# NONPARAMETRIC ESTIMATION UNDER SHAPE RESTRICTIONS

by

Joel L. Horowitz
Northwestern University

and

Sokbae Lee
Seoul National University

# INTRODUCTION

- This talk is about nonparametric estimation of the unknown function $g$ in the model

$$Y = g(X) + \varepsilon; \quad E(\varepsilon \mid X) = 0.$$

- $X$ is continuously distributed.
- $g$ is assumed to satisfy a shape restriction
- Examples of shape restrictions:
  - Monotonicity or convexity
  - The Slutsky restriction of consumer theory

# FORM OF THE SHAPE RESTRICTION

- Write the restriction as $(Ag)(x) \leq 0$, where $A$ is an operator.
- $A$ is a linear operator if the shape restriction is monotonicity or convexity.
- $A$ is a nonlinear operator if the shape restriction is the Slutsky condition.
- There is a large statistics literature on nonparametric estimation under monotonicity or convexity but not under a nonlinear shape restriction.

# THE ESTIMATION PROBLEM

- Let $\{Y_i, X_i : i = 1, \ldots, n\}$ be an independent random sample of $(Y, X)$.

- In principle, $g$ can be estimated by solving the problem

$$\hat{g}(x) = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} [Y_i - f(x)]^2$$

subject to:

$$(Af)(x) \leq 0 \quad \text{for all } x \in \operatorname{supp}(X).$$

  - $\mathcal{F}$ is a class of estimators, such as Nadaraya-Watson, local linear, or series estimators.

# WHY INFERENCE IS DIFFICULT

- There are two sources of difficulty
  - The estimation problem has uncountably many constraints
  - The values of $x$ for which the shape constraint is or is not binding are unknown.
    - The asymptotic distribution of $\hat{g}$ depends on where the constraint binds.
- The first problem could be dealt with easily if it were known where the shape constraint is binding.
  - Therefore, the second problem is more fundamental.

# OUTLINE OF APPROACH

- To minimize the complexity, assume $X$ is scalar and its density is bounded away from 0 on its support [0,1].
  - Extension to a multi-dimensional $X$ with compact support involves mainly notational complications.
- Let $x_j = j/(J+1);\ \ j = 1,...,J$ be a grid of $J$ points in [0,1].
  - Let $J \to \infty$ at a suitable rate as $n \to \infty$.
  - Define $g_j = g(x_j)$
- We estimate and impose the shape restriction only on the $g_j$'s, thereby obtaining finitely many constraints.

# EXAMPLES OF SHAPE RESTRICTIONS ON THE GRID

- Monotonicity

$$g_j - g_{j+1} \leq 0; \quad j = 1, \ldots, J-1$$

- A Slutsky-like nonlinear constraint (two-dimensional covariate: $X, Z$)

$$[g(x_{j+1}, z_k) - g(x_j, z_k)]$$

$$+ g(x_j, z_k)[g(x_j, z_{k+1}) - g(x_j, z_k)] \leq 0$$

# REPRESENTING CONSTRAINTS ON THE GRID

- Define the vector $\boldsymbol{g} = (g_1, ..., g_J)'$.

- Write the shape restrictions on the grid as

$$A_k(\boldsymbol{g}) \leq 0; \quad k = 1, ..., K,$$

where $A_k$ is a real-valued function $\mathbb{R}^J$ and there are $K$ constraints.

- Example: monotonicity

$$A_k(\boldsymbol{g}) = g_k - g_{k+1}; \; k = 1, ..., J-1 = K$$

# ESTIMATION ON THE GRID

- With $\boldsymbol{g} = [g(x_1), \ldots, g(x_J)]'$, the estimation problem is

$$\hat{\boldsymbol{g}} = \arg \min_{f_1, \ldots, f_J} \sum_{i=1}^{n} \left[ Y_i - \sum_{j=1}^{J} f_j I\left( \frac{|X_i - x_j|}{h} \leq 1 \right) \right]^2$$

subject to:

$$A_k(f_1, \ldots f_J) \leq 0; \quad k = 1, \ldots, K$$

- If the bandwidth, $h$, is less than half the distance between grid points, the unconstrained estimator of $g_j$ is the average of the $Y_i$'s for which $X_i$ is within $h$ of $x_j$.

# INFERENCE

- Define $\mathcal{S} = \{k: \ A_k(\boldsymbol{g}) = 0\}$.
- The estimation problem is equivalent to:

$$\hat{\boldsymbol{g}} = \arg \min_{f_1,\ldots,f_J} \sum_{i=1}^{n} \left[ Y_i - \sum_{j=1}^{J} f_j I \left( \frac{|X_i - x_j|}{h} \le 1 \right) \right]^2$$

subject to:

$$A_k(f_1,\ldots f_J) = 0; \quad k \in \mathcal{S}$$

- Inference about $\boldsymbol{g}$ is infeasible because $\mathcal{S}$ is unknown.

# FEASIBLE INFERENCE

- Let $\hat{\mathcal{S}}$ be a known random set such that:
  - $P(\mathcal{S} \subset \hat{\mathcal{S}}) = 1$ for all sufficiently large $n$.
  - If $A_k(\boldsymbol{g}) \neq 0$, then $P(k \in \hat{\mathcal{S}}) \to 0$.

- Define the "substitute" estimation problem

$$\hat{\boldsymbol{g}} = \arg \min_{f_1,\ldots,f_J} \sum_{i=1}^{n} \left[ Y_i - \sum_{j=1}^{J} f_j I\left( \frac{|X_i - x_j|}{h} \leq 1 \right) \right]^2$$

subject to:

$$A_k(f_1,\ldots f_J) = 0; \quad k \in \hat{\mathcal{S}}$$

# FEASIBLE INFERENCE (cont.)

- In the substitute estimation problem:
  - $\hat{\boldsymbol{g}}$ is consistent for $\boldsymbol{g}$.
  - The asymptotic distribution of suitably scaled $\hat{\boldsymbol{g}} - \boldsymbol{g}$ can be derived because $\hat{\mathcal{S}}$ is known.
- But $\hat{\boldsymbol{g}} - \boldsymbol{g}$ has an asymptotic bias arising from the fact that $\hat{\mathcal{S}} \neq \mathcal{S}$
  - This is in addition to the usual asymptotic bias of nonparametric estimation.

# PROBLEMS TO BE SOLVED TO DO INFERENCE ABOUT $g$

- How to choose $\hat{\mathcal{S}}$
- Removing asymptotic biases from $\hat{g} - g$.
- Constructing a uniform confidence band for $g$ that satisfies the shape constraints.
- Remainder of talk describes methods for solving these problems.
  - I assume that $|\mathcal{S}|$ is bounded away from 0 as $n \to \infty$.
  - I do not discuss empirical choice of tuning parameters

# METHOD FOR A LINEAR SHAPE RESTRICTION

- A linear shape restriction is simpler and provides more intuition than a nonlinear one.
- The extension to a nonlinear restriction is done later.
- A linear shape restriction has the form $A\boldsymbol{g} \le 0$, where $A$ is a $K \times J$ matrix.
- Assume there are at most $J_0$ values of $j$ for which $A_{kj} \ne 0$ and $J_0$ values of $k$ for which $A_{kj} \ne 0$ (e.g., $J_0 = 2$ for monotonicity and $J_0 = 3$ for convexity).

# FINDING $\hat{\mathcal{S}}$

- The unconstrained estimator of $\boldsymbol{g}$ is the Nadaraya-Watson estimator with a uniform kernel:

  - Define

$$w_{ij} = \frac{I\left(|X_i - x_j| \le h_1\right)}{\sum_{i=1}^{n} I\left(|X_i - x_j| \le h_1\right)}; \;\; i = 1,\ldots,n; \;\; j = 1,\ldots,J$$

  - Then the unconstrained estimator of $g_j$ is

$$\tilde{g}_j = \sum_{i=1}^{n} w_{ij} Y_i = g_j + \sum_{i=1}^{n} w_{ij}[g(X_i) - g_j] + \sum_{i=1}^{n} w_{ij}\varepsilon_i$$

# FINDING $\hat{\mathcal{S}}$ (2)

- Let $h_1$ undersmooth, so $(nh_1)^{1/2}(\tilde{g}_j - g_j)$ has no asymptotic bias.

- Then

$$\tilde{g}_j - g_j = \sum_{i=1}^{n} w_{ij}\varepsilon_i + o_p[(nh_1)^{-1/2}]$$

- If $h_1 \propto n^{-b}$ for some $b \geq 1/5$, and $J \propto n^a$ for some $a < b$, then probability 1 for all sufficiently large $n$,

$$\max_{1 \leq j \leq J} \left| [A(\tilde{g}_j - g_j) \right| \leq const \left( \frac{\log n}{nh_1} \right)^{1/2}.$$

# FINDING $\hat{\mathcal{S}}$ (3)

- This implies that for all $j$ such that $(A\boldsymbol{g})_j = 0$

$$|(A\tilde{\boldsymbol{g}})_j| \le const\left(\frac{\log n}{nh_1}\right)^{1/2}.$$

- Set

$$\hat{\mathcal{S}} = \left\{ j : |(A\tilde{\boldsymbol{g}})_j| \le const\left(\frac{\log n}{nh_1}\right)^{1/2} \right\}$$

- Then $P(\mathcal{S} \subset \hat{\mathcal{S}}) = 1$ for all sufficiently large $n$.
- Methods for choosing the constant are not yet developed.

# INFERENCE ABOUT $\boldsymbol{g}$

- The estimator of $\boldsymbol{g}$ is

$$\hat{\boldsymbol{g}} = \arg \min_{f_1,\ldots,f_J} \sum_{i=1}^{n} \left[ Y_i - \sum_{j=1}^{J} f_j I\left( \frac{|X_i - x_j|}{h} \le 1 \right) \right]^2$$

subject to:

$$\sum_{j=1}^{J} A_{kj} f_j = 0; \quad k \in \hat{\mathcal{S}}$$

- Consider only $g_j$'s with $A_{kj} \neq 0$ for some $k \in \hat{\mathcal{S}}$.
    - The constraints are not binding for the other $g_j$'s.

# ANALYTIC EXPRESSION FOR $\hat{\boldsymbol{g}}$

- Redefine $\boldsymbol{g}$ and $\hat{\boldsymbol{g}}$ to be vectors consisting only of $g_j$'s that are affected by the constraints.

- With linear constraints, the estimation problem is constrained least squares and has an analytic solution.

# ANALYTIC SOLUTION

- Define $A$ as constraint matrix $[A_{kj}]$, $p_X$ = prob. dens. of $X$,

$$Q_n = diag[p_X(x_1),\ldots,p_X(x_J)],$$

$$R_{n1} = [I - Q_n^{-1}A'(AQ_n^{-1}A')^{-1}A]Q_n^{-1}$$

$$R_{n2} = Q_n^{-1}A'(AQ_n^{-1}A')^{-1}$$

$$B_{ij} = I(|X_i - x_j| \le h)$$

- Define $\beta$ as asymptotic bias of kernel estimator. Then

$$(nh)^{1/2}(\hat{\boldsymbol{g}} - \boldsymbol{g}) = R_{n1}(nh)^{-1/2}B'\varepsilon$$
$$+ R_{n1}(nh)^{-1/2}\beta - R_{n2}(nh)^{-1/2}A\boldsymbol{g}$$

# INTERPRETATION

$$(nh)^{1/2}(\hat{\boldsymbol{g}} - \boldsymbol{g}) = R_{n1}(nh)^{-1/2}B'\varepsilon + R_{n1}(nh)^{-1/2}\beta - R_{n2}(nh)^{1/2}A\boldsymbol{g}$$

- First term on right-hand side is asymptotically multivariate normal with mean 0 uniformly over $j \le J$.
- Second term is usual asymptotic bias of nonparametric estimation.
- Third term is additional bias because not all components of $A\boldsymbol{g}$ are zero.

# POINTWISE CONFIDENCE INTERVALS

- The two bias terms are asymptotically negligible under smoothness conditions and with suitable choices of $J$ and $h$
- Therefore, pointwise confidence intervals for the $g_j$'s are

$$\gamma_{j1} \leq (nh)^{1/2}(\hat{g}_j - g_j) \leq \gamma_{n2}$$

- The critical values can be obtained from the asymptotic normal distribution of $[R_{n1}(nh)^{-1/2}B'\varepsilon]_j$.
- Asymptotically, there is no selectivity bias from the randomness of $\hat{\mathcal{S}}$.

# UNIFORM CONFIDENCE BAND

- Define $J$-vector $Z \sim N(0, \Sigma_n)$, where $\Sigma_n$ is the covariance matrix of the asymptotic distribution of $R_{n1}(nh)^{-1/2} B'\varepsilon$.
- Asymptotic normality is uniform over $j \leq J$.
- Therefore, a uniform $(1-\alpha)$ confidence band that satisfies the shape restrictions can be obtained by solving

$$\underset{\gamma_1,\ldots,\gamma_j}{\text{minimize}} : \left\{ P\left[ \bigcap_{j=1}^{J} (|Z_j| \leq \gamma_j) \right] - (1-\alpha) \right\}^2$$

$$\text{subject to: } \sum_{j=1}^{J} A_{kj}\gamma_j \leq 0; \qquad k = 1,\ldots,K$$

# NONLINEAR CONSTRAINTS

- The estimation problem is

$$\hat{\boldsymbol{g}} = \arg \min_{f_1,\ldots,f_J} \sum_{i=1}^{n} \left[ Y_i - \sum_{j=1}^{J} f_j I\left( \frac{|X_i - x_j|}{h} \le 1 \right) \right]^2$$

subject to:

$$A_k(f_1,\ldots f_J) \le 0; \quad k = 1,\ldots,K$$

- Approach is same as with linear constraints
  - Find set $\hat{\mathcal{S}}$ such that $P(\mathcal{S} \subset \hat{\mathcal{S}}) \to 1$ and $P(k \subset \hat{\mathcal{S}}) \to 0$ if $A_k(\boldsymbol{g}) \ne 0$
    - Solve estimation problem with $A_k(\boldsymbol{f}) = 0; k \in \hat{\mathcal{S}}$

# FINDING $\hat{\mathcal{S}}$

- Arguments like those made for linear constraints give

$$|[A(\tilde{\boldsymbol{g}})]_j| \le const\left(\frac{\log n}{nh_1}\right)^{1/2}$$

- With a nonlinear constraint, the estimation problem does not have an analytic solution.

- If the $A_k$'s are twice differentiable, the theory of sensitivity analysis in nonlinear programming (Fiacco 1983) gives an approximation to the solution whose error is $o_p[(nh)^{-1/2}]$.

# CONFIDENCE BAND WITH NONLINEAR CONSTRAINTS

- Define $A_g$ as the $K \times J$ matrix whose $(k, j)$ is

$$(A_g)_{jk} = \frac{\partial A_k(\boldsymbol{g})}{\partial g_j}.$$

- Define $R_{n1}$ and $R_{n2}$ as before but with $A_g$ in place of $A$.

- Then

$$(nh)^{1/2}(\hat{\boldsymbol{g}} - \boldsymbol{g}) = R_{n1}(nh)^{-1/2} B'\varepsilon$$
$$+ R_{n1}(nh)^{-1/2}\beta - R_{n2}(nh)^{1/2} A(\boldsymbol{g})$$

# CONFIDENCE BAND WITH NONLINEAR CONSTRAINTS (cont.)

- $(nh)^{1/2}(\hat{g}_j - g_j)$ with nonlinear constraints has the same form as in the linear case.

- Confidence bands for the nonlinear case can be obtained using the same methods as in the linear case.
  - In particular, as with linear constraints, the two bias terms are asymptotically negligible and the random term is asymptotically multivariate normal uniformly over $j \le J$.

# CONFIDENCE BANDS FOR $g(x)$ AT POINTS BETWEEN GRID POINTS

- Let $x_j \le x \le x_{j+1}$.

- Obtain $\hat{g}(x)$ and confidence limits $\gamma(x)$ by interpolation. If $x = \lambda x_j + (1-\lambda)x_{j+1}$, set

  - $\hat{g}(x) = \lambda \hat{g}_j + (1-\lambda)\hat{g}_{j+1}$

  - $\gamma(x) = \lambda \gamma_j + (1-\lambda)\gamma_{j+1}$

- Then

$$P\left[(nh)^{1/2} \,|\, \hat{g}(x) - g(x) \,|\, \le \gamma(x) \ \ \forall x\right] \to 1-\alpha.$$

# ADDITIONAL PROPERTIES OF INTERPOLATION ESTIMATOR

- $\hat{g}(x)$ and $\gamma(x)$ satisfy the shape restrictions if $A_k(\cdot)$ is a convex function for every $k = 1, ..., K$.

- If $(A_k g)(\cdot)$ is non-convex, then

  - $(A_k \hat{g})(x) \leq 0$, and $(A_k \gamma)(x) \leq 0$ if $(A_k g)(x) \leq -r_n$, where $(nh)^{1/2} r_n \geq \varepsilon$ for some $\varepsilon > 0$.

  - $(A_k \hat{g})(x) \leq o_p[(nh)^{-1/2}]$ and $(A_k \hat{\gamma}) \leq o_p[(nh)^{-1/2}]$ if $(A_k g)(x) = o[(nh)^{-1/2}]$.

# CONCLUSIONS

- This talk has been about nonparametric inference of a conditional mean function under a possibly nonlinear shape restriction
  - The main result is a uniform confidence band for the unknown, shape-restricted conditional mean function.
- Remaining work includes:
  - Treating a shape restriction that holds only at one point.
  - Extension to multidimensional $X$
  - Methods for choosing tuning parameters.
  - Numerical studies.