
Data-guided Treatment Recommendation with Feature Scores

Zhongyuan Chen, Ziyi Wang, Qifan Song, and Jun Xie

Department of Statistics, Purdue University,

150 N. University Street, West Lafayette, IN 47907

Supplementary Material

We report the data and code information in Section 1. The proofs of Lemma 1 and Theorem 1 are provided in Section 2 and 3. In addition, we present a consistent theorem for an extended situation where the dimension $p \to \infty$ and $p/n \to 0$. This result is reported in Theorem 2 in Section 4.

S1. Data and code information

Both the data and code used for the paper are available on public websites as specified below. Readers can download them to reproduce the results in the paper. The real example data (Section 5 in the main paper) is originally from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9782). We have created an R format data file that can be directly loaded to the R workplace. The R code and partial data can be accessed on GitHub:

`https://github.com/chenstatistics/Treatment-Recommendation`

Due to the limit space of GitHub, we cannot store all data there. Instead, we have the complete set of R code and all data for the simulation studies and the real example on the following website:

`https://www.stat.purdue.edu/~chen3490/work/upload_rev2/`

Instructions for running the R code are also provided on the websites.

S2. Proof of Lemma 1

According to Section 2.1 in the main text, we have a random vector $(\mathbf{X}, A, Y)$, where $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$ denotes clinical covariates plus a big set of genetic variables, $A \in \mathcal{A} = \{1, \ldots, M\}$ denotes the treatment index, and Y is the treatment response with larger values for better treatment outcome. A treatment recommendation rule d is a deterministic decision rule from $\mathcal{X}$ into $\mathcal{A}$. The Value of d is defined as $V(d) \triangleq E^d(Y)$, where the expectation is with respect to P^d as the distribution of $(\mathbf{X}, A, Y)$ when d is used to assign the treatment. The optimal treatment recommendation rule, d_0, is defined as

$$d_0 \in \operatorname*{argmax}_d V(d).$$

We have shown $d_0(\mathbf{X}) \in \operatorname{argmax}_a Q_0(\mathbf{X}, a)$, where $Q_0(\mathbf{x}, a) \triangleq E(Y|\mathbf{X} = \mathbf{x}, A = a)$. Let $Q(\mathbf{X}, A)$ be an estimate of the true condition mean $Q_0(\mathbf{X}, A)$ and the corresponding treatment recommendation rule $d(\mathbf{X}) \in \operatorname{argmax}_a Q(\mathbf{X}, a)$. We make the following assumption on the margin of the treatment effect.

A. 1. Let $T_0(\mathbf{X}, A) = Q_0(\mathbf{X}, A) - E\left[Q_0(\mathbf{X}, A)|\mathbf{X}\right]$. There exists some constant $C > 0$ and $\alpha > 0$ such that

$$P\left(\max_a T_0(\mathbf{X}, a) - \max_{a \in \mathcal{A} \setminus \text{argmax}_a T_0(\mathbf{X}, a)} T_0(\mathbf{X}, a) \leq \epsilon\right) \leq C\epsilon^\alpha$$

for any $\epsilon > 0$.

Lemma 1. *Suppose $p(a|\mathbf{x}) \geq S^{-1}$ for a positive constant S for all $(\mathbf{x}, a)$ pairs and assume A.1. For any treatment rule $d : \mathcal{X} \mapsto \mathcal{A}$ and square integrable function $Q : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$ such that $d(\mathbf{X}) \in \text{argmax}_{a \in \mathcal{A}} Q(\mathbf{X}, a)$, we have*

$$V(d_0) - V(d) \leq C' \left[E(Q(\mathbf{X}, A) - Q_0(\mathbf{X}, A))^2\right]^{(1+\alpha)/(2+\alpha)}$$

where $C' = (2^{2+3\alpha} S^{1+\alpha} C)^{1/(2+\alpha)}$.

Proof. Let $T(\mathbf{X}, A) = Q(\mathbf{X}, A) - E[Q(\mathbf{X}, A)|\mathbf{X}]$ and $T_0[\mathbf{X}, A] = Q_0(\mathbf{X}, A) - E[Q_0(\mathbf{X}, A)|\mathbf{X}]$. Then

$$\begin{aligned}
E\left[(T(\mathbf{X}, A) - T_0(\mathbf{X}, A))^2\right] &= E\left[(Q(\mathbf{X}, A) - Q_0(\mathbf{X}, A) - E[Q(\mathbf{X}, A) - Q_0(\mathbf{X}, A)|\mathbf{X}])^2\right] \\
&= E[(Q(\mathbf{X}, A) - Q_0(\mathbf{X}, A))^2] \\
&\quad - 2E[(Q(\mathbf{X}, A) - Q_0(\mathbf{X}, A))E[Q(\mathbf{X}, A) - Q_0(\mathbf{X}, A)|\mathbf{X}]] \\
&\quad + E[(E[Q(\mathbf{X}, A) - Q_0(\mathbf{X}, A)|\mathbf{X}])^2] \\
&= E[(Q(\mathbf{X}, A) - Q_0(\mathbf{X}, A))^2] - E[(E[Q(\mathbf{X}, A) - Q_0(\mathbf{X}, A)|\mathbf{X}])^2] \\
&\leq E[(Q(\mathbf{X}, A) - Q_0(\mathbf{X}, A))^2]
\end{aligned}$$

Then Lemma 1 follows from Theorem 3.1 in [Qian and Murphy 2011]. □

S3. Proof of Theorem 1

Besides the margin condition (A.1), we require additional assumptions from SIR and for the nonparametric LOESS estimator. We first rewrite (A.2) by denoting the treatment index as $i \in \mathcal{A} = \{1, \ldots, M\}$ and the projection directions β's as $\mathbf{B}_i \in \mathbb{R}^{k\times p}, k < p$.

A. 2. There exist some full-rank matrices $\mathbf{B}_i \in \mathbb{R}^{k\times p}, k < p$, such that $E[Y|\mathbf{X}, A = i] = E[Y|\mathbf{B}_i\mathbf{X}, A = i] = \eta_i(\mathbf{B}_i\mathbf{X})$, where $\eta_i(\cdot)$'s are ρ-Lipschitz continuous and have continuous second derivatives. Furthermore, for any row vector $\xi \in \mathbb{R}^p$, $E[\xi\mathbf{X}|\mathbf{B}_i\mathbf{X}]$ is a linear function of $\mathbf{B}_i\mathbf{X}$. Besides, the dimension of the central inverse curve $E[\mathbf{X}|y, A = i]$ equals to the dimension of the space spanned by the columns of $\mathbf{B}_i$, $col(\mathbf{B}_i)$, and the variance $v_i(u) = Var[Y|\mathbf{B}_i\mathbf{X} = u, A = i]$ is a continuous function.

A. 3. Denote the kernel function of LOESS by $K_H(u) = |H|^{-1/2}K(H^{-1/2}u)$, where $u \in \mathbb{R}^k$ and the bandwidth matrix $H \in \mathbb{R}^{k\times k}$. Assume the kernel function $K(\cdot)$ is ρ-Lipschitz, compactly supported, and satisfies $\int uu^\top K(u)du = \mu_2(K)\mathbf{I}$, where $\mathbf{I}$ is the identity matrix and $\mu_2(K)$ is a constant depending on K. Moreover, all odd-order moments of K equal to zero, that is, $\int u_1^{l_1}\cdots u_d^{l_d}K(u)du = 0$ for all non-negative $l_1\cdots l_d$ when their sum is odd. Additionally, the bandwidth matrix H is symmetric and positive definite with each entry, as well as $n^{-1}|H|$, tending to 0 as $n \to \infty$, and the ratio of the largest and the smallest eigenvalue of H is uniformly bounded for all n.

A. 4. For all $i \in \mathcal{A}$, let $f_i(\cdot)$ be the conditional density function of $\mathbf{B}_i\mathbf{X}$ given $A = i$. Assume

that $f_i(\cdot)$ is uniformly bounded away from 0 and has a continuous gradient function $D_{f_i}(\cdot)$.

A. 5. Denote $n_i = |\{j : A_j = i\}|$ as the number of observations in the treatment group $A = i$. Assume $\min_{i\in\mathcal{A}} P(A = i) > c$ for some positive constant c and the support set of $\mathbf{X}$ is bounded.

As represented in Formula (2) in Section 2.3 in the main text, we write the treatment recommendation rule as $d(\mathbf{x}) \in \underset{i\in\mathcal{A}}{\operatorname{argmax}} Q(\mathbf{x}, i)$, where $Q(\mathbf{x}, i) = \tilde{g}_i(\widehat{\mathbf{B}_i}\mathbf{x})$ with $\widehat{\mathbf{B}_i}$ as the estimated projection directions from SIR and $\tilde{g}_i(\cdot)$ the LOESS function from the training data $\{\widehat{\mathbf{B}_i}\mathbf{x}_j, y_j\}_{\{j:A_j=i\}}$.

Theorem 1. *Assume (A.1)-(A.5). Then the difference between the optimal Value, $V(d_0)$, and $V(d)$ of our treatment recommendation rule converges to* 0 *in probability:*

$$V(d_0) - V(d) \leq \left(|H|^{-1}\|H^{-1/2}\|_F^2 \mathcal{O}_p(\frac{1}{n}) + \mathcal{O}_p\left(\frac{|H|^{-1/2}}{n} + \|H\|_1^2\right)\right)^{\frac{1+\alpha}{2+\alpha}}, \tag{S3.1}$$

where $\|H\|_1$ denotes the maximum column absolute sum and $\|\cdot\|_F^2$ denotes the Frobenius norm. When the bandwidth matrix $H = diag\{h, \cdots, h\}$ with $h = n^{-\frac{1}{k+3}}$, the upper bound on the right hand side becomes $\mathcal{O}_p(n^{-\frac{2(1+\alpha)}{(k+3)(2+\alpha)}})$.

Proof. Using the notation of Lemma 1, we have $Q_0(\mathbf{x}, i) = \eta_i(\mathbf{B}_i\mathbf{x})$, and the LOESS regression yields $Q(\mathbf{x}, i) = \tilde{g}_i(\widehat{\mathbf{B}_i}\mathbf{x})$, where $\widehat{\mathbf{B}_i}$'s are obtained by the SIR algorithm.

We can rewrite $Q(\mathbf{x}, A) = \sum_{i\in\mathcal{A}} \tilde{g}_i(\widehat{\mathbf{B}_i}\mathbf{x})\mathbb{1}_{A=i}$ and rewrite Q_0 in similar form as $Q_0(\mathbf{x}, A) = \sum_{i\in\mathcal{A}} \eta_i(\mathbf{B}_i\mathbf{x})\mathbb{1}_{A=i}$, where the function $\tilde{g}_i$ is the LOESS estimate of the true fucntion $\eta_i(\cdot)$ training from $\{\widehat{\mathbf{B}_i}X_j, Y_j\}_{\{j:A_j=i\}}$. Furthermore, let $g_i(\cdot)$ be the LOESS function with the

training data $\{\mathbf{B}_i\mathbf{X}_j, Y_j\}_{\{j:A_j=i\}}$. Then

$$\begin{aligned}
&E(Q(\mathbf{x}) - Q_0(\mathbf{x}))^2 = E\left[\sum_{i\in\mathcal{A}}(\tilde{g}_i(\widehat{\mathbf{B}_i}\mathbf{x}) - \eta_i(\mathbf{B}_i\mathbf{x}))^2\mathbb{1}_{A=i}\right] \\
\leq &\sum_{i\in\mathcal{A}} E[\tilde{g}_i(\widehat{\mathbf{B}_i}\mathbf{x}) - \eta_i(\mathbf{B}_i\mathbf{x})]^2 \\
= &\sum_{i\in\mathcal{A}} E[\tilde{g}_i(\widehat{\mathbf{B}_i}\mathbf{x}) - g_i(\widehat{\mathbf{B}_i}\mathbf{x}) + g_i(\widehat{\mathbf{B}_i}\mathbf{x}) - \eta_i(\widehat{\mathbf{B}_i}\mathbf{x}) + \eta_i(\widehat{\mathbf{B}_i}\mathbf{x}) - \eta_i(\mathbf{B}_i\mathbf{x})]^2 \\
\leq &3\sum_{i\in\mathcal{A}}\Big(E|g_i(\widehat{\mathbf{B}_i}\mathbf{x}) - \tilde{g}_i(\widehat{\mathbf{B}_i}\mathbf{x})|^2 + E|g_i(\widehat{\mathbf{B}_i}\mathbf{x}) - \eta_i(\widehat{\mathbf{B}_i}\mathbf{x})|^2 + E|\eta_i(\widehat{\mathbf{B}_i}\mathbf{x}) - \eta_i(\mathbf{B}_i\mathbf{x})|^2\Big)
\end{aligned} \tag{S3.2}$$

where the expectation is w.r.t. random variable $\mathbf{x}$ only. Note that due to the estimation error of SIR, there is a discrepancy between the true low dimensional projection $\mathbf{B}_iX$ and the SIR estimated projection $\widehat{\mathbf{B}_i}X$. To explicitly represent the dependency between the LOESS nonparametric regression function estimator and the training data for LOESS, we rewrite:

$$\begin{aligned}
\tilde{g}_i(\widehat{\mathbf{B}_i}\mathbf{x}) &= g(\{\widehat{\mathbf{B}_i}X_j, Y_j\}_{\{j:A_j=i\}}, \widehat{\mathbf{B}_i}\mathbf{x}) \\
g_i(\widehat{\mathbf{B}_i}\mathbf{x}) &= g(\{\mathbf{B}_iX_j, Y_j\}_{\{j:A_j=i\}}, \widehat{\mathbf{B}_i}\mathbf{x}).
\end{aligned} \tag{S3.3}$$

For notation simplicity, when not causing confusion, we denote the index set $\{j : A_j = i\}$ as $\{1, \ldots, n_i\}$.

It is sufficient to study $E|g_i(\widehat{\mathbf{B}_i}\mathbf{x}) - \tilde{g}_i(\widehat{\mathbf{B}_i}\mathbf{x})|^2 + E|g_i(\widehat{\mathbf{B}_i}\mathbf{x}) - \eta_i(\widehat{\mathbf{B}_i}\mathbf{x})|^2 + E|\eta_i(\widehat{\mathbf{B}_i}\mathbf{x}) - \eta_i(\mathbf{B}_i\mathbf{x})|^2$ for some fixed i. In what follows, when causing no confusion, we will drop the subscript i, e.g., we write $\mathbf{B}_i$, f_i and D_{f_i} as $\mathbf{B}$, f and D_f respectively.

We first study $E|g(\widehat{\mathbf{B}}\mathbf{x}) - \tilde{g}(\widehat{\mathbf{B}}\mathbf{x})|^2$. By [Ruppert and Wand 1994], we have $g(\{\mathbf{B}\mathbf{X_j}, Y_j\}_{j=1}^{n_i}, \widehat{\mathbf{B}}\mathbf{x}) = e_1^\top(\mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}^\top\mathbf{W}_{\widehat{\mathbf{B}}\mathbf{x}}\mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}})^{-1}\mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}^\top\mathbf{W}_{\widehat{\mathbf{B}}\mathbf{x}}\mathbf{Y}$, where $\mathbf{W_z} = diag\{K_H(\mathbf{B}\mathbf{X}_1 - \mathbf{z}), \cdots, K_H(\mathbf{B}\mathbf{X}_{n_i} - \mathbf{z})\}$,

$\mathbf{Y} = (Y_1, \cdots, Y_{n_i})^\top$,

$$X_{\mathbf{z}} = \begin{bmatrix} 1 & (\mathbf{B}\mathbf{X}_1 - \mathbf{z})^\top \\ \vdots & \vdots \\ 1 & (\mathbf{B}\mathbf{X}_{n_i} - \mathbf{z})^\top \end{bmatrix} \tag{S3.4}$$

and $e_1 = (1, 0, \ldots, 0)$ is a $d \times 1$ vector with only the first entry being 1. Similarly, we can rewrite $\tilde{g}(\{\widehat{\mathbf{B}}\mathbf{X}_j, Y_j\}_{j=1}^{n_i}, \widehat{\mathbf{B}}\mathbf{x}) = e_1^\top (\mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}'^\top \mathbf{W}'_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{X}'_{\widehat{\mathbf{B}}\mathbf{x}})^{-1} \mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}'^\top \mathbf{W}'_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{Y}$, with $\mathbf{X}'_{\mathbf{z}}, \mathbf{W}'_{\mathbf{z}}$ are defined accordingly. By the same argument used in [Ruppert and Wand 1994], we have

$$(n_i^{-1} \mathbf{X}_{\mathbf{z}}^\top \mathbf{W}_{\mathbf{z}} \mathbf{X}_{\mathbf{z}})^{-1} = \begin{pmatrix} f(\mathbf{z})^{-1} + o_p(1) & -D_f(\mathbf{z})^\top f(\mathbf{z})^{-2} + o_p(1) \\ -D_f(\mathbf{z}) f(\mathbf{z})^{-2} + o_p(1) & \{\mu_2(K) f(\mathbf{z}) H\}^{-1} + o_p(H^{-1}) \end{pmatrix}. \tag{S3.5}$$

Using first-order Taylor approximation, we have the following results:

$$\begin{aligned}
& E|g(\widehat{\mathbf{B}}\mathbf{x}) - \tilde{g}(\mathbf{B}\mathbf{x})|^2 \\
=& E[e_1^\top (\mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}^\top \mathbf{W}_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}})^{-1} \mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}^\top \mathbf{W}_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{Y} - (\mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}'^\top \mathbf{W}'_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{X}'_{\widehat{\mathbf{B}}\mathbf{x}})^{-1} \mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}'^\top \mathbf{W}'_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{Y})]^2 \\
=& E[e_1^\top (\Delta(\mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}^\top \mathbf{W}_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}})^{-1} \mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}^\top \mathbf{W}_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{Y} + (\mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}^\top \mathbf{W}_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}})^{-1} \Delta \mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}'^\top \mathbf{W}'_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{Y} \\
& \quad + (\mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}^\top \mathbf{W}_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}})^{-1} \mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}^\top \Delta \mathbf{W}_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{Y})]^2 \\
\leq & \sup_{\mathbf{z}} 3 \left([e_1^\top \Delta(\mathbf{X}_{\mathbf{z}}^\top \mathbf{W}_{\mathbf{z}} \mathbf{X}_{\mathbf{z}})^{-1} \mathbf{X}_{\mathbf{z}}^\top \mathbf{W}_{\mathbf{z}} \mathbf{Y}]^2 + [e_1^\top (\mathbf{X}_{\mathbf{z}}^\top \mathbf{W}_{\mathbf{z}} \mathbf{X}_{\mathbf{z}})^{-1} \Delta \mathbf{X}_{\mathbf{z}}^\top \mathbf{W}_{\mathbf{z}} \mathbf{Y}]^2 \right. \\
& \left. + [e_1^\top (\mathbf{X}_{\mathbf{z}}^\top \mathbf{W}_{\mathbf{z}} \mathbf{X}_{\mathbf{z}})^{-1} \mathbf{X}_{\mathbf{z}}^\top \Delta \mathbf{W}_{\mathbf{z}} \mathbf{Y}]^2 \right),
\end{aligned} \tag{S3.6}$$

where $\Delta(\mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}^\top \mathbf{W}_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}})^{-1} = (\mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}'^\top \mathbf{W}'_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{X}'_{\widehat{\mathbf{B}}\mathbf{x}})^{-1} - (\mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}^\top \mathbf{W}_{\widehat{\mathbf{B}}\mathbf{x}} \mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}})^{-1}$, $\Delta \mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}} = \mathbf{X}'_{\widehat{\mathbf{B}}\mathbf{x}} - \mathbf{X}_{\widehat{\mathbf{B}}\mathbf{x}}$ and $\Delta \mathbf{W}_{\widehat{\mathbf{B}}\mathbf{x}} = \mathbf{W}'_{\widehat{\mathbf{B}}\mathbf{x}} - \mathbf{W}_{\widehat{\mathbf{B}}\mathbf{x}}$. These delta values are introduced by the estimation error between $\widehat{\mathbf{B}}$ and $\mathbf{B}$ which converges to 0 due to the consistency of SIR. Note that due to compactness of $\mathcal{X}$, the supremum in the above equation is taken over a compact set as well.

The first term of the right hand side (RHS) of (S3.6) can be bounded by

$$\begin{aligned}
&\sup_{\mathbf{z}}(e_1^\top(\Delta(\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z})^{-1}\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{Y})^2 = \sup_{\mathbf{z}}\left[e_1^\top\Delta(\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}/n_i)^{-1}\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{Y}/n_i\right]^2\\
=&\sup_{\mathbf{z}}\left[e_1^\top(\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}/n_i)^{-1}\Delta(\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}/n_i)(\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}/n_i)^{-1}\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{Y}/n_i\right]^2\\
\le&\sup_{\mathbf{z}}\left\|e_1^\top(\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}/n_i)^{-1}\Delta(\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}/n_i)\right\|_2^2\left\|(\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}/n_i)^{-1}\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{Y}/n_i\right\|_2^2\\
\le&\sup_{\mathbf{z}}C\left\|e_1^\top(\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}/n_i)^{-1}\Delta(\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}/n_i)\right\|_2^2,
\end{aligned}\tag{S3.7}$$

where we use the fact that $\Delta A^{-1} = A^{-1}(\Delta A)A^{-1}$, and $(\frac{\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}}{n_i})^{-1}\frac{\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{Y}}{n_i} = (\eta_i(\mathbf{z}), \nabla\eta_i(\mathbf{z}))+$ $o(1)$ which is bounded due to the smoothness of η_i and compactness of $\mathcal{X}$.

To simplify the notations, K_j' and K_j is used to denote $K_H(\widehat{\mathbf{B}}X_j-\mathbf{z})$ and $K_H(\mathbf{B}X_j-\mathbf{z})$ respectively. Then, by (S3.5), (S3.7) can be further bounded by:

$$\begin{aligned}
&\sup_{\mathbf{z}}C\left\|e_1^\top(\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}/n_i)^{-1}(\Delta\mathbf{X}_\mathbf{z}{}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}/n_i + \mathbf{X}_\mathbf{z}{}^\top\Delta\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}/n_i + \mathbf{X}_\mathbf{z}{}^\top\mathbf{W}_\mathbf{z}\Delta\mathbf{X}_\mathbf{z}/n_i)\right\|_2^2\\
\le&\sup_{\mathbf{z}}C\left(\left\|f(\mathbf{z})^{-2}D_f(\mathbf{z})^\top\frac{\sum_j K_j(\widehat{\mathbf{B}}-\mathbf{B})\mathbf{X}_j}{n_i}\right\|_2^2 + f(\mathbf{z})^{-4}\left\|D_f(\mathbf{z})^\top\frac{\sum_j K_j(\widehat{\mathbf{B}}-\mathbf{B})\mathbf{X}_j(\mathbf{B}\mathbf{X}_j-\mathbf{z})^\top}{n_i}\right\|_2^2\right.\\
&+\left\|f(\mathbf{z})^{-1}\frac{\sum_j\Delta K_j}{n_i} - f(\mathbf{z})^{-2}D_f(\mathbf{z})^\top\frac{\sum_j\Delta K_j(\mathbf{B}\mathbf{X}_j-\mathbf{z})}{n_i}\right\|_2^2\\
&+\left\|f(\mathbf{z})^{-1}\frac{\sum_j\Delta K_j(\mathbf{B}\mathbf{X}_j-\mathbf{z})}{n_i} - f(\mathbf{z})^{-2}D_f(\mathbf{z})^\top\frac{\sum_j\Delta K_j(\mathbf{B}\mathbf{X}_j-\mathbf{z})(\mathbf{B}\mathbf{X}_j-\mathbf{z})^\top}{n_i}\right\|_2^2\\
&\left.+\left\|f(\mathbf{z})^{-1}\frac{\sum_j K_j(\widehat{\mathbf{B}}-\mathbf{B})\mathbf{X}_j}{n} - f(\mathbf{z})^{-2}D_f(\mathbf{z})\frac{\sum_j K_j(\mathbf{B}\mathbf{X}_j-\mathbf{z})((\widehat{\mathbf{B}}-\mathbf{B})\mathbf{X}_j)^\top}{n_i}\right\|_2^2\right)
\end{aligned}\tag{S3.8}$$

By the assumption of kernel K, we have $K_j \le \mathcal{O}(|H|^{-1})$, which further implies that $\left\|\frac{\sum K_j\mathbf{X}_j}{n_i}\right\|_2^2 \le \mathcal{O}\left(\left\||H|^{-1}\frac{\sum_j\mathbf{X}_j}{n_i}\right\|_2^2\right)$ and $\left\|\frac{\sum K_j\mathbf{X}_j(\mathbf{B}\mathbf{X}_j-\widehat{\mathbf{B}}\mathbf{x})^\top}{n_i}\right\|_2^2 \le \mathcal{O}\left(\left\||H|^{-1}\frac{\sum\mathbf{X}_j(\mathbf{B}\mathbf{X}_j-\widehat{\mathbf{B}}\mathbf{x})^\top}{n_i}\right\|_2^2\right)$.

By Assumption 4, $|f(\mathbf{z})|^{-1}$ and $D_f(\mathbf{z})$ are bounded. Moreover, by the Law of Large number, and the fact that $EX(\mathbf{BX}-\widehat{\mathbf{B}}\mathbf{x})^\top$, $E\mathbf{X}$, $E\|\mathbf{X}\|(\mathbf{B}X-\widehat{\mathbf{B}}\mathbf{x})$ are $E\|\mathbf{X}\|(\mathbf{BX}-\widehat{\mathbf{B}}\mathbf{x})(\mathbf{BX}-\widehat{\mathbf{B}}\mathbf{x})^\top$ bounded (due to Assumption 5, (S3.8) is then bounded, in probability, by

$$\begin{aligned}&\sup_{\mathbf{z}}(e_1^\top(\Delta(\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z})^{-1}\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{Y})^2\\ =&\mathcal{O}_p\left(|H|^{-1}\|H^{-1/2}\|_2^2\left\|\widehat{\mathbf{B}}-\mathbf{B}\right\|_F^2+|H|^{-1}\left\|\widehat{\mathbf{B}}-\mathbf{B}\right\|_F^2\right)=\mathcal{O}_p\left(|H|^{-1}\|H^{-1/2}\|_2^2\left\|\widehat{\mathbf{B}}-\mathbf{B}\right\|_F^2\right),\end{aligned} \tag{S3.9}$$

where we use the fact that $\|H^{-1/2}\|_2\to\infty$.

Next, we bound the second term of the RHS of equation (S3.6) by similar argument.

$$\begin{aligned}&\sup_{\mathbf{z}}\left[e_1^\top(\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}/n_i)^{-1}\Delta\mathbf{X}_\mathbf{z}{}^\top\mathbf{W}_\mathbf{z}\mathbf{Y}/n_i\right]^2\le\sup_{\mathbf{z}}\left[-f(\mathbf{z})^{-2}D_f(\mathbf{z})^\top\frac{\sum_j K_jY_j(\widehat{\mathbf{B}}-\mathbf{B})\mathbf{X}_j}{n_i}\right]^2\\ \le&\sup_{\mathbf{z}}f(\mathbf{z})^{-4}\|D_f(\mathbf{z})\|_2^2\left\|\widehat{\mathbf{B}}-\mathbf{B}\right\|_F^2\left\|\frac{\sum_j K_jY_j\mathbf{X}_j}{n_i}\right\|_2^2=\mathcal{O}_p\left(|H|^{-1}\left\|\widehat{\mathbf{B}}-\mathbf{B}\right\|_F^2\right).\end{aligned} \tag{S3.10}$$

Then we bound the third term on the RHS of equation (S3.6).

$$\begin{aligned}&\sup_{\mathbf{z}}\left[e_1^\top(\mathbf{X}_\mathbf{z}^\top\mathbf{W}_\mathbf{z}\mathbf{X}_\mathbf{z}/n_i)^{-1}\mathbf{X}_\mathbf{z}{}^\top\Delta\mathbf{W}_\mathbf{z}\mathbf{Y}/n_i\right]^2\\ \le&\sup_{\mathbf{z}}C\left[f(\mathbf{z})^{-1}\frac{\sum_j\Delta K_jY_j}{n_i}-f(\mathbf{z})^{-2}D_f(\mathbf{z})^\top\frac{\sum_j\Delta K_jY_j(\mathbf{BX}_j-\mathbf{z})}{n_i}\right]^2\\ \le&\sup_{\mathbf{z}}C'\left(\left[|H|^{-1/2}\|H^{-1/2}\|_2\left\|\widehat{\mathbf{B}}-\mathbf{B}\right\|_F\frac{\sum_j\|\mathbf{X}_j\|Y_j}{n_i}\right]^2+\left\|\frac{\sum_j\Delta K_jY_j(\mathbf{BX}_j-\mathbf{z})}{n_i}\right\|_2^2\right)\\ \le&C'\left(|H|^{-1}\|H^{-1/2}\|_2^2\left\|\widehat{\mathbf{B}}-\mathbf{B}\right\|_F^2+|H|^{-1}\|H^{-1/2}\|_2^2\|\widehat{\mathbf{B}}-\mathbf{B}\|_F^2\sup_{\mathbf{z}}\left\|\frac{\sum_j\|X_j\|Y_j(\mathbf{BX}_j-\mathbf{z})}{n_i}\right\|_2^2\right)\\ =&\mathcal{O}_p\left(|H|^{-1/2}\|H^{-1/2}\|_2^2\|\widehat{\mathbf{B}}-\mathbf{B}\|_F^2\right).\end{aligned} \tag{S3.11}$$

By the convergence theorem of SIR, (see, e.g., [Lin et al. 2021], we have $\|\widehat{\mathbf{B}}-\mathbf{B}\|_F^2=$

$\mathcal{O}_p(\frac{1}{n_i})$. Combined with (S3.9), (S3.10), (S3.11), we derive that

$$E|g(\widehat{\mathbf{B}}\mathbf{x}) - \tilde{g}(\widehat{\mathbf{B}}\mathbf{x})|^2 \leq |H|^{-1}\|H^{-1/2}\|_F^2\mathcal{O}_p(\frac{1}{n_i}) = |H|^{-1}\|H^{-1/2}\|_F^2\mathcal{O}_p(\frac{1}{n}). \tag{S3.12}$$

Due to Theorem 1 of [Ruppert and Wand 1994] and the compactness of $\mathcal{X}$, the second term of the RHS of equation (S3.2) can be bounded by,

$$\begin{aligned} E\left[g(\widehat{\mathbf{B}}\mathbf{x}) - \eta(\widehat{\mathbf{B}}\mathbf{x})\right]^2 =& \mathcal{O}_p\left(|H|^{-1/2}/n + E(tr\{H\mathcal{H}_\eta(\widehat{\mathbf{B}}\mathbf{x})\}^2)\right) + o_p(tr\{H\}^2) \\ \leq& \mathcal{O}_p\left(|H|^{-1/2}/n + \|H\|_1^2\right). \end{aligned} \tag{S3.13}$$

where $\mathcal{H}_\eta(\widehat{\mathbf{B}}\mathbf{x})$ denotes the Hessian matrix of function η at $\widehat{\mathbf{B}}\mathbf{x}$, and $\|H\|_1$ denote the maximum column absolute sum. Note that the last inequality is due to the boundedness of $\mathcal{H}_\eta(\cdot)$.

In the next step, we bound the third term of equation (S3.2). By assumption 2, we can bound the term by the difference of estimate dimension reduction matrix with the true one, which is controlled by the SIR method.

$$E|\eta_i(\widehat{\mathbf{B}}\mathbf{x}) - \eta_i(\mathbf{B}\mathbf{x})|^2 \leq E\left[\rho\left\|\widehat{\mathbf{B}} - \mathbf{B}\right\|_F \|\mathbf{x}\|_2\right]^2 = \mathcal{O}_p\left(\left\|\widehat{\mathbf{B}} - \mathbf{B}\right\|_F^2\right), \tag{S3.14}$$

which is dominated by the term $\mathcal{O}_p\left(|H|^{-1}\|H^{-1/2}\|_2^2\left\|\widehat{\mathbf{B}} - \mathbf{B}\right\|_F^2\right)$.

Combining the above three inequality, finally, the equation (S3.2) can be bounded by:

$$\begin{aligned} &E(Q - Q_0)^2 \\ \leq& |H|^{-1}\|H^{-1/2}\|_F^2\mathcal{O}_p(\frac{1}{n}) + \mathcal{O}_p\left(|H|^{-1/2}/n + \|H\|_1^2\right). \end{aligned} \tag{S3.15}$$

To be more specific, there exists some full rank matrix A, $\|\widehat{\mathbf{B}}_i - A\mathbf{B}_i\|_F^2 = \mathcal{O}_p(\frac{1}{n_i})$ due to the nonidentifiability $\mathbf{B}_i$. W.O.L.G, we ignore this matrix A in the proof.

If we choose diagonal bandwidth matrix $H = diag\{h, \cdots, h\}$, then

$$\begin{aligned} &|H|^{-1}\|H^{-1/2}\|_F^2\mathcal{O}_p(\frac{1}{n}) + \mathcal{O}_p\left(|H|^{-1/2}/n + \|H\|_1^2\right) \\ =&\mathcal{O}_p(\frac{1}{nh^{k+1}} + \frac{1}{nh^{k/2}} + h^2) = \mathcal{O}_p(\frac{1}{nh^{k+1}} + h^2), \end{aligned} \tag{S3.16}$$

where the last equation is because $k > 1$ then $\frac{1}{nh^{k+1}} \geq \frac{1}{nh^{k/2}}$ for $h < 1$. To minimize the RHS of (S3.16) w.r.t. h, we choose $h \asymp n^{-1/(3+k)}$ which leads to that $E(Q - Q_0)^2 = \mathcal{O}(n^{-2/(3+k)})$. This concludes our theorem by combining Lemma 1. □

S4. Extension to high dimension

In the section, we study the asymptotic behavior of the proposed algorithm when $\lim p = \infty$ and $\lim p/n = 0$. Note that we will still keep k as an constant. Technically, we can still explore the convergence as k grows, however we believe this matter is of less interest, since kernel-based regression estimation usually performs unsatisfactorily under high dimensional setting in practice.

To establish consistency beyond fixed-p scenario, we need the following additional assumption.

A. 6. $\mathbf{X}_{\{A=i\}}$ is sub-Gaussian, and there exists positive constants C_1 and C_2 for each subscript i such that,

$$C_1 \leq \lambda_{min}(\mathbf{\Sigma}_{\mathbf{X}_{\{A=i\}}}) \leq \lambda_{max}(\mathbf{\Sigma}_{\mathbf{X}_{\{A=i\}}}) \leq C_2$$

where $\mathbf{\Sigma}_{\mathbf{X}_{\{A=i\}}}$ is the covariance matrix of $\mathbf{X}_{\{A=i\}}$ and λ_{min}, λ_{max} refer to the minimum and maximum eigenvalue respectively. The central curve $\omega_i(y) = E[\mathbf{X}|y, A = i]$ has finite fourth

moment and is v-sliced stable with respective to y and $\omega_i(y)$. The term v-sliced stable function is defined in Definition 1 shown as below.

Definition 1. Let C_1 and C_2 be any two positive constants and $\mathcal{B}_L(C_1, C_2)$ be a collection $\mathbb{R}$ partitions with size N which any partition $-\infty = a_0 < a_1 < \cdots < a_{L-1} = \infty$ in that set satisfying

$$\frac{C_1}{L} \leq P(a_i \leq y < a_{i+1}) \leq \frac{C_2}{L}$$

The central curve $\omega(y) = E[\mathbf{X}|y]$ is v-sliced stable for some constant v if there exist positive constants C_1, C_2 and C_3 such that for any ϑ in $\mathbb{R}^p$ and any partition in $\mathcal{B}_L(C_1, C_2)$,

$$\frac{1}{L}|\sum_{l=0}^{L-1} var(\vartheta^\top \omega(y)|a_l \leq y \leq a_{h+1})| \leq \frac{C_3}{L^v} var(\vartheta^\top \omega(y))$$

for sufficiently large L.

Remark 1. Assumption 6 is a technical condition due to [Lin et al. 2018] which ensures the consistency of SIR under the growing p situation. It is a refined version of the smoothness and tail conditions proposed by [Hsing and Carrol 1992].

Theorem 2. *If we allow the dimension p grows with n satisfying $\lim p/n = 0$ but the dimension k of the projection space is fixed, and set the number of slices in the SIR procedure to be a sufficiently large constant, then the difference between Values of d_0 and d converges to 0 in probability:*

$$V(d_0) - V(d) \leq \left(|H|^{-1} \|H^{-1/2}\|_F^2 \mathcal{O}_p(\frac{p}{n}) + \mathcal{O}_p\left(\frac{|H|^{-1/2}}{n} + \|H\|_1^2 \right) \right)^{\frac{1+\alpha}{2+\alpha}}, \tag{S4.17}$$

where $\|H\|_1$ *denote the maximum column absolute sum. To simplify, we set the bandwidth matrix* $H = diag\{h, \cdots, h\}$ *with* $h = (\frac{p}{n})^{\frac{1}{k+3}}$. *Then the convergence rate on the right hand side becomes* $\mathcal{O}_p((\frac{p}{n})^{\frac{2(1+\alpha)}{(k+3)(2+\alpha)}})$.

Proof. By the same arguments used in the proof of Theorem 1, we can conclude that

$$E[\tilde{g}_i(\widehat{\mathbf{B}_i}\mathbf{x}) - \eta_i(\mathbf{B}_i\mathbf{x})]^2 \leq \mathcal{O}_p(|H|^{-1}\|H^{-1/2}\|_F^2)\|\widehat{\mathbf{B}}_i - \mathbf{B}_i\|^2 + \mathcal{O}_p\left(|H|^{-1/2}/n + \|H\|_1^2)\right).$$

By Theorem 1, Remark 3, Lemma 13 of [Lin et al. 2018], Lemma 22 of [Lin et al. 2021], we have that when the number of slice in SIR procedure is a sufficiently large constant, we have that $\|\widehat{\mathbf{B}}_i - \mathbf{B}_i\|^2 = \mathcal{O}(p/n)$.

If we choose diagonal bandwidth matrix $H = diag\{h, \cdots, h\}$, then

$$\begin{aligned} &|H|^{-1}\|H^{-1/2}\|_F^2\mathcal{O}_p(\frac{p}{n}) + \mathcal{O}_p\left(|H|^{-1/2}/n + \|H\|_1^2)\right) \\ =&\mathcal{O}_p\left(h^{-k-1}\frac{p}{n} + h^{-k/2}\frac{1}{n} + h^2\right) = \mathcal{O}_p\left(h^{-k-1}\frac{p}{n} + h^2\right) \end{aligned} \tag{S4.18}$$

To minimize the RHS of above inequality w.r.t. h, we choose $h \asymp (\frac{p}{n})^{1/(k+3)}$ which leads to that $E(Q - Q_0)^2 = \mathcal{O}((\frac{p}{n})^{2/(k+3)})$. This concludes our theorem by combining Lemma 1. □

Bibliography

[Li 1991] Li KC (1991), Sliced inverse regression for dimension reduction, Journal of the American Statistical Association, 86(414):316-327

[Qian and Murphy 2011] Qian M and Murphy SA (2011), Performance guarantees for individualized treatment rules, The Annals of Statistics, 39: 1180-1210

[Ruppert and Wand 1994] D. Ruppert, M. P. Wand (1994) Multivariate Locally Weighted Least Squares Regression. The Annals of Statistics, 22(3), 1346-1370.

[Zhu et al. 2006] Zhu L, Miao B, and Peng H. (2006) On Sliced Inverse Regression with High-Dimensional Covariates. Journal of the American Statistical Association, 101(474), 630-643.

[Lin et al. 2018] Lin Q, Zhao Z, Liu J S. (2018) On consistency and sparsity for sliced inverse regression in high dimensions. The Annals of Statistics, 46(2): 580-610.

[Lin et al. 2021] Lin Q, Li X, Huang D, et al. (2021) On the optimality of sliced inverse regression in high dimensions. The Annals of Statistics, 49(1): 1-20.

[Lin et al. 2019] Lin Q, Zhao Z, Liu J S. (2019) Sparse sliced inverse regression via lasso. Journal of the American Statistical Association, 114(528): 1726-1739.

[Hsing and Carrol 1992] Hsing T, Carroll R J. (1992) An Asymptotic Theory for Sliced Inverse Regression, The Annals of Statistics, 20(2): 1040-1061.