Biostatistics (2020), **0**, 0, *pp.* 1–16
doi:10.1093/biostatistics/Supp˙Material˙Additive˙Models˙Interactions˙revised

Supplementary Materials for "A sparse additive model for treatment effect-modifier selection"

HYUNG G. PARK*, EVA PETKOVA, THADDEUS TARPEY,

R. TODD OGDEN

Division of Biostatistics, Department of Population Health, New York University,

180 Madison Ave., New York, USA

Department of Biostatistics, Columbia University, 168st, New York, USA

parkh15@nyu.edu

In Section S.1, we provide proof of Theorem 1 of the main manuscript, and Section S.2 provides the supplementary information referenced in Section 3.1 of the main manuscript. In Section S.3, we report the supplementary simulation results referenced in Section 4.1 of the main manuscript. In Section S.4, we consider the variable selection performance when there is no interaction effect between treatment and pretreatment covariates. In Section S.5, we provide additional simulation results supplemental to the results reported in Section 4.2 of the main manuscript. In Section S.6, we consider the computation time for the proposed regression approach.

*To whom correspondence should be addressed.

S.1. Proof of Theorem 1

Proof. The squared error criterion on the right-hand side of (2.6) of the main manuscript is

$$\begin{aligned}\mathbb{E}\Big[\big\{Y - \sum_{j=1}^{p} g_{j,A}(X_j)\big\}^2\Big] &\propto \mathbb{E}\Big[Y \sum_{j=1}^{p} g_{j,A}(X_j) - \big\{\sum_{j=1}^{p} g_{j,A}(X_j)\big\}^2/2\Big] \quad \text{(with respect to } \{g_j\}\text{)} \\ &= \mathbb{E}\Big[\big\{\mu^*(\boldsymbol{X}) + \sum_{j=1}^{p} g^*_{j,A}(X_j)\big\} \sum_{j=1}^{p} g_{j,A}(X_j) - \big\{\sum_{j=1}^{p} g_{j,A}(X_j)\big\}^2/2\Big] \\ &= \mathbb{E}\Big[\mu^*(\boldsymbol{X}) \sum_{j=1}^{p} g_{j,A}(X_j)\Big] + \mathbb{E}\Big[\big\{\sum_{j=1}^{p} g^*_{j,A}(X_j)\big\}\big\{\sum_{j=1}^{p} g_{j,A}(X_j)\big\} - \big\{\sum_{j=1}^{p} g_{j,A}(X_j)\big\}^2/2\Big] \\ &= \mathbb{E}\Big[\big\{\sum_{j=1}^{p} g^*_{j,A}(X_j)\big\}\big\{\sum_{j=1}^{p} g_{j,A}(X_j)\big\} - \big\{\sum_{j=1}^{p} g_{j,A}(X_j)\big\}^2/2\Big],\end{aligned} \tag{S.1}$$

where the last equality follows from the constraint $\mathbb{E}[g_{j,A}(X_j)|X_j] = 0$ $(j = 1, \ldots, p)$ in (2.6) of the main manuscript imposed on $\{g_j\}$, and by noting $\mathbb{E}\big[\mu^*(\boldsymbol{X}) \sum_{j=1}^{p} g_{j,A}\big(X_j\big)\big] = \mathbb{E}\big[\mathbb{E}\big[\mu^*(\boldsymbol{X}) \sum_{j=1}^{p} g_{j,A}(X_j)|\boldsymbol{X}\big]\big] = \mathbb{E}\big[\mu^*(\boldsymbol{X}) \sum_{j=1}^{p} \mathbb{E}\big[g_{j,A}(X_j)|X_j\big]\big] = 0$. From (S.1), the squared error criterion in (2.6) of the main manuscript can be expressed as:

$$\operatorname*{argmin}_{\{g_j \in \mathcal{H}_j\}} \mathbb{E}\Big[\big(Y - \sum_{j=1}^{p} g_{j,A}\big(X_j\big)\big)^2\Big] = \operatorname*{argmin}_{\{g_j \in \mathcal{H}_j\}} \mathbb{E}\Big[\big(\sum_{j=1}^{p} g^*_{j,A}\big(X_j\big) - \sum_{j=1}^{p} g_{j,A}\big(X_j\big)\big)^2\Big]. \tag{S.2}$$

In the following, we closely follow the proof of Theorem 1 in Ravikumar *and others* (2009). The constrained objective function in (2.6) of the main manuscript can be rewritten in Lagrangian form as:

$$Q(\{g_j\}; \lambda) := \mathbb{E}\Big[\big(\sum_{j=1}^{p} g^*_{j,A}(X_j) - \sum_{j=1}^{p} g_{j,A}(X_j)\big)^2\Big] + \lambda \sum_{j=1}^{p} \|g_j\| \tag{S.3}$$

For the notational simplicity, let us write $g_j = g_{j,A}(X_j)$. For each j, consider the minimization of (S.3) with respect to the component function $g_j \in \mathcal{H}_j$, holding the other component functions $\{g_{j'}, j' \neq j\}$ fixed. The stationary condition is obtained by setting its Fréchet derivative to 0. Denote by $\partial_j Q(\{g_j\}; \lambda; \eta_j)$ the directional derivative with respect to g_j $(j = 1, \ldots, p)$ in the direction, say, $\eta_j \in \mathcal{H}_j$. Then, the stationary point of the Lagrangian (S.3) can be formulated as:

$$\partial_j Q(\{g_j\}; \lambda; \eta_j) = 2\mathbb{E}\Big[(g_j - \tilde{R}_j + \lambda \nu_j)\eta_j\Big] = 0, \tag{S.4}$$

where

$$\tilde{R}_j := \sum_{j=1}^{p} g_{j,A}^*(X_j) - \sum_{j' \neq j} g_{j',A}(X_j) \tag{S.5}$$

is the partial residual for g_j, and ν_j is an element of the subgradient $\partial \|g_j\|$, which satisfies $\nu_j = g_j/\|g_j\|$ if $\|g_j\| \neq 0$, and $\nu_j \in \{s \in \mathcal{H}_j \mid \|s\| \leqslant 1\}$, otherwise. Using iterated expectations conditional on X_j and A, (S.4) can be rewritten as

$$2\mathbb{E}\left[\left(g_j - \mathbb{E}\left[\tilde{R}_j | X_j, A\right] + \lambda \nu_j\right) \eta_j\right] = 0. \tag{S.6}$$

Since $g_j - \mathbb{E}\left[\tilde{R}_j | X_j, A\right] + \lambda \nu_j \in \mathcal{H}_j$, we can evaluate (S.4) (i.e., (S.6)) in the direction: $\eta_j = g_j - \mathbb{E}\left[\tilde{R}_j | X_j, A\right] + \lambda \nu_j$, implying $\mathbb{E}\left[\left(g_j - \mathbb{E}\left[\tilde{R}_j | X_j, A\right] + \lambda \nu_j\right)^2\right] = 0$. This implies:

$$g_j + \lambda \nu_j = \mathbb{E}\left[\tilde{R}_j | X_j, A\right] \quad \text{(almost surely).} \tag{S.7}$$

Let f_j denote the right-hand side of (S.7), i.e., $f_j (= f_{j,A}(X_j)) := \mathbb{E}\left[\tilde{R}_j | X_j, A\right]$. If $\|g_j\| \neq 0$, then $\nu_j = g_j/\|g_j\|$. Therefore, by (S.7), we have $\|f_j\| = \|g_j + \lambda g_j/\|g_j\|\| = \|g_j\| + \lambda \geqslant \lambda$. On the other hand, if $\|g_j\| = 0$, then $g_j = 0$ (almost surely) and $\|\nu_j\| \leqslant 1$ which, together with condition (S.7), implies that $\|f_j\| \leqslant \lambda$. This gives us the equivalence between $\|f_j\| \leqslant \lambda$ and the statement $g_j = 0$ (almost surely). Therefore, condition (S.7) leads to the following expression:

$$\left(1 + \lambda/\|g_j\|\right) g_j = f_j \quad \text{(almost surely)}$$

if $\|f_j\| > \lambda$; otherwise, and $g_j = 0$ (almost surely). This gives the soft thresholding update rule for g_j.

The underlying model (2.1) of the main manuscript indicates that $\sum_{j=1}^{p} g_{j,A}^*(X_j) = \mathbb{E}[Y|X, A] - \mu^*(\boldsymbol{X})$. Thus, (S.5) can be equivalently written as: $\tilde{R}_j = \mathbb{E}[Y|X, A] - \mu^*(\boldsymbol{X}) - \sum_{j' \neq j} g_{j',A}(X_{j'})$.

Therefore, the function $f_{j,A}(X_j) = \mathbb{E}\left[\tilde{R}_j | X_j, A\right]$ can be written by:

$$
\begin{aligned}
f_{j,A}(X_j) &= \mathbb{E}\big[\mathbb{E}[Y|X,A] - \mu^*(\boldsymbol{X}) - \sum_{j' \neq j} g_{j',A}(X_{j'}) \mid X_j, A\big] \\
&= \mathbb{E}\big[\mathbb{E}[Y|X,A] - \sum_{j' \neq j} g_{j',A}(X_{j'}) | X_j, A\big] - \mathbb{E}\big[\mu^*(\boldsymbol{X}) | X_j, A\big] \\
&= \mathbb{E}\big[Y - \sum_{j' \neq j} g_{j',A}(X_{j'}) | X_j, A\big] - \mathbb{E}\big[\mu^*(\boldsymbol{X}) | X_j\big] \\
&= \mathbb{E}\big[Y - \sum_{j' \neq j} g_{j',A}(X_{j'}) | X_j, A\big] - \mathbb{E}\big[\mu^*(\boldsymbol{X}) + \sum_{j=1}^{p} g^*_{j,A}\big(X_j\big) | X_j\big] \\
&= \mathbb{E}\big[Y - \sum_{j' \neq j} g_{j',A}(X_{j'}) | X_j, A\big] - \mathbb{E}\big[Y | X_j\big] \\
&= \mathbb{E}\big[Y - \sum_{j' \neq j} g_{j',A}(X_{j'}) | X_j, A\big] - \mathbb{E}\big[Y - \sum_{j' \neq j} g_{j',A}(X_{j'}) | X_j\big] \\
&= \mathbb{E}\big[R_j | X_j, A\big] - \mathbb{E}\big[R_j | X_j\big],
\end{aligned}
$$

where the fourth equality follows from the identifiability constraint (2.2) of the underlying model (2.1) of the main manuscript, and the sixth equality follows from the optimization constraint $\mathbb{E}[g_{j,A}(X_j)|X_j] = 0$ $(j = 1, \ldots, p)$ in (2.6) of the main manuscript imposed on $\{g_j\}$; this gives the desired expression (3.8) of the main manuscript.

□

S.2. Supplementary Materials for Section 3.1

The restriction of the function g_j to the form (3.10) of the main manuscript restricts also the minimizer g_j^* in (3.7) of the main manuscript (note, $g^*_{j,A}(X_j) = s_j^{(\lambda)} f_{j,A}(X_j)$, where $s_j^{(\lambda)} = [1 - \lambda / \|f_j\|]_+]$ to the form (3.10). In particular, we can express the function f_j in (3.8) of the main manuscript as:

$$
\begin{aligned}
f_{j,A}(X_j) &= \mathbb{E}[R_j | X_j, A] - \sum_{a=1}^{L} \pi_a \mathbb{E}[R_j | X_j, A = a] \\
&= \boldsymbol{\Psi}_j(X_j) \boldsymbol{\theta}^*_{j,A} - \boldsymbol{\Psi}_j(X_j) \big(\sum_{a=1}^{L} \pi_a \boldsymbol{\theta}^*_{j,a} \big)
\end{aligned}
\tag{S.8}
$$

where $\{\boldsymbol{\theta}^*_{j,a}\}_{a \in \{1,\ldots,L\}} := \underset{\{\boldsymbol{\theta}_{j,a} \in \mathbb{R}^{d_j}\}}{\text{argmin}} \; \mathbb{E}\big[\{R_j - \boldsymbol{\Psi}_j(X_j)^\top \boldsymbol{\theta}_{j,A}\}^2\big]$. The first term $\mathbb{E}[R_j | X_j, A]$ in (S.8) corresponds to the L^2 projection of the jth partial residual R_j onto the class of functions of the form (3.10) of the main manuscript (without the imposition of the constraint (3.11)), whereas the second term $-\sum_{a=1}^{L} \pi_a \mathbb{E}[R_j | X_j, A = a]$ centers the first term to satisfy the linear constraint (3.11). Then it follows that f_j, as given in (S.8), corresponds to the L^2 projection of R_j onto the subspace of measurable functions of the form (3.10) subject to the linear constraint (3.11) of the main manuscript.

S.3. Supplementary Materials for Section 4.1

In this section we provide additional details on the simulation experiment reported in Section 4.1 of the main manuscript, illustrating the performance of the treatment effect-modifier selection. The data generating model from Section 4.1 is:

$$Y = \sum_{j=1}^{10} \cos(X_j) \; + \; g^*_{1,A}(X_1) + g^*_{2,A}(X_2) + g^*_{3,A}(X_3) \; + \; \epsilon \quad A \in \{1, 2\}, \tag{S.9}$$

where X_j $(j = 1, \ldots, p)$, $p \in \{50, 200\}$ are generated from independent Unif$[-\pi/2, \pi/2]$, and the treatment variable $A \in \{1, 2\}$ is generated independently of $\boldsymbol{X}$ and the error term $\epsilon \sim \mathcal{N}(0, 0.5^2)$, such that $\Pr(A = 1) = \Pr(A = 2) = 1/2$. We set $g^*_{1,A}(X_1) = (A - 1.5)X_1$, $g^*_{2,A}(X_2) = (A - 1.5)\left\{I_{(X_2 \leqslant 1.3)} 0.05 e^{(X_2 - 1.3)} + I_{(X_2 > 1.3)} e^{4(X_2 - 1.3)} - 1\right\}$ and $g^*_{3,A}(X_3) = (A - 1.5)\{2e^{-X_3^2} - 1\}$, which are displayed in Figure S.1, given the treatment condition $A = 2$.

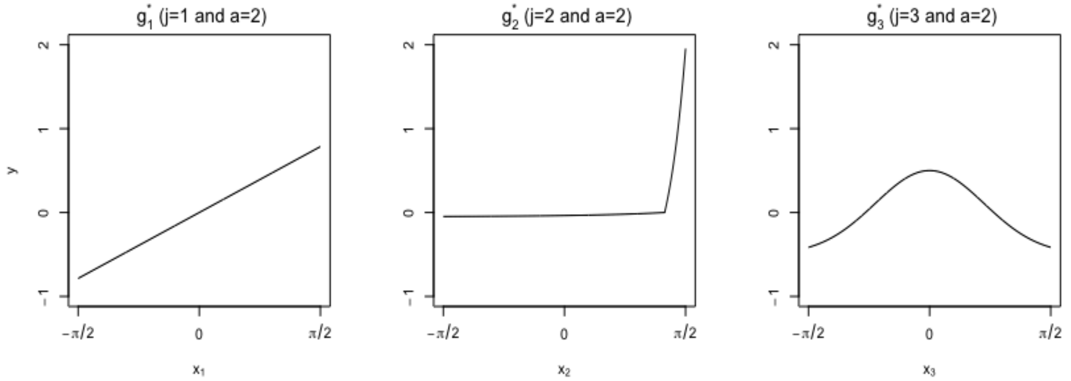

Fig. S.1. Given the treatment $a = 2$, the three component functions $g^*_{1,a}(x_1) = (a-1.5)x_1$, $g^*_{2,a}(x_2) = (a-1.5)\left\{I_{(x_2 \leqslant 1.3)} 0.05 e^{(x_2 - 1.3)} + I_{(x_2 > 1.3)} e^{4(x_2 - 1.3)} - 1\right\}$ and $g^*_{3,a}(x_3) = (a - 1.5)\{2e^{-x_3^2} - 1\}$ are displayed.

For a fixed A, the 1st component function $g^*_{1,A}(X_1)$ is a linear function indicating that the treatment effect varies linearly with X_1; the 2nd component function $g^*_{2,A}(X_2)$ is a monotone piece-wise exponential function indicating that the treatment effect varies monotonically but non-linearly with X_2; the 3rd component function $g^*_{3,A}(X_3)$ is a Gaussian function indicating that the treatment effect non-monotonically varies with X_3. See Figure S.1 for the graphs of the component functions.

There were 3 "signal" covariates (X_1, X_2 and X_3) and $p-3$ "noise" covariates ($X_4, X_5, \ldots, X_p$). Figure 1 of the main manuscript summarizes the results of the treatment effect-modifier selection performance with respect to the true/false positive rates (the left/right two panels, respectively) for $p \in \{50, 200\}$ under setting (S.9), comparing the proposed additive regression to the linear regression approach, which are reported as the averages (and ±1 standard deviation bars around the averages) across the 200 simulation runs.

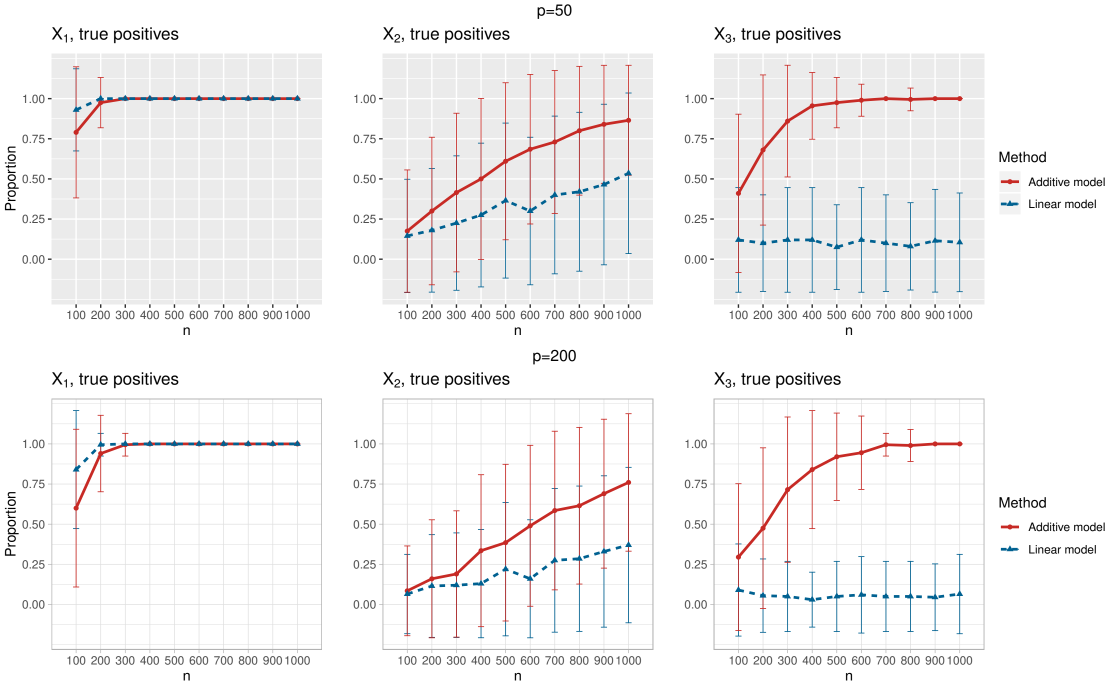

Fig. S.2. The proportions of each individual covariate (X_1, X_2 and X_3, respectively) *correctly selected* (i.e., the "true positives") under setting (S.9), as the sample size n varies from 100 to 1000, for each $p \in \{50, 200\}$.

In addition, we have examined the true positive rates reported in Figure 1 of the main manuscript, by separately displaying the true positive rates associated with selection of X_1, X_2 and X_3 (respectively). Figure S.2 displays those individual true positive rates for the case of $p = 50$ in the top panels and $p = 200$ in bottom panels. (Note, both the $p = 50$ and $p = 200$ cases appear to be qualitatively similar.)

In the left panels of Figure S.2, with n increasing, both of the additive and linear approaches tend to easily identify X_1 as a treatment effect-modifier, since $g^*_{1,a}$ is a linear function (see the left panel of Figure S.1). In the middle panels of Figure S.2, although the true positive rate of both methods associated with selection of X_2 increase with sample size n, the more flexible additive regression approach significantly outperforms the linear regression approach (see the middle panel of Figure S.1 for the shape of the associated component function, $g^*_{2,a}(X_2)$).

Although not displayed in Figure S.2, when $p = 50$ and $n = 2000$, the true positive rate associated with selection of X_2 was 0.93 (sd: 0.25) for the additive regression, and 0.74 (sd: 0.43) for the linear regression; when $p = 200$ and $n = 2000$, it was 0.86 (sd: 0.34) for the additive regression, and 0.64 (sd: 0.48) for the linear regression. Overall, the additive regression approach significantly outperforms the linear regression approach in terms of correctly identifying X_2 as a treatment effect-modifier. (Note, the contribution of $g^*_{2,A}(X_2)$ to the variance of Y is much smaller than those of the components $g^*_{1,A}(X_1)$ and $g^*_{3,A}(X_3)$, and therefore correctly identifying X_2 as a treatment effect-modifier is more difficult compared to that of X_1 or X_3.)

In the right panels of Figure S.2, the proposed additive regression clearly outperforms the linear regression, in terms of correctly selecting X_3 which has a nonlinear and non-monotone association with the treatment effect (see the right panel of Figure S.1 for the shape of the associated component function, $g^*_{3,a}(X_3)$).

S.4. Variable selection performance when there is no A-by-$\boldsymbol{X}$ interaction effect

In this section, we consider a null case, i.e., one without any A-by-$\boldsymbol{X}$ interactions. To consider such a null case of no A-by-$\boldsymbol{X}$ interaction effects on Y, we perform a set of simulations using the data generating model (S.9), but without the A-by-$\boldsymbol{X}$ interaction effect components $g^*_{1,A}(X_1)$, $g^*_{2,A}(X_2)$ and $g^*_{3,A}(X_3)$. Since there is no "signal" covariates in such a "null" case, we shall only report the "false positive" rates (i.e., the proportions out of the p covariates *incorrectly selected* as treatment effect-modifiers) associated with the considered selection approaches.

In particular, to examine the performance behavior, we also include a $p = 10$ case, in which all the 10 covariates are related to the outcome Y, but none of them is related as a treatment effect-modifier (thus, there is no "signal" covariate). In Figure S.3, we report the selection performance given each case $p \in \{10, 20, 50, 200\}$, as we vary the sample size n from $n = 100$ to $n = 1000$. (Note also, when p is large, the false positive rates tend to be relatively small, since the number of incorrectly selected covariates is divided by the total number of the covariates p.)

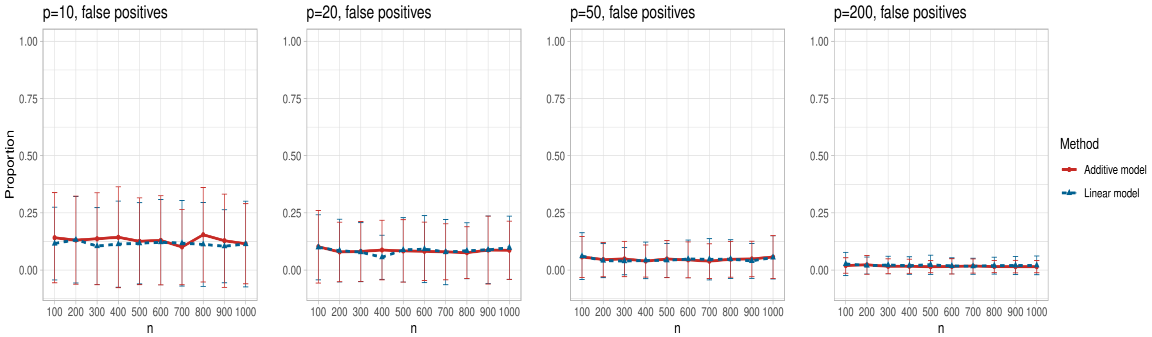

Fig. S.3. The proportion of the p irrelevant covariates (i.e., $X_1, X_2, \ldots, X_p$) incorrectly selected (i.e., the "false positives") (and ±1 standard deviation bars) obtained from 200 simulation runs, as the sample size n varies from 100 to 1000, for each $p \in \{10, 20, 50, 200\}$.

In Figure S.3, when $p = 10$ (the left panel), both the additive and linear model approaches have false positive rates of 12% approximately, i.e., about 88% of the 10 covariates exhibiting the main effect (and no interactions with A) only is *correctly unselected*. The false positive rates of the both selection approaches are also relatively small, for the other cases of $p = 20, 50$ and 200.

S.5. Supplementary Materials for Section 4.2

S.5.1 *Simulation results for a larger p and n*

For the "*moderately-nonlinear*" A-by-$\boldsymbol{X}$ interaction effect scenario, we have also considered the case with a larger number of covariates ($p = 100$) and sample size ($n = 1000$) in addition to the cases with $n \in \{250, 500\}$. The simulation results are shown in Figure S.4.

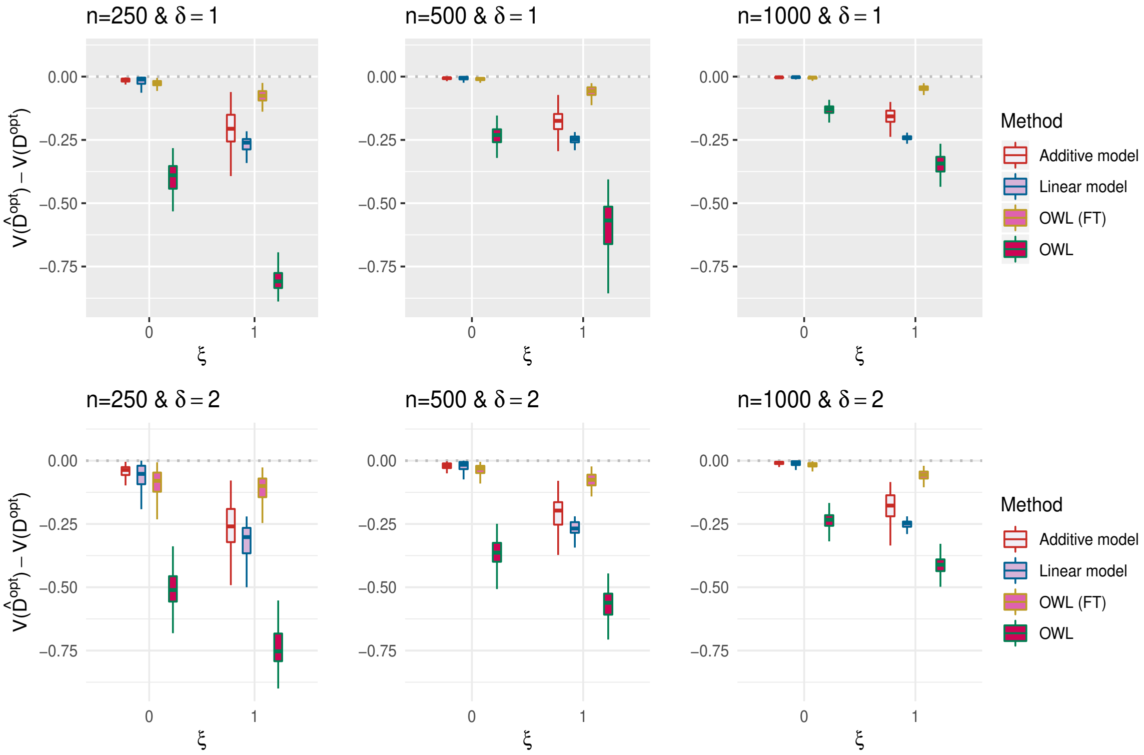

Fig. S.4. Boxplots based on 100 simulation runs, comparing the 4 approaches to estimating $\mathcal{D}^{opt}$, with respect to the (normalized) value $V(\hat{\mathcal{D}}^{opt}) - V(\mathcal{D}^{opt})$, given each scenario indexed by $\xi \in \{0, 1\}$, $\delta \in \{1, 2\}$ and a varying sample size $n \in \{250, 500, 1000\}$, for the *moderately-nonlinear* A-by-$\boldsymbol{X}$ interaction effect case. The dotted horizontal line represents the optimal value corresponding to $\mathcal{D}^{opt}$.

The results remain basically the same, in this increased number of covariates and sample size setting, as in the ones appearing in Figure 2 of the main manuscript. The additive model either performs at a similar level compared to the linear model or outperforms the linear model.

S.5.2 *Simulation results for a linear A-by-$\boldsymbol{X}$ interaction effect scenario*

In this subsection, as an extension of the simulation example in Section 4.2 of the main manuscript, we consider a case where the treatment effect varies linearly in the covariates, i.e., a "linear" A-by-$\boldsymbol{X}$ interaction effect scenario and assess the ITR estimation performance of the methods. Again, we generate a vector of covariates $\boldsymbol{X} = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ $(p = 50)$ based on a multivariate normal distribution with each component having the marginal distribution $\mathcal{N}(0, (\pi/2)^2)$ with the correlation between the components $\mathrm{corr}(X_j, X_k) = 0.1^{|j-k|}$. Given the same parametrization of the data model with $\delta \in \{1, 2\}$ and $\xi \in \{0, 1\}$ as in Section 4.2 of the main manuscript, we generate the response Y from:

$$Y = \delta \sum_{j=1}^{5} \sin(X_j) \; + \; (A - 1.5)\big\{X_1 - X_2 + \xi X_1 X_2\big\} \; + \; \epsilon \quad A \in \{1, 2\}, \tag{S.10}$$

where the treatment variable $A \in \{1, 2\}$ is generated independently from the covariates $\boldsymbol{X}$ and the error term $\epsilon \sim \mathcal{N}(0, 0.5^2)$, such that $\Pr(A = 1) = \Pr(A = 2) = 1/2$. For each combination of $n \in \{250, 500\}$, $\xi \in \{0, 1\}$ and $\delta \in \{1, 2\}$, we perform 100 simulation runs and compare the four approaches considered in the main manuscript. In Figure S.5, as in the simulation example of Section 4.2 of the main manuscript, we report the performance of the four approaches to estimating $\mathcal{D}^{opt}$ in terms of the (normalized) value $V(\hat{\mathcal{D}}^{opt}) - V(\mathcal{D}^{opt})$.

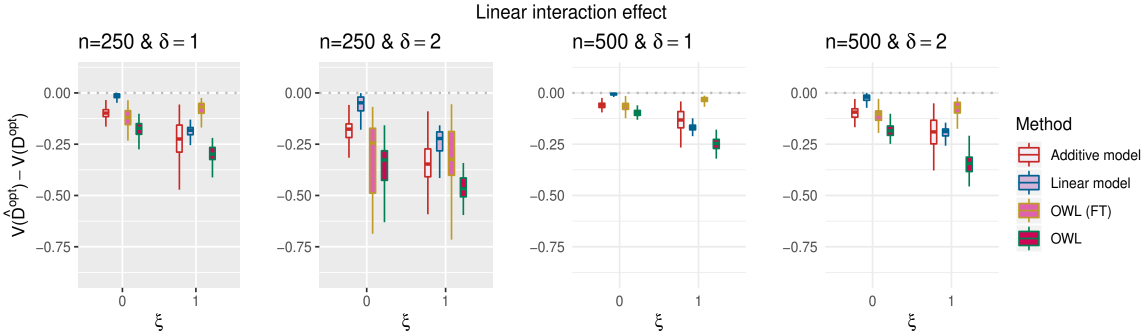

Fig. S.5. Boxplots based on 100 simulation runs, comparing the 4 approaches to estimating $\mathcal{D}^{opt}$, given each scenario indexed by $\xi \in \{0, 1\}$, $\delta \in \{1, 2\}$ and $n \in \{250, 500\}$, for the "*linear*" A-by-$\boldsymbol{X}$ interaction effect case. The dotted horizontal line represents the optimal value corresponding to $\mathcal{D}^{opt}$.

When $\xi = 0$ (i.e., when the linear interaction model is correctly specified), the linear regression outperforms the additive regression, but not much, whereas if the underlying model deviates from the exact linear structure (i.e., $\xi = 1$ in model (S.10)) and $n = 500$, the more flexible additive model tends to outperform the linear model. Given the outstanding performance of the additive model compared to the linear model in the nonlinear A-by-$\boldsymbol{X}$ interaction effect scenarios considered in the main manuscript, this result suggests that, in the absence of prior knowledge about the form of the interaction effect, flexible modeling of the interaction effect using the proposed additive regression can lead to good results in comparison to the linear regression.

S.5.3 *Comparison of OWL (FT) and OWL with only feature selection (FS) conducted*

We have additionally considered the OWL with only feature selection (FS) conducted (i.e., without any specific feature transformation), which we denote as OWL (FS). As in OWL (FT), we use the proposed additive regression for conducting variable selection (i.e., feature selection); however, unlike OWL (FT), we do not perform the transformation $X_j \mapsto g^*_{j,1}(X_j)$ on the selected features. For OWL (FS), the selected features, among $\{X_1, X_2, \ldots, X_p\}$, are used as inputs to the OWL.

We consider the same simulation setting (with $p = 50$) as in Section 4.2 of the main manuscript (see Figure 2 of the main manuscript for the results), but for the clarity of presentation, we focus on the $\delta = 2$ case (the case with $\delta = 1$ is qualitatively similar). The results are shown in Figure S.6.

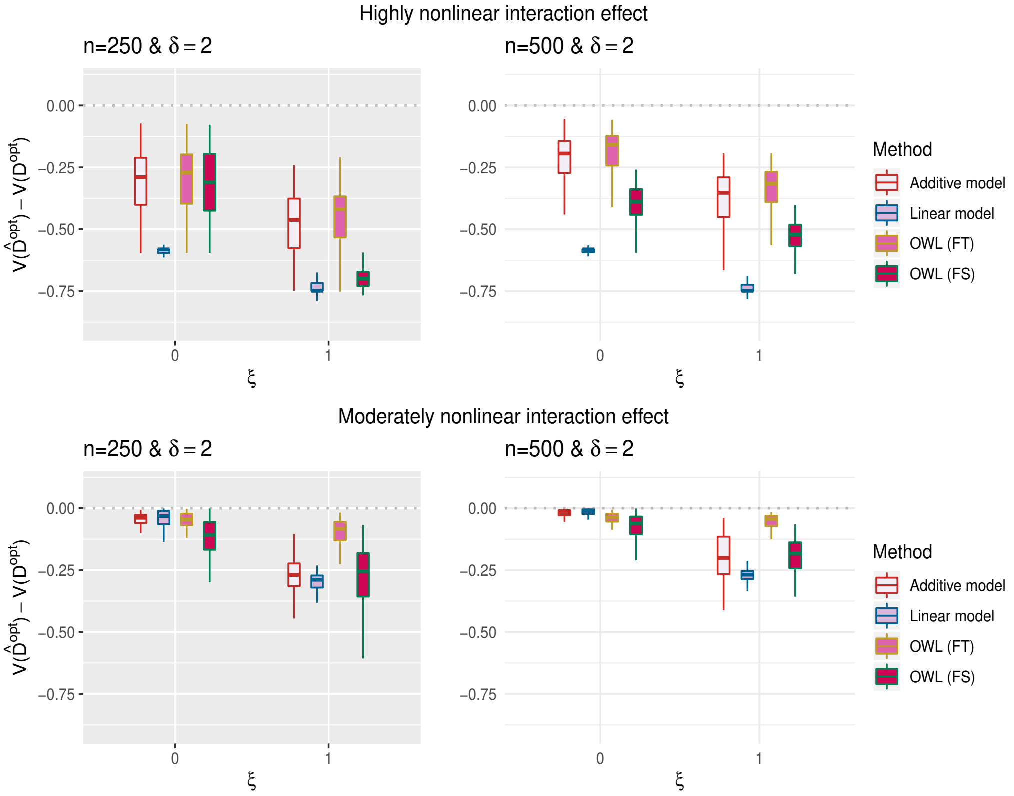

Fig. S.6. Boxplots based on 100 simulation runs, comparing: 1) additive model; 2) linear model; 3) OWL (FT); and 4) OWL (FS), given each scenario indexed by $\xi \in \{0, 1\}$ and $n \in \{250, 500\}$ (and $\delta = 2$), for the *highly-nonlinear* A-by-$\boldsymbol{X}$ interaction effect case in the top panels and the *moderately-nonlinear* A-by-$\boldsymbol{X}$ interaction effect case in the bottom panels.

The results in Figure S.6 indicate that OWL (FT) significantly improves OWL (FS) in the ITR estimation performance. This is especially true when the underlying data model is relatively more complex (i.e., when $\xi = 1$, in comparison to the case when $\xi = 0$). When $\xi = 1$, OWL (FT) generally outperforms the additive model; on the other hand, OWL (FS), which does not perform any feature transformation on the selected features, is often outperformed by the additive model. This suggests that the data-driven feature transformation via the component functions, g_j^*, of the proposed additive model can lead to useful representations for the input data used in OWL, particularly when the underlying model is complex.

S.5.4 *Simulation results for a case where there is a relatively large number of active treatment effect-modifiers but with treatment small effect-modifications*

We have performed a set of simulation experiments under a setting similar to the setting (S.10), with $\delta = 1$ and $\xi = 0$ (with $p = 50$), but using a relatively large number of nonzero A-by-$\boldsymbol{X}$ component functions (30 nonzero component functions associated with the A-by-$\boldsymbol{X}$ interaction effect): $Y = \sum_{j=1}^{5} \sin(X_j) + (A - 1.5) \sum_{j=1}^{30} \frac{e^{5X_j}}{1+e^{5X_j}} (-1)^j + \epsilon$. Although in this setting there were 30 covariates that modify the treatment effect, the magnitude of each covariate's modifying effect was set to be only one-fourth of that of X_1 (or X_2) of the setting (S.10), and therefore each of these 30 treatment effect-modifiers had a relatively small treatment modifying effect.

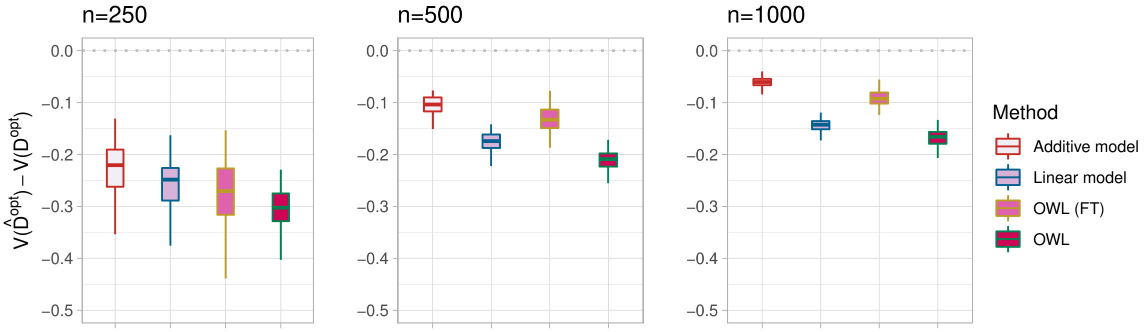

Fig. S.7. Boxplots based on 100 simulation runs given each scenario $n \in \{250, 500, 1000\}$, comparing the 4 approaches to estimating $\mathcal{D}^{opt}$, with respect to the (normalized) value $V(\hat{\mathcal{D}}^{opt}) - V(\mathcal{D}^{opt})$.

The results are given in Figure S.7. The results illustrate that, as the sample size increases from $n = 250$ to $n = 500$ and $n = 1000$, the performance level of the proposed additive model approaches the optimal performance. In particular, the relative advantage of using the additive model to the linear model is more prominent in comparison to the results reported in Figure 2 of the main manuscript, where there were only 2 active treatment effect-modifiers (see the "*moderately-nonlinear*" interaction effect scenario with $\xi = 0$ in Figure 2 of the main manuscript). The additive model-based feature-transformed OWL (i.e., OWL (FT)) outperforms the original

OWL as well as the linear model. On the other hand, the additive model outperforms the OWL (FT), which is not surprising, since the A-by-$\boldsymbol{X}$ interaction effect structure is of an "additive" form.

We further note that, in the binary treatment ($L = 2$) case, the proposed additive modeling approach can be viewed as creating an optimal 1-dimensional data-driven *index* $h(\boldsymbol{X}) = \sum_{j=1}^{p} \hat{g}^{*}_{j,1}(X_j)$ (see Figure 5 of the main manuscript) that can be used to determine an optimal treatment decision. The *index* is an additive combination of the p baseline covariates, that acts as a composite treatment effect-modifier that collectively exhibits a stronger, and possibly non-linear, interaction effect with the treatment. In such situations where there are potentially many treatment effect-moderators that individually contribute little, combining multiple biomarkers to generate a single, stronger composite treatment effect-modifier (as is done in the proposed additive model approach) is a clinically significant endeavor to optimizing treatment decisions.

S.6. Computation time

In Figure S.8, the computation time of the proposed additive regression approach implemented through the R-package `samTEMsel` (Park *and others*, 2020) is compared to that of the linear regression (MC) approach implemented through the R-package `glmnet` (Friedman *and others*, 2010). Even when $p = 200$ and $n = 500$, implementation of the proposed approach, including a 10-fold cross-validation, can be done within about 3 seconds. This is in sharp contrast to the computation time of the outcome-weighted learning considered in this paper, which is at least on the order of several minutes. (The computation time is measured on a MacBook computer running 64-bit, 2.5 GHz Intel Core i7, with 16 GB random access memory, and we report the average over 200 simulation runs.)

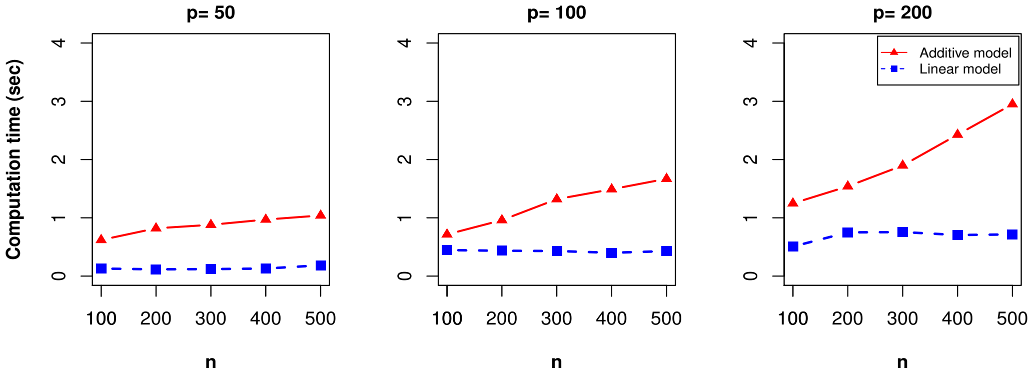

Fig. S.8. The averaged computation time (in seconds) (including the time to conduct 10-fold cross-validations for choosing the sparsity tuning parameters) with a varying $n \in \{100, 200, 300, 400, 500\}$ and $p \in \{50, 100, 200\}$, comparing the proposed sparse additive regression approach to the lasso-based linear regression approach.

One observation from Figure S.8 is that for the additive model, the computation time tends to increase linearly with n, whereas the computation time stays relatively flat with respect to n for the linear model. The major difference between these two approaches comes from a fact that the additive model needs to represent each observation in terms of a (spline) basis whereas the linear model does not involve such basis expansion/evaluation. The computation time for such a process typically scales linearly with the training sample size n.

References

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1), 1–22.

Park, H., Petkova, E., Tarpey, T. and Ogden, R.T. (2020). samTEMsel: Sparse additive models for treatment effect-modifier selection. *R package version 0.1.0. https://github.com/syhyunpark/samTEMsel.*

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *Journal of Royal Statistical Society: Series B* **71**, 1009–1030.

[*Received January 1, ****; revised January 1, ****; accepted for publication January 1, *****]