incidence of concavity increases and the constrained and unconstrained estimator appear to be more similar. Recall that the distance metric reaches its minimum of 0 when each weight is set equal to $1/n$, or in other words, the estimated function is *de facto* concave.

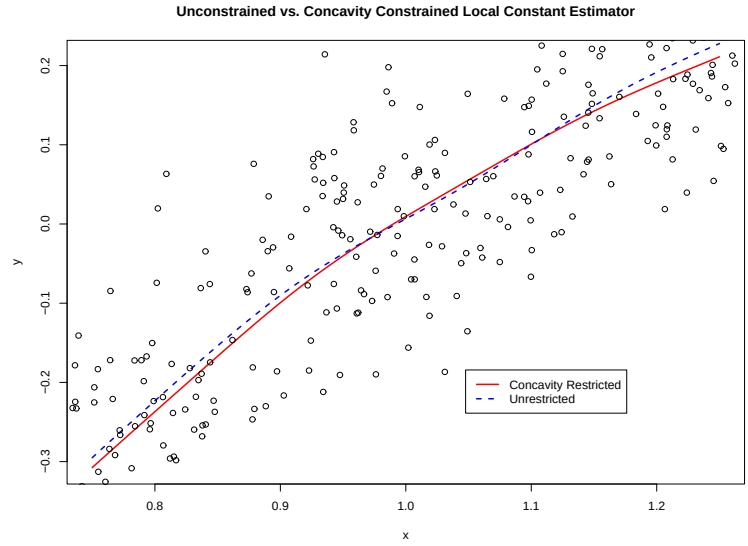
In panel (b) of each figure is the corresponding set of weights. The unconstrained estimator sets each of the weights equal to $1/n$. It is obvious that the unconstrained estimators show regions where the second derivative is positive. Our constrained estimator corrects for these non-concavities by changing the probability weights. Where the weights are larger than $1/n$ these points are given a greater influence in the construction of the estimate and where the weights are less than $1/n$ these observations are given a lesser influence in the construction of the estimate.

4.2. **Empirical Application.** The seminal work of Jacob Mincer on human capital suggested that the logarithm of a worker's earnings is concave in her age (potential work experience). Concavity is consistent with the investment behavior implied by the optimal distribution of human capital investment over a worker's life cycle. A voluminous literature within labor economics has generally specified age-earnings profiles as quadratic (Heckman & Polachek (1974)), consistent with concavity. Murphy & Welch (1990) challenged the conventional empirical strategy of specifying a quadratic in age for an age-earnings profile. Their work suggests that a quadratic specification in age understates early career earnings growth by 30-50% and overstates midcareer earnings growth by 20-50%. An analysis of residual plots from their estimated quadratic relationships (as well as several statistical tests) reveal patterns suggesting substantial differences from this specification. They advocate on behalf of a quartic age-earnings profile and find that this specification yields a substantial improvement in fit relative to the common quadratic relationship.

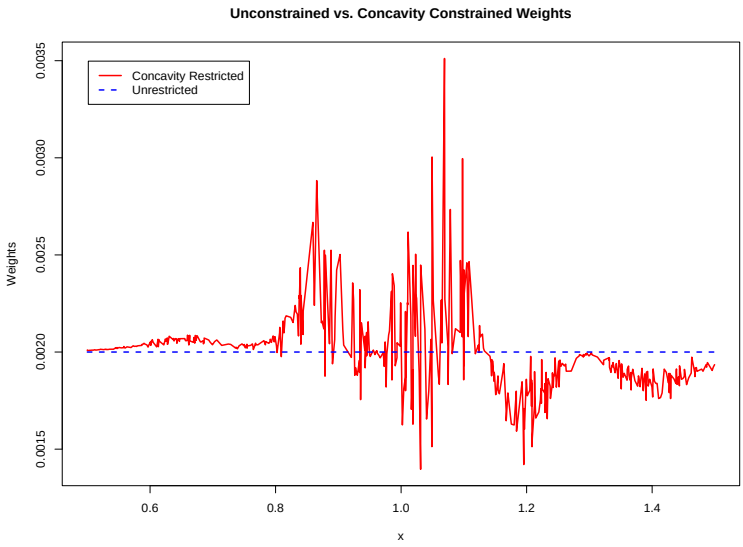
Given that the human capital theory of Mincer does not suggest a particular empirical relationship, Pagan & Ullah (1999, Section 3.14.2) considered the use of nonparametric regression techniques to shed light on the appropriate link between income and ages. They provided an example using the 1971 Canadian Census Public Use Tapes consisting of 205 individuals who had 13 years of education. Fitting a local-constant kernel regression function (see their Figure 3.4) they found a visually substantial difference between the common quadratic specification and their nonparametric estimates. A 'dip' in the age-earnings profile around age 40 suggested that the

[18]It should be noted that the number of times that the unconstrained estimator was concave over the grid was very small. Specifically, out of the 999 Monte Carlo simulations for each scenario, the unconstrained estimator was concave 20 and 37 times when $n = 100$ and 500 observations, respectively.

FIGURE 2. Simulation for $n = 500$ corresponding to 95^{th} percentile of $D_{1/2}(p)$ for 999 simulations.



(a)



(b)

relationship was neither quadratic nor concave. Pagan & Ullah (1999) argue that this 'dip' may occur because of generational effects present in the cross-section, specifically, pooling workers who have differing earnings trajectories.

Given the need to conform to theory in applied work, partnered with the findings of Murphy & Welch (1990) and Pagan & Ullah (1999) we fit a concavity restricted age-earnings profile. This

approach will adopt the theoretical restrictions but relax the functional form specifications primarily used in the empirical labor economics literature. Panel (a) of figure 3, plots the unrestricted nonparametric regression estimator of Pagan & Ullah (1999) (using bandwidth $h = \hat{\sigma}_{Age} n^{-1/5}$), the concave restricted estimator with identical bandwidth and the common quadratic specification.[19] The corresponding weights are provided in panel (b).

We see that the concavity restricted estimator still has a visually distinct difference from the quadratic specification around age 40 (as does the unrestricted nonparametric estimator), yet the concave restricted estimator does not have the 'dip' found in Pagan & Ullah (1999), consistent with the core interpretation of Mincer's human capital theory. Additionally, the unrestricted estimator appears to have a slight nonconcavity around age 25, further highlighting the need to impose concavity.

To focus on the importance of the bandwidth in examining this relationship we plot our the unrestricted estimator of Pagan & Ullah (1999) using their bandwidth as well as the optimal bandwidth found using least-squares cross-validation along with their corresponding concavity restricted fits. These plots are provided in Figure 4, panel (a). The 'dip' presented in Pagan & Ullah (1999) now takes on the appearance of a trough. Again, both unconstrained estimators are nonconcave. The estimator using the cross-validated bandwidths produces a distance metric value of 0.005272, almost double that found using the rule-of-thumb bandwidth. In addition to the nonconcave area around age 40, the cross-validated curve has a region of nonconcavity around age 33 which is more distinct that that for the curve of Pagan & Ullah (1999) which has a slight area of nonconcavity around age 25. The constraint weights, presented in panel (b) of Figure 4, bear this out as well. An interesting feature of this comparison is that the constraint weights for the cross-validated curve appear to be rougher than those for the rule-of-thumb curve whereas the cross-validated bandwidth is smaller than the rule-of-thumb bandwidth (1.89 vs. 4.22).

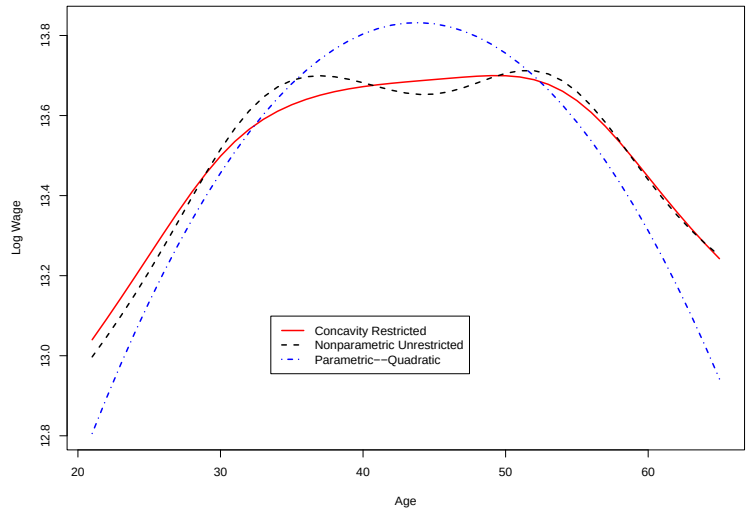
While we have not statistically tested for a difference between our concave restricted nonparametric estimator and the unconstrained estimator, our example shows that we can think more soundly about the implementation of nonparametric estimators in the presence of economic smoothness conditions. We mention again that the ability to impose theoretically consistent smoothness constraints on an economic relationship paired with the ability to relax restrictive functional form requirements provides the researcher with a serious set of tools with which to investigate substantive economic questions.

5. Conclusion

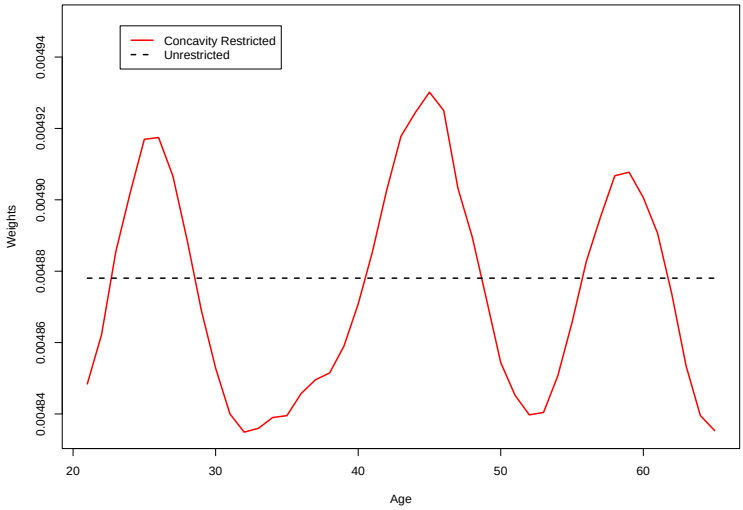
This chapter has surveyed the existing literature on imposing constraints in nonparametric regression, described a plethora of methods and discussed computational implementation. This survey included recent research that has not been discussed previously in the literature. We also described a novel method to impose general nonlinear constraints in nonparametric regression that can be implemented using only a standard quadratic programming solver. We illustrated this method

[19]Our restricted estimator was calculated using $\rho = 1/2$ and at the optimum we had $D_{1/2}(\hat{p}) = 0.003806$.

FIGURE 3. Unrestricted, Restricted and Quadratic Fits of the Age-Earnings Profile, CPS 1971 data.



(a)

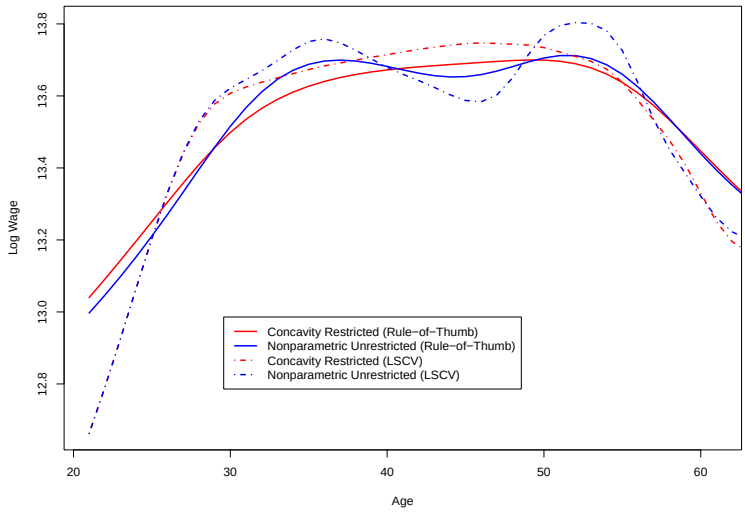


(b)

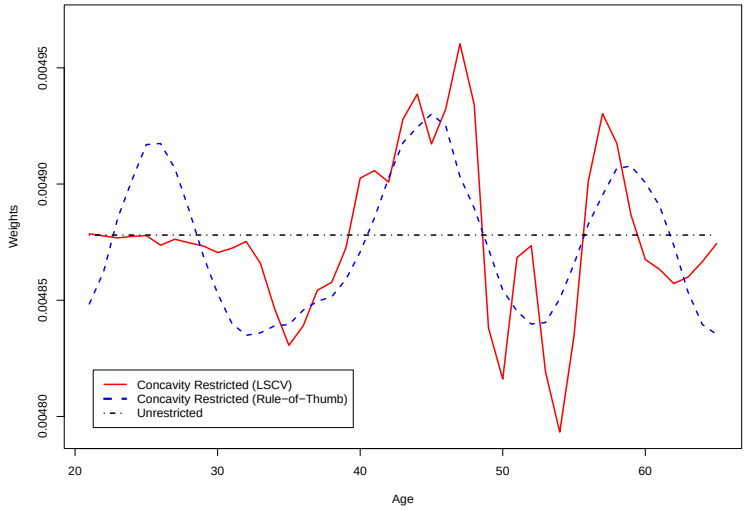
with a small simulated example focusing on concavity and a detailed example from the empirical labor economics literature. Our results showcased that constrained nonparametric methods can still uncover detail in the data overlooked by rigid parametric models while maintaining theoretical consistency.

Overall future research should determine the relevant merits of each of the methods described here to narrow the set of potential methods down to a few which can be easily and successfully

FIGURE 4. Unrestricted and Restricted with Differing Bandwidths of the Age-Earnings Profile, CPS 1971 data.



(a)



(b)

used in applied nonparametric settings. Given the dearth of detailed simulation studies comparing the available methods highlighted here (notwithstanding Dette & Pilz 2006), an interesting topic for future research would be to compare the varying methods (kernel, spline, series) across various constraints to discover under which settings which methods perform the best. Additionally, we feel that our description of the available methods should help to further research in extending these ideas to additional nonparametric settings, most notably in the estimation of quantile functions

(Li & Racine 2008), conditional densities, treatment effects (Li, Racine & Wooldridge 2008), and structural estimators (Henderson et al. 2008).

References

Beresteanu, A. (2004), Nonparametric estimation of regression functions under restricitons on partial derivatives. *Mimeo*, Duke University.

Braun, W. J. & Hall, P. (2001), 'Data sharpening for nonparametric inference subject to constraints', *Journal of Computational and Graphical Statistics* **10**, 786–806.

Brunk, H. D. (1955), 'Maxmimum likelihood estimates of monotone parameters', *Annals of Mathematical Statistics* **26**, 607–616.

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

Choi, E. & Hall, P. (1999), 'Data sharpening as a prelude to density estimation', *Biometrika* **86**, 941–947.

Cressie, N. A. C. & Read, T. R. C. (1984), 'Multinomial goodness-of-fit tests', *Journal of the Royal Statistical Society, Series B* **46**, 440–464.

Delecroix, M. & Thomas-Agnan, C. (2000), Spline and kernel regression under shape restrictions, *in* M. G. Schimek, ed., 'Smoothing and Regression: Approaches, Computation, and Application', Wiley Series in Probability and Statistics, John Wiley & Sons, chapter 5, pp. 109–133.

Dette, H., Neumeyer, N. & Pilz, K. F. (2006), 'A simple nonparametric estimator of a strictly monotone regression function', *Bernoulli* **12**(3), 469–490.

Dette, H. & Pilz, K. F. (2006), 'A comparative study of monotone nonparametric kernel estimates', *Journal of Statistical Computation and Simulation* **76**(1), 41–56.

Dierckx, P. (1980), 'An algorithm for cubic spline fitting with convexity constraints', *Computing* **24**, 349–371.

Dykstra, R. (1983), 'An algorithm for restricted least squares', *Journal of the American Statistical Association* **78**, 837–842.

Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*, Dekker, New York.

Friedman, J., Tukey, J. W. & Tukey, P. (1980), Approaches to analysis of data that concentrate near intermediate-dimensional manifolds, *in* E. Diday et al., ed., 'Data Analysis and Informatics', North-Holland: Amsterdam.

Gallant, A. R. (1981), 'On the bias in flexible functional forms and an essential unbiased form: The fourier flexible form', *Journal of Econometrics* **15**, 211–245.

Gallant, A. R. (1982), 'Unbiased determination of production technologies', *Journal of Econometrics* **20**, 285–323.

Gallant, A. R. & Golub, G. H. (1984), 'Imposing curvature restrictions on flexible functional forms', *Journal of Econometrics* **26**, 295–321.

Goldman, S. & Ruud, P. (1992), Nonparametric multivariate regression subject to constraint, Technical report, University of California, Berkeley, Department of Economics.

Hall, P. & Huang, H. (2001), 'Nonparametric kernel regression subject to monotonicity constraints', *The Annals of Statistics* **29**(3), 624–647.

Hall, P. & Presnell, B. (1999), 'Intentionally biased bootstrap methods', *Journal of the Royal Statistical Society Series B* **61**, 143–158.

Hansen, D. L., Pledger, G. & Wright, F. T. (1973), 'On consistency in monotonic regression', *Annals of Statistics* **1**(3), 401–421.

Heckman, J. & Polachek, S. (1974), 'Empirical evidence of functional form of the earnings-schooling relationship', *Journal of the American Statistical Association* **69**(346), 350–354.

Henderson, D. J., List, J. L., Millimet, D. L., Parmeter, C. F. & Price, M. K. (2008), Imposing monotonicity nonparametrically in first price auctions. Virginia Tech AAEC working paper.

Hildreth, C. (1954), 'Point estimates of ordinates of concave functions', *Journal of the American Statistical Association* **49**, 598–619.

Holm, S. & Frisen, M. (1985), Nonparametric regression with simple curve characteristics, Technical Report 4, Department of Statistics, University of Goteborg, Goteborg, Sweden.

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

Li, Q. & Racine, J. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Li, Q. & Racine, J. S. (2008), 'Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data', *Journal of Business & Economic Statistics* **26**(4), 423–434.

Li, Q., Racine, J. S. & Wooldridge, J. M. (2008), 'Estimating average treatment effects with continuous and discrete covariates: The case of swan-ganz catherization', *American Economic Review* **98**(2), 357–362.

Mammen, E. (1991*a*), 'Estimating a smooth monotone regression function', *Annals of Statistics* **19**(2), 724–740.

Mammen, E. (1991*b*), 'Nonparametric regression under qualitative smoothness assumptions', *Annals of Statistics* **19**(2), 741–759.

Mammen, E., Marron, J. S., Turlach, B. A. & Wand, M. P. (2001), 'A general projection framework for constrianed smoothing', *Statistical Science* **16**(3), 232–248.

Matzkin, R. L. (1991), 'Semiparametric estimation of monotone and concave utility functions for polychotomous choice models', *Econometrica* **59**, 1315–1327.

Matzkin, R. L. (1992), 'Nonparametric and distribution-free estimation of the binary choice and the threshold-crossing models', *Econometrica* **60**, 239–270.

Matzkin, R. L. (1993), 'Nonparametric identification and estimation of polychotomous choice models', *Journal of Econometrics* **58**, 137–168.

Matzkin, R. L. (1994), Restrictions of economic thoery in nonparametric methods, *in* D. L. McFadden & R. F. Engle, eds, 'Handbook of Econometrics', Vol. 4, North-Holland: Amsterdam.

Matzkin, R. L. (1999), Computation of nonparametric concavity restricted estimators. *Mimeo*.

Mukerjee, H. (1988), 'Monotone nonparametric regression', *Annals of Statistics* **16**, 741–750.

Murphy, K. M. & Welch, F. (1990), 'Empirical age-earnings profiles', *Journal of Labor Economics* **8**(2), 202–229.

Nocedal, J. & Wright, S. J. (2000), *Numerical Optimization*, 2nd edn, Springer.

Pagan, A. & Ullah, A. (1999), *Nonparametric Econometrics*, Cambridge University Press, New York.

Racine, J. S. & Li, Q. (2004), 'Nonparametric estimation of regression functions with both categorical and continuous data', *Journal of Econometrics* **119**(1), 99–130.

Racine, J. S., Parmeter, C. F. & Du, P. (2009), Constrained nonparametric kernel regression: Estimation and inference. Working Paper.

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

Robinson, S. M. (1974), 'Perturbed kuhn-tucker points and rates of convergence for a class of nonlinear-programming algorithms', *Mathematical Programming* **7**, 1–16.

Ruud, P. A. (1995), Restricted least squares subject to monotonicity and concavity restraints. Presented at the 7^{th} World Congress of the Econometric Society.

Schumaker, L. (1981), *Spline Functions: Basic T*, Wiley, New York.

Wahba, G. (1990), *Spline models for observational data*, CBMS-NSF Conference Series in Applied Mathematics 59, SIAM, Philadelphia, PA.

Wheelock, D. C. & Wilson, P. W. (2001), 'New evidence on returns to scale and product mix among U.S. commercial banks', *Journal of Monetary Economics* **47**(3), 653–674.

Yatchew, A. & Bos, L. (1997), 'Nonparametric regression and testing in economic models', *Journal of Quantitative Economics* **13**, 81–131.

Yatchew, A. & Härdle, W. (2006), 'Nonparametric state price density estimation using constrained least squares and the bootstrap', *Journal of Econometrics* **133**, 579–599.