NONPARAMETRIC ESTIMATION OF ADDITIVE MODELS

by

Joel L. Horowitz
Department of Economics
Northwestern University
Evanston, IL

INTRODUCTION

- Single-index model achieves dimension reduction by assuming that $\boldsymbol{E}(Y \mid X = x) = F(\beta' X)$ for some unknown F and β.
  - Can estimate β with $n^{-1/2}$ rate of convergence and F with $n^{-2/5}$ rate if it is twice differentiable.
- A nonparametric additive model is alternative way to achieve dimension reduction.
  - It has form

    $$\boldsymbol{E}(Y \mid X = x) = \mu + \sum_{j=1}^{d} m_j(x^j),$$

    where $\dim(X) = d$, x^j is j'th component of x, and μ and the m_j's are unknown.
- Additive models are non-nested with single-index models
  - A single-index model is not additive unless F is the identity function.
  - An additive model is not single-index unless the m_j's are linear.

PROPERTIES OF ADDITIVE MODELS

- Additive components m_j can be estimated with one-dimensional nonparametric rate of convergence ($n^{-2/5}$ if the components are twice differentiable)

- Asymptotically normal estimators are available

- Each component can be estimated with same accuracy that it would have if other components were known

  - This is called "oracle property."

- Three kinds of estimators are available:

  - Marginal integration yields asymptotically normal estimators but is not oracle-efficient.

  - Backfitting yields asymptotically normal, oracle efficient estimators.

  - Two-step estimator based on series-approximation first step is asymptotically normal and oracle-efficient.

MARGINAL INTEGRATION

- Model:

$$\boldsymbol{E}(Y \mid X = x) = \mu + \sum_{j=1}^{d} m_j(x^j)$$

- Need location normalization to identify the m_j's.
  - Achieve this by setting $\boldsymbol{E}[m_j(X^j)] = 0$.
- Get identifying relations
  - $\mu = \boldsymbol{E}(Y)$
  -

$m_1(x^1) =$

$$\int \boldsymbol{E}(Y \mid X^1 = x^1, X^{(-1)} = x^{(-1)}) p_{-1}(x^{(-1)}) dx^{(-1)}$$

$- \mu,$

where $X^{(-1)}$ is vector consisting of all components of X except X^1, and p_{-1} is density of $X^{(-1)}$.

ESTIMATION

- Estimate μ and m_1 by replacing population quantities with sample analogs in identifying relations

- This gives estimator of μ: $\hat{\mu} = n^{-1}\sum_{i=1}^{n} Y_i$.

- Let $\hat{g}(x^1, x^{(-1)})$ be nonparametric estimator of $\boldsymbol{E}(Y \mid X^1 = x^1, X^{(-1)} = x^{(-1)})$
  - Example: Kernel or local linear estimator

- Estimator of m_1 is

$$\hat{m}_1(x^1) = n^{-1}\sum_{i=1}^{n} \hat{g}(x^1, X_i^{(-1)}) - \hat{\mu}.$$

- Under regularity conditions:

$$n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)] \to^d N[b_1(x^1), V_1(x^1)]$$

for suitable b_1 and V_1

ASYMPTOTIC DISTRIBUTION

- If $\hat{g}$ is local-linear estimator with bandwidth $h = c_h n^{-2/5}$ and kernel K in x^1 direction and other conditions hold, then

  $$b_1(x^1) = 0.5 c_h^2 R_K m_1''(X^1)$$

  $$V_1(x^1) = c_h^{-1} \nu_K \int Var(U \mid x^1, x^{(-1)}) \frac{p_{-1}^2(x^{(-1)})}{p(x)} dx^{(-1)},$$

  where p is density of X,

  $$R_K = \int v^2 K(v) dv,$$

  $$\nu_K = \int K(v)^2 dv.$$

- In homoskedastic case

  $$V_1(x^1) = c_h^{-1} \nu_K \sigma_U^2 \int \frac{p_{-1}^2(x^{(-1)})}{p(x)} dx^{(-1)}$$

  - Oracle estimator gives $V_1(x^1) = c_h^{-1} \nu_K \sigma_U^2 / p_1(x^1)$, which is smaller.
  - Marginal integration estimator is not oracle efficient.

PROPERTIES (cont.)

- Need m_j's and p to have at least d continuous derivatives
  - So marginal integration estimator has curse of dimensionality
  - This is caused by full-dimensional non-parametric estimation in first step.
- Marginal integration estimator is hard to compute.
  - Computing $\hat{m}_1(x^1)$ requires n nonparametric regressions for each value of x^1.
- Marginal integration estimator can be modified to overcome the curse of dimensionality.

MODIFIED MI ESTIMATOR

- Write model as

$$m(x) = \mu + m_1(x^1) + m_{-1}(x^{(-1)})$$

- Let q_1 and q_{-1}, respectively, be "smooth" density functions on $\mathbb{R}$ and $\mathbb{R}^{d-1}$, respectively.
  - Define $q = q_1 q_{-1}$

- Use location normalization

$$\int m_1(x^1) q_1(x^1) dx^1 = 0$$

$$\int m_{-1}(x^{(-1)}) q_{-1}(x^{(-1)}) dx^{(-1)} = 0$$

  - This normalization makes it possible to use smoothness of q to reduce bias of estimator instead of using smoothness of m.

ESTIMATOR (cont.)

- Let h_1 and h_2 be bandwidths, and K and L be kernel functions.

- Let $\hat{p}$ be kernel estimator of density of X.

- Define

  $\tilde{m}_n(x) =$

  $$\frac{1}{nh_1h_2^{d-1}} \sum_{i=1}^{n} \frac{Y_i}{\hat{p}(X_i)} K\left(\frac{x^1 - X_i^1}{h_1}\right) L\left(\frac{x^{(-1)} - X_i^{(-1)}}{h_2}\right)$$

  This is form of kernel estimator of $\boldsymbol{E}(Y \mid X = x)$.

- Define

  $\eta_1(x^1) = \int m(x) q_{-1}(x^{(-1)}) dx^{(-1)} - \mu$.

  - Under location normalization $\eta_1 = m_1$

- Estimator of η_1 is

  $\tilde{\eta}_1(x^1) = \int \tilde{m}_n(x) q_{-1}(x^{-1}) dx^{-1}$

  $$-\int \tilde{m}_n(x) q(x) dx$$

PROPERTIES

- Hengartner and Sperlich (2005) give conditions under which

  $$n^{2/5}[\hat{\eta}_1(x^1)-\eta_1(x^1)] \to^d N[b_{\eta 1}(x^1), V_\eta(x^1)]$$

  where b_{h1} and V_η are the bias and variance functions.

- Conditions require m to be only twice differentiable, regardless of $\dim(X)$.
  - Therefore, curse of dimensionality is avoided
  - But modified estimator is not oracle efficient.

- Computation can be simplified by letting q_{-1} be Dirac δ function centered at some $x^{(-1)}$ value.
  - This gives

    $$\tilde{\eta}(x^1) = \tilde{m}_n(x^1, x^{(-1)}) - \int \tilde{m}_n(z^1, x^{(-1)}) q_1(z^1) dz^1$$

  - Asymptotic normality and rate result still holds
  - Hengartner and Sperlich do not investigate extent to which this causes loss of asymptotic efficiency

ACHIEVING ORACLE EFFICIENCY

- Oracle efficiency means: Estimator of each additive component has asymptotic distribution it would have if the other components were known
  - Asymptotically, there is no penalty for having to estimate other components.
- Marginal integration estimators are not oracle efficient but can be made so by taking one "backfitting" step.
- Main idea: Suppose $m_2,...,m_d$ and μ were known.
  - Define $W_i = Y_i - \mu - m_2(X_i^2) - ... - m_d(X_i^d)$
  - Then model is

$$W_i = m_1(X_i^1) + U_i$$

  - Can estimate m_1 by, for example, kernel or local-linear regression of W on X^1
  - Estimator is oracle efficient by definition.

ACHIEVING ORACLE EFFICIENCY (cont.)

- In applications, replace $\mu, m_2, ..., m_d$ with preliminary (possibly marginal integration) estimates $\tilde{\mu}, \tilde{m}_2, ..., \tilde{m}_d$.
  - Define $\tilde{W}_i = Y_i - \tilde{\mu} - \tilde{m}_2(X_i^2) - ... - \tilde{m}_d(X_i^d)$
  - Estimate m_1 by kernel or local-linear regression of $\tilde{W}$ on X^1
- For case $d = 2$, Linton (1997) gives conditions under which resulting estimator of m_1 is asymptotically normal with same mean and variance as estimator from regression of W_i on X_i^1.
  - Conditions include undersmoothing in estimating the $\tilde{m}_j$'s ($j = 2, ..., d$).
  - This makes the bias of preliminary estimator asymptotically negligible
  - Variance increases but is reduced by the averaging entailed in second estimation step.
- Is unknown whether oracle efficiency for $d > 2$ can be achieved by starting with Hengartner-Sperlich estimator

ACHIEVING ORACLE EFFICIENCY (cont.)

- Other methods are available for achieving oracle efficiency with $d > 2$
- Two-step estimation can be used in more general settings to achieve oracle efficiency.

BACKFITTING

- For $j = 1,...,d$, define

$$W_j = Y_i - \mu - \sum_{k \neq j} m_k(X_i^k)$$

- Write model as

$$W_j = m_j(X_i^j) + U_i$$

- Let $\hat{\mu}^0, \hat{m}_2^0,...,\hat{m}_d^0$ be preliminary estimates, and set

$$\hat{W}_1^0 = Y_i - \hat{\mu}^0 - \sum_{j=2}^{d} \hat{m}_j^0(X_i^j)$$

- Backfitting consists of:
  - Estimate m_1 by nonparametric regression of $\hat{W}_1^0$ on X^1. Let $\hat{m}_1^1$ denote resulting estimate.
  - Set $\hat{W}_2^1 = Y_i - \hat{\mu}^0 - \hat{m}_1^1(X_i^1) - \sum_{j=3}^{d} \hat{m}_j^0(X_i^j)$
  - Estimate m_2 by nonparametric regression of $\hat{W}_2^1$ on X^2. Let $\hat{m}_2^1$ denote resulting estimate.

BACKFITTING (cont.)

- Set

  $$\hat{W}_3^1 = Y_i - \hat{\mu}^0 - \hat{m}_1^1(X_i^1) - \hat{m}_2^1(X_i^2) - \sum_{j=3}^{d} \hat{m}_j^0(X_i^j)$$

- Iterate procedure to convergence, thus obtaining estimators of all additive components and μ

- This version of backfitting is hard to analyze theoretically
  - Little known about its convergence or distributional properties

- Modified versions of backfitting are easier to analyze
  - Mammen et al. (1999) have found conditions under which a suitably modified version is asymptotically normal and oracle efficient

MODIFIED BACKFITING

- Notation
  - $\breve{m}_j(x^j)$ denotes Nadaraya-Watson kernel estimator of $\boldsymbol{E}(Y \mid X^j = x^j)$.
  - $\hat{p}_j$ and $\hat{p}_{jk}$, respectively are kernel estimators of density of X^j and joint density of (X^j, X^k)
  - $\tilde{m}_j^0$ is initial guess at estimator of m_j, possibly $\breve{m}_j$ or a marginal integration estimator
  - $$\hat{p}_{k,[j+]}(x^k) = \int \hat{p}_{jk}(x^j, x^k)dx^j \left[\int \hat{p}_j(x^j)dx^j\right]^{-1}$$
  - $$\tilde{m}_{0,j} = \frac{\int \hat{m}_j(x^j)\hat{p}_j(x^j)dx^j}{\int \hat{p}_j(x^j)dx^j}$$
- Location normalization: $\boldsymbol{E}m_j(X^j) = 0$.

ITERATIVE SCHEME AND ASYMPTOTICS

- In r'th iteration, estimate of m_j is

$$\tilde{m}_j^r(x^j) = \breve{m}_j(x^j) - \tilde{m}_{0,j}$$

$$-\sum_{k<j}\int \tilde{m}_k^r(x^k)\left[\frac{\hat{p}_{jk}(x^j, x^k)}{\hat{p}_j(x^j)} - \hat{p}_{k,[j+]}(x^k)\right]dx^k$$

$$-\sum_{k>j}\int \tilde{m}_k^{[r-1]}(x^k)\left[\frac{\hat{p}_{jk}(x^j, x^k)}{\hat{p}_j(x^j)} - \hat{p}_{k,[j+]}(x^k)\right]dx^k$$

- Mammen, Linton, and Nielsen show that if the m_j's are twice continuously differentiable and some other conditions are satisfied, then
  - The iterative scheme converges to limiting estimators $\tilde{m}_j$
  - $n^{1/2}[\tilde{m}_j - m_j(x^j)]$ are asymptotically normally distributed for any finite d (no curse of dimensionality).
  - The mean and variance of the asymptotic distribution are oracle

COMMENTS ON BACKFITTING

- Modified backfitting estimator avoids curse of dimensionality and is oracle efficient but is analytically and computationally complicated

- Taking one backfitting step from Hengartner-Sperlich estimator may produce simpler oracle-efficient estimator, but this is not yet proved.

- Next lecture will present approach that uses series estimation in first step followed by a backfitting step

  - This method is simpler computationally than marginal integration or modified backfitting

  - It is oracle efficient

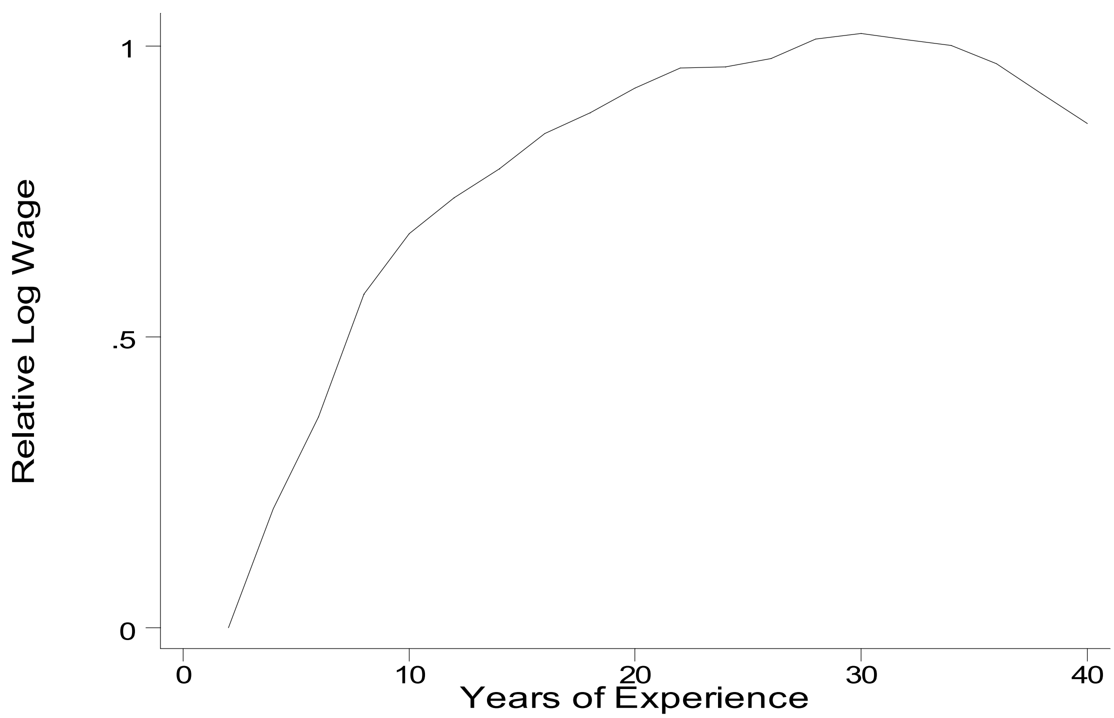
  - Can be applied to additive quantile regressions and models with link functions.

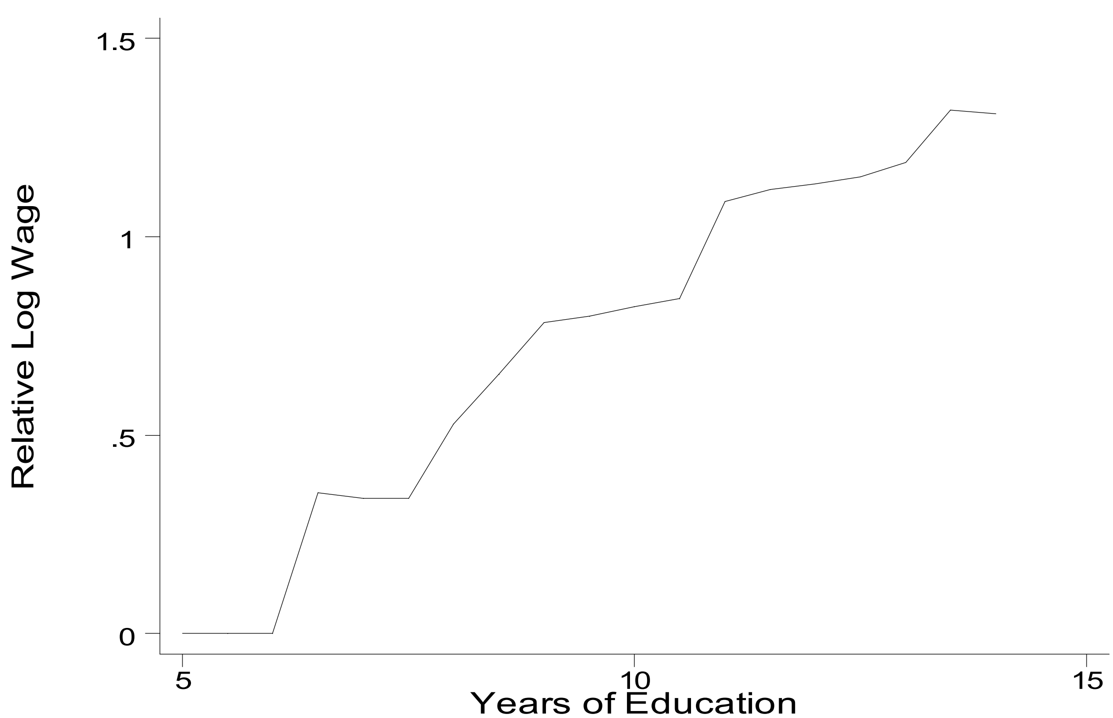
EMPIRICAL EXAMPLE

- Use data from Current Population Survey to estimate wage function

$$\boldsymbol{E}(\log W \mid EXP, EDUC) = \mu + f_{EXP}(EXP) + f_{EDUC}(EDUC),$$

  - *EXP* and *EDUC* are years of experience and education.
  - Population is white males with 14 or fewer years of education who work full time and live in urban areas in North Central U.S.


Relative Log Wage
1
.5
0
0
10
20
30
40
Years of Experience


Relative Log Wage
1.5
1
.5
0
5
10
15
Years of Education

COMMENTS ON ESTIMATION RESULTS

- Estimates of f_{EXP} and f_{EDUC} are nonlinear and differently shaped
- Functions f_{EXP} and f_{EDUC} with different shapes cannot be produced by a single-index model
- A lengthy specification search might be needed to find a parametric model that produces the shapes shown in the figure
- Some of the fluctuations of the estimates of f_{EDUC} and f_{EDUC} may be artifacts of random sampling errors.
- But a more elaborate analysis rejects the hypothesis that either function is linear.

CONCLUSIONS

- Nonparametric additive model

$$\boldsymbol{E}(Y \mid X = x) = \mu + \sum_{j=1}^{d} m_j(x^j)$$

- Additive components m_j can be estimated so as to:
  - Achieve one-dimensional nonparametric rate of convergence (dimension reduction)\
  - Have asymptotical normal limiting distributions
  - Achieve oracle efficiency