Semiparametric Regression Models with Missing Data: the Mathematics in the Work of Robins et al.

Menggang Yu*        Bin Nan†

*University of Michigan, menggang@umich.edu
†University of Michigan, bnan@umich.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/umichbiostat/paper17

Copyright ©2003 by the authors.
Semiparametric Regression Models with Missing Data: the Mathematics in the Work of Robins et al.

Menggang Yu and Bin Nan

Abstract

This review is an attempt to understand the landmark papers of Robins, Rotnitzky, and Zhao (1994) and Robins and Rotnitzky (1992). We revisit their main results and corresponding proofs using the theory outlined in the monograph by Bickel, Klaassen, Ritov, and Wellner (1993). We also discuss an illustrative example to show the details of applying these theoretical results.
Semiparametric Regression Models with Missing Data: the Mathematics in the Work of Robins et al.

Menggang Yu and Bin Nan

*University of Michigan*

*May 3, 2003*

**Abstract**

This review is an attempt to understand the landmark papers of Robins, Rotnitzky, and Zhao (1994) and Robins and Rotnitzky (1992). We revisit their main results and corresponding proofs using the theory outlined in the monograph by Bickel, Klaassen, Ritov, and Wellner (1993). We also discuss an illustrative example to show the details of applying these theoretical results.

*Keywords and phrases:* Efficient score, influence function, missing at random, regression models, scores, tangent set, tangent space.

1 **Introduction**

Improving efficiency for the estimates in semiparametric regression models with missing data has been an interesting and active research subject. Robins, Rotnitzky, and Zhao (1994) (hereafter RRZ) provided profound calculations of efficient score functions and information bounds for models with data Missing At Random (MAR, a terminology of Little and Rubin (1987)). Part of their calculations can also be found in Robins and Rotnitzky (1992) (hereafter RR). Their basic idea is to bridge the model with missing data and the corresponding model without missing data (full model), if certain properties of the full model
are known or easily obtained. The results are fundamental and can be applied to a variety of regression models. But it is very difficult to read these comprehensive abstract results since the authors only supplied very condensed proofs, and the whole material, including notation, was organized in a way that is hard to follow. We feel that it is necessary to revisit these very important results. The purpose of this study is to explicate RRZ and RR using the theory (and notation) in Bickel, Klaassen, Ritov, and Wellner (1993) (hereafter BKRW). The desired result is that a wider audience of statisticians interested in semiparametric models with missing data feel more comfortable following the recent developments in the area and applying the results in RRZ and RR. We begin by introducing the general semiparametric model with data MAR and the notation that will be used in the following sections. In Section 3, we introduce the main results of calculations of efficient score functions for data MAR with arbitrary missing patterns, monotonic missing patterns, and the two-phase sampling designs which are special cases of monotonic missingness. In Section 4 we discuss a simple example of the mean regression model with surrogate outcome to show the details of applying the general results in RRZ and RR. The detailed rigorous proofs of the main results can be found in Section 5, which is followed by the properties of influence functions and corresponding proofs in Section 6. We wrap up with a brief discussion in Section 7.

2 A General Model and Notation

We will adopt the notation mostly from Bickel, Klaassen, Ritov, and Wellner (1993) and Nan, Emond, and Wellner (2000). Suppose the underlying full data are i.i.d. copies of the $m$-dimensional random vector $X = (X_1, \ldots, X_m)$. We denote the model for $X$ as $Q = \{Q_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^d, \eta \in \mathcal{H}\}$ where $Q_{\theta,\eta}$ is a distribution function, $\theta$ is the parameter of interest and $\eta$ is an infinite-dimensional nuisance parameter or a vector of several infinite-dimensional nuisance parameters.

Let $R = (R_1, \ldots, R_m)$ be a random vector with $R_j = 1$ if $X_j$ is observed and $R_j = 0$ if $X_j$ is missing, $j = 1, \ldots, m$. Let $r$ be the realized value of $R$. For some $R$ we observe the data

$$X_{(R)} = (R_1 \ast X_1, \ldots, R_m \ast X_m),$$
where

\[ R_j \ast X_j \equiv \begin{cases} X_j, & R_j = 1; \\ \text{Missing}, & R_j = 0. \end{cases} \quad j = 1, \ldots, m. \]

Thus the observed data are i.i.d. copies of \((R, X_{(R)})\). Throughout the paper we will assume that the data are MAR, i.e.,

\[ \pi(r) \equiv P(R = r | X) = P(R = r | X_{(r)}) \equiv \pi(r, X_{(r)}); \quad (2.1) \]

and the probability of observing full data is bounded away from zero, i.e.,

\[ \pi(1_m) \geq \sigma > 0, \quad (2.2) \]

where \(1_m\) is the \(m\)-dimensional vector of 1’s. So \(R = 1_m\) means that we observe full data \(X = X_{(1_m)}\). It is obvious that \(\sum_r \pi(r) = 1\).

We will also assume that \(\pi(r)\) is unknown. It is easily seen that the final results still hold when \(\pi(r)\) is known by going through a simplified version of the derivations in this article.

The induced model for the observed data \((R, X_{(R)})\) is denoted as \(\mathcal{P} = \{P_{\theta,\eta,\pi} : \theta \in \Theta \subset \mathbb{R}^d, \eta \in \mathcal{H}\}\) where \(P_{\theta,\eta,\pi}\) is a distribution function with an additional nuisance parameter \(\pi\).

Let \(q_{\theta,\eta}\) be the density function of the probability measure \(Q_{\theta,\eta}\), and \(p_{\theta,\eta,\pi}\) the density function of the probability measure \(P_{\theta,\eta,\pi}\). By the MAR assumption in equation (2.1), we have the following relationship between the two density functions:

\[ p_{\theta,\eta,\pi}(r, x_{(r)}) = \pi(r) \int q_{\theta,\eta}(x) \prod_{j=1}^{m} (d\mu_j(x_j))^{1-r_j}, \quad (2.3) \]

where \(\mu_j\) are dominating measures for \(x_j, j = 1, \ldots, m\).

Our goal is to derive efficient score functions for \(\theta\) in model \(\mathcal{P}\) under different missing patterns: arbitrary missingness, monotonic missingness, and two-phase sampling design where some random variables are always observed and others are either observed or missing simultaneously. For arbitrary missingness, the patterns of 1’s and 0’s in vector \(r\) can be arbitrary. When we say monotonic missingness, we mean that \(r \in \{1_j : j = 1, \ldots, m\}\), where \(1_j\) are \(m\)-dimensional vectors with the first \(j\) components being all 1’s and the rest being all 0’s, i.e.,
\[ 1_j = (1, \ldots, 1, \underbrace{0, \ldots, 0}_j, \ldots, 0), \quad j = 1, \ldots, m. \] (2.4)

A natural example of monotonic missingness is the longitudinal study with dropouts. Sometimes monotonic missingness can be obtained by rearranging the order of random variables \( X_1, \ldots, X_m \). For example, if we put all the fully observed random variables in front of the variables with missing data in a two-phase sampling design, then the data structure becomes monotonically missing with \( r \in \{1_t, 1_m\} \), where \( t \) is a fixed integer. It is clearly seen that the two-phase sampling designs are special cases of monotonic missingness.

Now we introduce the other notation that we use in the paper. We refer to BKRW for definitions and detailed discussions.

**Full data model** \( \mathcal{Q} \):

1. \( \hat{\mathcal{Q}}_\eta^0 \): Tangent set for the nuisance parameter \( \eta \) in model \( \mathcal{Q} \).
2. \( \hat{\mathcal{Q}}_\eta \): Tangent space for the nuisance parameter \( \eta \) in model \( \mathcal{Q} \), which is the closed linear span of the tangent set \( \hat{\mathcal{Q}}_\eta^0 \).
3. \( \hat{\mathcal{Q}}_\eta^\perp \): Orthogonal complement of the nuisance tangent space \( \hat{\mathcal{Q}}_\eta \) with respect to \( L_2^0(\mathcal{Q}) \).
4. \( \hat{l}^0_\theta \): Score function for \( \theta \) in model \( \mathcal{Q} \).
5. \( l^0_\theta \): Efficient Score function for \( \theta \) in model \( \mathcal{Q} \).
6. \( \Psi^0_\theta \): The space of influence functions for any regular asymptotically linear estimators for \( \theta \) in model \( \mathcal{Q} \).
7. \( \langle \cdot, \cdot \rangle_0 \) and \( \| \cdot \|_0 \) are inner product in \( L_2(\mathcal{Q}) \) and \( L_2(\mathcal{Q}) \)-norm, respectively.

**Observed data model** \( \mathcal{P} \):

1. \( \hat{\mathcal{P}}_\eta^0 \): Tangent set for the nuisance parameter \( \eta \) in model \( \mathcal{P} \).
2. \( \mathcal{P}_{\eta, \pi}, \mathcal{P}_\eta, \) and \( \mathcal{P}_\pi \): Tangent spaces for the nuisance parameters \( (\eta, \pi), \eta, \) and \( \pi \) in model \( \mathcal{P} \).
3. $\mathcal{P}_{\eta,\pi} \perp$, $\hat{\mathcal{P}}_\eta$, and $\hat{\mathcal{P}}_\pi$: Orthogonal complements of the nuisance tangent spaces $\hat{\mathcal{P}}_{\eta,\pi}$, $\hat{\mathcal{P}}_\eta$, and $\hat{\mathcal{P}}_\pi$, respectively, with respect to $L_2^0(P)$.

4. $\hat{l}_\theta$: Score function for $\theta$ in model $\mathcal{P}$.

5. $l_\theta^*$: Efficient Score for $\theta$ in model $\mathcal{P}$.

6. $\Psi_{\theta}$: The space of influence functions for any regular asymptotically linear estimators for $\theta$ in model $\mathcal{P}$.

7. $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ are inner product in $L_2(P)$ and $L_2(P)$-norm, respectively.

According to BKRW, the efficient score function $l_\theta^*$ can be written as

$$l_\theta^* = \hat{l}_\theta - \Pi(\hat{l}_\theta|\hat{\mathcal{P}}_{\eta,\pi}) = \Pi(\hat{l}_\theta|\hat{\mathcal{P}}_{\eta,\pi}).$$

Here $\Pi$ is the projection operator. The calculation of the above projection is often extremely difficult. RRZ and RR are able to relate $l_\theta^*$ to the full data efficient score function $l_{\theta}^{0*} = \hat{l}_\theta^0 - \Pi(\hat{l}_\theta^0|\hat{\mathcal{Q}}_\eta)$ that may be easily computed and thus make the calculation of $l_\theta^*$ possible.

Now we define the following three important operators that will be used throughout the derivations:

**Definition 2.1.**

1. For $g^0 \in L_2^0(Q)$, define $A: L_2^0(Q) \rightarrow L_2^0(P)$ by

$$A(g^0) \equiv E[g^0(X) | R, X(R)] = \sum_r I(R = r)E[g^0(X) | R = r, X(r)].$$

2. For $g^0 \in L_2^0(Q)$, define $U(g^0): L_2^0(Q) \rightarrow L_2^0(P)$ by

$$U(g^0) \equiv \frac{I(R = 1_m)}{\pi(1_m)} g^0.$$  

Note that $U$ is not well defined if $\pi(1_m)$ is not bounded away from 0.

3. For $g^0 \in L_2^0(Q)$ and $a \in L_2^0(P)$, define $V(g^0, a): L_2^0(Q) \times L_2^0(P) \rightarrow L_2^0(P)$ by

$$V(g^0, a) \equiv U(g^0) + a - \Pi[U(g^0) + a | \hat{\mathcal{P}}_\pi] = \Pi[U(g^0) + a | \hat{\mathcal{P}}_\pi].$$
The operator $\mathbf{A}$ makes nice connections between models $\mathcal{P}$ and $\mathcal{Q}$. The following properties of the operator $\mathbf{A}$ can be easily verified via direct calculations.

**Proposition 2.1.**

1. $\mathbf{A}(\dot{i}_\theta^0) = \dot{i}_\theta$ and $\mathbf{A}(\dot{i}_\eta^0) = \dot{i}_\eta$.

2. The adjoint $\mathbf{A}^T: L^2_0(\mathcal{P}) \rightarrow L^2_0(\mathcal{Q})$ of $\mathbf{A}$ is given by $\mathbf{A}^T(g) = E[g \mid \mathbf{X}]$ for $g \in L^2_0(\mathcal{P})$. It is obvious that $\mathbf{A}^T U(g^0) = g^0$.

3. $\mathbf{A}^T A(g^0) = \sum_r \pi(r) E[g^0(\mathbf{X}) \mid \mathbf{R} = r, \mathbf{X}(r)]$. Notice that $\mathbf{A}^T A$ is self-adjoint.

### 3 Main Results

We introduce the fundamental results of efficient score calculations in RRZ and RR here in this section. Detailed proofs are deferred to Section 5.

#### 3.1 Arbitrary Missingness

The Proposition 8.1 in RRZ includes the fundamental results for models with data missing in arbitrary patterns. We first define $\mathcal{N}(\mathbf{A}^T)$ as the null space of $\mathbf{A}^T$, i.e.,

$$\mathcal{N}(\mathbf{A}^T) \equiv \{ a(\mathbf{R}, \mathbf{X}(\mathbf{R})) \in \mathbb{R}^k : E[a \mid \mathbf{X}] = 0, \ a \in L^2_0(\mathcal{P}) \},$$

the space of functions of the observed data with conditional mean 0 given full data $\mathbf{X}$. For the two-phase sampling designs studied by Nan, Emond, and Wellner (2000), it reduces to their $\mathcal{J}^{(2)}$. By rearranging the material of Proposition 8.1 in RRZ to emphasize the calculation of efficient score function, we obtain the following theorem:

**Theorem 3.1.** The efficient score function for $\theta$ in model $\mathcal{P}$ has the following form

$$l^*_\theta = \mathbf{U}(h^0) - \Pi \left( \mathbf{U}(h^0) \bigg| \mathcal{N}(\mathbf{A}^T) \right) = \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1}(h^0), \quad (3.1)$$

where $h^0$ is the unique function in $\dot{\mathcal{Q}}^\perp_\eta$ satisfying the following operator equation

$$\Pi \left( \mathbf{A}^T \mathbf{A}^{-1}(h^0) \mid \dot{\mathcal{Q}}^\perp_\eta \right) = l^*_\theta. \quad (3.2)$$
Since $h^0$ is an estimating function of complete data, by the definition of operator $U$ we know that the leading term on the right hand side of equation (3.1) is a Horvitz and Thompson type of inverse sampling probability weighted estimating function of completely observed data (see e.g. Horvitz and Thompson (1952)).

We can see from Theorem 3.1 that for any specific full data model $Q$, we will be able to derive the efficient score function for the missing data model $P$ once we have the following three ingredients from model $Q$: (1) The efficient score function $l^*_0 \theta$; (2) The characterization of space $\hat{\mathcal{Q}}_\eta^\perp$; and (3) The calculation of projecting functions in $L^2_0(Q)$ to space $\hat{\mathcal{Q}}_\eta^\perp$. However, an explicit form of $(A^T A)^{-1}$ is not available for arbitrary missing patterns. We will see in the next subsection that the explicit form of $(A^T A)^{-1}$ exists for monotonic missingness.

### 3.2 Monotonic Missingness

We know that for monotonic missingness, we have $r \in \{1_j : j = 1, \ldots, m\}$. If $r = 1_k$, then $X(r) = (X_1, X_2, \ldots, X_k)$. Instead of using the whole vector $R$ or $r$, we can actually work on individual observing indicators for every random variables in $X$. Let $R_k$ be the $k$-th element of $R$ and $R_0 = 1$ for convenience. One fact used constantly is that $R_k = 1$ implies $R_j = 1$ whenever $k \geq j$. We define

$$\pi_k = P(R_k = 1 \mid R_{k-1} = 1, X(1_{k-1}))$$

and

$$\bar{\pi}_k = \prod_{j=1}^{k} \pi_j .$$

Let $\pi_0 = 1$ and $\bar{\pi}_0 = 1$. Then we have the following result for monotonic missingness from Proposition 8.2 in RRZ:

**Theorem 3.2.** When data are missing in monotonic patterns, the efficient score function for $\theta$ in model $P$ has the following form

$$l^*_0 = \frac{R_m}{\bar{\pi}_m} h^0 - \sum_{k=1}^{m} \frac{R_k - \pi_k R_{k-1}}{\bar{\pi}_k} E(h^0 \mid X(1_{k-1})) ,$$

where $h^0$ is the unique function in $\hat{\mathcal{Q}}_\eta^\perp$ satisfying the following operator equation

$$\Pi \left( \frac{1}{\bar{\pi}_m} h^0 - \sum_{k=1}^{m} \frac{1 - \pi_k}{\bar{\pi}_k} E(h^0 \mid X(1_{k-1})) \bigg| \hat{\mathcal{Q}}_\eta^\perp \right) = l^*_0 .$$
Notice that \( I(R = 1_m) = R_m \), and from the identity (5.4) that we will show in Section 5 we have \( \pi(1_m) = \bar{\pi}_m \). So we see that the leading term on the right hand side of equation (3.3) is also an inverse sampling probability weighted estimating function of completely observed data as in Theorem 3.1.

### 3.3 Two-Phase Sampling Designs

Consider the two-phase sampling scheme where we have either \( R = 1_m \) or \( R = 1_t \) for a known integer \( t < m \). Hence \( \pi(1_t) = 1 - \pi(1_m) \), which means that \( \pi(1_m) \) is a function of \( X_{(1_t)} \), the always observed variables. Then we have the following theorem which is actually a corollary of Theorem 3.2:

**Theorem 3.3.** For two-phase sampling designs, the efficient score function for \( \theta \) in model \( \mathcal{P} \) has the following form

\[
l_\theta^* = \frac{I(R = 1_m) h^0}{\pi(1_m)} - \frac{I(R = 1_m) - \pi(1_m)}{\pi(1_m)} E(h^0 \mid X_{(1_t)}) ,
\]

where \( h^0 \) is the unique function in \( \hat{\mathcal{Q}}_\eta \) satisfying the following operator equation

\[
\Pi \left( \frac{1}{\pi(1_m)} h^0 - \frac{1 - \pi(1_m)}{\pi(1_m)} E(h^0 \mid X_{(1_t)}) \right) \hat{\mathcal{Q}}_\eta = l_\theta^* .
\]

This is the same result as that in **Nan, Emond, and Wellner (2000)** which was derived independently using alternative method.

### 4 An Illustration of Applications

It is not unusual in medical research that the outcome variables of interest are difficult or expensive to obtain. Often in these settings, surrogate outcome variables can be easily ascertained (see e.g. **Pepe (1992)**). Suppose \( Y \) is the outcome of interest that is not always observable. Let \( Z \) be a surrogate variable of \( Y \), which is always available. The association of \( Y \) and \( d \)-dimensional covariate \( X \) (always observable) is of the major interest. We assume
that the conditional expectation of \( Y \) given \( X \) is known up to a parameter \( \theta \in \mathbb{R}^d \), i.e.

\[
E[Y|X = x] = g(x; \theta),
\]

(4.1)

where \( g(\cdot) \) is a known function. Let \( \epsilon = Y - g(X; \theta) \), then \( E[\epsilon|X] = 0 \).

Model (4.1) is semiparametric in the sense that there are three unknown functions in the underlying density function of \((Z, Y, X)\): \( f_1 \), the conditional density function of \( Z \) given \((Y, X)\); \( f_2 \), the conditional density function of \( Y \), or equivalently \( \epsilon \), given \( X \); and \( f_3 \), the marginal density function of \( X \). We can write the full data density functions of the form

\[
q(z, y, x; \theta, f_1, f_2, f_3) = f_1(z|y, x)f_2(y - g(x; \theta)|x)f_3(x).
\]

(4.2)

The regression parameter \( \theta \) is the parameter of interest, and the nuisance parameter \( \eta \) is a vector of the densities \((f_1, f_2, f_3)\). No assumption is made for \( \eta = (f_1, f_2, f_3) \) other than that they are density functions.

Let \( R \) be the observing indicator taking value either 1 when \( Y \) is observed or 0 when \( Y \) is missing. Let \( \pi(z, x) = P(R = 1|Z = z, X = x) \). Then the observed data density function is

\[
p(z, ry, x, r; \theta, f_1, f_2, f_3) = \left\{ \pi(z, x)q(z, y, x; \theta, f_1, f_2, f_3) \right\}^r \cdot \left\{ (1 - \pi(z, x)) \int q(z, y, x; \theta, f_1, f_2, f_3)\nu(dy) \right\}^{1-r},
\]

(4.3)

where \( r \in \{0, 1\} \) and \( \nu \) is a dominating measure. Obviously, this is a two-phase design problem.

It can be verified easily that for model (4.2), the three components of the tangent space \( \dot{Q}_1, \dot{Q}_2, \) and \( \dot{Q}_3 \) corresponding to \( f_1, f_2, \) and \( f_3 \), respectively, are mutually orthogonal. Direct calculations show that these three components have the following structures:

\[
\dot{Q}_1 = \{a_1(Z, Y, X) : E[a_1|Y, X] = 0, Ea_1^2 < \infty \},
\]

(4.4)

\[
\dot{Q}_2 = \{a_2(Y, X) : E[a_2|X] = 0, E\epsilon a_2[X] = 0, Ea_2^2 < \infty \},
\]

(4.5)

\[
\dot{Q}_3 = \{a_3(X) : Ea_3 = 0, Ea_3^2 < \infty \}.
\]

(4.6)

The nuisance tangent space is thus the sum of the three: \( \dot{Q}_\eta = \dot{Q}_1 + \dot{Q}_2 + \dot{Q}_3 \), according to BKRW. The equality (4.5) may not be exactly true since that \( \dot{Q}_2 \) contains the right side
in (4.5) is hard to prove. However, the equality assumption works for our purpose. See the discussion in BKRW, page 76.

The following Theorems 4.1 and 4.2 supply all three ingredients of Theorem 3.3, which is reduced from Theorem 3.1 for two-phase sampling designs. Theorem 4.3 gives us the efficient score for $\theta$ in model (4.3) based on Theorem 3.3 and Theorems 4.1 and 4.2. Similar results in Theorems 4.1 and 4.2 can be found in Chamberlain (1987), RRZ, and van der Vaart (1998). The proofs can also be found in Nan, Emond, and Wellner (2000). Notice that $\nabla \theta \equiv \partial / \partial \theta$.

**Theorem 4.1.** Suppose model $Q \in Q$ is as described in (4.2). Then for any $b \in L_2^0(Q)$,

$$\Pi(b|\hat{Q}_n^+) = \frac{E[b(Z, Y, X)e|X]}{E[e^2|X]} \epsilon.$$  \hspace{1cm} (4.7)

Thus the efficient score for $\theta$ in the full model is

$$l^*_{\theta} = \Pi(\hat{l}_{\theta}^0|\hat{Q}_n^+) = \frac{\nabla g(X; \theta)}{E[e^2|X]} \epsilon,$$  \hspace{1cm} (4.8)

where $\hat{l}_{\theta}^0$ is the usual score function for $\theta$ in the full model.

**Proof:** Let

$$r_b = \frac{E[eb(Z, Y, X)|X]}{E[e^2|X]} \epsilon.$$

To prove (4.7), we will show that $r_b \in \hat{Q}^+_n = (\hat{Q}_1 + \hat{Q}_2 + \hat{Q}_3)^\perp$ and that $b - r_b \in (\hat{Q}_1 + \hat{Q}_2 + \hat{Q}_3)$. For any $a_1 \in \hat{Q}_1$:

$$\langle r_b, a_1 \rangle_{L_2^0(Q)} = E(r_b a_1) = E \left\{ \frac{E(\epsilon b|X)E(\epsilon a_1|X)}{E[\epsilon^2|X]} \right\} = 0$$

by (4.4). For any $a_2 \in \hat{Q}_2$:

$$\langle r_b, a_2 \rangle_{L_2^0(Q)} = E(r_b a_2) = E \left\{ \frac{E(\epsilon b|X)E(\epsilon a_2|X)}{E[\epsilon^2|X]} \right\} = 0$$

by (4.5). And, for any $a_3 \in \hat{Q}_3$:

$$\langle r_b, a_3 \rangle_{L_2^0(Q)} = E(r_b a_3) = E \left\{ \frac{E(\epsilon b|X)a_3(X)E(\epsilon|X)}{E[\epsilon^2|X]} \right\} = 0$$

by (4.6).
since \( E(\epsilon|X) = 0 \). Hence \( r_b \in \hat{\mathcal{Q}}_\eta^\perp \).

Let \( b - r_b = b - E[b|X] - r_b + E[b|X] \). Since \( E\{b - E[b|X] - r_b|X\} = 0 \) and \( E\{(b - E[b|X] - r_b)\epsilon|X\} = 0 \), we know that \( b - E[b|X] - r_b \in \hat{\mathcal{Q}}_2 \). The other part has zero mean since \( b \in L_2^0(Q) \), so \( E[b|X] \in \hat{\mathcal{Q}}_3 \). Thus \( b - r_b \in (\hat{\mathcal{Q}}_2 + \hat{\mathcal{Q}}_3) \subset \hat{\mathcal{Q}}_\eta \), which shows the desired result.

The efficient score \( l_\theta^{*0} \) can be obtained via direct calculation from (4.7) and using the fact that \( E[-\epsilon(f_2^*/f_2)(\epsilon|X)|X] = 1 \). \( \square \)

**Theorem 4.2.** \( \hat{\mathcal{Q}}_\eta^\perp = \{\zeta(X)\epsilon : E[\zeta^2(X)\epsilon^2] < \infty\} \).

**Proof:** Take \( a_1 \in \hat{\mathcal{Q}}_1 \), \( a_2 \in \hat{\mathcal{Q}}_2 \), and \( a_3 \in \hat{\mathcal{Q}}_3 \). Then we have \( E[a_1\zeta(X)\epsilon|X] = 0 \), \( E[a_2\zeta(X)\epsilon|X] = 0 \), and \( E[a_3\zeta(X)\epsilon|X] = 0 \), as in the proof of Theorem 4.1, which shows \( \{\zeta(X)\epsilon : E[\epsilon^2\zeta^2(X)] < \infty\} \subset \hat{\mathcal{Q}}_\eta^\perp \). Equation (4.7) shows the reverse inclusion, since

\[
E \left\{ \frac{E^2(b|X)}{E^2(\epsilon^2|X)} \right\} \leq Eb^2 < \infty
\]

by the Cauchy inequality. \( \square \)

**Theorem 4.3.** The efficient score \( l_\theta^* \) for the observed model (4.3) is given by

\[
l_\theta^* = \frac{\nabla_\theta g(X;\theta)}{E\left[ \frac{1}{\pi} \epsilon^2 - \frac{1-\pi}{\pi} E^2(\epsilon|Z,X) \right]|X} \left\{ \frac{R}{\pi} Y - \frac{R - \pi}{\pi} E[Y|Z,X] - g(X;\theta) \right\}. \tag{4.9}
\]

**Proof:** From Theorem 3.3 and Theorem 4.2 we have

\[
\Pi \left( \frac{1}{\pi} \zeta \epsilon - \frac{1-\pi}{\pi} E[\zeta\epsilon|Z,X] \right) = l_\theta^{*0}.
\]

Applying Theorem 4.1 we obtain:

\[
\frac{\nabla_\theta g(X;\theta)}{E[\epsilon^2|X]} \epsilon = \frac{1}{E[\epsilon^2|X]} E \left[ \frac{1}{\pi} \zeta \epsilon^2 - \frac{1-\pi}{\pi} E(\zeta\epsilon|Z,X) \right]|X \epsilon
\]

\[
= \frac{1}{E[\epsilon^2|X]} E \left[ \frac{1}{\pi} \epsilon^2 - \frac{1-\pi}{\pi} E^2(\epsilon|Z,X) \right]|X \zeta \epsilon.
\]

Simplifying the above equality yields

\[
\zeta(X) = \frac{\nabla_\theta g(X;\theta)}{E \left[ \frac{1}{\pi} \epsilon^2 - \frac{1-\pi}{\pi} E^2(\epsilon|Z,X) \right]|X}.
\]
Thus from Theorem 3.3 we have

\[ l_{\theta}^* = \frac{R}{\pi} \zeta(X) \epsilon - \frac{R - \pi}{\pi} E[\epsilon \zeta(X)|Z, X] \]

\[ = \zeta(X) \left\{ \frac{R}{\pi} \epsilon - \frac{R - \pi}{\pi} E[\epsilon|Z, X] \right\} \]

\[ = \zeta(X) \left\{ \frac{R}{\pi} Y - \frac{R - \pi}{\pi} E[Y|Z, X] - g(X; \theta) \right\} , \]

which yields (4.9).

Let

\[ Y' = \frac{R}{\pi} Y - \frac{R - \pi}{\pi} E[Y|Z, X] . \] (4.10)

Using nested conditioning, we can easily verify that

\[ E[Y'|X] = E[Y|X] = g(X; \theta) \]

and

\[ E[(Y' - g(X; \theta))^2|X] = E \left[ \frac{1}{\pi} \epsilon^2 - \frac{1 - \pi}{\pi} E^2(\epsilon|Z, X) \right| X \right] . \]

Hence the efficient score \( l_{\theta}^* \) is actually the efficient score for the “full” data \((Y', X)\) applying “transformation” (4.10) to the response variable, i.e.,

\[ l_{\theta}^* = \nabla_{\theta} g(X; \theta) \frac{E[\epsilon'^2|X]^{-1}}{E[\epsilon'^2|X]} \epsilon' , \] (4.11)

where \( \epsilon' = Y' - g(X; \theta) \). So analyzing the observed data \((Z, RY, X, R)\) with the outcome \( Y \) missing at random and the availability of surrogate outcome \( Z \) is actually equivalent to analyzing the “full” data \((Y', X)\) with the same conditional mean structure as that of \((Y, X)\).

The interpretation of the parameter \( \theta \) does not change at all, even though the scale of \( Y' \) may not be the same as \( Y \). We refer Nan (2003) for detailed discussions of estimating \( \theta \) from equation (4.11). The proof of Theorem 4.3 is taken from the the Appendix of the same paper.

There are many applications of the main results in Section 3 in literature. Very often in practice the operator equation (3.2), or (3.4), or (3.6) is some kind of integral equation, so the efficient score usually does not have a simple closed form as (4.9). The computing of efficient estimates may have to involve solving integral equations. Among those applications, we refer Robins, Rotnitzky, and Zhao (1995) for longitudinal studies with dropouts, Holcroft, Rotnitzky, and Robins (1997) for multistage studies, Nan, Emond, and Wellner (2000) for classical and mean regression models with missing covariates, and Nan, Emond, and Wellner (2002) for Cox model with missing data.
5 Proofs of Main Results

5.1 Preliminaries

In this section we first introduce some preliminary results that we will use for the proofs of main results in Section 3 and for the proofs in Section 6. All these results appeared in RRZ and RR in a variety of ways.

For any (one-dimensional) regular parametric submodel \( \pi(R, X(R); \gamma) \) passing through the true parameter \( \pi = \pi(R, X(R)) \) at \( \gamma = 0 \), we can calculate the score operator for \( \pi \) as

\[
\dot{l}_\pi a = a(R, X(R)) \equiv \left( \frac{\partial \log \pi(R, X(R); \gamma)}{\partial \gamma} \right)_{\gamma=0}.
\]

Thus \( \dot{P}_\pi \subset [\dot{l}_\pi a] \) for all \( a \in L^0_2(P) \), where \([ \cdot ]\) means closed linear span.

**Lemma 5.1.** \( \dot{P}_\pi \subset N(A^T) \).

(Briefly described in the proof of Lemma 8.2 in RRZ)

**Proof:** For any regular parametric submodel \( \pi(R, X(R); \gamma) \) described above, we have

\[
E[\dot{l}_\pi a \mid X] = E \left[ \left( \frac{\partial \log \pi(R, X(R); \gamma)}{\partial \gamma} \right)_{\gamma=0} \mid X \right]
\]

\[
= \sum_r \left( \frac{\partial \log \pi(r, X(r); \gamma)}{\partial \gamma} \right)_{\gamma=0} P(R = r \mid X)
\]

\[
= \sum_r \frac{\pi(r, X(r); \gamma)}{\pi(r, X(r); \gamma)_{\gamma=0}} \pi(r, X(r))
\]

\[
= \sum_r \frac{\partial \pi(r, X(r); \gamma)}{\partial \gamma} \bigg|_{\gamma=0}
\]

\[
= \left( \frac{\partial}{\partial \gamma} \sum_r \pi(r, X(r); \gamma) \right)_{\gamma=0}
\]

\[
= 0. \quad \Box
\]

**Remark:** RRZ claimed equality of the two spaces when \( \pi \) is totally unspecified. We only show the inclusion in Lemma 5.1 since this is enough for the purpose.

**Lemma 5.2.** The operator \( A \) is a continuous linear operator, and satisfies \( \|A(g^0)\|^2 \geq \sigma \|g^0\|_0^2 \) for all \( g^0 \in L^0_2(Q) \). Hence \( A \) has a continuous inverse \( A^{-1} \).
Proof: Linearity follows from the definition of $A$. To show the continuity of $A$, we only need to show that $A$ is bounded (BKRW A.1.2). For all $g^0 \in L^2_0(Q)$,

$$\|A(g^0)\|^2 = E \{(A(g^0))^2\} = E \left\{ \left( E(g^0 \mid R, X_{(R)}) \right)^2 \right\} \leq E \{E(g^0)^2 \mid R, X_{(R)}\} = \|g^0\|^2_0$$

So the norm of the operator $A$ is bounded by 1, which is a finite number.

As for the second part,

$$\|A(g^0)\|^2 = \| \sum_r I(R = r) E[g^0 \mid R = r, X_{(r)}] \|^2 = \sum_r \|I(R = r) E[g^0 \mid R = r, X_{(r)}]\|^2 \geq \|I(R = 1_m) E[g^0 \mid X]\|^2 = E \left[ I(R = 1_m)(g^0)^2 \right] = \pi(1_m) \|g^0\|^2_0 \geq \sigma \|g^0\|^2_0.$$

The invertibility of $A$ follows from BKRW, A.1.7 (a consequence of the inverse mapping theorem). \qed

Lemma 5.3 $\hat{P}_\eta = \{ A(a) = E[a \mid R, X_{(R)}] : a \in \hat{Q}_\eta \}$, or in simplified notation, $\hat{P}_\eta = A\hat{Q}_\eta$.

(Briefly described in the proof of Lemma A.4 in RRZ)

Proof: From lemma 5.2, we know that $A$ has a continuous inverse $A^{-1}$ from $R(A)$ to $D(A)$, the range and domain of $A$. Since $\hat{Q}_\eta$ is a closed linear space, $\{ A(a) : a \in \hat{Q}_\eta \}$ is a closed set and hence a closed linear space. By score calculation (BKRW A.5.5), we have the nuisance tangent set in model $P$ that satisfies $\hat{P}^0_\eta = \{ A(a) : a \in \hat{Q}^0_\eta \} \subset \{ A(a) : a \in \hat{Q}_\eta \}$. Thus the nuisance tangent space, the closed linear span of $\hat{P}^0_\eta$, satisfies $\hat{P}_\eta \subset \{ A(a) : a \in \hat{Q}_\eta \}$. As for the other direction, i.e., $\hat{P}_\eta \supset \{ A(a) : a \in \hat{Q}_\eta \}$, we can quote the conclusion from BKRW, equation (1) on page 144. Yet we provide the following brief proof. We need to show that $A[\hat{Q}_\eta] \subset [A\hat{Q}_\eta] \equiv \hat{P}_\eta$ where $[\cdot]$ means the closed linear span. Since $A\hat{Q}_\eta \subset [A\hat{Q}_\eta]$, we only need to show that for any limit point $a$ in $[\hat{Q}_\eta]$, $Aa \in [A\hat{Q}_\eta]$. Since $a$ is a limit point in $[\hat{Q}_\eta]$, 

\[\text{(Appeared in the proof of Lemma A.4 in RRZ)}\]
there exists a sequence \( a_n \in \langle \hat{Q}_\eta^0 \rangle \) such that \( \lim_n a_n = a \) where \( \langle \cdot \rangle \) denotes the linear span. By the continuity of \( A \), we have \( \lim_n Aa_n = Aa \). Since \( Aa_n \in \langle A \hat{Q}_\eta^0 \rangle \), we have \( Aa \in [A \hat{Q}_\eta^0] \). \( \square \)

**Corollary 5.1.** \( \hat{P}_\eta \perp \mathcal{N}(A^T) \).

*(Part of Lemma A.3 in RRZ)*

**Proof:** For all \( a \in \mathcal{N}(A^T) \) and all \( g^0 \in \hat{Q}_\eta \), it is easy to see that \( \langle A(g^0), a \rangle = \langle g^0, A^T(a) \rangle_0 = \langle g^0, 0 \rangle_0 = 0 \). We obtain the conclusion from Lemma 5.3. \( \square \)

From Lemma 5.1 and Corollary 5.1 we know that \( \hat{P}_\pi \) and \( \hat{P}_\eta \) are orthogonal, so we can write the nuisance tangent space for model \( \mathcal{P} \) as the following:

**Corollary 5.2.** \( \hat{P}_{\eta,\pi} = \hat{P}_\eta + \hat{P}_\pi \).

*(Part of Lemma A.3 in RRZ)*

**Corollary 5.3** Let \( g \in L_2^0(P) \). Then \( g \in \hat{P}_\eta^\bot \) if and only if \( A^T g \in \hat{Q}_\eta^\bot \).

*(Lemma A.6 in RRZ)*

**Proof:** This is obvious since \( \hat{P}_\eta = A \hat{Q}_\eta \) from Lemma 5.3 and for any \( h^0 \in \hat{Q}_\eta \), \( \langle A^T g, h^0 \rangle_0 = \langle g, A h^0 \rangle = 0 \). \( \square \)

**Lemma 5.4** Operator \( A^T A \) is invertible.

*(Appeared in the proof of Proposition 8.1 part d in RRZ)*

**Proof:** Write \( A^T A = I - [I - A^T A] \), where \( I \) is the identity operator. To see that \( A^T A \) is invertible, we need only to show that \( \|I - A^T A\|_0 < 1 \).

\[
\|I - A^T A\|_0^2 = \sup_{\|g^0\|=1} \langle [I - A^T A](g^0), [I - A^T A](g^0) \rangle_0 \\
= \sup_{\|g^0\|=1} \langle g^0, [I - A^T A](g^0) \rangle_0 \\
= 1 - \inf_{\|g^0\|=1} \langle g^0, A^T A(g^0) \rangle_0 \\
= 1 - \inf_{\|g^0\|=1} \langle A(g^0), A(g^0) \rangle \\
\leq 1 - \inf_{\|g^0\|=1} \sigma \langle g^0 \rangle_0^2 \\
< 1 - \sigma ,
\]
where the second equality holds because $I - A^T A$ is self-adjoint (see e.g. CONWAY (1990), Proposition 2.13 on page 34).

\[\]  

5.2 Proof of Theorem 3.1

**Proof:** Define $g^0 = \hat{h}_0^\theta - a^0$, where $a^0 \in \hat{Q}_\eta$ satisfies $A(a^0) = \Pi(\hat{l}_\eta | \hat{P}_\eta)$. Note that $a^0$ is unique by Lemma 5.2. Let $h^0 = A^T g^0$. Since $\hat{l}_\theta = A(\hat{l}_0^\theta)$ by Proposition 2.1, we have $\hat{l}_\theta \perp N(A^T)$ because for all $a \in N(A^T)$ we have $\langle A(\hat{l}_0^\theta), a \rangle = \langle l_0^\theta, A^T a \rangle = 0$. Thus we have $\hat{l}_\theta \perp \hat{P}_\pi$ by Lemma 5.1. Then by Corollary 5.2 we have

$$A(A^T A)^{-1}(h^0) = A(g^0) = A\hat{l}_0^\theta - Aa_0^\theta = \hat{l}_\theta - \Pi(\hat{l}_\theta | \hat{P}_\eta) = \hat{l}_\theta - \Pi[\hat{l}_\theta | \hat{P}_{\eta, \pi}] = l_\theta^* .$$

The other equality in equation (3.1) can be argued as follows. Since for all $a \in N(A^T)$, we have $\langle A(A^T A)^{-1}(h^0), a \rangle = \langle (A^T A)^{-1}(h^0), A^T a \rangle = 0$. Thus $A(A^T A)^{-1}(h^0) \perp N(A^T)$. Now we need to only show that $U(h^0) - A(A^T A)^{-1}(h^0) \in N(A^T)$, which follows from

$$A^T \left\{ U(h^0) - A(A^T A)^{-1}(h^0) \right\} = A^T U(h^0) - A^T A(A^T A)^{-1}(h^0) = h^0 - h^0 = 0 .$$

Thus $\Pi \left( U(h^0) \mid N(A^T) \right) = U(h^0) - A(A^T A)^{-1}(h^0)$.

It can be shown that $h^0 \in \hat{Q}_\eta^\perp$, since for all $b^0 \in \hat{Q}_\eta$ we have $A b^0 \in \hat{P}_\eta$ by Lemma 5.3 and thus $\langle h^0, b^0 \rangle_0 = \langle A^T A g^0, b^0 \rangle_0 = \langle A g^0, A b^0 \rangle = \langle l_\theta^*, A b^0 \rangle = 0$. Equation (3.2) can be obtained by the following calculation:

$$\Pi \left( (A^T A)^{-1}(h^0) \mid \hat{Q}_\eta^\perp \right) = \Pi \left( g^0 \mid \hat{Q}_\eta^\perp \right) = \Pi \left( \hat{l}_0^\theta - a^0 \mid \hat{Q}_\eta^\perp \right) = \Pi \left( \hat{l}_0^\theta \mid \hat{Q}_\eta^\perp \right) = l_\theta^* ,$$

since $a^0 \in \hat{Q}_\eta$.

Since $A^T A$ is a one-to-one mapping, uniqueness follows if we can prove that $g^0 = (A^T A)^{-1}(h^0)$ is the unique solution to (3.2). Suppose both $g_1^0 = (A^T A)^{-1}(h_1^0)$ and $g_2^0 = (A^T A)^{-1}(h_2^0)$ satisfy equation (3.2), where $h_1^0, h_2^0 \in \hat{Q}_\eta^\perp$. Let $\Delta g^0 = g_1^0 - g_2^0$. Thus $\Delta g^0 = (A^T A)^{-1}(h_1^0 - h_2^0)$ and $A^T A(\Delta g^0) = h_1^0 - h_2^0 \in \hat{Q}_\eta^\perp$. But $\Delta g^0 \perp \hat{Q}_\eta^\perp$ since $\Pi(\Delta g^0 \mid \hat{Q}_\eta) = \Pi(g_1^0 \mid \hat{Q}_\eta) - \Pi(g_2^0 \mid \hat{Q}_\eta) = \hat{l}_\theta^0 - \hat{l}_\theta^0 = 0$. Then we have $0 = \langle A^T A(\Delta g^0), \Delta g^0 \rangle_0 = \langle A(\Delta g^0), A(\Delta g^0) \rangle \geq \sigma \| \Delta g^0 \|_0^2$ by Lemma 5.2. Thus we must have $\Delta g^0 = 0$ with probability 1. \[\]
5.3 Proof of Theorem 3.2

The proof of Theorem 3.2 will involve a lot of algebra for the $\pi$’s defined in Subsection 3.2. From the MAR assumption in (2.1) we have

$$P(R_k = 1 \mid \mathbf{X}) = 1 - \sum_{j<k} P(R = 1_j \mid \mathbf{X})$$

$$= 1 - \sum_{j<k} P(R = 1_j \mid \mathbf{X}_{(1_{k-1})})$$

$$= P(R_k = 1 \mid \mathbf{X}_{(1_{k-1})}) . \quad (5.1)$$

Hence,

$$P(R_k = 1 \mid R_{k-1} = 1, \mathbf{X}) = \frac{P(R_k = 1 \mid \mathbf{X})}{P(R_{k-1} = 1 \mid \mathbf{X})}$$

$$= \frac{P(R_k = 1 \mid \mathbf{X}_{(1_{k-1})})}{P(R_{k-1} = 1 \mid \mathbf{X}_{(1_{k-1})})}$$

$$= \pi_k , \quad (5.2)$$

and

$$P(R_{k+t} = 1 \mid R_{k-1} = 1, \mathbf{X}) = P(R_{k+t} = 1, R_{k+t-1} = 1, \ldots, R_k = 1 \mid R_{k-1} = 1, \mathbf{X})$$

$$= P(R_{k+t} = 1 \mid R_{k+t-1} = 1, \mathbf{X}) \cdots P(R_k = 1 \mid R_{k-1} = 1, \mathbf{X})$$

$$= \pi_{k+t} \cdots \pi_k$$

$$= \frac{\bar{\pi}_{k+t}}{\bar{\pi}_{k-1}} . \quad (5.3)$$

From (5.2) we also have

$$P(R_k = 1 \mid \mathbf{X}) = \pi_k P(R_{k-1} = 1 \mid \mathbf{X}) = \bar{\pi}_k . \quad (5.4)$$

Thus for $j < m$ we obtain

$$\pi(1_j) = P(R_{j+1} = 0, R_j = 1 \mid \mathbf{X})$$

$$= P(R_{j+1} = 0 \mid R_j = 1, \mathbf{X}) P(R_j = 1 \mid \mathbf{X})$$

$$= (1 - \pi_{j+1}) \bar{\pi}_j$$

$$= \bar{\pi}_j - \bar{\pi}_{j+1} . \quad (5.5)$$
Now we show the following two additional identities that we will apply to the proofs in this subsection.

**Lemma 5.5.** For any \( l \leq m \),

\[
\sum_{k=1}^{l} \frac{1 - \pi_k}{\bar{\pi}_k} + 1 = \frac{1}{\bar{\pi}_l}.
\]

**Proof:** When \( l = 1 \), we have

\[
\sum_{k=1}^{l} \frac{1 - \pi_k}{\bar{\pi}_k} + 1 = \frac{1 - \pi_1}{\bar{\pi}_1} + 1 = 0.
\]

When \( l > 1 \), we have

\[
\sum_{k=1}^{l} \frac{1 - \pi_k}{\bar{\pi}_k} + 1 - \frac{1}{\bar{\pi}_l} = \sum_{k=1}^{l-1} \frac{1 - \pi_k}{\bar{\pi}_k} + 1 - \frac{1}{\bar{\pi}_{l-1}} = \ldots
\]

\[
= \sum_{k=1}^{1} \frac{1 - \pi_k}{\bar{\pi}_k} + 1 - \frac{1}{\bar{\pi}_1}
\]

\[
= \frac{1 - \pi_1}{\bar{\pi}_1} + 1 - \frac{1}{\bar{\pi}_1} = 0.
\]

\[\square\]

**Corollary 5.4** For any \( 1 \leq l \leq m \),

\[
\sum_{k=l}^{m} \frac{1 - \pi_k}{\bar{\pi}_k} = \frac{1}{\bar{\pi}_m} - \frac{1}{\bar{\pi}_{l-1}}.
\]

**Proof:** By Lemma 5.5, we have

\[
\sum_{k=l}^{m} \frac{1 - \pi_k}{\bar{\pi}_k} = \sum_{k=1}^{m} \frac{1 - \pi_k}{\bar{\pi}_k} - \sum_{k=1}^{l-1} \frac{1 - \pi_k}{\bar{\pi}_k} = \left( \frac{1}{\bar{\pi}_m} - 1 \right) - \left( \frac{1}{\bar{\pi}_{l-1}} - 1 \right) = \frac{1}{\bar{\pi}_m} - \frac{1}{\bar{\pi}_{l-1}}.
\]

\[\square\]

Theorem 3.2 is a direct consequence of Theorem 3.1 once we obtain the following calculations that are based on Proposition 8.2 in RRZ. We use Lemma 5.5 and Corollary 5.4 several times without referring in the proof of the following Proposition 5.1.

**Proposition 5.1** For monotonic missingness we have
1. \( A(h^0) = R_m h^0 + \sum_{k=1}^{m} (R_{k-1} - R_k) E(h^0 | X_{(1_{k-1})}) \).

2. \( A^T A(h^0) = \tilde{\pi}_m h^0 + \sum_{k=1}^{m} (1 - \pi_k) \tilde{\pi}_{k-1} E(h^0 | X_{(1_{k-1})}) \).

3. \( (A^T A)^{-1}(h^0) = \frac{1}{\tilde{\pi}_m} h^0 - \sum_{k=1}^{m} \left( 1 - \frac{\pi_k}{\tilde{\pi}_k} \right) h^0 E(h^0 | X_{(1_{k-1})}) \).

4. \( A(A^T A)^{-1}(h^0) = U h^0 - \sum_{k=1}^{m} R_k \pi_k R_k - 1 \bar{\pi}_{k-1} E(h^0 | X_{(1_{k-1})}) \).

**Proof:** 1. First we show that
\[
E(h^0 | R = r, X_{(r)}) = E(h^0 | X_{(r)}) . \tag{5.6}
\]
This holds since
\[
f(X | R = r, X_{(r)}) = \frac{f(X, R = r, X_{(r)})}{f(R = r, X_{(r)})} = \frac{P(R = r | X, X_{(r)}) f(X, X_{(r)})}{P(R = r | X_{(r)}) f(X_{(r)})} = f(X | X_{(r)}) ,
\]
here \( f \) denotes density function. Also notice that \( R_{k-1} - R_k = I(R = 1_{k-1}) \), then by the definition of \( A \), we have the conclusion.

2. Since \( A^T A(h^0) = E(A(h^0) | X) \), we take the conditional expectation of the righthand side of part 1 given \( X \). Notice that \( \tilde{\pi}_m = \pi(1_m) \), and by (5.5) we have
\[
E(R_{k-1} - R_k | X) = P(R = 1_{k-1} | X) = \pi(1_{k-1}) = (1 - \pi_k) \bar{\pi}_{k-1} .
\]

3. Let
\[
B(h^0) = \frac{1}{\tilde{\pi}_m} h^0 - \sum_{k=1}^{m} \frac{1 - \pi_k}{\tilde{\pi}_k} E(h^0 | X_{(1_{k-1})}) .
\]
We want to show that \( B = (A^T A)^{-1} \), i.e., we want verify both \( A^T A B(h^0) = h^0 \) and \( B A^T A(h^0) = h^0 \).

From part 2 we have
\[
A^T A B(h^0) = \tilde{\pi}_m B(h^0) + \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_{k-1} E(B(h^0) | X_{(1_{k-1})}) . \tag{5.7}
\]
By the above definition of $\mathbf{B}$, the first term of RHS of (5.7) can be written as

$$\bar{\pi}_m \mathbf{B}(h^0) = h^0 - \sum_{k=1}^{m} \frac{\bar{\pi}_m (1 - \pi_k)}{\bar{\pi}_k} E(h^0 | \mathbf{X}_{(1_{k-1})}) ,$$  \hspace{1cm} (5.8)

and the second term is

$$\sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_{k-1} E(\mathbf{B}(h^0) | \mathbf{X}_{(1_{k-1})})$$

$$= \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_{k-1} E(h^0 / \bar{\pi}_m | \mathbf{X}_{(1_{k-1})})$$

$$- \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_{k-1} \left[ \sum_{j=1}^{m} (1 - \pi_j) \bar{\pi}_j^{-1} \{ E(h^0 | \mathbf{X}_{1_{j-1}}) \} \right] \mathbf{X}_{(1_{k-1})}$$

$$= \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_{k-1} E(h^0 / \bar{\pi}_m | \mathbf{X}_{(1_{k-1})}) - \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_{k-1} \left\{ \sum_{j=1}^{k} (1 - \pi_j) \bar{\pi}_j^{-1} E(h^0 | \mathbf{X}_{1_{j-1}}) \right\}$$

$$- \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_{k-1} E \left[ \sum_{j=k+1}^{m} (1 - \pi_j) \bar{\pi}_j^{-1} h^0 \right] \mathbf{X}_{(1_{k-1})}$$

$$= \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_{k-1} E(h^0 / \bar{\pi}_m | \mathbf{X}_{(1_{k-1})}) - \sum_{k=1}^{m} \sum_{j=1}^{k} (1 - \pi_k) \bar{\pi}_{k-1} (1 - \pi_j) \bar{\pi}_j^{-1} E(h^0 | \mathbf{X}_{1_{j-1}})$$

$$- \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_{k-1} E \left[ \left( \frac{1}{\bar{\pi}_m} - \frac{1}{\bar{\pi}_k} \right) h^0 \right] \mathbf{X}_{(1_{k-1})}$$

$$= \sum_{j=1}^{m} (1 - \pi_j) \bar{\pi}_j^{-1} E(h^0 | \mathbf{X}_{1_{j-1}}) \left\{ \sum_{k=j}^{m} (1 - \pi_k) \bar{\pi}_{k-1} \right\}$$

$$+ \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_{k-1} E \left[ \frac{1}{\bar{\pi}_k} h^0 \right] \mathbf{X}_{(1_{k-1})}$$

$$= \sum_{j=1}^{m} (1 - \pi_j) \bar{\pi}_j^{-1} E(h^0 | \mathbf{X}_{1_{j-1}}) (\bar{\pi}_{j-1} - \bar{\pi}_m) + \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_{k-1} E \left[ \frac{1}{\bar{\pi}_k} h^0 \right] \mathbf{X}_{(1_{k-1})}$$

$$= \sum_{j=1}^{m} (1 - \pi_j) \bar{\pi}_j^{-1} E(h^0 | \mathbf{X}_{1_{j-1}}) + \sum_{j=1}^{m} \bar{\pi}_m (1 - \pi_j) \bar{\pi}_j^{-1} E(h^0 | \mathbf{X}_{1_{j-1}})$$

$$+ \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_{k-1} E \left[ h^0 \right] \mathbf{X}_{(1_{k-1})}$$

$$= \sum_{j=1}^{m} \bar{\pi}_m (1 - \pi_j) \bar{\pi}_j^{-1} E(h^0 | \mathbf{X}_{1_{j-1}}) .$$  \hspace{1cm} (5.9)

Now combine (5.8) and (5.9) into (5.7), we have $\mathbf{A}^T \mathbf{A} \mathbf{B}(h^0) = h^0$.  

20
By the definition of $B$ we have

$$BA^T A(h^0) = A^T A(h^0) / \bar{\pi}_m - \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_k^{-1} E(A^T A(h^0) | X_{(1_{k-1})}) .$$

(5.10)

From part 2, the first term of RHS of (5.10) is

$$A^T A(h^0) / \bar{\pi}_m = h^0 + \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_m^{-1} E(h^0 | X_{(1_{k-1})}) ,$$

(5.11)

and the second term, without the minus sign, is

$$\sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_k^{-1} E(A^T A(h^0) | X_{(1_{k-1})})$$

$$= \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_k^{-1} E \left[ \bar{\pi}_m h^0 + \sum_{j=1}^{m} (1 - \pi_j) \bar{\pi}_{j-1} E(h^0 | X_{1_{j-1}}) \right] \ X_{(1_{k-1})}$$

$$= \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_k^{-1} E(\bar{\pi}_m h^0 | X_{(1_{k-1})}) + \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_k^{-1} \left\{ \sum_{j=1}^{k} (1 - \pi_j) \bar{\pi}_{j-1} E(h^0 | X_{1_{j-1}}) \right\}$$

$$+ \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_k^{-1} \left\{ E \left[ \sum_{j=k+1}^{m} (1 - \pi_j) \bar{\pi}_{j-1} h^0 \right] X_{(1_{k-1})} \right\}$$

$$= \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_k^{-1} E(\bar{\pi}_m h^0 | X_{(1_{k-1})}) + \sum_{k=1}^{m} \sum_{j=1}^{k} (1 - \pi_k) \bar{\pi}_k^{-1} (1 - \pi_j) \bar{\pi}_{j-1} E(h^0 | X_{1_{j-1}})$$

$$+ \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_k^{-1} \left\{ E \left[ (\bar{\pi}_k - \bar{\pi}_m) h^0 \right] X_{(1_{k-1})} \right\}$$

$$= \sum_{j=1}^{m} (1 - \pi_j) \bar{\pi}_{j-1} E(h^0 | X_{1_{j-1}}) \left\{ \sum_{k=j}^{m} (1 - \pi_k) \bar{\pi}_k^{-1} \right\}$$

$$+ \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_k^{-1} E \left[ \bar{\pi}_k h^0 \right] X_{(1_{k-1})}$$

$$= \sum_{j=1}^{m} (1 - \pi_j) \bar{\pi}_{j-1} E(h^0 | X_{1_{j-1}}) \left\{ \frac{1}{\bar{\pi}_m} - \frac{1}{\bar{\pi}_{j-1}} \right\} + \sum_{k=1}^{m} (1 - \pi_k) \bar{\pi}_k^{-1} E \left[ \bar{\pi}_k h^0 \right] X_{(1_{k-1})}$$

$$= \sum_{j=1}^{m} (1 - \pi_j) \bar{\pi}_{j-1} \bar{\pi}_m^{-1} E(h^0 | X_{1_{j-1}}) - \sum_{j=1}^{m} (1 - \pi_j) E(h^0 | X_{1_{j-1}})$$

$$+ \sum_{k=1}^{m} (1 - \pi_k) E \left[ h^0 \right] X_{(1_{k-1})}$$

$$= \sum_{j=1}^{m} (1 - \pi_j) \bar{\pi}_{j-1} \bar{\pi}_m^{-1} E(h^0 | X_{1_{j-1}}) .$$

(5.12)
Now combine (5.11) and (5.12) into (5.10), we have
\[ \mathbf{B} \mathbf{A}^\top \mathbf{A}(h^0) = h^0. \]

4. From part 1 we have
\[
\mathbf{A} (\mathbf{A}^\top \mathbf{A})^{-1}(h^0) = R_m (\mathbf{A}^\top \mathbf{A})^{-1}(h^0) + \sum_{k=1}^{m} (R_{k-1} - R_k) E((\mathbf{A}^\top \mathbf{A})^{-1}(h^0) \mid \mathbf{X}_{(1_{k-1})}).
\]  

(5.13)

From part 3, the first term of RHS of (5.13) can be written as
\[
R_m (\mathbf{A}^\top \mathbf{A})^{-1}(h^0) = R_m h^0 / \bar{\pi}_m - \sum_{k=1}^{m} R_m (1 - \pi_j) \bar{\pi}_j^{-1} E(h^0 \mid \mathbf{X}_{1_{j-1}}),
\]

and the second term is
\[
\sum_{k=1}^{m} (R_{k-1} - R_k) E((\mathbf{A}^\top \mathbf{A})^{-1}(h^0) \mid \mathbf{X}_{(1_{k-1})})
\]
\[
= \sum_{k=1}^{m} (R_{k-1} - R_k) E\left\{ \frac{h^0}{\bar{\pi}_m} - \sum_{j=1}^{m} (1 - \pi_j) \bar{\pi}_j^{-1} E(h^0 \mid \mathbf{X}_{1_{j-1}}) \left\mid \mathbf{X}_{(1_{k-1})} \right\}
\]
\[
= \sum_{k=1}^{m} (R_{k-1} - R_k) E\left\{ \frac{h^0}{\bar{\pi}_m} \left\mid \mathbf{X}_{(1_{k-1})} \right\}
\]
\[
- \sum_{k=1}^{m} (R_{k-1} - R_k) E\left\{ \sum_{j=1}^{m} (1 - \pi_j) \bar{\pi}_j^{-1} E(h^0 \mid \mathbf{X}_{1_{j-1}}) \left\mid \mathbf{X}_{(1_{k-1})} \right\}
\]
\[
= \sum_{k=1}^{m} (R_{k-1} - R_k) E\left\{ \frac{h^0}{\bar{\pi}_m} \left\mid \mathbf{X}_{(1_{k-1})} \right\}
\]
\[
- \sum_{k=1}^{m} (R_{k-1} - R_k) E\left\{ \left( \frac{1}{\bar{\pi}_m} - \frac{1}{\bar{\pi}_k} \right) h^0 \left\mid \mathbf{X}_{(1_{k-1})} \right\}
\]
\[
= \sum_{k=1}^{m} (R_{k-1} - R_k) E\left\{ \frac{h^0}{\bar{\pi}_m} \left\mid \mathbf{X}_{(1_{k-1})} \right\}
\]
\[
- \sum_{k=1}^{m} (R_{k-1} - R_k) E\left\{ (1 - \pi_j) \bar{\pi}_j^{-1} E(h^0 \mid \mathbf{X}_{1_{j-1}}) \left\mid \mathbf{X}_{(1_{k-1})} \right\}
\]
\[
- \sum_{k=1}^{m} (R_{k-1} - R_k) E\left\{ \frac{h^0}{\pi_j} \left\mid \mathbf{X}_{(1_{k-1})} \right\}
\]
\[
+ \sum_{k=1}^{m} (R_{k-1} - R_k) E\left\{ h^0 / \pi_k \left\mid \mathbf{X}_{(1_{k-1})} \right\}
\]
\[
= - \sum_{j=1}^{m} (1 - \pi_j) \bar{\pi}_j^{-1} E(h^0 \mid \mathbf{X}_{1_{j-1}}) (R_{j-1} - R_m)
\]
\[
+ \sum_{k=1}^{m} (R_{k-1} - R_k) E\left\{ h^0 / \pi_k \left\mid \mathbf{X}_{(1_{k-1})} \right\}.
\]  

(5.15)
Now combine (5.14) and (5.15) into (5.13), we have

\[
\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}(h^0) = R_m h^0 / \bar{\pi}_m + \sum_{k=1}^{m} \left\{ -R_m (1 - \pi_k) \bar{\pi}_k^{-1} - (1 - \pi_k) \bar{\pi}_k^{-1}(R_{k-1} - R_m) \right. \\
+ \left. (R_{k-1} - R_k) \bar{\pi}_k^{-1} \right\} E(h^0 | \mathbf{X}_{(1_{k-1})}) \\
= R_m h^0 / \bar{\pi}_m + \sum_{k=1}^{m} \left\{ -R_m (1 - \pi_k) - (1 - \pi_k) R_{k-1} \right. \\
+ \left. (1 - \pi_k) R_m + (R_{k-1} - R_k) \bar{\pi}_k^{-1} \right\} E(h^0 | \mathbf{X}_{(1_{k-1})}) \\
= R_m h^0 / \bar{\pi}_m + \sum_{k=1}^{m} \left\{ \pi_k R_{k-1} - R_k \bar{\pi}_k^{-1} \right\} E(h^0 | \mathbf{X}_{(1_{k-1})}) \\
= U h^0 - \sum_{k=1}^{m} \frac{R_k - \pi_k R_{k-1}}{\bar{\pi}_k} E(h^0 | \mathbf{X}_{(1_{k-1})}) .
\]

**Proof of Theorem 3.2:** Theorem 3.2 follows directly from Theorem 3.1 using the results in Proposition 5.1 part 3 and part 4.

5.4 **Proof of Theorem 3.3**

**Proof:** Since for two-phase designs we have

\[
R_k = \begin{cases} 
1, & \text{if } k \leq t \\
R_m, & \text{if } k \geq t + 1,
\end{cases} \quad (5.16)
\]

and

\[
\pi_k = \begin{cases} 
1, & \text{if } k \neq t + 1 \\
\pi(1_m), & \text{if } k = t + 1,
\end{cases} \quad (5.17)
\]

then by Proposition 5.1 part 3 and (5.17) we obtain

\[
(\mathbf{A}^T\mathbf{A})^{-1}(h^0) = h^0 / \bar{\pi}_m - (1 - \pi_{t+1}) \bar{\pi}_{t+1}^{-1} E(h^0 | \mathbf{X}_{(1_t)}) = \frac{1}{\bar{\pi}(1_m)} h^0 - \frac{1 - \pi(1_m)}{\bar{\pi}(1_m)} E(h^0 | \mathbf{X}_{(1_t)}),
\]

and by Proposition 5.1 part 4, (5.16) and (5.17),

\[
\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}(h^0) = U h^0 - \sum_{k=1}^{m} \left( R_k - \pi_k R_{k-1} \right) \bar{\pi}_k^{-1} E(h^0 | \mathbf{X}_{(1_{k-1})}) \\
= U h^0 - \sum_{k=1}^{m} \frac{R_k - R_{k-1}}{\bar{\pi}_k} E(h^0 | \mathbf{X}_{(1_{k-1})}) + \frac{R_t - \pi(1_m) R_t}{\bar{\pi}_{t+1}} E(h^0 | \mathbf{X}_{(1_t)})
\]

23

Hosted by The Berkeley Electronic Press
$$= U h^0 - \frac{R_m - \pi(1_m)}{\pi(1_m)} E(h^0 \mid X(1_t))$$
$$= \frac{I(R = 1_m)}{\pi(1_m)} h^0 - \frac{I(R = 1_m) - \pi(1_m)}{\pi(1_m)} E(h^0 \mid X(1_t))$$.

Then applying Theorem 3.1 or Theorem 3.2 we obtain Theorem 3.3.

Another way of reducing the results of monotonic missingness (Theorem 3.2) to that of two-phase sampling design is to view the first $t$ components of $X$ as the first variable and the rest as the second. Then we have $m = 2$ and the first variable is always observable. Hence $R_1 = 1$ and $\pi_1 = 1$, and the reduction is obvious.

6 Space of Influence Functions / Estimating Functions

Since finding efficient estimates can be very challenging for missing data problems, we may be interested in some inefficient estimators that are relatively easy to compute and have satisfactory efficiency. As we know from BKRW, the space of all influence functions (or estimating functions) for regular asymptotically linear estimators of $\theta$ in model $P$ is a subset of $\dot{P}_{\eta, \pi}^{\perp}$. Hence characterizing the space $\dot{P}_{\eta, \pi}^{\perp}$ becomes important.

Since the influence function $\psi$ for any regular asymptotically linear estimator has the property that $\psi \perp \dot{P}_{\eta, \pi}$ and $\langle \psi, \dot{l}_\theta \rangle = 1$, we have the following representation of the space of influence functions:

Lemma 6.1.

1. $\Psi_\theta = \{ \langle g, \dot{l}_\theta \rangle^{-1} g : g \in \dot{P}_{\eta, \pi} \} = \{ g \in \dot{P}_{\eta, \pi}^{\perp} : \langle g, \dot{l}_\theta \rangle = 1 \} \subset \dot{P}_{\eta, \pi}^{\perp}$

2. $\Psi^0_\theta = \{ \langle g^0, \dot{l}_\theta^0 \rangle^{-1} g^0 : g^0 \in \dot{Q}_{\eta} \} = \{ g^0 \in \dot{Q}_{\eta}^{\perp} : \langle g^0, \dot{l}_\theta^0 \rangle = 1 \} \subset \dot{Q}_{\eta}^{\perp}$

3. For any $g \in \dot{P}_{\eta, \pi}^{\perp}$, $\langle g, \dot{l}_\theta^* \rangle = \langle g, \dot{l}_\theta \rangle$.

(Lemma 8.1 in RRZ)

Proof: The proof is trivial and thus omitted.

The following Propositions 6.1-6.3 can be found from Proposition 8.1 in RRZ.

Proposition 6.1. For any $g \in \dot{P}_{\eta, \pi}^{\perp}$, it can be decomposed as

$$g = U(A^T g) + \left( g - U(A^T g) \right),$$
where $A^T g \in \hat{Q}^\perp_\eta$ and $g - U(A^T g) \in N(A^T)$.

**Proof:** Since $g \in \hat{P}^\perp_{\eta,\pi} \subset \hat{P}^\perp_\eta$, we have $A^T g \in \hat{Q}^\perp_\eta$ by Corollary 5.3. The other part can be shown by

$$A^T(g - U(A^T g)) = A^T g - A^T g = 0$$

using the fact in Proposition 2.1, part 2. \hfill \Box

**Proposition 6.2.** $\hat{P}^\perp_{\eta,\pi} = \{ V(g^0, a) : g^0 \in \hat{Q}^\perp_\eta, a \in N(A^T) \}$

**Proof:** We first prove $\hat{P}^\perp_{\eta,\pi} \supset \{ V(g^0, a) : g^0 \in \hat{Q}^\perp_\eta, a \in N(A^T) \}$. By the definition of $V$, we know that $V(g^0, a) \in \hat{P}^\perp_\pi$. From Corollary 5.2 we have $\hat{P}^\perp_{\eta,\pi} = \hat{P}^\perp_\eta + \hat{P}^\perp_\pi$, so we need to only show that $V(g^0, a) \in \hat{P}^\perp_\eta$, i.e., $\langle V(g^0, a), A(h^0) \rangle = 0$ for all $h^0 \in \hat{Q}_\eta$. For any $a \in N(A^T)$, by Lemma 5.1 we have

$$A^T V(g^0, a) = A^T U(g^0) + A^T \left\{ a - \Pi \left( U(g^0) + a \right) \bigg| \hat{P}^\perp_\pi \right\} = A^T U(g^0) + 0 = g^0.$$

Thus $\langle V(g^0, a), A(h^0) \rangle = \langle A^T V(g^0, a), h^0 \rangle_0 = \langle g^0, h^0 \rangle_0 = 0$ since $g^0 \in \hat{Q}^\perp_\eta$ and $h^0 \in \hat{Q}_\eta$.

Now we show $\hat{P}^\perp_{\eta,\pi} \subset \{ V(g^0, a) : g^0 \in \hat{Q}^\perp_\eta, a \in N(A^T) \}$, i.e., for any $g \in \hat{P}^\perp_{\eta,\pi}$, we want to find some $g^0 \in \hat{Q}^\perp_\eta$ and $a \in N(A^T)$ such that $g = V(g^0, a)$. This is obviously true from Proposition 6.1 and $g \perp \hat{P}^\perp_\pi$. \hfill \Box

**Remark:** The above Proposition 6.2 characterizes the space $\hat{P}^\perp_{\eta,\pi}$, which contains all the estimating functions (or influence functions) for regular asymptotically linear estimators in model $P$. From the definition of operator $V$ in Definition 2.1 we see that any estimating function in model $P$ is an inverse probability weighted estimating function in model $Q$ plus a term which has expectation zero given full data $X$. This property gives us great opportunity and flexibility to develop estimating methods for missing data problems if we have enough knowledge about the complete data model. It can be easily shown that for a two-phase sampling design where $\pi$ is given, the second term in any estimating function has the form $-(R_m - \pi(1_m)\pi(1_m)^{-1}\phi(X(1)))$, where $\phi$ has finite second moment.

**Proposition 6.3.** If $g^0 \in \hat{Q}^\perp_\eta$ and $a \in N(A^T)$, then $\langle V(g^0, a), l^*_\theta \rangle = \langle g^0, l^*_{\theta^0} \rangle_0$.

**Proof:** By Lemma 6.1 part 3 and Proposition 6.2, we have $\langle V(g^0, a), l^*_\theta \rangle = \langle V(g^0, a), \hat{l}_\theta \rangle = \langle V(g^0, a), A l^*_\theta \rangle = \langle A^T V(g^0, a), l^*_{\theta^0} \rangle_0 = \langle g^0, l^*_{\theta^0} \rangle_0$. \hfill \Box
The following two propositions are properties for spaces of influence functions.

**Proposition 6.4.** If \( g \in \Psi_\theta \), then \( A^T g \in \Psi_\theta^0 \).

**Proof:** For any \( g \in \Psi_\theta \), we have \( g \in \hat{\mathcal{P}}_{\eta,\pi}^\perp \subset \hat{\mathcal{P}}_\eta^\perp \), and thus \( A^T g \in \hat{\mathcal{Q}}_\eta^\perp \) with \( \langle A^T g, \hat{l}_0^\theta \rangle_0 = \langle g, A\hat{l}_0^\theta \rangle = \langle g, \hat{l}_0 \rangle = 1 \). \( \Box \)

**Proposition 6.5.** \( \Psi_\theta = \{ V(g^0, a) : g^0 \in \Psi_0^0, a \in \mathcal{N}(A^T) \} = \{ \langle g^0, l^0 \rangle_0^{-1} V(g^0, a) : g^0 \in \hat{\mathcal{Q}}_\eta^\perp, a \in \mathcal{N}(A^T) \} \).

**Proof:** By Lemma 6.1 and Proposition 6.2, we need to only show for the first equality that \( \langle V(g^0, a), \hat{l}_\theta \rangle = 1 \) for \( g^0 \in \Psi_0^0, a \in \mathcal{N}(A^T) \), which is easily seen from Proposition 6.3, i.e., \( \langle V(g^0, a), l^0_\theta \rangle = \langle g^0, l^0_\theta \rangle_0 = 1 \) for \( g^0 \in \Psi_\theta^0 \). The second equality can be verified directly by using Proposition 6.3 and Lemma 6.1 part 1. \( \Box \)

**7 Discussion**

The major challenges in deriving efficient score functions for models with missing data can be the characterization of space \( \hat{\mathcal{Q}}_\eta^\perp \) and the calculation of the projection onto this space. Although these problems have been successfully solved for the models discussed in the papers cited at the end of Section 4, there are still many unsolved problems for other complicated models. Even though sometime we are able to obtain the efficient score function, finding efficient estimates can also be very challenging. We refer Nan (2002) for an example on this line. It is not unusual that we compromise efficiency (or make stronger assumptions) to obtain certain computable estimates. Thus the characterization of space \( \hat{\mathcal{P}}_{\eta,\pi}^\perp \) in Proposition 6.2 can be very helpful.

**Acknowledgements**

The authors would like to thank Mary Emond and Jon Wellner for their generous help in this research.

**Appendix**

Table 1 lists the two sets of corresponding notation used in RRZ and our paper.
References


University of Michigan
Department of Biostatistics
1420 Washington Heights
Ann Arbor, MI 48109-2029
U.S.A.
e-mail: menggang@umich.edu

University of Michigan
Department of Biostatistics
1420 Washington Heights
Ann Arbor, MI 48109-2029
U.S.A.
e-mail: bnan@umich.edu
Table 1: Corresponding notation in our paper and RRZ.

<table>
<thead>
<tr>
<th>Meaning</th>
<th>Our Paper</th>
<th>RRZ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full data model</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuisance tangent space</td>
<td>$\hat{\mathcal{Q}}_\eta$</td>
<td>$\Lambda^F$</td>
</tr>
<tr>
<td>Orthogonal complement to the nuisance tangent space</td>
<td>$\hat{\mathcal{Q}}_\eta^\perp$</td>
<td>$\Lambda_0^{F,\perp}$</td>
</tr>
<tr>
<td>Score for $\theta$</td>
<td>$\hat{l}_\theta^0$</td>
<td>$S_\theta^F$</td>
</tr>
<tr>
<td>Efficient Score for $\theta$</td>
<td>$l_\theta^{*0}$</td>
<td>$S_{\theta}^{F\text{eff}}$</td>
</tr>
<tr>
<td>The space of influence function</td>
<td>$\Psi_\theta^0$</td>
<td>$\Lambda_0^{F,\perp}$</td>
</tr>
<tr>
<td><strong>Observed data model</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuisance tangent space</td>
<td>$\hat{\mathcal{P}}_{\eta,\pi}$</td>
<td>$\Lambda$</td>
</tr>
<tr>
<td>Orthogonal complement to the nuisance tangent space</td>
<td>$\hat{\mathcal{P}}_{\eta,\pi}^\perp$</td>
<td>$\Lambda_0^{\perp}$</td>
</tr>
<tr>
<td>Nuisance tangent space for $\pi$</td>
<td>$\hat{\mathcal{P}}_\pi$</td>
<td>$\Lambda^{(3)}$</td>
</tr>
<tr>
<td>Score for $\theta$</td>
<td>$\hat{l}_\theta$</td>
<td>$S_\theta$</td>
</tr>
<tr>
<td>Efficient Score for $\theta$</td>
<td>$l_\theta^*$</td>
<td>$S_{\text{eff}}$</td>
</tr>
<tr>
<td>The space of influence function</td>
<td>$\Psi_\theta$</td>
<td>$\Lambda_0^{\perp}$</td>
</tr>
<tr>
<td>The null space of $A^T$</td>
<td>$\mathcal{N}(A^T)$</td>
<td>$\Lambda^{(2)}$</td>
</tr>
</tbody>
</table>

**Operators**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$g$</td>
</tr>
<tr>
<td>$U$</td>
<td>$u$</td>
</tr>
<tr>
<td>$V$</td>
<td>$v$</td>
</tr>
<tr>
<td>$A^TA$</td>
<td>$m$</td>
</tr>
</tbody>
</table>