Statistica Sinica **23** (2013), 1347-1371
doi:http://dx.doi.org/10.5705/ss.2012.024

NONPARAMETRIC KERNEL REGRESSION WITH MULTIPLE PREDICTORS AND MULTIPLE SHAPE CONSTRAINTS

Pang Du, Christopher F. Parmeter and Jeffrey S. Racine

Virginia Tech, University of Miami and McMaster University

Abstract: Nonparametric smoothing under shape constraints has recently received much well-deserved attention. Powerful methods have been proposed for imposing a single shape constraint such as monotonicity and concavity on univariate functions. In this paper, we extend the monotone kernel regression method in Hall and Huang (2001) to the multivariate and multi-constraint setting. We impose equality and/or inequality constraints on a nonparametric kernel regression model and its derivatives. A bootstrap procedure is also proposed for testing the validity of the constraints. Consistency of our constrained kernel estimator is provided through an asymptotic analysis of its relationship with the unconstrained estimator. Theoretical underpinnings for the bootstrap procedure are also provided. Illustrative Monte Carlo results are presented and an application is considered.

Key words and phrases: Hypothesis testing, multivariate kernel estimation, nonparametric regression, shape restrictions.

1. Introduction

Imposing shape constraints on a regression model is a necessary component of sound applied data analysis. For example, imposing monotonicity and concavity constraints is often required in a range of application domains. Early developments in the absence of smoothness restrictions can be found in Robertson, Wright, and Dykstra (1988) and the references therein. When data are believed to have nonlinear structure in addition to obeying shape constraints, nonparametric smoothing methods such as kernel smoothing and splines are often used. For example, restricted kernel smoothing has been considered by Mukerjee (1988), Mammen (1991), Hall and Huang (2001), Braun and Hall (2001), Hall and Kang (2005), Birke and Dette (2007), and Carroll, Delaigle, and Hall (2011), among others; restricted spline smoothing has been studied by Wright and Wegman (1980), Ramsay (1988), Mammen and Thomas-Agnam (1999), Pal and Woodroofe (2007), and Wang and Shen (2010), to name but a few. An example of restricted estimation by other smoothing methods is Xu and Phillips

(2012) where positivity is imposed on the estimation of a conditional covariance function by empirical likelihood.

As pointed out in the comprehensive review by Mammen et al. (2001), a common limitation of these approaches is that they only consider shape constraints on univariate functions and often only one constraint at a time is entertained. Extensions to either multivariate functions or multiple constraints has received far less attention. For example, Dette and Scheder (2006) consider the problem of estimating a multivariate regression function that is strictly monotone in all directions by successively applying one-dimensional isotonization procedures to an initial unconstrained kernel regression estimator, while Birke and Pilz (2009) impose monotonicity and convexity in the kernel estimation of a single dimension call price function.

In practice, however, many applications call for imposing multiple shape constraints on multivariate functions. Examples include dose-response studies where multiple drug combinations are applied and the response function is often assumed to be monotone in the amount of each drug, or economic studies where the response function must satisfy coordinate-wise monotonicity, coordinate-wise concavity, and constant returns to scale that essentially require first-order partial derivatives from a double-log model to sum to one. For such problems, Gallant (1982) and Gallant and Golub (1984) proposed a series-based estimator called the Flexible Fourier Form (Gallant (1981)) whose coefficients can be restricted to impose the relevant shape constraints. Although the method can handle multiple constraints, it is hard to incorporate certain common constraints such as monotonicity. Villalobos and Wahba (1987) proposed a constrained thin-plate spline estimator and applied it to the posterior probability estimation in classification problems where the probability function must lie between 0 and 1. But the asymptotic properties of their estimator are not well understood. Matzkin (1991, 1992) studied constrained maximum likelihood estimators using interpolation, but these estimators are not smooth and are hard to generalize beyond coordinate-wise monotonicity and concavity. Mammen et al. (2001) discussed extension of their projection framework for constrained smoothing to nonparametric additive models, which can be considered as a special case of the general constraints that we consider below. Also, none of the aforementioned approaches for multiple constraints propose hypothesis testing procedures for the constraints themselves.

In this paper, we propose a kernel smoothing method that can handle multiple general shape constraints for multivariate functions. Our method can be considered as a generalization of Hall and Huang (2001) where monotone regression was considered for a general class of kernel smoothers, as well as a partial

generalization of Carroll, Delaigle, and Hall (2011) where the monotonicity constraint is replaced by a general shape constraint with the complication of measurement errors. Similar to Hall and Huang (2001), our estimator is constructed by introducing weights for each response data point that can dampen or magnify the impact of any observation. In order to deliver an estimate satisfying the shape constraints, the weights are selected to minimize their distance from the uniform weights of the unconstrained estimator while obeying the constraints. In Hall and Huang (2001), this distance was measured by a power divergence metric introduced in Cressie and Read (1984); this has a rather complicated form and is hard to generalize. Instead, we resort to the well-known l_2-metric that is much simpler and has all the desired properties of the power divergence metric (Theorem 1). Under certain conditions, we generalize the consistency results of Hall and Huang (2001, Thm. 4.3) to our multivariate and multiple constraint setting. In essence, when the shape constraints are non-binding (i.e., strictly satisfied) on the domain, the restricted estimator is asymptotically and numerically equivalent to the unrestricted estimator. When the shape constraints are non-binding everywhere except for a certain area of measure 0, the restricted and unrestricted estimators are still very close to each other except in a neighborhood of the binding area.

Besides the multivariate and multi-constraint extension, we also propose a bootstrap procedure to test the validity of the shape constraints. Inference for shape restrictions in nonparametric regression settings has attracted much attention in the past decade. For example, Hall and Heckman (2000) developed a bootstrap test for monotonicity whose test statistic is formulated around the slope estimates of linear regression models fitted over small intervals. Ghosal, Sen, and van der Vaart (2000) introduced a statistic based on a U-process measuring the discordance between predictor and response. Both of these approaches are specifically designed for monotonicity constraints and it is difficult to extend them to our multivariate and multi-constraint setting. Yatchew and Härdle (2006) employed a residual-based test to check for monotonicity and convexity simultaneously in a regression setting, but their result is for univariate functions and requires that the constraints be strictly satisfied. Our bootstrap procedure originates from Hall et al. (2001), which tested the monotonicity of hazard functions but provided no theoretical justification for the procedure. Although the test statistic is difficult to analyze asymptotically, we are able to provide asymptotic results for its implementation when shape constraints are satisfied on a sufficiently dense grid of points. The derivation of our result takes advantage of the simple form of the l_2 metric that may not be easily available for the power divergence metric, a promising artifact of the l_2 metric we adopt. Another appealing aspect of our method is that it only involves quadratic programming and

can be readily implemented using standard sequential quadratic programming software.

The rest of this paper proceeds as follows. Section 2 proposes a general nonparametric regression estimator in the presence of linear shape constraints and presents a simple test of the validity of the constraints. Section 3 investigates the theoretical properties of the proposed method, including the existence and consistency of the constrained estimator as well as the asymptotic behavior of the test statistic. Section 4 considers examples treating monotonicity/concavity and global concavity. Section 5 considers a number of simulated applications, examines the finite-sample performance of the proposed test, and presents an empirical application involving technical efficiency on Indonesian rice farms. Section 6 presents some discussion and concluding remarks. The proofs of the theorems are presented in the online Supplementary Material.

2. A Constrained Kernel Regression Estimator

In the following two subsections we outline the mechanics of our estimation and inference procedures before launching into theoretical properties and proofs. Theoretical properties are then outlined in Section 3 while the detailed proofs are relegated to the online Supplementary Material.

2.1. The estimator

In what follows we let $\{Y_i, \mathbf{X}_i\}_{i=1}^n$ denote sample pairs of response and explanatory variables, where Y_i is a scalar, $\mathbf{X}_i$ is of dimension r, and n denotes the sample size. The goal is to estimate the unknown mean response $g(\mathbf{x}) \equiv E(Y|\mathbf{X} = \mathbf{x})$ from the regression model $Y_i = g(\mathbf{X}_i) + \varepsilon_i$, subject to constraints on $g^{(\mathbf{s})}(\mathbf{x})$, where $\mathbf{s}$ is an r-vector corresponding to the dimension of $\mathbf{x}$ and the ε_i's are independent and identically distributed errors with zero mean and variance σ^2. In what follows, the elements of $\mathbf{s}$ represent the order of the partial derivative corresponding to each element of $\mathbf{x}$, so for $\mathbf{s} = (s_1, s_2, \ldots, s_r)$, $g^{(\mathbf{s})}(\mathbf{x}) = [\partial^{s_1} g(\mathbf{x}) \cdots \partial^{s_r} g(\mathbf{x})]/[\partial x_1^{s_1} \cdots \partial x_r^{s_r}]$.

Given an estimate $\hat{g}(\mathbf{x})$, suppose one wishes constraints on $\hat{g}(\mathbf{x})$ of the form

$$l(\mathbf{x}) \leq \hat{g}^{(\mathbf{s})}(\mathbf{x}) \leq u(\mathbf{x}) \tag{2.1}$$

for arbitrary $l(\cdot)$, $u(\cdot)$, and $\mathbf{s}$. For some applications, $\mathbf{s} = (0, \ldots, 0, 1, 0, \ldots, 0)$ would be of particular interest, say for example when the partial derivative represents a budget share and therefore must lie in $[0, 1]$; $\mathbf{s} = (0, 0, \ldots, 0)$ might be of interest when an outcome must be bounded. Additional constraints that could be imposed in this framework are (log-) supermodularity, (log-) convexity, and quasiconvexity, all of which focus on second order or cross-partial derivatives. Variously, equality rather than inequality constraints might be required.

The two-sided constraint (2.1) can be considered as a special case of multiple simultaneous one-sided constraints. Hence for general purposes, we consider restrictions of the form

$$\sum_{\mathbf{s}\in\mathbf{S}_k} \alpha_{\mathbf{s},k}\hat{g}^{(\mathbf{s})}(\mathbf{x}) - c_k(\mathbf{x}) \geq 0, \quad k = 1,\ldots,T, \tag{2.2}$$

where T is the number of restrictions and, in each restriction, the sum is taken over all vectors in $\mathbf{S}_k$ that correspond to the constraints, with $\alpha_{\mathbf{s},k}$ a set of constants used to generate them. Note that (2.2) could be further generalized to contain more sophisticated constraints such as global concavity/convexity or homogeneity of degree R (Euler's theorem) by allowing the $\alpha_{\mathbf{s},k}$ to be functions of the covariates. See Section 4.2 for one such generalization. In what follows we presume, without loss of generality, that for all $\mathbf{s}$, $\alpha_{\mathbf{s},k} \geq 0$ and $c_k(\mathbf{x}) \equiv 0$, since the $c_k(\mathbf{x})$'s are known functions. The approach we describe is quite general; it admits arbitrary combinations of constraints subject to the obvious caveat that the constraints must be internally consistent.

Standard kernel regression smoothers can be written as linear combinations of the response Y_i,

$$\hat{g}(\mathbf{x}) = \sum_{i=1}^{n} A_i(\mathbf{x})Y_i, \tag{2.3}$$

where $A_i(\mathbf{x})$ is a local weighting matrix. This includes the Nadaraya-Watson estimator (Nadaraya (1965),Watson (1964)), the Priestley-Chao estimator (Priestley and Chao (1972)), the Gasser-Müller estimator (Gasser and Müller (1979)), and the local polynomial estimator (Fan (1992)), among others. Following Hall and Huang (2001), we consider a generalization of (2.3) to

$$\hat{g}(\mathbf{x}|p) = \sum_{i=1}^{n} p_i A_i(\mathbf{x})Y_i, \tag{2.4}$$

and where $\hat{g}^{(\mathbf{s})}(\mathbf{x}|p) = \sum_{i=1}^{n} p_i A_i^{(\mathbf{s})}(\mathbf{x})Y_i$.

As an example, we use (2.4) to generate an *unrestricted* Nadaraya-Watson estimator. Here we take $p_i = 1/n$, $i = 1,\ldots,n$, and set $A_i(\mathbf{x}) = nK_h(\mathbf{X}_i,\mathbf{x})/\sum_{j=1}^{n} K_h(\mathbf{X}_j,\mathbf{x})$, where $K_h(\cdot)$ is a product kernel and h is a vector of bandwidths; see Racine and Li (2004) for details. When $p_i \neq 1/n$ for some i, we have a *restricted* Nadaraya-Watson estimator; the selection of p satisfying particular restrictions is discussed below.

Let p_u be the n-vector of uniform weights and let p be the vector of weights to be selected. To impose our constraints, we choose p to minimize some distance measure from p to p_u as proposed by Hall and Huang (2001).

Whereas Hall and Huang (2001) consider probability weights and distance measures suitable for probability weights (e.g., Hellinger), we allow for both positive and negative weights while retaining $\sum_i p_i = 1$, and so require alternative distance measures.

We also forgo the power divergence metric of Cressie and Read (1984) that was used by Hall and Huang (2001) since it is only valid for probability weights. Instead we use the l_2 metric $D(p) = (p_u - p)'(p_u - p)$ that has a number of appealing features in this context, as will be seen. Our problem then is to select weights p that minimize $D(p)$ subject to $l(\mathbf{x}) \leq \hat{g}^{(\mathbf{s})}(\mathbf{x}|p) \leq u(\mathbf{x})$, and perhaps additional constraints of a similar form; this can be cast as a general nonlinear programming problem. Theoretical underpinnings of the constrained estimator are provided in Theorems 1 and 2 in Section 3. The explicit form of the quadratic programming problem is presented right before Theorem 3 in Section 3, where such form is needed for the theoretical development. Section 4 describes setup and implementation details for two (common) types of constraints: coordinate-wise monotonicity/concavity constraints and global concavity.

2.2. Hypothesis testing for shape constraints

In this section, we propose a test for the validity of arbitrary shape constraints using $D(\hat{p})$ as the test statistic. It is a bootstrap testing procedure that is simple to implement and extends the monotonicity testing procedure in Hall et al. (2001) to our multivariate and multiple constraints setting. Theoretical underpinnings are provided in Theorem 3 in Section 3.

This bootstrap approach involves estimating the constrained regression function $\hat{g}(\mathbf{x}|p)$ based on the sample realizations $\{Y_i, \mathbf{X}_i\}$ and then rejecting H_0 if the observed value of $D(\hat{p})$ is too large. We use a resampling approach for generating the null distribution of $D(\hat{p})$ which involves generating resamples for y drawn from the constrained model via *iid* residual resampling (i.e., conditional on the sample $\{\mathbf{X}_i\}$), which we denote $\{Y_i^*, \mathbf{X}_i\}$ (for non-*iid* data the dependent wild bootstrap residual resampling scheme could be used instead; see Shao (2010)). These resamples are generated under H_0, hence we recompute $\hat{g}(\mathbf{x}|p)$ for the bootstrap sample $\{Y_i^*, \mathbf{X}_i\}$, denoted $\hat{g}(\mathbf{x}|p^*)$, which then yields $D(p^*)$. We repeat this process B times. Finally, we compute the empirical P value, P_B, the proportion of the B bootstrap resamples $D(p^*)$ that exceed $D(\hat{p})$,

$$P_B = 1 - \hat{F}(D(\hat{p})) = \frac{1}{B}\sum_{j=1}^{B} I(D(p^*) > D(\hat{p})),$$

where $I(\cdot)$ is the indicator function and $\hat{F}(D(\hat{p}))$ is the empirical distribution function of the bootstrap statistics; one rejects the null hypothesis if P_B is less than α, the level of the test.

We note three situations here that can occur in practice.

(i) Impose non-binding constraints (they are 'correct' de facto).

(ii) Impose binding constraints that are correct.

(iii) Impose binding constraints that are incorrect.

If one encounters (i) in practice, $D(\hat{p}) = 0$, we recommend following the advice of Hall et al. (2001, p.609): "For those datasets with $D(\hat{p}) = 0$, no further bootstrapping is necessary [...] and so the conclusion (for that dataset) must be to not reject H_0."

3. Theoretical Properties of the Estimator and Test Statistic

In the following subsections we consider the theoretical properties of the estimator and test procedure. We denote the domain of interest by $\mathcal{J} \equiv [\mathbf{a}, \mathbf{b}] = \prod_{i=1}^{r}[a_i, b_i]$. And to simplify the notation, we define a differential operator $m \mapsto m^{\mathcal{D}}$ such that $m^{\mathcal{D}}(\mathbf{x})$ is a length-T vector with kth entry $\sum_{\mathbf{s}\in\mathbf{S}_k} \alpha_{\mathbf{s},k} m^{(\mathbf{s})}(\mathbf{x})$. We take $|\mathbf{s}| = \sum_{i=1}^{r} s_i$ as the *order* for a derivative vector $\mathbf{s} = (s_1, \ldots, s_r)$, and say a derivative $\mathbf{s}_1$ has a *higher order* than $\mathbf{s}_2$ if $|\mathbf{s}_1| > |\mathbf{s}_2|$. Let $\mathbf{S} = \cup_{k=1}^{T}\mathbf{S}_k$ and $\mathbf{d}_{\mathbf{S}}$ be the derivative of the "maximum order" among all the derivatives in $\mathbf{S}$; for simplicity, we drop the subscript $\mathbf{S}$ from $\mathbf{d}_{\mathbf{S}}$. With $c_k(\mathbf{x})$'s set to 0, we plug (2.4) into (2.2) to yield

$$\sum_{i=1}^{n} p_i A_i^{\mathcal{D}}(\mathbf{x}) Y_i \geq 0. \tag{3.1}$$

3.1. Existence of the constrained estimator

The theorem here shows the existence of a set of weights that satisfy the constraints in (3.1). Its proof is in Section S1 of the online Supplementary Material.

Theorem 1. *Assume that the set $\{1, \ldots, n\}$ contains a sequence $\{i_1, \ldots, i_k\}$ with the following properties.*

(i) *For each k, $A_{i_k}^{\mathcal{D}}(\mathbf{x})$ is strictly positive and continuous on an open set $\mathbf{O}_{i_k} \subset \mathbb{R}^r$, and vanishes on $\mathbb{R}^r \setminus \mathbf{O}_{i_k}$,*

(ii) *Every $\mathbf{x} \in \mathcal{J}$ is contained in at least one open set $\mathbf{O}_{i_k}$,*

(iii) *For $1 \leq i \leq n$, $A_{i_k}^{\mathcal{D}}(\mathbf{x})$ is continuous on $(-\infty, \infty)^r$.*

Then there exists a vector $p = (p_1, \ldots, p_n)$ such that the constraints are satisfied for all $\mathbf{x} \in \mathcal{J}$.

Conditions (i) and (ii) are to ensure the existence of an open cover of the domain $\mathcal{J}$ by the open sets $\mathbf{O}_i$ on which $A_i^{\mathcal{D}}$ is positively supported for some i. We note that the above conditions are sufficient but not necessary for the existence of a set of weights that satisfy the constraints for all $\mathbf{x} \in \mathcal{J}$. For example, if $\operatorname{sgn} A_{j_n}^{\mathcal{D}}(\mathbf{x}) = 1\ \forall \mathbf{x} \in \mathcal{J}$ for some sequence j_n in $\{1, \ldots, n\}$ and $\operatorname{sgn} A_{l_n}^{\mathcal{D}}(\mathbf{x}) = -1$ $\forall \mathbf{x} \in \mathcal{J}$ for another sequence l_n in $\{1, \ldots, n\}$, then for those observations that switch signs, p_i may be set equal to zero, while $p_{j_n} > 0$ and $p_{l_n} < 0$ is sufficient to ensure existence of a set of p's satisfying the constraints.

3.2. Consistency of the constrained estimator

Here we detail the consistency of our constrained estimator. To begin, define a *hyperplane subset* of $\mathcal{J}$ to be a subset of the form $\mathcal{S} = \left\{ x_{0k} \times \prod_{i \neq k} [a_i, b_i] \right\}$ for some $1 \leq k \leq r$ and some $x_{0k} \in [a_k, b_k]$. We call $\mathcal{S}$ an *interior hyperplane subset* if $x_{0k} \in (a_k, b_k)$. For what follows, $g(\cdot)$ (or $g^{\mathcal{D}}(\cdot)$) is the true conditional mean (or its derivative), $\hat{p}$ is the optimal weight vector satisfying the constraints, $\hat{g}(\cdot|\hat{p})$ (or $\hat{g}^{\mathcal{D}}(\cdot|\hat{p})$) is the constrained estimator defined in (2.4), and $\tilde{g}(\cdot)$ (or $\tilde{g}^{\mathcal{D}}(\cdot)$) is the unconstrained estimator defined in (2.3).

Assumption A1.

(i) *The sample* $\mathbf{X}_i$ *either form a regularly spaced grid on a compact set* $\mathcal{I} \equiv [\mathbf{c}, \mathbf{e}] = \prod_{i=1}^{r} [c_i, e_i]$ *or constitute independent random draws from a distribution whose density* f *is continuous and nonvanishing on* $\mathcal{I}$*; the* ε_i *are independent and identically distributed with zero mean and variance* σ^2*, and are independent of the* $\mathbf{X}_i$*; the kernel function* $K(\cdot)$ *is a symmetric, compactly supported density such that* $K^{\mathcal{D}}$ *is Hölder-continuous on* $\mathcal{J} \subset \mathcal{I}$*.*

(ii) $E\left(|\varepsilon_i|^t\right)$ *is bounded for sufficiently large* $t > 0$*.*

(iii) $g^{\mathcal{D}}$ *is continuous on* $\mathcal{J}$*. For random* X_i*'s, their density function* f *is also continuous on* $\mathcal{J}$*.*

(iv) *The bandwidth associated with each explanatory variable,* h_j*, satisfies* $h_j \propto n^{-1/(3r+2|\mathbf{d}|)}$*,* $1 \leq j \leq r$*, where* $|\mathbf{d}|$ *is the maximum order of the derivative vector* $\mathbf{d}$*.*

Assumption A1(i) is standard in the kernel regression literature. Assumption A1(ii) is a sufficient condition required for the application of a strong approximation result that we invoke in Lemma S2.2 in the online Supplementary Material, while Assumption A1(iii) assures requisite smoothness of $f^{\mathcal{D}}$ and $g^{\mathcal{D}}$ (f is the design density). Note that the bandwidth rate in Assumption A1(iv) is generally higher than the standard optimal rate $n^{-1/(r+4)}$. However, this is not surprising for our restricted problem. The optimal rate only guarantees the convergence of our unrestricted function estimator $\tilde{g}$. But the restricted problem also requires the convergence of the derivative $\tilde{g}^{\mathcal{D}}$, which often needs a higher

bandwidth rate. In the single-predictor monotone regression problem considered in Hall and Huang (2001), this rate happens to coincide with the optimal rate $n^{-1/5}$. Furthermore, when the bandwidths all share the same rate, one can rescale each component of $\mathbf{x}$ to ensure a uniform bandwidth $h \propto n^{-1/(3r+2|\mathbf{d}|)}$ for all components. This simplification is made without loss of generality. Thus we use h^r rather than $\prod_{j=1}^r h_j$ for notational simplicity.

Theorem 2. *Suppose that Assumption* A1(i)−(iv) *holds.*

(i) *If $g^{\mathcal{D}} > 0$ on $\mathcal{J}$ then, with probability 1, $\hat{p} = 1/n$ for all sufficiently large n and $\hat{g}^{\mathcal{D}}(\cdot|\hat{p}) = \tilde{g}^{\mathcal{D}}$ on $\mathcal{J}$ for all sufficiently large n. Hence, $\hat{g}(\cdot|\hat{p}) = \tilde{g}$ on $\mathcal{J}$ for all sufficiently large n.*

(ii) *Suppose that $g^{\mathcal{D}} > 0$ except on an interior hyperplane subset $\mathcal{X}_0 \subset \mathcal{J}$ where we have $g^{\mathcal{D}}(\mathbf{x}_0) = 0, \forall \mathbf{x}_0 \in \mathcal{X}_0$. Also, for any $\mathbf{x}_0 \in \mathcal{X}_0$, suppose that $g^{\mathcal{D}}$ has second order continuous derivatives in the neighborhood of $\mathbf{x}_0$ with $\frac{\partial g^{\mathcal{D}}}{\partial \mathbf{x}}(\mathbf{x}_0) = \mathbf{0}$ and $\frac{\partial^2 g^{\mathcal{D}}}{\partial \mathbf{x} \partial \mathbf{x}^T}(\mathbf{x}_0)$ nonsingular; then $|\hat{g}(\cdot|\hat{p}) - \tilde{g}| = O_p\left(h^{|\mathbf{d}|+(r+1)/2}\right)$ uniformly on $\mathcal{J}$.*

(iii) *Under the conditions in* (ii), *there exist random variables $\Theta = \Theta(n)$ and $Z_1 = Z_1(n) \geq 0$ satisfying $\Theta = O_p\left(h^{|\mathbf{d}|+r+1}\right)$ and $Z_1 = O_p(1)$, such that $\hat{g}(\mathbf{x}|\hat{p}) = (1+\Theta)\tilde{g}(\mathbf{x})$ uniformly for $\mathbf{x} \in \mathcal{J}$ with $\inf_{\mathbf{x}_0 \in \mathcal{X}_0} |\mathbf{x} - \mathbf{x}_0| > Z_1 h^{(r+1)/4}$.*

In Theorem 2, part (i) suggests that when the constraint is strictly satisfied by the true function, the constrained estimator $\hat{g}(\cdot|\hat{p})$ and the unconstrained estimator $\tilde{g}$ are essentially the same and thus share the same rate of convergence. Part (ii) gives the order of difference between $\hat{g}(\cdot|\hat{p})$ and $\tilde{g}$ when $g^{\mathcal{D}} = 0$ on an interior hyperplane. Note that the order in (ii) indicates a different convergence rate of $\hat{g}(\cdot|\hat{p})$ from that of $\tilde{g}$ in such a case. Part (iii) is concerned with the asymptotic behavior of the weights $\hat{p}$ in such a case. Also note that the results are easily extendable to the case of $g^{\mathcal{D}} \leq 0$ with a switch of sign in g.

The proof of Theorem 2 is relegated to the online Supplementary Material. Theorem 2 is the multivariate, multi-constraint, hyperplane subset generalization of the univariate, single constraint, single point violation setting considered in Hall and Huang (2001) having dispensed with probability weights and power divergence distance measures of necessity. The theory in Hall and Huang (2001) lays the foundation for the multivariate analogues in Theorem 2 and (iii) of the rates in their univariate setting (Hall and Huang (2001, Thm. 4.3(c))).

A further issue with the imposition of constraints is the choice of the distance metric used to select the optimal weights. Hall and Huang (2001) use the power divergence metric, $D_\rho(p) = \rho^{-1}(1-\rho)^{-1}\{n - \sum_{i=1}^n (np_i)^\rho\}$ that depends on a parameter ρ; they impose the condition that ρ lies between 0 and 1 for their technical arguments. In a sense, if $0 \leq \rho \leq 2$ then the l_1 and l_2 norms can be

viewed as limiting cases for the Hall and Huang analysis; see their equation (5.26). However, a key difference here is the relative ease with which the constraints can be implemented in practice if one forgoes power divergence and uses either a linear or quadratic program to solve for the optimal weights. Furthermore, in our proof of Theorem 3 the use of the l_2 norm delivers simplifications that are not available when using the power divergence metric. Additionally, under the condition $\sum_{i=1}^{n} p_i = 1$, $D_2(p)$ is equivalent to the l_2 norm.

3.3. Asymptotic properties of $D(\hat{p})$

Let $\psi_i(\mathbf{x}) = A_i^{\mathcal{D}}(\mathbf{x})Y_i$, $i = 1, \ldots, n$, so that our minimization problem is

$$\min_{p_i,\ldots,p_n} \sum_{i=1}^{n}(n^{-1} - p_i)^2, \quad \text{s.t.} \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i\psi_i(\mathbf{x}) \geq 0, \forall \mathbf{x}. \tag{3.2}$$

In practice, this can be carried out by taking a fine grid $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, N large, and solving

$$\min_{p_i,\ldots,p_n} \sum_{i=1}^{n}(n^{-1} - p_i)^2, \text{ s.t. } \sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i\psi_i(\mathbf{x}_j) \geq 0, 1 \leq j \leq N. \tag{3.3}$$

Assumption A2.

(i) *$N \to \infty$ as $n \to \infty$ and $N = O(n)$.*

(ii) *If $d_N = \inf_{1\leq j_1, j_2 \leq N} |\mathbf{x}_{j_1} - \mathbf{x}_{j_2}|$, then $d_N \to 0$ and $h^{-1}d_N \to \infty$.*

Assumption A2(ii) requires that the minimum distance between grid points decreases at a rate slower than h such that the correlation between derivative estimates at these grid points is zero when n is sufficiently large.

Let $\hat{p}_i, i = 1, \ldots, n$, solve (3.3). The proof of the following theorem is in Section S3 of the online Supplemental Material.

Theorem 3. *Suppose Assumptions* A1(i)−(iv) *and* A2(i)−(ii) *hold. Then, as $n \to \infty$, we have*

$$\frac{n^2\sigma_K^2}{h^{2|\mathbf{d}|+r}\left(\sum_{j=1}^{M} g^{\mathcal{D}}(\mathbf{x}_j^*)\right)^2} D(\hat{p}) \sim \chi^2(n), \tag{3.4}$$

where $\sigma_K^2 = \sigma^2 \int \left[K^{(\mathbf{d})}(y)\right]^2 dy$, and $\{\mathbf{x}_1^*, \ldots, \mathbf{x}_M^*\} \subset \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ are the slack points defined in Section S3 of the online Supplementary Material.

The diverging degrees of freedom in the asymptotic distribution here is of no surprise since both the null and alternative hypotheses are nonparametric and reside on infinite dimensional parameter spaces. A similar phenomenon was

observed by Fan, Zhang, and Zhang (2001) for their generalized likelihood ratio test. Theoretically, the asymptotic distribution of $D(\hat{p})$ in Theorem 3 can be used in determining the p-value of our test. However, this may pose difficulties in practice: the asymptotic distribution may not be a good approximation for finite sample sizes; the normalizing constant in (3.4) requires the determination of slack points. A bootstrap approach, like that proposed in Section 2.2, is an alternative.

Under the power divergence metric, Carroll, Delaigle, and Hall (2011) showed the consistency of the hypothesis test on monotonicity using $D_\rho(\hat{p})$ as the test statistic, which implies consistency of the bootstrap version. A similar result here consists of two parts:

(i) If the true function g satisfies the shape constraint, then as $n \to \infty$,

$$P\{D(\hat{p}) \le n\epsilon\} \to 1 \quad \text{for all } \epsilon > 0. \tag{3.5}$$

(ii) If the true function g does not satisfy the shape constraint on $\mathcal{J}$, then

$$\lim_{\epsilon \to 0} \liminf_{n \to \infty} P\{D(\hat{p}) \ge n\epsilon\} = 1. \tag{3.6}$$

In particular, (i) can be proved by a similar argument to that of (iii) of Theorem 2. The proof of (ii) follow the steps listed in Carroll, Delaigle, and Hall (2011): provide an almost-sure lower bound of an integrated distance between a given constraint-violating function and all the shape-constrained functions; use this to derive an almost-sure lower bound for the distance between the constrained estimator $\hat{g}(\cdot|\hat{p})$ and the unconstrained estimator $\tilde{g}$; show that the latter distance is of a lower order in probability than $D(\hat{p})$. A formal treatment would be interesting but lies beyond the scope of this paper.

Even if the true function g does not satisfy the shape constraints, our bootstrap test procedure simulates resamples based on the constrained estimator $\hat{g}(\cdot|\hat{p})$ that does satisfy them. For each resample with its constrained estimator $\hat{g}^*(\cdot|p^*)$ we can show, similar to (3.5), that for all $\epsilon > 0$, $P\{D(p^*) \le n\epsilon|\text{Data}\} \to 1$ in probability. This implies that the empirical critical point $\hat{\xi}_\alpha$ used in the test satisfies $P(\hat{\xi}_\alpha \le n\epsilon) \to 1$. Combining this with (3.6) yields the validity of our bootstrap test procedure, that is, $P\{D(\hat{p}) > \hat{\xi}_\alpha\} \to 1$ as $n \to \infty$.

4. Two Illustrations

In this section, we give two examples of the implementation of the quadratic programming outlined above for our method. The first example incorporates monotonicity or/and concavity in each dimension. The second example on global concavity takes up a generalized version of the constraints in (2.2) that is more

challenging numerically. The constraints are enforced on the sample realizations, but it is straightforward to also enforce them on non-sample regions if desired.

4.1. Coordinate-wise monotonicity and concavity

To impose coordinate-wise monotonicity and concavity at each $\mathbf{x}_i$, we use, respectively, the constraints

$$\sum_{j=1}^{n} p_j A_j^{(\mathbf{s})}(\mathbf{x}_i) Y_j \geq 0, \text{ for } \mathbf{s} \in \mathbf{S}_1 \tag{4.1}$$

$$\sum_{j=1}^{n} p_j A_j^{(\mathbf{s})}(\mathbf{x}_i) Y_j \leq 0, \text{ for } \mathbf{s} \in \mathbf{S}_2, \tag{4.2}$$

where $\mathbf{S}_1 = \{(1,0,\ldots,0),(0,1,\ldots,0),\ldots,(0,0,\ldots,1)\}$ and $\mathbf{S}_2 = \{(2,0,\ldots,0),(0,2,\ldots,0),\ldots,(0,0,\ldots,2)\}$. To enforce either (4.1) or (4.2) at all data points consists of rn conditions in total, r the dimension of the covariates. Even with thousands of observations, the constraints are easy to construct and implement via quadratic programming procedures.

4.2. Global concavity

A popular, if challenging, shape constraint is global concavity in a multiple-dimension setting. In this section we highlight some results from nonlinear programming that can be utilized in such an implementation.

While there are a variety of ways to enforce concavity (convexity), we require that the constraints be linear in p. And so using the Hessian matrix does not fit into our framework. Instead, we use the Afriat condition (Afriat (1967)) that states that a function g is (globally) concave if and only if

$$g(\mathbf{z}) - g(\mathbf{x}) \leq \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})'(\mathbf{z}-\mathbf{x}), \forall \mathbf{z}, \mathbf{x}.$$

To impose this first-order condition at a given point $\mathbf{x}_i$, our constraint set is

$$\sum_{j=1}^{n} p_j \left[\frac{\partial A_j}{\partial \mathbf{x}}(\mathbf{x}_i)'(\mathbf{x}_\ell - \mathbf{x}_i) - \{A_j(\mathbf{x}_\ell) - A_j(\mathbf{x}_i)\} \right] Y_j \geq 0, \ \forall 1 \leq i \neq \ell \leq n. \tag{4.3}$$

Hence, enforcing concavity at all points results in $n(n-1)$ overall constraints. Note that the number of constraints does not depend on the dimension r of the covariates, indicating that the program scales well with respect to r. However, the computational burden can be overwhelming when n is large.

To address such computational issues, generally, we briefly describe the *constraint generation* approach (Dantzig, Fulkerson, and Johnson (1954, 1959)) that

can be used to construct and enforce constraints more efficiently. Rather than imposing concavity at all sample realizations, impose concavity on some sizable subset of observations, and then check which observations do not satisfy global concavity. Call this set $\mathbf{V}$. Take observations from $\mathbf{V}$ and add them to the original set of observations where concavity is enforced and re-solve the quadratic program. Repeat the procedure until there is a subset of observations large enough that imposing concavity on these points is sufficient to ensure concavity at all points. The approach is widespread and an excellent example is in Lee et al. (2012). A large literature on constraint complexity bounds for linear and quadratic programming problems exists, see Potra and Wright (2000) for an in-depth review.

For illustrative constraints we consider that the necessary (in)equalities are linear in p, which can be solved using standard quadratic programming methods and off-the-shelf software, using the `quadprog` package in R, for example.

5. Numerical Properties and an Application

We demonstrate the flexibility and simplicity of the proposed method through a series of numerical studies, and provide a data set on which one imposes the economic constraint known as 'constant returns to scale'.

5.1. Visualization of bivariate estimates

In what follows we simulate data from two nonlinear bivariate relationships and consider imposing a range of restrictions. We demonstrate the method by imposing restrictions on the surface and also on its first and second partial derivatives using the locally constant kernel estimator.

5.1.1. Visualization of bivariate estimates

Consider the bivariate surface

$$Y_i = \frac{\sin\left(\sqrt{X_{1i}^2 + X_{2i}^2}\right)}{\sqrt{X_{1i}^2 + X_{2i}^2}} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{5.1}$$

where x_1 and x_2 are independent drew from the uniform [-5,5]. We draw $n = 10,000$ observations from this DGP with $\varepsilon \sim N(0, \sigma^2)$ and $\sigma = 0.1$. The large sample size speaks to the feasibility of the approach in moderate/large sample settings. Our simulations with sample sizes in the hundreds resulted in reasonable although slightly rougher estimates. Figure 1 shows the unrestricted regression estimate with bandwidths chosen via least squares cross-validation.

We first imposed the constraint that the regression function lies in the range $[0, 0.5]$. A plot of the restricted surface appears in Figure 2. We next imposed

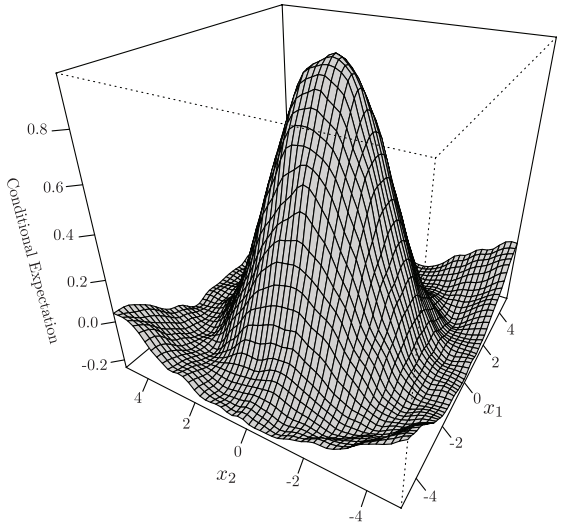

Figure 1. Unrestricted kernel estimate of (5.1).

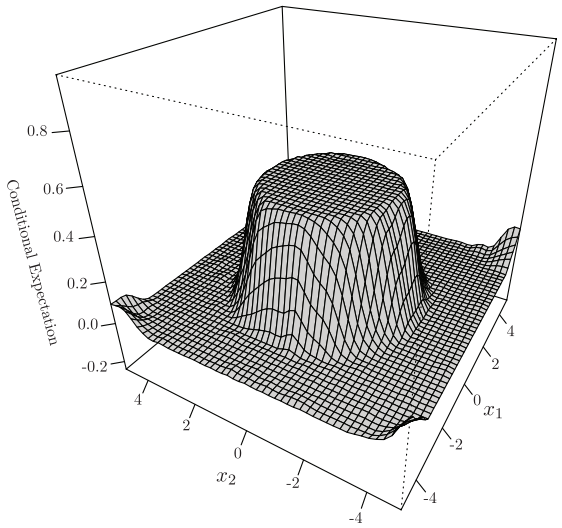

Figure 2. Restricted kernel estimate of (5.1) with the restriction $0 \leq \hat{g}(\mathbf{x}|p) \leq 0.5$.

the constraint that the first derivatives with respect to both x_1 and x_2 lie in the range [-0.1,0.1]. A plot of the restricted surface appears in Figure 3.

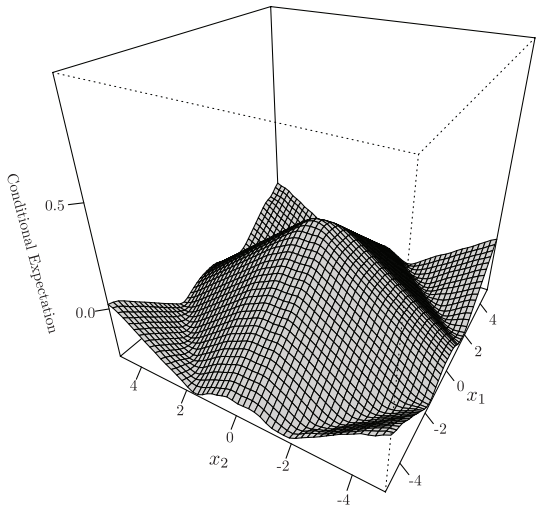

Figure 3. Restricted kernel estimate of (5.1) with the restrictions $-0.1 \leq \partial\hat{g}(\mathbf{x}|p)/\partial x_1 \leq 0.1$, $-0.1 \leq \partial\hat{g}(\mathbf{x}|p)/\partial x_2 \leq 0.1$.

5.1.2. Imposing true constraints on a shape-constrained function

Consider then the bivariate surface

$$Y_i = (X_{1i}X_{2i})^{0.4} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{5.2}$$

where x_1 and x_2 are independent draws from the uniform [1,10]. Figure 4 shows the surface. We drew $n = 100$ observations with $\varepsilon \sim N(0, \sigma^2)$ and $\sigma = 0.7$. This DGP is positive, monotonic in both x_1 and x_2, globally concave, symmetric in x_1 and x_2, and homogeneous of degree 0.8. Any or all of these constraints could be imposed on the estimated surface. Here we imposed negativity of the second derivatives of both x_1 and x_2 on a grid of 250 points equally spaced over the support of X.

Figure 5 presents the unrestricted local constant kernel regression estimate with bandwidths chosen via least squares cross-validation. Figure 6. presents the restricted local constant kernel regression estimates. Here the bumps that were present in the unrestricted estimator have been removed by the enforcement of the constraints.

5.2. Inference with the constrained estimator

5.2.1. Testing Inequality Restrictions

Consider testing the inequality restriction $H_0 : g(\mathbf{x}) \geq 0$ versus $H_1 : g(\mathbf{x}) < 0$ when

$$Y_i = g(X_i) + \varepsilon_i = X_i + \varepsilon_i, \tag{5.3}$$

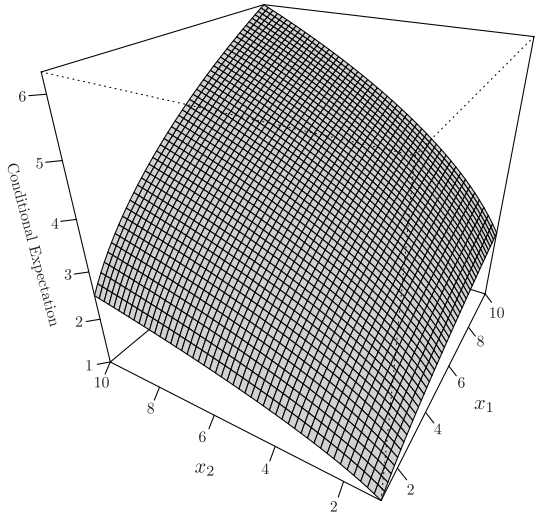

Figure 4. Actual unknown DGP in (5.2).

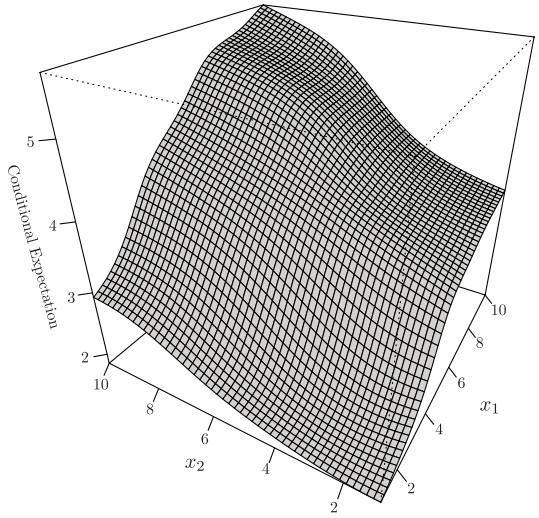

Figure 5. Unrestricted kernel estimate of (5.2).

where X_i is uniform $[-a, 4-a]$, a is a parameter that determines whether the constraint is binding ($a > 0$) or not ($a = 0$) and the length of interval over which the constraint binds, and $\varepsilon \sim N(0, 1/4)$.

We constructed power curves based on $M = 1,000$ Monte Carlo replications, and we computed $B = 99$ bootstrap replications. The power curves corresponding to $\alpha = 0.05$ appear in Figure 7. This reveals that the empirical rejection frequencies are in line with nominal size while power increases with n.

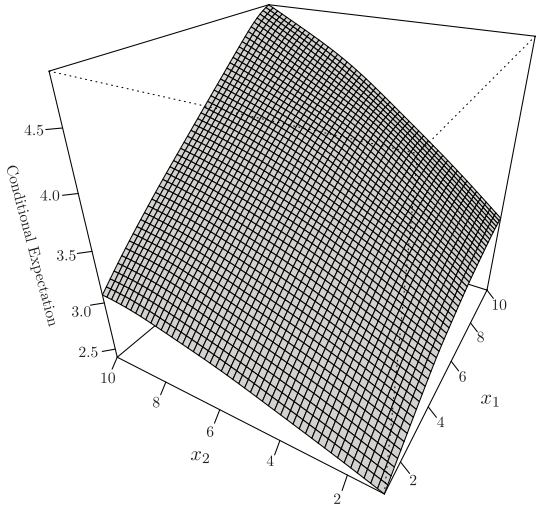

Figure 6. Restricted kernel estimate of (5.2) with the restrictions $\partial^2 \hat{g}(\mathbf{x}|p)/\partial x_1^2 \leq 0$, $\partial^2 \hat{g}(\mathbf{x}|p)/\partial x_2^2 \leq 0$.

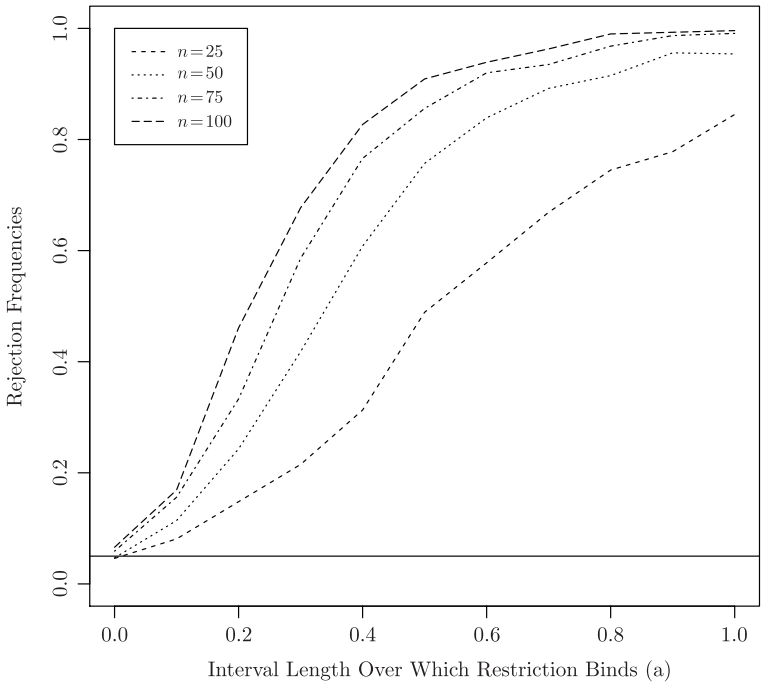

Figure 7. Power curves for $\alpha = 0.05$ for sample sizes $n = (25, 50, 75, 100)$ based upon the DGP given in (5.3). The solid horizontal line represents the test's nominal level (α).

It is known (Andrews (2000)) that standard bootstrap inference procedures may not be consistent when inequality constraints are involved, Galindo-Garre and Vermunt (2004). For example, one approach for dealing with size distortions in such instances is the 'double-bootstrap' (van de Schoot, Hoijtink, and Deković (2010)). Carroll, Delaigle, and Hall (2011) use calibration, not a double-bootstrap, and attain excellent finite sample properties for their bootstrap inequality test, while Andrews (2000) suggests a number of alternatives including subsampling (Politis and Romano (1994)), among others. Though there are no discernible size distortions present in the power curve summarized in Figure 7, it is prudent to verify results based on the simple bootstrap with these alternatives.

5.2.2. Testing equality restrictions

We consider testing the restriction that a nonparametric model $g(\mathbf{x})$ has a specific parametric functional form. Starting with

$$Y_i = g(X_{i1}, X_{i2}) + \varepsilon_i = 1 + X_{i1}^2 + X_{i2} + \varepsilon_i,$$

where the X_{ij}, $j = 1, 2$, are uniform $[-2, 2]$ and $\varepsilon \sim N(0, 1/2)$.

When we generated data from this DGP and imposed the correct model as a restriction we could assess the test's rejection frequencies under H_0, and when we generated data from this DGP and imposed an incorrect model that is in fact linear in variables we could assess the test's power.

We conducted $M = 1,000$ Monte Carlo replications from our DGP, and took $B = 99$ bootstrap replications. Results are in Table 1 in the form of empirical rejection frequencies for nominal size $\alpha = (0.10, 0.05, 0.01)$, for samples of size $n = (25, 50, 75, 100, 200)$. Table 1 indicates that the tests' empirical rejection frequency appears to be in line with nominal size while power increases with n.

5.3. An empirical assessment of implementation issues

We conducted simulations to gauge several implementation issues for the constrained estimator. We focused on how long it takes the quadratic program to solve the problem based on a fixed grid of points, how the method performs across different bandwidths, and how the method performs across alternative nonparametric methods. For all simulations we took $n = (100, 200, 300, 400, 500)$ observations drawn from the DGP in Lee et al. (2012),

$$Y_i = (X_{i1}X_{i2}X_{i3}X_{i4})^{0.2} + \varepsilon_i, \tag{5.4}$$

where $X_{\cdot j} \sim \mathcal{U}[1, 10]$ for $1 \leq j \leq 4$ and ε is normal with mean zero and variance 0.49. For the simulations we focused on enforcing coordinate-wise monotonicity.

Table 1. Test for correct parametric functional form. Values are empirical rejection frequencies over the $M = 1,000$ Monte Carlo replications.

n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
		Size	
25	0.100	0.049	0.010
50	0.074	0.043	0.011
75	0.086	0.034	0.008
100	0.069	0.031	0.006
200	0.093	0.044	0.007
		Power	
25	0.391	0.246	0.112
50	0.820	0.665	0.356
75	0.887	0.802	0.590
100	0.923	0.849	0.669
200	0.987	0.970	0.903

Table 2. Performance of imposing constraints on a rough grid in a 4-dimension space. The numbers are the median percentages of the observations that ended up satisfying the constraints and the median execution times for the quadratic program solver in seconds.

n	% satisfied	Time (secs)
100	99.00	0.42
200	99.50	0.64
300	99.42	0.93
400	99.38	1.40
500	99.28	1.97

5.3.1. Enforcing constraints over a grid versus sample realizations

Table 2 gives the percentage of observations for which the constraints were violated when we enforced the constraints over an equi-spaced grid of 5 points on $[1, 10]$ in each dimension (a total of $5^4 = 625$ grid points). We give median run times and the median percentage of observations where the constraints were satisfied, using the optimal weights determined by the grid points. In all the simulations the percentage of observations satisfying monotonicity was over 99%, suggesting that imposing constraints on a rough grid we can still achieve monotonicity almost everywhere.

5.3.2. An assessment of bandwidth selection

We used least squares cross-validation to determine optimal bandwidths, then used these bandwidths, cross-validated bandwidths divided by 2, and cross-validated bandwidths multiplied by 2,d and enforced the constraints. We compared performance of the estimators based on the median ratio of average squared error (ASE) taken at the observations.

Table 3. Bandwidth selection performance. The values are median ratios of ASEs for the estimators with cross-validated smoothing (ASE_2), undersmoothing (ASE_3), and oversmoothing (ASE_4) (ASE_1 is unrestricted with cross-validated smoothing). Numbers greater than one indicate superior performance of the restricted cross-validated estimator.

n	ASE_1/ASE_2	ASE_3/ASE_2	ASE_4/ASE_2
100	1.30	2.23	2.34
200	1.28	2.44	2.99
300	1.31	2.58	3.32
400	1.31	2.65	3.57
500	1.29	2.76	3.58

Table 4. Comparison of kernel versus B-spline and unrestricted versus restricted estimation. The numbers are the median ratios of the ASEs: unrestricted (ASE_1) and restricted (ASE_3) local constant kernel methods; unrestricted (ASE_2) and restricted (ASE_4) B-splines.

n	ASE_1/ASE_3	ASE_2/ASE_4	ASE_3/ASE_4
100	1.30	1.11	1.28
200	1.28	1.00	1.18
300	1.31	1.03	1.23
400	1.31	1.01	1.25
500	1.29	1.02	1.34

As seen in Table 3, the constrained estimator with cross-validated bandwidths outperforms both the undersmoothed and the oversmoothed constrained estimator. We also see that generically imposing valid constraints results in improved in-sample fit of the unknown function.

5.3.3. Implementation with alternative smoothers

While the theory is provided for kernel estimators, the estimation procedure constrained by (2.2) is applicable to any (local) linear smoother. Here we compare implementations involving local constant kernel regression and involving B-splines (both with cross-validated smoothing parameter selection), and compare their ASEs. Table 4 gives the median ratios of the ASEs, comparing the restricted and the unrestricted estimators for each method against the constrained estimators enforcing monotonicity in each dimension. As expected, the restricted methods outperformed the unrestricted methods. Moreover, it appears that the unrestricted regression B-splines satisfy the constraints more often as the sample size increases (note that $ASE_2/ASE_4 \approx 1$). We also see that the regression B-splines outperformed the local kernel methods.

Table 5. Summary Statistics for the Data

Variable	Mean	StdDev
log(rice)	6.9170	0.9144
log(seed)	2.4534	0.9295
log(urea)	4.0144	1.1039
log(TSP)	2.7470	1.4093
log(labor)	5.6835	0.8588
log(land)	-1.1490	0.9073

5.4. Application: Imposing constant returns to scale for indonesian rice farmers

We consider a data set studied by Horrace and Schmidt (2000) who analyzed technical efficiency for Indonesian rice farms. We examine the issue of returns to scale, focusing on one growing season's worth of data, in 1977, acknowledged to be a particularly wet season. 171 farmers were selected from six villages in the rice production area of the Cimanuk River Basin in West Java by the Center for Agro Economic Research, Ministry of Agriculture, Indonesia. Output was measured as kilograms (kg) of rice produced, with inputs of seed (kg), urea (kg), trisodium phosphate (TSP) (kg), labour (hours), and land (hectares). Table 5 presents several summary statistics for the data. We use log transformations throughout.

Of interest here is whether or not the production technology exhibits constant returns to scale, whether or not the sum of the first order partial derivatives is one. Constant returns to scale implies that output increases by exactly the amount that all the inputs are increased if all the inputs are doubled, then output doubles. Given the primitive nature of the production of rice, one expects the existence of constant returns to scale to be present in the underlying technology. The constraint set for imposing returns to scale can be written as

$$\sum_{k=1}^{5} \frac{\partial \hat{g}(\mathbf{x}_i)}{\partial x_{ik}} x_{ik} = 1, \quad \forall i. \tag{5.5}$$

We impose this constraint for each observation, as opposed to over a grid. This results in $n = 171$ total constraints.

We estimate the production function using a nonparametric local linear estimator with least squares cross-validated bandwidth selection. Figure 8 presents the unrestricted and restricted partial derivative sums for each observation (i.e., farm), where the restriction is that the sum of the partial derivatives equals one. The horizontal line represents the restricted partial derivative sum (1.00) and the points represent the unrestricted sums for each farm. An examination of Figure 8 reveals that the estimated returns to scale lie in the interval $[0.98, 1.045]$.

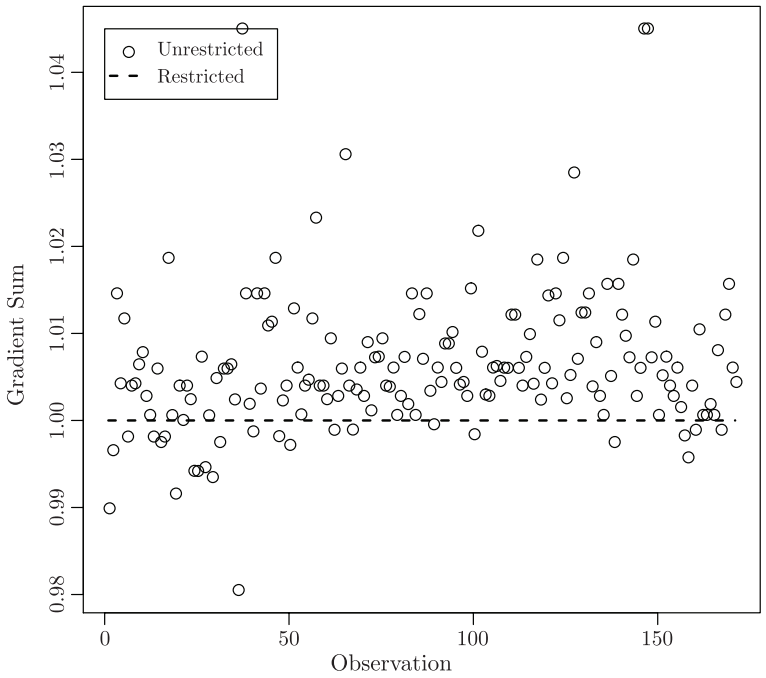

Figure 8. The sum of the partial derivatives for observation i appear on the vertical axis, and each observation appears on the horizontal axis.

In order to test whether the restriction is valid we apply the test outlined in Section 2.2. We conducted $B = 999$ bootstrap replications and tested the null that the technology exhibits constant returns to scale. The empirical P value is $P_B = 0.122$, hence we fail to reject the null at all conventional levels. We are encouraged by this nonparametric application as it involves a fairly large number of predictors (five) and a fairly small number of observations ($n = 171$).

6. Discussion

We present a framework for imposing and testing the validity of conventional constraints on the partial derivatives of a multivariate nonparametric kernel regression function. The proposed approach covers imposing monotonicity and concavity while delivering a seamless framework for general restricted nonparametric kernel estimation and inference. Simulations are run and the method is applied to a data set. An open implementation in the R language (R Core Team (2012)) is available from the authors.

We note that our procedure is valid for a range of kernel estimators as well as for estimation and testing in the presence of categorical data. Our constrained smoothing approach can be used in a wide variety of settings. Future work on the

theoretical side could focus on the importance of the choice of distance metric, the asymptotic behavior of the bootstrap testing procedure, and the relative merits of the alternative data tilting methods that exists. These are subjects for future research.

Acknowledgements

We thank an associate editor and the referees for their insightful comments that have significantly improved the paper. We would also like to thank but not implicate Daniel Wikström for inspiring conversations and Li-Shan Huang and Peter Hall for their insightful comments and suggestions. The research of Du is supported by NSF DMS-1007126. Racine would like to gratefully acknowledge support from Natural Sciences and Engineering Research Council of Canada (www.nserc.ca), the Social Sciences and Humanities Research Council of Canada (www.sshrc.ca), and the Shared Hierarchical Academic Research Computing Network (www.sharcnet.ca).

References

Afriat, S. N. (1967). The construction of utility functions from expenditure data. *Internat. Econom. Rev.* **8**, 67-77.

Andrews, D. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68**, 399-405.

Birke, M. and Dette, H. (2007). Estimating a convex function in nonparametric regression. *Scand. J. Statist.* **34**, 384-404.

Birke, M. and Pilz, K. F. (2009). Nonparametric option pricing with no-arbitrage constraints. *J. Finan. Econom.* **7**, 53-76.

Braun, W. J. and Hall, P. (2001). Data sharpening for nonparametric inference subject to constraints. *J. Comput. Graph. Statist.* **10**, 786-806.

Carroll, R. J., Delaigle, A. and Hall, P. (2011). Testing and estimating shape-constrained nonparametric density and regression in the presence of measurement error. *J. Amer. Statist. Assoc.* **106**, 191-202.

Cressie, N. A. C. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46**, 440-464.

Csörgő, M. and Révész, P. (1981). *Strong Approximations in Probability and Statistics.* Academic Press, New York.

Dantzig, G. B., Fulkerson, D. R. and Johnson, S. M. (1954). Solution of a large-scale traveling-salesman problem. *Operations Research* **2**, 393-410.

Dantzig, G. B., Fulkerson, D. R. and Johnson, S. M. (1959). On a linear-programming combinatorial approach to the traveling-salesman problem. *Operations Research* **7**, 58-66.

Dette, H. and Scheder, R. (2006). Strictly monotone and smooth nonparametric regression for two or more variables. *Canad. J. Statist.* **34**, 535-561.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87**, 998-1004.

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29**, 153-193.

Galindo-Garre, F. and Vermunt, J. K. (2004). The order-restricted association model: two estimation algorithms and issues in testing. *Psychometrika* **69**, 641-654.

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

Gallant, A. R. (1982). Unbiased determination of production technologies. *J. Econometrics* **20**, 285-323.

Gallant, A. R. and Golub, G. H. (1984). Imposing curvature restrictions on flexible functional forms. *J. Econometrics* **26**, 295-321.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, 23-68. Springer-Verlag, New York.

Gasser, T. and Müller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11**, 171-185.

Ghosal, S., Sen, A. and van der Vaart, A. W. (2000). Testing monotonicity of regression. *Ann. Statist.* **28**, 1054-1082.

Hall, P. and Heckman, N. E. (2000). Testing monotonicity of a regression mean by calibrating for linear functionals. *Ann. Statist.* **28**, 20-39.

Hall, P. and Huang, H. (2001). Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.* **29**, 624-647.

Hall, P., Huang, H., Gifford, J. and Gijbels, I. (2001). Nonparametric estimation of hazard rate under the constraint of monotonicity. *J. Comput. Graph. Statist.* **10**, 592-614.

Hall, P. and Kang, K. H. (2005). Unimodal kernel density estimation by data sharpening. *Statist. Sinica* **15**, 73-98.

Horrace, W. and Schmidt, P. (2000). Multiple comparisons with the best, with economic applications. *J. Appl. Econometrics* **15**, 1-26.

Komlós, J., Major, P. and Tusnády, G. (1975). An approximation of partial sums of independent random variables and the sample distribution function, part I. *Z. Wahrsch. Verw. Gebiete* **32**, 111-131.

Lee, C.-Y., Johnson, A. L., Moreno-Ceteno, E. and Kuosmanen, T. (2012). A more efficient algorithm for convex nonparametric least squares. Texas A&M University Technical Report.

Mammen, E. (1991). Estimating a smooth monotone regression function. *Ann. Statist.* **19**, 724-740.

Mammen, E., Marron, J. S., Turlach, B. A., and Wand, M. P. (2001). A general projection framework for constrained smoothing. *Statist. Sci.* **16**, 232-248.

Mammen, E. and Thomas-Agnam, C. (1999). Smoothing splines and shape restrictions. *Scand. J. Statist.* **26**, 239-252.

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.

Mukerjee, H. (1988). Monotone nonparametric regression. *Ann. Statist.* **16**, 741-750.

Nadaraya, E. A. (1965). On nonparametric estimates of density functions and regression curves. *Theory Probab. Appl.* **10**, 186-190.

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

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22**, 2031-2050.

Potra, F. A. and Wright, S. J. (2000). Interior-point methods. *J. Comput. Appl. Math.* **124**, 281-302.

Priestley, M. B. and Chao, M. T. (1972). Nonparametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34**, 385-392.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

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

Ramsay, J. O. (1988). Monotone regression splines in action (with comments). *Statist. Sci.* **3**, 425-461.

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

Shao, X. (2010). The dependent wild bootstrap. *J. Amer. Statist. Assoc.* **105**, 218-235.

van de Schoot, R., Hoijtink, H. and Deković, M. (2010). Testing inequality constrained hypotheses in SEM models. *Structural Equation Modeling* **17**, 443-463.

Villalobos, M. and Wahba, G. (1987). Inequality-constrained multivariate smoothing splines with application to the estimation of posterior probabilities. *J. Amer. Statist. Assoc.* **82**, 239-248.

Wang, X. and Shen, J. (2010). A class of grouped Brunk estimators and penalized spline estimators for monotone regression. *Biometrika* **97**, 585-601.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā* **26:15**, 175-184.

Wright, I. W. and Wegman, E. J. (1980). Isotonic, convex and related splines. *Ann. Statist.* **8**, 1023-1035.

Xu, K.-L. and Phillips, P. C. B. (2012). Tilted nonparametric estimation of volatility functions with empirical applications. *J. Bus. Econom. Statist.* **29**, 518-528.

Yatchew, A. and Härdle, W. (2006). Nonparametric state price density estimation using constrained least squares and the bootstrap. *J. Econometrics* **133**, 579-599.

Department of Statistics, Virginia Tech, Blacksburg, VA 24061, U.S.A.

E-mail: pangdu@vt.edu

Department of Economics, University of Miami, Coral Gables, FL 33124, U.S.A.

E-mail: cparmeter@bus.miami.edu

Department of Economics, Department of Mathematics and Statistics (Graduate Program in Statistics), McMaster University, Hamilton, Ontario L8S 4M4, Canada.

E-mail: racinej@mcmaster.ca

(Received January 2012; accepted September 2012)