Inference for additive model at response random missing based on principal component imputation

Xuchen Dong$^{1,3}$ and Simin Chen$^2$

$^1$ College of Science, University of Nottingham Ningbo China, Ningbo 315100, P. R. China
$^2$ College of Science, Northeastern University, Shenyang 110189, P. R. China
$^3$ Corresponding author’s e-mail, Hmyxd1@nottingham.edu.cn

Abstract. In this paper, we consider statistical inference for the additive model when the response is missing at random. The paper proposes a framework for statistical inference under missing data, which integrates basis function approximation and principal component imputed estimation equation (PCIEE). Inverse probability weighting is used to construct robust estimator. Under mild assumptions, the convergence rate of component function estimator is proved. In addition, when the sample size $n$ tends to infinity, the PCIEE-based variable selection can select the real predictor. We also demonstrated the performance of PCIEE through numerical experiments.

1. Introduction
The additive model is flexible and avoids the “curse of dimensionality”, the expression is as follows

$$Y = \mu + \sum_{l=1}^{d} m_l(X_l) + \varepsilon,$$  \hspace{1cm} (1)

where $Y$ is the response variable, $X = (X_1, \cdots, X_d)^T$ is vector of predictor variables, and $m(x) = \sum_{l=1}^{d} m_l(x_l)$ is unknown nonparametric function. Assume $E[m_l(X_l)] = 0, l = 1, \cdots, d$. and error $E(\varepsilon|X) = 0$.

Additive models are special cases of more general projection pursuit regression model of Friedman and Stuetzle[1]. Two algorithms closely related to backfitting are proposed by Burg and Breiman respectively[2][3]. Wang proposed a framework of spline backfitting kernel to inference the component function in additive model and obtain the asymptotic normality[4]. Ma extended spline-backfitted kernel (SBK) smoothing to partially linear additive models[5]. This studies on additive models above mainly focused on clean data.

In the study of statistical models, there is missing data in the data collection process, and imputation is one of the most common methods to deal with missing data. Rubin distinguished between item and unit non-response, and proposed a multiple imputation method[6]. Chen applied a random imputation method to regression model with incomplete data[7]. The above method is divided into two steps when dealing with missing data, the missing data is imputed first, and then the complete data after imputed is used for statistical inference. In addition, there are other methods for statistical inference under missing data. Zhou adopted the method
of imputation estimating equations when response variable is missing at random[8]. More recent work on additive models is focused on nonparametric estimation with missing data, Seaman proposed the penalized splines of the propensity score under the missing at random assumption, and the result of inference is robust [9]. Boente estimated the component function of additive model based on marginal integration to deal with missing data[10]. The statistical inference of additive model with missing data mainly adopted the technology of data imputation, Wu studied the statistical inference of the additive model with missing data, they used the method of imputed estimation equation, but encountered curse of dimensionality in statistical inference[11]. Inspired by the work of [11], we proposed a framework to inference additive model with response variable missing at random. Principal component analysis solves the problem “curse of dimensionality”, and the technique of inverse probability weighting is used to construct a robust statistical inference. The simulation results show that the proposed method is better than the two methods proposed by [11], and this method is closer to Oracle estimation, that assumes the missing data is known in advance, it is only used as a comparison benchmark and cannot be obtained in practical applications.

This paper proposes a framework for robust estimation of additive models under missing data. The framework integrates principal component analysis, basis expansion, and techniques such as imputed estimation equations. The imputation is an important method of dealing with missing data. Principal component analysis(PCA) is a good tool to reduce the dimension of covariances and extract main information. Inverse probability weighting is used to construct robust estimator. We have made some novel contributions in this paper: (i) We proposed a method to effectively reduce the dimensional of the robust estimation equation to inference the robust estimator. We have made some novel contributions in this paper: (i) We proposed a covariances and extract main information. Inverse probability weighting is used to construct such as imputed estimation equations. The imputation is an important method of dealing with data. The framework integrates principal component analysis, basis expansion, and techniques cannot be obtained in practical applications.

The paper is organized as follows. The framework of PCIEE and it’s theoretical results are in section 2. Tuning parameters are selected in section 3. And section 4 is Monte Carlo simulation, section 5 is the conclusion. Proofs are deferred to the Appendix.

2. Methodology and Asymptotic Properties
2.1. Principal Component Imputed Estimation Equation(PCIEE)
Frist we describe the method of principal component imputed estimation equation (PCIEE). Let dataset \( \{Y_i, X_i\}_{i=1}^{n} \) following model (1). Without loss of generality, the support set of covariance available \( X_i \) is \([a_l, b_l] = [0, 1], l = 1, \cdots, d \). For identifiability purpose, both component functions are centered. Let \( R_i \) is defined as a missing indicator, \( R_i = 1 \) if \( Y_i \) is observed and \( R_i = 0 \) if \( Y_i \) is missing. The conditional distribution of \( R_i \) given \( X_i \) is bernoulli distribution

\[
EE(R_i = 1|X_i, Y_i) = E(R_i = 1|X_i) = g(X_i, \eta) = \{1 + (\exp\{1, X_i^\tau \})\eta\}^{-1}.
\]  

(2)

\( \eta \) can be consistently estimate by maximum likelihood estimation with observed data. We use a centralized q-order spline basis to approximate the unknown function \( m_l(\cdot) \)'s, and the number of knots is \( N_n = O(n^{1/3} \log(n)) \), let \( B(x) = \{B_{s,l}(x_l) : 1 \leq l \leq d, 1 \leq s \leq J \} \) is a basis system of functional space \( G, x = (x_l)_{l=1}^d, J = N_n + q + 1 \). Suppose that \( m_l(\cdot) \in G \), that is

\[
m_l(x) = \sum_{j=1}^J \beta_j B_{s,l}(x), l = 1, \cdots, d.
\]

Substituting them into model (1), we can get linear model

\[
Y_i \approx B_1(\beta + \varepsilon_i), i = 1, \cdots, n, \text{where } \beta = (\mu, \beta_{11}, \cdots, \beta_{j1}, \cdots, \beta_{1d}, \cdots, \beta_{jd})^\tau \text{ is the collection of the coefficients, and } B_i = (1, B_{1,1}(X_i), \cdots, B_{1,l}(X_i1), \cdots, B_{d,l}(X_{id}))^\tau,
\]

we use linear combination of basis functions to express functions \( m_l(x) \). Therefore, consider using the estimation equation
method of imputation, the expression is as follows

$$\sum_{i=1}^{n} \phi_i(\beta, X_i) = \sum_{i=1}^{n} \left\{ R_i \phi_i(\beta) + \{1 - R_i\} m_{\phi}(\beta, X_i) \right\} = 0, \quad (3)$$

where $\phi_i(\beta) = B_i(Y_i - B_i^T \beta)$, and $m_{\phi}(\beta, x) = E[\{\phi_i(\beta)\} | X_i = x]$ is a $d$-dimensional score function vector, dimensional reduction or parameterization is needed for estimate $m_{\phi}(\beta, x)$, [11] adopted the marginal imputation(MI) and the maximum correlation imputation method (MRI) to solve the equation (3). They used two different methods instead of $m_{\phi}(\beta, X_i)$, $m_{\phi}^{(MRI)}(\beta, x) = E[\phi_i(\beta)|X_{iu} = x]$ and $m_{\phi}^{(MI)}(\beta, x) = \frac{1}{d} \sum_{i=1}^{d} E[\phi_i(\beta)|X_{id} = x_l]$, and $u$ is the variable index with the highest correlation with $Y$ among the covariates, N-W kernel estimation was adopted to estimate the conditional expectation. But the two methods can only use a small amount of marginal variable information, which affects the efficiency of model inference. Therefore, we introduce the method of principal component analysis and inverse probability weighting to ensure that important information can be extracted. At the same time, our method can also ensure that the consistent estimator even if response mechanism model (2) is not true.

The method is called principal component imputed estimation equation (PCIEE), that solves the following estimation equation,

$$\sum_{i=1}^{n} \phi_i^{(PCI)}(\beta, Z_i) = \sum_{i=1}^{n} \left\{ \frac{R_i}{\pi(X_i, \eta)} \phi_i(\beta) + \{1 - \frac{R_i}{\pi(X_i, \eta)}\} \hat{m}_{\phi}^{(PCI)}(\beta, Z_i) \right\} = 0, \quad (4)$$

where $m_{\phi}^{(PCI)}(\beta, x) = E[\phi_i(\beta)|Z_i = x]$ and $Z_i$ is the first principal component, $m_{\phi}^{(PCI)}(\beta, x)$ can be estimated by N-W kernel estimation, $\hat{m}_{\phi}^{(PCI)}(\beta, x) = \sum_{i=1}^{n} \omega_i(x_i) \phi_i(\beta)$, where $\omega_i(x) = R_i K_h(Z_i - x)/\sum_{j=1}^{n} R_j K_h(Z_j - x)$, $K_h(\cdot) = 1/h K(\cdot/h)$, $K(\cdot)$ is a density function of normal distribution, a bandwidth $h$ is used to balance smoothing and fitting. By solving the principal component imputation estimating equation (4), $\hat{\beta}^{(PCI)} = (\hat{\mu}^{(PCI)}, \hat{\beta}_1^{(PCI)}, \ldots, \hat{\beta}_d^{(PCI)})^T$ are obtained. Thus, $\hat{m}_l^{(PCI)}(x) = \sum_{s=1}^{J} \hat{\beta}_{sl}^{(PCI)} B_{s,l}(x), \ l = 1, \ldots, d, \ \eta$ is consistent estimator of $\eta$.

In equation (4), we use the singular value decomposition method to solve the principal components, and there is a software package in the R program that can directly perform principal component analysis. In next section, a bandwidth $h$ is selected by a data-driven approach, so the nonparametric function $m_{\phi}^{(PCI)}(\beta, x)$ is obtained. Next, variable selection are considered to screen the important covariates with incompletely data. By [11] and [12], the smooth-threshold estimating equations are adopted to select influential variable. Based on the principal component imputed estimating equation (4), our framework of variable selection as follows,

$$\left(I_{dJ} - \hat{\Delta}\right) \sum_{i=1}^{n} \phi_i^{(PCI)}(\beta, Z_i) + \hat{\Delta} \beta = 0 \quad (5)$$

where $I_{dJ}$ is the $dJ$-dimensional identity matrix and $\hat{\Delta} = \text{diag}\{\hat{\delta}_1, \ldots, \hat{\delta}_1, \ldots, \hat{\delta}_d, \ldots, \hat{\delta}_d\}$ is the $dJ \times dJ$ diagonal matrix. equation (5) can yield a sparse solution, where $\hat{\delta}_l = \min\{1, \lambda/\|\hat{\beta}_l(0)\|\}$, where $\hat{\beta}_l^{(0)}$ is the least squares estimation of $\beta_k$ under complete data, and $\lambda$ is the regularization parameter.

By solving (5), the estimator $\hat{\beta}^{(SPCI)}$ of $\beta$ can be obtain. Then, the estimator of nonparametric function of additive model is obtained as follows $\hat{m}_l^{(SPCI)}(x) = \sum_{s=1}^{J} \hat{\beta}_{sl}^{(SPCI)} B_{s,l}(x), l = 1, \ldots, d$. 


2.2. Error Bound
The large sample properties of the imputation estimators (4) and (5) are proofed in this subsection. Let $m_l^0(\cdot)$ be the true function of $m_l(\cdot)$, and the coefficient $\beta$ of basis expansion is denoted by $\beta^0$. We take that $m_l^0(\cdot) \equiv 0$, $l = p+1, \cdots, d$ and $m_l^0(\cdot) \neq 0$, $l = 1, \cdots, p$, the theorem is as follows.

**Theorem 1** Assume (C1)–(C5) is hold in the Appendix, the data are following model (1), the parameter $K = O_p(n^{1/(2r+1)})$. As $n \to \infty$, then error bounds of the estimators defined by (4) and (5) are as follows

\[
\|\hat{m}_l^{(PCI)}(\cdot) - m_l^0(\cdot)\| = O_p(n^{-r/(2r+1)}),
\]
\[
\|\hat{m}_l^{(SPCI)}(\cdot) - m_l^0(\cdot)\| = O_p(n^{-r/(2r+1)}),
\]

$l = 1, \cdots, p$.

Theorem 1 means the estimation of the nonparametric function obtained by (4) and the estimators defined by (5) achieved the optimal convergence rate. missing data has no effect on the error bounds of estimators.

**Theorem 2** Suppose conditions (C1)–(C5) are hold in the Appendix, $K = O_p(n^{1/(2r+1)})$. If the tuning parameter $\lambda \to 0$ as $n \to \infty$, the model (2) is not correct, then the estimators defined by (4) are consistent estimator, $l = 1, \cdots, p$

\[
\hat{m}_l^{(PCI)}(\cdot) \overset{p}{\longrightarrow} m_l^0(\cdot), \quad \hat{m}_l^{(SPCI)}(\cdot) \overset{p}{\longrightarrow} m_l^0(\cdot)
\]

Theorem 2 shows that even the model (2) is not hold, the proposed method is consistent estimation, and this procedure is robust to the wrong assumption of the model. The error bound of the estimators defined by (5) based on smooth threshold estimating equations are showed in following theorem.

**Theorem 3** Assume conditions (C1)–(C5) are hold in the Appendix and, $K = O_p(n^{1/(2r+1)})$. If tuning parameter $\lambda \to 0$ and $n^{r/(2r+1)} \lambda \to \infty$, the estimators of (5) must satisfy $\hat{m}_l^{(SPCI)}(\cdot) \equiv 0$, $l = p+1, \cdots, d$ with probability tending to 1 when $n \to \infty$.

If the regularization parameters are selected correctly, Theorem 1 and Theorem 3 means our method of sparse estimation SPCI can obtain the Oracle properties proposed by [13]. In a other word, with proper select regularization parameters, the framework of variable selection based on principal imputation estimation equation is consistent.

3. Tuning Parameters Selection
In our framework of variable selection, The correct selection of parameters is the guarantee for the establishment of consistency, we select $K$, $\lambda$ and $h$ by method of data driven. To simplify, $K$ was chose by minimizing the cross-validation loss function with complete data,

\[
CV(K) = \sum_{i=1}^{n} r_i(Y_i - B_i^T \hat{\beta}_{[i]}^2),
\]

where $\hat{\beta}_{[i]}$ is the solution of estimation equation $\sum_{i=1}^{n} r_i \phi_i(\beta) = 0$ after deleting the $i$th data. When $K$ is determined, the parameter $\lambda$ can be selected by minimizing the BIC criterion.

\[
BIC(\lambda) = \sum_{i=1}^{n} r_i(Y_i - B_i^T \hat{\beta}_i^2) + DF(\hat{\beta}_i^2) \log n,
\]

$\hat{\beta}_i^2$ denotes the estimators given $\lambda$ defined by (4), $DF(\cdot)$ means The number of nonzero elements. As in [8], use optimal bandwidth conditions, $h = C n^{-1/3}$, and $h_l = \hat{\sigma}_X h$, where $\hat{\sigma}_X$ is the sample standard deviation of covariate $X_l$. 


4. Experiments

Though empirical experiments, we compare our method and other benchmark method the dataset from model (1), and take \( d = 10, \) \( m_i(x) = 4x - 1, \) \( m_2(x) = \cos(2\pi x) \) and \( m_3(x) = \sin(2\pi x), \) \( m_i(x) = 0, \) \( i = 4, \cdots, d. \) For generality, to ensure that the \( d \) dimensional covariates are dependent, we make the following settings, take \( Z = (Z_1, \cdots, Z_{10})^T \sim N(0, \Sigma_z), \) and the covariates \( X_k = \Phi(Z_k), \) \( k = 1, \cdots, d, \) where \( \Phi(\cdot) \) denotes the cumulative distribution function of standard normal distribution. \( \varepsilon \sim N(0, 0.5), \) \( r_i \sim B(1, g(\cdot)) \). Two missing mechanism (MM) models of (2.1) are considered, MM1: we select \( \eta = (1, 2, 3, 0, 0, 0, 0, 0, 0, 0.5)^T, \) and the missing rate is 40%.

MM2: we select \( \eta = (0.3, 0.6, 1, 1, 0, 0, 0, 0, 0, 0, 0.5)^T, \) and the missing rate is 20%.

When the model (2.1) does not hold, the relation of between response index and covariates mean square errors (RMSE),

\[
P(R_i = 1|X_i) = \frac{\exp\{(1, \sin(X_i^T)\eta\}}{1 + \exp\{(1, \sin(X_i^T)\eta\}}.
\]

In order to verify the efficiency of the method under different sample sizes, \( n = 150, 300 \) and 500, the order of B-splines \( q = 3. \) the simulation can evaluate the performance of PCIEE and other benchmark when the sample size is small. The performances of estimators for \( m(\cdot) \) is assessed by the square root of average square errors (RASE) and the parameter \( \mu \) is assessed by root of mean square errors (RMSE),

\[
RASE = \left\{ \frac{1}{T} \sum_{s=1}^{T} \sum_{l=1}^{p} [\tilde{m}_l(x_s) - m_l^0(x_s)]^2 \right\}^{1/2}, \quad \text{RMSE} = \left\{ \frac{1}{N} \sum_{s=1}^{N} (\hat{\theta}_s - \theta)^2 \right\}^{1/2},
\]

where \( x_s, s = 1, \cdots, T, \) are the grid points. where \( \hat{\theta}_s, s = 1, \cdots, N, \) are estimation of \( \theta \) using \( s \)th data-set. The average RMSE of the nonparametric function in \( N \)-th simulation and the MSE of the intercept parameter \( \mu \) are reported, \( N = 200. \) We first compare the performances of our procedure (PCIEE) and marginal imputation estimator by [11], and we also compute the oracle estimator as a benchmark in table 1, that is not feasible in practical applications, and only available in simulation studies.

The results are reported in Table 1, that demonstrate the performance of PCIEE. With sample size increasing, our method is very close to the Oracle method, so we can say that our approach has Oracle efficiency, and the result of MM3 demonstrate our procedure is robust, no matter whether the response mechanism model (2) is hold.

With 1000 simulation, the result of the variable selection are showed in Table 2, in which the column labeled “C” denotes the average number of the component functions \( m^0_l(\cdot) \equiv 0, l = p + 1, \cdots, d \) correctly set to zero, and the column labeled “T” gives the average number of the function \( m^0_l(\cdot) \equiv 0, l = 1, \cdots, p \) incorrectly set to zero. Furthermore, the median of RASE of function \( \mu + \sum_{i=1}^{n} m_i(x_i) \) is showed in table 2. For different missing probabilities, our procedure is close to the performance of Oracle, and the framework of variable selection can select the real model with probability 1, therefore it not only has Oracle efficiency but also is robust select procedure.

5. Concluding Remark

In summary, with response variable missing at random, we propose a new additive model statistical inference method. Our algorithm is a combination of singular value decomposition and solving linear equations, avoiding optimization problems, inverse probability weighting is used to construct robust estimator and improving the efficiency of procedure. At the same time, a robust variable selection of the additive model based on the PCIEE and smooth threshold estimating equations was proposed. The simulation research shows that our estimation is better...
Table 1. The 100*RMSE of $\mu$ and 100*RASE of component functions of our procedure and benchmarks

| n   | Method  | $\mu$ | $m_1(x)$ | $m_2(x)$ | $\mu$ | $m_1(x)$ | $m_2(x)$ | $\mu$ | $m_1(x)$ | $m_2(x)$ |
|-----|---------|-------|----------|----------|-------|----------|----------|-------|----------|----------|
| 150 | NAIVE   | 17.06 | 32.55    | 31.89    | 13.12 | 25.56    | 24.68    | 18.25 | 35.22    | 34.66    |
|     | MRIEE   | 15.36 | 26.22    | 25.31    | 13.19 | 22.25    | 21.38    | 17.34 | 34.34    | 33.26    |
|     | PCIEE   | 15.21 | 25.22    | 24.14    | 12.36 | 22.16    | 21.25    | 16.93 | 26.23    | 25.75    |
|     | ORACLE  | 14.51 | 24.50    | 23.51    | 11.35 | 18.86    | 18.34    | 14.06 | 22.64    | 23.51    |
| 300 | NAIVE   | 15.13 | 26.22    | 25.31    | 11.18 | 20.30    | 19.31    | 15.03 | 28.84    | 26.38    |
|     | MRIEE   | 11.24 | 19.48    | 19.44    | 10.25 | 15.17    | 15.64    | 12.96 | 18.55    | 18.69    |
|     | PCIEE   | 10.60 | 17.56    | 17.49    | 9.66  | 15.47    | 18.27    | 13.37 | 20.58    | 21.49    |
|     | ORACLE  | 10.21 | 15.72    | 15.53    | 8.84  | 13.63    | 13.86    | 10.04 | 14.54    | 14.86    |
| 500 | NAIVE   | 8.44  | 14.19    | 13.83    | 8.44  | 13.38    | 12.75    | 10.24 | 14.43    | 14.64    |
|     | MRIEE   | 8.32  | 13.33    | 12.01    | 8.33  | 13.22    | 12.12    | 9.05  | 13.64    | 13.47    |
|     | PCIEE   | 8.07  | 12.98    | 12.84    | 8.07  | 11.31    | 11.26    | 8.65  | 13.14    | 13.03    |
|     | ORACLE  | 7.87  | 11.91    | 11.41    | 5.32  | 9.63     | 9.45     | 7.96  | 12.53    | 12.84    |

Table 2. The performance comparison of variable selection between our procedure and other methods

| n   | NAIVE     | SPICI     | MRIVS     | ORACLE    |
|-----|-----------|-----------|-----------|-----------|
|     | $C$ | $I$ | RASE | $C$ | $I$ | RASE | $C$ | $I$ | RASE | $C$ | $I$ | RASE |
| M1  | 150 | 5.463 | 0   | 0.251 | 6.613 | 0   | 0.176 | 6.479 | 0   | 0.186 | 6.785 | 0   | 0.172 |
|     | 300 | 6.828 | 0   | 0.145 | 6.986 | 0   | 0.122 | 6.952 | 0   | 0.129 | 6.996 | 0   | 0.124 |
|     | 500 | 7    | 0   | 0.087 | 7     | 0   | 0.085 | 7     | 0   | 0.085 | 7     | 0   | 0.084 |
| M2  | 150 | 6.479 | 0   | 0.186 | 6.867 | 0   | 0.153 | 6.860 | 0   | 0.156 | 6.982 | 0   | 0.136 |
|     | 300 | 6.974 | 0   | 0.126 | 6.997 | 0   | 0.103 | 6.992 | 0   | 0.110 | 6.998 | 0   | 0.096 |
|     | 500 | 7    | 0   | 0.089 | 7     | 0   | 0.080 | 7     | 0   | 0.080 | 7     | 0   | 0.079 |
| M3  | 150 | 6.503 | 0   | 0.184 | 6.834 | 0   | 0.168 | 6.801 | 0   | 0.198 | 6.982 | 0   | 0.159 |
|     | 300 | 6.843 | 0   | 0.140 | 6.987 | 0   | 0.114 | 6.961 | 0   | 0.128 | 6.997 | 0   | 0.104 |
|     | 500 | 7    | 0   | 0.092 | 7     | 0   | 0.086 | 7     | 0   | 0.098 | 7     | 0   | 0.084 |

than Naive estimation and two estimation methods proposed by [11]. our procedure is close to Oracle estimation, which is a comparative benchmark and cannot be obtained in practical application.

6. Appendix: Proofs of Theorems

some regularity conditions are listed that are used.

C1. $\theta(u)$ is $r$th continuously differentiable on $(0,1)$, where $r > 1/2$.

C2. The density function of $X$ says $f(X)$, is bounded away from 0 and infinity on $[0,1]$. Furthermore, we assume that $f(x)$ is continuously differentiable on $(0,1)$.

C3. Let $\sigma^2 = E\{e^2|U = u\}$ and $G(u) = E\{XX^T|U = u\}$. Then, $\sigma^2(u)$ and $G(u)$ are continuous with respect to $u$. Furthermore, for given $u$, $G(u)$ is a positive definite matrix, and the eigenvalues of $G(u)$ are bounded.
C4. Let $c_1, \ldots, c_K$ be the interior knots of $[0,1]$. Furthermore, let $c_0 = 0, c_{K+1} = 1$, $h_i = c_i - c_{i-1}$. There exists a constant $C_0$ such that $\frac{\max(h_i)}{\min(h_i)} \leq C_0, \quad \max \{h_{i+1} - h_i\} = o(K^{-1})$

C5. The tuning parameter $\lambda$ satisfies $n^{r/2(2r+1)} \lambda \to \infty, n^{1/2} \lambda \to 0, r > 1/2$.

**Proof of Theorem 1.** From $\sum_{i=1}^{n} \left\{ \frac{r_i}{\pi(X_i, U_i, \eta)} Z_i (Y_i - Z_i^\tau \beta) + (1 - \frac{r_i}{\pi(X_i, U_i, \eta)}) \hat{m}_\phi(U_i) \right\} = 0$ and $j = 1, \ldots, p$, we can get $\hat{\beta}_j^{(0)} = O_p\left(n^{-r/(2r+1)}\right)$ for any $\varepsilon > 0$, $p(r\hat{\delta}_j > n^{-1/2} \varepsilon) = pr\left(\frac{\lambda}{||\hat{\beta}_j^{(0)}||} > n^{-\frac{1}{2}} \varepsilon\right) = pr\left(||\hat{\beta}_k^{(0)}|| < n^{\frac{1}{2}} \varepsilon\right) = \lambda_{\beta_{0}} = o_p\left(n^{-1/2}\right)$, for $j = 1, \ldots, p$. A slight symbol abuse $\hat{\Delta} = o_p\left(n^{-1/2}\right)$

$$Q(\beta) = (I_{pL} - \hat{\Delta}) \sum_{i=1}^{n} \frac{r_i}{\pi(X_i, U_i, \eta)} Z_i (Y_i - Z_i^\tau \beta) + (I_{pL} - \hat{\Delta}) \sum_{i=1}^{n} \left(1 - \frac{r_i}{\pi(X_i, U_i, \eta)} \right) \hat{m}_\phi(U_i)$$

$$+ (I_{pL} - \hat{\Delta}) \sum_{i=1}^{n} \frac{r_i}{\pi(X_i, U_i, \eta)} \{Z_i (Y_i - Z_i^\tau \beta) - \hat{m}_\phi(U_i)\} + \hat{\Delta} \beta$$

$$= I_1 + I_2 + I_3 + I_4.$$

In summary,$I_1 = O_p\left(n^{1/2} + O_p\left(n^{1-r/(2r+1)} + O_p\left(n (\beta - \beta_0)\right)\right)\right)$, $I_2 = (I_{pL} - \hat{\Delta}) \sum_{i=1}^{n} \left(1 - \frac{r_i}{\pi(X_i, U_i, \eta)} \right) \hat{m}_\phi(U_i) = o_p\left(K^{-r} + n^{-1/2}\right)$ and $I_4 = o_p\left(n^{-1/2}\right)$, $I_3 = o_p\left((nh)^{-1/2}\right)$, using $Q(\beta) = 0$, SPCI’s proof is similar, we complete the proof of Theorem 1.

**Proof of Theorem 2.** We just proof (4) is a estimation equation, that can produce a consistent estimation.

$$E\left(\sum_{i=1}^{n} \phi_i^{(PCI)}(\beta, Z_i)\right) = E\left(\frac{R}{\pi(X, \eta)} \phi(\beta) + \left(1 - \frac{R}{\pi(X, \eta)}\right) \hat{m}_\phi^{(PCI)}(\beta, Z)\right)$$

$$= E\left(\hat{m}_\phi^{(PCI)}(\beta, Z)\right) + E\left(\frac{R}{\pi(X, \eta)} (\phi(\beta) - \hat{m}_\phi^{(PCI)}(\beta, Z))\right)$$

$$= E\left(E(\phi(\beta)|Z)\right) + E\left(E\left(\frac{R}{\pi(X, \eta)} (\phi(\beta) - \hat{m}_\phi^{(PCI)}(\beta, Z))|X, Y\right)\right)$$

$$= E(\phi(\beta)) = 0.$$
Reference

[1] Friedman J H and Stuetzle W 1981 Projection pursuit regression J. Am. Statist. Ass. 76 817-23.
[2] Buja A, Hastie T and Tibshirani R 1989 Linear smoothers and additive models Ann. Statist. 453-510.
[3] Breiman L and Friedman J H Estimating optimal transformations for multiple regression and correlation J. Am. Statist. Ass. 80 580-98.
[4] Wang L and Yang L 2007 Spline-backfitted kernel smoothing of nonlinear additive auto-regression model Ann. Statist. 35 2474-503.
[5] Ma S and Yang L 2011 Spline-backfitted kernel smoothing of partially linear additive model204-19.
[6] Rubin D B 2004 Multiple imputation for nonresponse in surveys (New York: John Wiley and Sons).
[7] Chen L and Sun J. 2010 A multiple imputation approach to the analysis of interval-censored failure time data with the additive hazards model Comput. Stat. Data. An. 54 1109-16.
[8] Zhou Y, Wan A T K and Wang X 2008 Estimating equations inference with missing data J. Am. Statist. Ass. 103 1187-99.
[9] Seaman S R and Vansteelandt S 2018 Introduction to double robust methods for incomplete data Stat. Sci. 33 184-97.
[10] Boente G and Martinez A M 2016 Estimating additive models with missing responses Commun. Stat. Theor. M. 45 413-29.
[11] Wu J, Zhang J and Li G R 2017 Variable selection for additive models with missing response at random[J]. J. Adv. Stat. 2 1-9.
[12] Ueki M 2009 A note on automatic variable selection suing smooth-threshold estimating equations Biometrika 96 1005–11.
[13] Fan J Q and Li R 2001 Variable selection via nonconcave penalized likelihood and its oracle properties J. Am. Statist. Ass. 96 1348-60.