NONPARAMETRIC ESTIMATION

OF THE PROPENSITY SCORE

WITH LOCAL LIKELIHOOD METHOD

By
Burç Kayahan

SEPTEMBER 2005

DEPARTMENT OF ECONOMICS

UNIVERSITY OF GUELPH

1. Introduction

Objective of this paper: To explore the extent of the performance gains from nonparametric estimation of the propensity scores

Estimation technique: Local Likelihood Method

Problem: Curse of dimensionality due to need of large number of covariates required for the propensity score matching

Suggested solution: Principal Components as Regressors

Contribution of this paper:

i) To explore the consequences of misspecification of functional form in the propensity scores on the estimation of treatment measure

ii) To evaluate the performance of a recent nonparametric estimation framework: Local Likelihood Method

iii) To investigate the role of principal components as regressors in nonparametric estimation

2. THE METHODOLOGY

2.1 THE PROGRAM EVALUATION PROBLEM

$Y_1 \rightarrow$ Outcome in the treated state

$Y_0 \rightarrow$ Outcome in the untreated state

$X \rightarrow$ Covariates possessing the information about individual

$\mathrm{D} \rightarrow$ binary treatment variable:

$$\mathrm{D} = \begin{cases} 1 & \text{Did participate in the program} \\ 0 & \text{Did not participate in the program} \end{cases}$$

- Parameter of interest: Average Treatment Effect on the Treated (ATET)

$$\text{ATET} = E\,(Y_1 - Y_0 \mid D=1)$$

Problem: $E\,(Y_0 \mid D=1)$ is an unobserved counterfactual mean

- Randomized experiments: Randomly exclude D=1 persons who would have been treated

- Non-experimental estimators make assumptions so that $E\,(Y_0 \mid D=0)$ identifies $E\,(Y_0 \mid D=1)$

- Assumption needed for selection on observable estimators:
  The Conditional Independence Assumption, (CIA):

$$Y_1, Y_0 \perp D \mid X$$

Rosenbaum and Rubin (1983) → Propensity Score Matching

$Y_0 \perp D \mid X \rightarrow Y_0 \perp D \mid P(X)$ provided that $0 < \Pr(D=1 \mid X) < 1$

- Advantages of Matching over Regression Methods:

  i) Matching does not make linear functional form assumption that regression does.

  ii) Matching addresses common support problem in a way regression does not

- Two steps in implementing matching estimation

  i) Estimate the Propensity Score via parametric or nonparametric methods → Focus and the Contribution of this paper

  ii) Match treated outcomes with weighted average of comparison group outcomes using propensity scores.

2.2 LOCAL MAXIMUM LIKELIHOOD METHOD

Assume that (X_i, D_i) are i.i.d and that the conditional distribution of D_i given X_i is a Bernoulli distribution:

$$P(D_i=1 \mid X_i = x) = p(x) \qquad P(D_i=0 \mid X_i = x) = 1 - p(x) = q(x)$$

Mean regression function is $p(x) = E(Y \mid X = x) = F(g(x))$

$F(.) \rightarrow$ Cumulative density function of choice

$g(x) \rightarrow$ Unknown parameterized function of interest

Assume that error term from mean regression function is exponentially distributed $\rightarrow$ F(.) to be cumulative logistic function:

$$P(Y_i = 1 \mid X_i = x) = \frac{\exp[g(x)]}{1 + \exp[g(x)]} \text{ where } g(x) = \log \frac{p(x)}{1 - p(x)}$$

Estimating p (.) is equivalent to estimating g (.); preferable to work in logit domain since:

i) Log-likelihood is concave

ii) Logistic linear regression corresponds to the local case with $h = \infty$.

The log-likelihood in this context can be written as:

$$\ell(g(X), D_i) = \log\left\{ p(x)^{D_i} \; q(x)^{1-D_i} \right\} = D_i \; g(x) - \log\left[1 + \exp\{g(x)\}\right]$$

Functional form of g (.) is unknown. Assume that g(.) has a $(p+1)^{th}$ continuous derivative at point x_o; then $g(X_i)$ is approximated via p^{th} order Taylor series expansion:

$$g(X_i) \cong g(x_o) + g'(x_o)(X_i - x_o) + ... + \frac{g^p(x_o)}{p!}(X_i - x_o)^p = X_i^T\beta_o \quad (5.1)$$

Then the conditional local log - likelihood is given by :

$$L_p^*(\beta;h,x_o) = \sum\left\{ D_i(X_i^T\beta + r_i) - \log\left(1 + \exp(X_i^T\beta + r_i)\right)\right\} K_h \quad (5.2)$$

$$\text{The weighting Kernel} \rightarrow K_h = K\left(\frac{X_i - x_o}{h}\right)/h$$

Maximizing (5.2) with respect to β yields the vector of estimators: $\hat{\beta} = (\hat{\beta}_0,......,\hat{\beta}_p)$ $\hat{\beta}$ vector that maximizes (5.2) can not be expressed explicitly. It is estimated using Newton-Raphson Optimization algorithm.

The main obstacle in conducting such estimation:

LARGE # OF COVARIATES → CURSE OF DIMENSIONALITY

Suggested solution is using Principal Components as Regressors!

2.3 PRINCIPAL COMPONENT ANALYSIS (PCA)

- Pearson (1901) & Hotelling (1933): PCA linearly transforms an original set of variables into a smaller set of uncorrelated variables that represents most of the information in the original set.

$$
\begin{array}{llll}
Pc^1 = p_1'X & \max \mathrm{Var}(Pc^1) & \text{subject to} & \| p_1 \| = 1 \\
Pc^2 = p_2'X & \max \mathrm{Var}(Pc^2) & \text{subject to} & \| p_2 \| = 1 \;\; \& \; \mathrm{Cov}(Pc^1, Pc^2) = 0 \\
\vdots & & \vdots \quad\quad \vdots & \quad\quad \vdots \\
Pc^j = p_j'X & \max \mathrm{Var}(Pc^j) & \text{subject to} & \| p_j \| = 1 \;\& \; \mathrm{Cov}(Pc^k, Pc^j) = 0 \text{ for } k < j
\end{array}
$$

- In practice: p_i → Eigenvector corresponding to the i^{th} largest eigenvalue of Correlation matrix of X

- Picking all pc's → Curse of Dimensionality
  How to pick pc → literature is divided
  Pick the pc's that have the highest absolute sample correlation with Y following Hwang (2000)

- Using pc's less than the # of covariates introduces bias, gain the flexibility of functional form in logit framework! Tradeoff!

- Since extracted pc's have zero correlation, in theory, estimation with pc's without information loss is possible via partially additive estimation!

3. MONTE CARLO SIMULATION

➢ 3.1 MONTE CARLO DESIGN # 1

Objective: Explore performance gains from estimating probability nonparametrically

Method: Mean Squared Error and Count correct prediction of D_i via estimated $\hat{D}_i$:

$$\hat{D}_i = \begin{cases} 1 & \text{if } \hat{P}_i \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \quad \text{where } \hat{P}_i \text{ is the estimated probability}$$

(Step 1) Generate D* using various specifications:

a) Univariate Nonlinear model:

$D^* = \sin(2X) + U$ X~ Uniform (-2, 2) and U~ Exponential (1)

Set D=1 if $D^*>0$ and D=0 if $D^*\leq 0$

b) Multivariate Nonlinear model:

$D^* = \sin(4Z_1 - 2Z_2) + U$ Z_1~Exp (1) Z_2~Unif (0, 1) and U~ Exp (1)

Set D=1 if $D^*>0$ and D =0 otherwise

(Step 2) Compute estimated probabilities using both specifications

Results from Univariate DGP:

Table 1 - MSE and Correct Prediction Count for 300 rep, n=300		
LOGIT	**MSE**	**Count**
Parametric	0.0268	182
Nonparametric	0.0160	220
Performance Gain/Loss	40 %	21 %
Dominance Ratio	88 %	97 %

Figure 1 – Plot of Probabilities for n=100

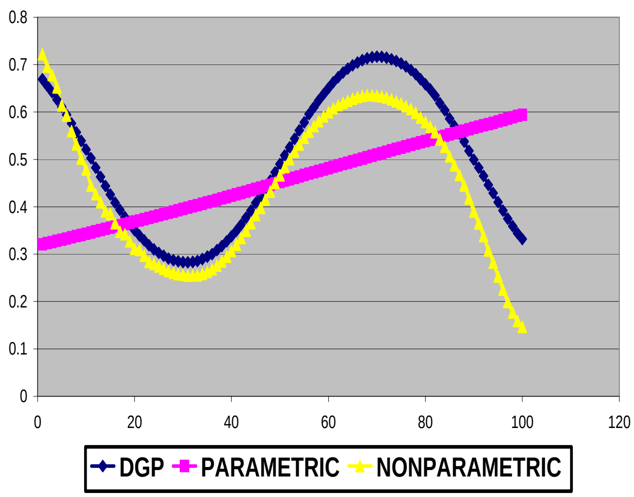
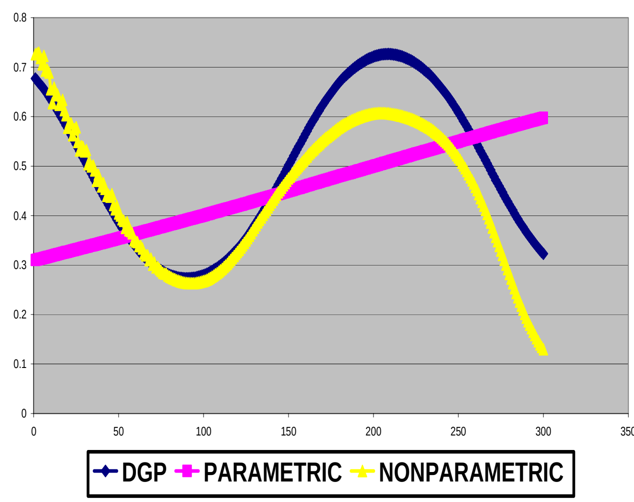


Figure 2 - Plot of Probabilities for n=300



Results from Multivariate DGP:

Table 2 – MSE and Count for 300 rep, n=300 with 1 pc		
LOGIT	**MSE**	**Count**
Parametric	0.0256	180
Nonparametric	0.0247	186
Performance Gain	3.5 %	3.3 %
Dominance Ratio	56%	73 %

- **3.2 MONTE CARLO DESIGN # 2**

Objective: Explore performance gains from estimating impact of the program with nonparametrically estimated propensity scores

Following Heckman, Lalonde and Smith (2000):

a) Outcome Equation:

$$Y_{it} = \beta + \alpha_i D_i + u_{it} \qquad u_{it} = \rho\, u_{it-1} + \varepsilon_{it} \qquad \varepsilon_{it} \text{ iid } N(0,\sigma_\varepsilon^2)$$

$\alpha_i \longrightarrow$ Impact of the program for individual i : $(Y_1 - Y_0 \mid D=1)$

b) Participation Equation:

$$D_i \rightarrow \text{Participation Dummy such that } D_i = \begin{cases} 1 & \text{if } NPB_i > 0 \\ 0 & \text{if } NPB_i \leq 0 \end{cases}$$

$$NPB \rightarrow \text{Net Present Benefit of the Program: } NPB_i = \frac{\alpha_i}{r} - Y_{ik} - c_i$$

$\frac{\alpha_i}{r} \rightarrow$ Discounted lifetime gain from training where r is interest rate

$Y_{ik} \rightarrow$ Foregone income in the training period

$c_i \rightarrow$ Cost of training for an individual: $c_i = f(Z_i)$

$Z_i \rightarrow$ Observed characteristics that effect the cost of training

- In practice, the parameters are chosen as follows:

$$\beta = 1000, \quad \alpha_i = 100 + N(0,300) \text{ and } u_{it} = 0.78\, u_{it-1} + \varepsilon_{it} \quad \varepsilon_{it} \sim N(0,450)$$

The parameters of the Participation Equation are:

$c_i = \omega + 0.05Z_i^4 - Z_i^3 + 6Z_i^2 + (10Z_i) + v_i$ where $v_i = N(0,450)$

ω is set to ensure 10% participation rate and $Z_i = 10 + N(0,1)$

Interest rate is set to be 10 %

- In total there are 10 periods in this Monte Carlo design:

  5 Pre-program periods, Program period, 4 Post-Program Periods

- Recall that $NPB_i = \frac{\alpha_i}{r} - Y_{ik} - c_i$ and Di = 1 if NPBi is >0

  α_i is unknown and the parameter of interest

  $Y_{\mathbf{ik}}$ is counterfactual so proxy it with $Y_{\mathbf{ik-1}}$

  $C_i = f(Z_i)$ therefore Z_i is the second covariate

  Use two covariates, $Z_{\mathbf{i}}$ and $Y_{\mathbf{ik-1}}$, to estimate $P(D_{\mathbf{i=1}})$

- The ATET parameter is estimated by matching the outcomes of individuals in the 4th Post-Program Period, Y_{it+4}, with nearest neighbor matching.

- **<u>Results from Monte Carlo Design #2</u>**:

N=750, Rep=300	Nearest Neighbor Matching			Nearest 5 Neighbors Matching		
	1 PC	**2 PCs**	**1 PC & 1 Var**	**1 PC**	**2 PCs**	**1 PC + 1 Var**
DGP	111.80	112.02	112.14	111.80	112.02	112.14
Parametric	108.53	109.50	110.75	109.24	110.42	110.29
Nonparametric	109.67	110.49	111.21	109.68	110.27	110.98
Dominance Ratio	67%	70%	69%	60%	65%	69%
MSE - Parametric	160.06	174.66	147.17	75.29	74.96	66.49
MSE - Nonpar.	48.51	38.86	35.30	33.88	27.03	24.33
Reduction	70%	78%	76%	55%	64%	63%
MAPE parametric	9.84	10.27	9.59	6.65	6.88	6.56
MAPE nonparamet.	5.40	4.87	4.75	4.68	4.22	3.98
Reduction	45%	53%	50%	30%	39%	39%
$MAPE_{MAX}$ - param.	39.24	42.13	42.73	43.29	29.99	24.56
$MAPE_{MAX}$ - nonpar	23.94	21.76	19.98	20.87	15.96	14.29
Reduction	39%	48%	53%	52%	47%	42%
$MAPE_{MIN}$ - param.	0.045	0.050	0.061	0.009	0.045	0.013
$MAPE_{MIN}$ - nonpar.	0.044	0.009	0.007	0.002	0.001	0.032

✓ In the average, both specifications do a good job in estimating Treatment effect on the Treated parameter.

✓ MSE of the parameter can be reduced by at least 70% with Nearest Neighbor Matching and by 55% with Nearest 5 Neighbors.

✓ <u>Conclusion</u>: Efficiency gains are possible with nonparametric estimation of the propensity scores.

4. Empirical Application

Data: Job Training and Partnership Act (JTPA)

Sample: Adult Males

Sub sample: Control group & Eligible Nonparticipants (ENPs)

Sample Size: 937 observations, 536 Controls and 401 ENP's

Methodology: Estimate bias between Controls and ENPs for post program periods via matching on the propensity scores using nearest neighbor method.

- Two principal components that maximize the percentage of the variation explained in the data are chosen for nonparametric estimation.

- Use weights to take choice-based sampling into account.

➢ **Results from Empirical Estimation**:

Table 3 - Results from Weighted Parametric Logistic Regression		
Variable	**Coefficient**	**P-value**
Intercept	-3.686	0.000
Site - Fort Wayne (IN)	1.622	0.000
Site - Jersey City (NJ)	0.709	0.051
Site - Providence (RI)	1.161	0.000
Black	0.496	0.093
Unemployed in last 6 months	1.527	0.000
Transitions in last 24 months	0.452	0.000
Age	-0.025	0.053
Earnings in last year (in 000's)	-0.0001	0.000

Table 4 - Estimated Bias with Nearest Neighbor Matching Method			
Outcome Variable: Quarterly Earnings in Monthly Dollars			
Time Period	**Raw Mean Differences**	**Parametric Logit**	**Nonparametric Logit**
t+1	-395.84	-35.45	-105.85
t+2	-343.61	-127.91	-8.32
t+3	-309.88	-138.6	-17.36
t+4	-269.57	-67.43	55.34
t+5	-266.59	-65.17	23.97
t+6	-338.57	-141.12	-63.28
Average Earnings	-122.65	-34.47	-7.32

✓ Raw mean differences between two groups are large considering the estimated monthly returns for JTPA are around $30!

✓ Matching on the nonparametric estimates of the propensity scores with principal components reduces bias.

✓ Empirical results support the findings of Monte Carlo experiments

5. Conclusion

a) Monte Carlo Design # 1: Estimation of Propensity Score

For a Nonlinear DGP: MSE can be decreased and the prediction of Y_i can be increased significantly in finite samples.

b) Monte Carlo Design # 2: Estimation of ATET

For a Nonlinear Multivariate DGP: In the average, the bias from estimating ATET is approximately the same for both specifications. However, variation in estimated ATET can be reduced significantly by nonparametric logit as shown by considerable reductions in MSE of the parameter. Further reductions in MSE are possible via increasing number of principal components in nonparametric regression.

c) Empirical Application: JTPA Results

Bias between Control Group and ENPs is reduced when outcomes are matched using nonparametrically estimated propensity scores.

Conclusion: Performance gains from Nonparametric Estimation of Propensity Scores are possible with the use of Principal Components