NONPARAMETRIC ESTIMATION OF AN ADDITIVE MODEL WITH A LINK FUNCTION

by

Joel L. Horowitz
Northwestern University
Evanston, IL
USA

INTRODUCTION

- Problem: Estimate $H(x) = E(Y|X = x)$ under weak assumptions about its functional form when X is a continuous random variable

- Fully nonparametric estimation is unattractive when X is multidimensional because of the curse of dimensionality.

- Dimension reduction methods reduce effective dimension of estimation problem and mitigate or eliminate curse of dimensionality

    - They make assumptions about the form of $H(x)$ that are stronger than those of a fully nonparametric model but weaker than those of a parametric model

DIMENSION REDUCTION METHODS

- Semiparametric single-index model
- Additive model with known link function

$$H(x) = F\left[\mu + \sum_{j=1}^{d} m_j(x^j)\right],$$

where F is known, and μ and m_j's are unknown.

- Partially linear model with known link function (Robinson 1988, Golubev and Härdle 1997, Severini and Staniswalis 1994)

$$E(Y \mid X = x, W = w) = G[\beta' x + f_w(w)],$$

where G is known but β and f_w are not.

- Additive model with unknown link function

$$H(x) = F\left[\sum_{j=1}^{d} m_j(x^j)\right],$$

where F and the m_j's are unknown.

PURPOSE OF THIS PAPER

Paper is concerned with estimating nonparametric additive model with known link function.

- Marginal integration estimator (Linton and Härdle 1996) has curse-of-dimensionality
  - Smoothness of the m_j's must increase as dimension of X increases to achieve $n^{-2/5}$ rate of convergence of nonparametric estimator of the m_j's.
- If F is identity function, this problem can be overcome by use of backfitting
  - Methods for achieving $n^{-2/5}$ rate of convergence with no curse of dimensionality not available with non-identity F.
- This paper develops method for avoiding curse of dimensionality in estimating nonparametric additive model with known link function.
  - Estimator is pointwise $n^{2/5}$-consistent and asymptotically normal when F and the m_j's are twice differentiable, regardless of dimension of X.

MARGINAL INTEGRATION ESTIMATOR
(Linton and Härdle 1996)

- Define $G = F^{-1}$ and $H(x) = \boldsymbol{E}(Y \mid X = x)$.

- Linton and Härdle (1996) write model in form

  $$G[H(x^1,...,x^d)] = \mu + m_1(x^1) + ... + m_d(x^d),$$

  where $G = F^{-1}$ and $\boldsymbol{E}[m_j(X^j)] = 0$.

- Therefore

  $$\mu + m_1(x^1) = \boldsymbol{E}G[H(x^1, X^2,..., X^d)].$$

- Estimate $m_1(x^1)$ up to additive constant by replacing H with kernel estimator and $\boldsymbol{E}$ with sample average.

- This creates curse-of-dimensionality effect because a d-dimensional nonparametric regression is needed to estimate H.

  - More smoothness needed as d increases to insure bias and variance of full-dimensional estimator are sufficiently small.

SOLUTION TO PROBLEM

- Avoid curse of dimensionality by replacing kernel estimator with estimator that does not require full-dimensional nonparametric regression.

- Nonparametric series approximation can be used to impose additive structure from outset, thereby avoiding need for full-dimensional estimation.

- Getting pointwise rates of convergence and asymptotic normality with series estimator is difficult

- Use two-step procedure to obtain estimator with tractable asymptotics:

  - Step 1: Use nonparametric series estimation to obtain pilot estimates $\tilde{\mu}, \tilde{m}_1, ..., \tilde{m}_d$

  - Step 2: Take one Newton step from pilot estimates toward local constant or local linear least squares estimator of (say) m_1

  - Second-stage estimator has structure of kernel estimator, so its asymptotic distribution can be obtained easily.

FURTHER MOTIVATION

- If μ and $m_2,...,m_d$ were known, could estimate $m_1(x^1)$ by (say) local nonlinear least squares:

$$\hat{m}_1(x^1) = \arg\min_{m_1} \sum_{i=1}^{n} \{Y_i - F[\mu + m_1(x^1)$$

$$+ m_2(X_i^2) + ... + m_d(X_i^d)]\}^2 K_h(x^1 - X_i^1)$$

where $K_h(x^1 - X_i^1) = K[(x^1 - X_i^1)/h]$, K is kernel.

- Replace unknown μ and $m_2,...,m_d$ with pilot estimates to get kernel-like estimator of $m_1(x^1)$.
  - Undersmooth pilot estimates to reduce bias
  - Resulting $\hat{m}_1(x^1)$ is asymptotically equivalent to estimator that would be obtained if μ and $m_2,...,m_d$ were known.
  - So there is (asymptotically) no penalty for not knowing μ and $m_2,...,m_d$ and no curse of dimensionality.

AVOIDING NONLINEAR OPTIMIZATION

- Nonparametric series estimation yields estimate $\tilde{m}_1$ of m_1.

- Avoid nonlinear optimization by taking one Newton step from pilot estimate toward solution of local least squares problem.

  - Resulting estimator is asymptotically equivalent to solution of full nonlinear optimization.

- Define $\tilde{m}_{-1}(x_{-1}) = \tilde{m}_2(x^2) + \ldots + \tilde{m}_d(x^d)$,

  $S_{n1}(x^1, \tilde{m}) =$

  $$\sum_{i=1}^{n} \{Y_i - F[\tilde{\mu} + \tilde{m}_1(x^1) + \tilde{m}_{-1}(\tilde{X}_i)]\}^2 K_h(x^1 - X_i^1)$$

  $S_{n1}{}'(x^1, \tilde{m}), S_{n1}{}''(x^1, \tilde{m})$ are first and second derivatives of S_{n1} with respect to $\tilde{m}_1$

SECOND-STAGE ESTIMATOR

- Second-stage estimator is

$$\hat{m}_1(x^1) = \tilde{m}_1(x^1) - S'_{n1}(x^1, \tilde{m}) / S''_{n1}(x^1, \tilde{m}).$$

NONPARAMETRIC SERIES ESTIMATOR

- Define $m(x) = m_1(x^1) + \ldots + m_d(x^d)$
- Let support of X be $[-1,1]^d$.
- Normalize m_j's by $\int_{-1}^{1} m_j(v)dv = 0 \;\; (j = 1,\ldots,d)$.
- Let $\{p_k : k = 1,2,\ldots\}$ denote basis for smooth functions on $[-1,1]$ that satisfy normalization condition and

$$\int_{-1}^{1} p_k(v)dv = 0$$

$$\int_{-1}^{1} p_j(v)p_k(v)dv = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise} \end{cases}$$

$$m_j(x^j) = \sum_{k=1}^{\infty} \theta_{jk} p_k(x^j); \;\; j = 1,\ldots,d; \;\; x^j \in [0,1]$$

- For any positive integer $\kappa > 0$ define

$$P_\kappa(x) = [1, p_1(x^1),\ldots, p_\kappa(x^1),\ldots, p_1(x^d),\ldots, p_\kappa(x^d)]'$$

- Then for $\theta_\kappa \in \mathbb{R}^{\kappa d+1}$, $P_\kappa(x)'\theta_\kappa$ is series approximation to $\mu + m(x)$.

FIRST-STEP ESTIMATOR

- Let $\{Y_i, X_i : \ i = 1,...,n\}$ be random sample of (Y, X)
- Let $\hat{\theta}_{n\kappa}$ be solution to

  $$\underset{\theta \in \Theta_\kappa}{\text{minimize}}: \ n^{-1}\sum_{i=1}^{n}\{Y_i - F[P_\kappa(X_i)'\theta]\}^2$$

  where Θ_κ is compact parameter set.

- Series estimator of $\mu + m(x)$ is

  $$\tilde{\mu} + \tilde{m}(x) = P_\kappa(x)'\hat{\theta}_{n\kappa},$$

  where $\tilde{\mu}$ is first component of $\hat{\theta}_{n\kappa}$.

- First-step estimator of $m_j(x^j)$ is product of $[p_1(x^j),...,p_\kappa(x^j)]$ with appropriate subvector of $\hat{\theta}_{n\kappa}$.

ASSUMPTIONS

- Data are random sample of (Y, X), support of X is $\mathcal{X} \equiv [-1,1]^d$, and $\boldsymbol{E}(Y \mid X = x) = F[\mu + m(x)]$.

- Density of X is bounded, bounded away from zero, and twice differentiable.

- Set $U \equiv Y - F[\mu + m(X)]$. Then:

  - $Var(U \mid X = x)$ is bounded and bounded away from zero.

  - U has finite unconditional moments of all orders

- The m_j's are bounded and twice continuously differentiable

Only two derivatives needed regardless of dimension of *X*.

- F'' satisfies Lipschitz condition

  $$| F''(v_2) - F''(v_1) | \leq C | v_2 - v_1 |^s$$

  for some $s > 5/7$.

- Conditions insuring that covariance matrix of $\hat{\theta}_{n\kappa}$'s is bounded and non-singular.

MORE ASSUMPTIONS

- Basis functions satisfy

$$\sup_{x \in \mathcal{X}} \|P_\kappa(x)\| = O(\kappa^{1/2})$$

$$\sup_{x \in \mathcal{X}} | \mu + m(x) - P_\kappa(x)'\theta_{\kappa 0} | = O(\kappa^{-2})$$

  for some $\theta_{\kappa 0} \in \Theta_\kappa$

These conditions are satisfied by spline and (for periodic functions) Fourier bases.

- Smoothing parameters satisfy:
  - $\kappa = C_\kappa n^{4/15+\nu}$ for some $\nu < 1/30$
  - $h_n = C_h n^{-1/5}$

The L_2 rate of convergence of series estimator is maximized by setting $\kappa \propto n^{1/5}$, so the series estimator here is undersmoothed to reduce asymptotic bias.

- Kernel function K of second-stage estimator is a bounded, continuous probability density function on $[-1,1]$ and is symmetrical about 0.

MAIN RESULTS: FIRST-STAGE ESTIMATOR

- Uniform consistency:

$$\sup_{x \in \mathcal{X}} | \tilde{m}(x) - m(x) | = O_p(\kappa / n^{1/2} + \kappa^{-3/2})$$

- Decomposition: Define

$$Q_\kappa = \boldsymbol{E}\{F'[\mu + m(X)]^2 P_\kappa(X) P_\kappa(X)'\}$$

Then

$$\hat{\theta}_{n\kappa} - \theta_{\kappa 0} = n^{-1} Q_\kappa^{-1} \sum_{i=1}^{n} F'[\mu + m(X_i)] P_\kappa(X_i) U_i$$

$$+ n^{-1} Q_\kappa^{-1} \sum_{i=1}^{n} F'[\mu + m(X_i)]^2 P_\kappa(X_i) b_\kappa(X_i) + R_n,$$

where $\|R_n\| = O_p(\kappa^{3/2} / n + n^{-1/2})$

MAIN RESULTS: SECOND-STAGE ESTIMATOR

- Asymptotic representation: Define

$$D(x^1) = \operatorname*{plim}_{n\to\infty} S''_{n1}(x^1, \tilde{m})$$

Then

$$(nh_n)^{1/2}[\hat{m}_1(x^1) - m_1(x^1)] =$$

$$-(nh_n)^{1/2} S'_{n1}(x^1, m) / D(x^1) + o_p(1)$$

This is representation that would be obtained by linearizing first-order condition for local least-squares estimation of m_1 with known $m_2, \ldots, m_d$.

So asymptotically there is no penalty for not knowing $m_2, \ldots, m_d$.

Structure of right-hand side is same as with kernel estimator.

RESULTS (cont.)

- Asymptotic normality

$$n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)] \to^d N[\beta_1(x^1), V_1(x^1)]$$

  This holds when the m_j's are twice continuously differentiable, regardless of dimension of X.

  So there is no curse of dimensionality.

- If $j \neq 1$, then $n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)]$ and $n^{2/5}[\hat{m}_j(x^j) - m_j(x^j)]$ are asymptotically independently normally distributed.

INTUITION FOR SECOND-STAGE RESULT

- Second-stage estimator is

$$\hat{m}_1(x^1) = \tilde{m}_1(x^1) - S'_{n1}(x^1, \tilde{m}) / S''_{n1}(x^1, \tilde{m}).$$

- This can be written:

$$(nh_n)^{1/2}[\hat{m}_1(x^1) - m_1(x^1)] =$$

$$= (nh_n)^{1/2}[\tilde{m}_1(x^1) - m_1(x^1)]$$

$$-(nh_n)^{1/2} S'_{n1}(x^1, \tilde{m}) / D(x^1) + o_p(1).$$

- Use Taylor series approximation to write

$$(nh_n)^{1/2} S'_{n1}(x^1, \tilde{m}) =$$

$$(nh_n)^{1/2} S'_{n1}(x^1, m) + T_{n1} + T_{n2} + o_p(1)$$

INTUITION (cont.)

- $T_{n1} = D(x^1)(nh_n)^{1/2}[\tilde{m}_1(x^1) - m_1(x^1)] + o_p(1)$
  - So

$$(nh_n)^{1/2}[\hat{m}_1(x^1) - m_1(x^1)] =$$

$$-(nh_n)^{1/2} S'_{n1}(x^1, m) / D(x^1) + T_{n2} + o_p(1)$$

- T_{n2} consists of
  - Bias term arising from asymptotic bias of $\tilde{m}_1$
  - Sum of mean-zero stochastic terms arising from random component of $\hat{\theta}_{n\kappa} - \theta_{\kappa 0}$

- Because first-stage estimator is undersmoothed

$$(nh_n)^{1/2}[\text{Bias Term}] = o_p(1)$$

  - Contribution of bias term to T_{n2} is asymptotically negligible.

INTUITION (cont.)

- Stochastic terms have slower than $n^{-2/5}$ rates of convergence but are averaged in T_{n2}.
  - First-stage estimator has no curse of dimensionality, so rate of convergence of variance of stochastic term does not increase with increasing dimension of X.
  - Averaged stochastic term converges faster than $n^{-2/5}$.
  - So contribution of stochastic term to T_{n2} is negligible.
- Consequently, T_{n2} is asymptotically negligible.

BANDWIDTH SELECTION

- Asymptotic integrated mean-square error of $\hat{m}_1$ is

  $$AIMSE_1 = n^{4/5}\int_{-1}^{1} w(x^1)[\beta_1(x^1)^2 + V_1(x^1)]dx^1,$$

  where w is a weight function.

- $AIMSE_1$ minimized by setting $h = C_{h1}n^{-1/5}$, where

  $$C_{h1} = \left[(1/4)\frac{\int_{-1}^{1} w(x^1)\tilde{V}_1(x^1)dx^1}{\int_{-1}^{1} w(x^1)\tilde{\beta}_1(x^1)^2 dx^1}\right]^{1/5},$$

  $\tilde{\beta}_1(x^1) = \beta_1(x^1)/C_h^2$ and $\tilde{V}_1(x^1) = C_h V_1(x^1)$.

  - Plug-in estimator of C_{h1} can be obtained by replacing $\tilde{\beta}_1$ and $\tilde{V}_1$ with kernel estimates.

- The asymptotically optimal bandwidths for all the m_j's can be estimated simultaneously by penalized least squares.

  - This minimizes empirical analog of asymptotic squared error:

MONTE CARLO EXPERIMENTS

- Compare finite-sample performance of new estimator with that of Linton and Härdle (1996)

- New estimator implemented using local constant and local linear smoothing in second stage.

- Experiments carried out with $d = 2$ and $d = 5$.

  - L-H estimator is $O_p(n^{-2/5})$ if $d = 2$, not $d = 5$.

- Sample size is $n = 500$

- With $d = 2$ estimate m_1 and m_2 in logit model

  - $\boldsymbol{P}(Y = 1 | X = x) = L[m_1(x^1) + m_2(x^2)]$

  - $L(v) = e^v / (1 + e^v)$

  - $m_1(x^1) = \sin(\pi x^1)$

  - $m_2(x^2) = \Phi(3x^2)$, where Φ is normal CDF

- With $d = 5$ estimate m_1 and m_2 in logit model

$$\boldsymbol{P}(Y = 1 | X = x) = L[m_1(x^1) + m_2(x^2) + \sum_{j=3}^{5} x^j]$$

- Components of X are independently $U[-1,1]$.

MONTE CARLO EXPERIMENTS (cont.)

- B-splines used for first-stage series estimator
- Second-order kernel used for second-stage estimator
- Tuning parameters chosen to minimize empirical integrated mean-square errors.
- 1000 replications with 2-stage estimator but only 500 with Linton-Härdle estimator

RESULTS

Estimator	Empirical IMSE f_1	Empirical IMSE f_2
$d = 2$		
FHS	.116	.015
2-Stage LC	.052	.015
2-Stage LL	.052	.023
$d = 5$		
FHS	.145	.019
2-Stage LC	.060	.018
2-Stage LL	.057	.029

- Local constant and local linear estimators both dominate Linton-Härdle for estimating f_1
- For estimating f_2 Local constant and Linton-Härdle estimators have roughly same IMSE
  - Local linear estimator is worse

CONCLUSIONS

- Paper has considered additive model with known link function

  $$\boldsymbol{E}(Y \mid X = x) = F[\mu + m^1(x^1) + ... + m_j(x^j)]$$

- Marginal integration estimator of Linton and Härdle (1996) has curse of dimensionality

- Backfitting method of Mammen *et al*. (1999) avoids curse of dimensionality if F is identity function

- This paper has proposed two-step method for avoiding curse of dimensionality with non-identity link function.

  - First step uses nonparametric series estimator that imposes additive structure

  - Second step takes a Newton step from series estimate toward a local least squares estimator.

  - Second-stage estimator has structure of kernel estimator and is pointwise asymptotically normal with $n^{-2/5}$ rate of convergence regardless of dimension of X.