Econometric Theory, **35**, 2019, 1089–1110.
doi:10.1017/S0266466618000397

COMBINING ESTIMATES OF CONDITIONAL TREATMENT EFFECTS

CRAIG A. ROLLING
Saint Louis University

YUHONG YANG
University of Minnesota

DAGMAR VELEZ
Saint Louis University

Estimating a treatment's effect on an outcome conditional on covariates is a primary goal of many empirical investigations. Accurate estimation of the treatment effect given covariates can enable the optimal treatment to be applied to each unit or guide the deployment of limited treatment resources for maximum program benefit. Applications of conditional treatment effect estimation are found in direct marketing, economic policy, and personalized medicine. When estimating conditional treatment effects, the typical practice is to select a statistical model or procedure based on sample data. However, combining estimates from the candidate procedures often provides a more accurate estimate than the selection of a single procedure. This article proposes a method of model combination that targets accurate estimation of the treatment effect conditional on covariates. We provide a risk bound for the resulting estimator under squared error loss and illustrate the method using data from a labor skills training program.

1. INTRODUCTION

Estimating the causal effect of a treatment on a response is a primary goal of many empirical investigations, particularly in fields such as business, medicine, and public policy. Imbens and Wooldridge (2009) provide a review of treatment effect estimation in the context of program evaluation in econometrics. Accurate estimation of treatment effects is often challenging because of confounding and because each observed unit is assigned only one of the possible treatments; the latter issue was famously deemed the "fundamental problem of causal inference" in Holland (1986). Much causal inference research has focused on estimation of the treatment's average effect across a population. Imbens (2004) reviews a variety of methods for estimating the average treatment effect (ATE) from an observational study; Abadie and Imbens (2011) and Zhang (2016) provide two recently developed methods to estimate the ATE.

The authors would like to thank Arthur Lewbel and four anonymous referees for their helpful comments and suggestions that have improved the article. Address correspondence to Craig A. Rolling, Department of Epidemiology and Biostatistics, Saint Louis University, St. Louis, Missouri, USA; e-mail: craig.rolling@slu.edu.

© Cambridge University Press 2018

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

However, treatment effects are often heterogeneous within the population of interest. In the presence of treatment effect heterogeneity, when a treatment can be applied at the level of individual units (as in medicine or direct marketing), accurate estimation of the treatment's effect on different individuals can be used to increase the effectiveness of the treatment program in maximizing the outcome of interest. For example, a retailer with a limited marketing budget would be able to optimize a seasonal catalog mailing if it knew the effect of the catalog on the purchasing behavior of each household. In the public sector, an economic development agency often needs to decide which applicants will create the best utilization of limited grant dollars. Accurate estimation of individual treatment effects is central to the success of personalized medicine.

With the increasing volume of data becoming available to many organizations, estimating heterogeneous treatment effects has become more feasible. As a result, this topic has attracted much recent attention, particularly in the areas of business and econometrics. Much of the work on heterogeneous treatment effects has targeted the conditional average treatment effect (CATE). The CATE, formally defined in Section 2, describes the average treatment effect conditional on the value of a (scalar or vector-valued) covariate that is observed before the treatment is applied. Hsu (2016) illustrates a test of the hypothesis that the CATE is non-negative for every possible value of a subset of covariates. Chang, Lee, and Whang (2015) develop a class of nonparametric tests involving the CATE. Abrevaya, Hsu, and Lieli (2015) propose estimators for the CATE conditional on a subset of covariates. Taddy, Gardner, Chen, and Draper (2016) propose a nonparametric Bayesian approach for estimating heterogeneous treatment effects. Recently proposed methods for estimating the CATE in applications of biostatistics and political science include Cai, Tian, Wong, and Wei (2011), Green and Kern (2012), and Imai and Ratkovic (2013).

For data analysts, there are an ever-increasing number of different methods available for estimating conditional treatment effects. Therefore, for an analyst with a dataset in hand and treatment effect estimation as the goal, two questions naturally follow. First, which of these procedures will produce the best treatment effect estimates? Second, can the estimates from the procedures be combined to produce more accurate estimates?

Rolling and Yang (2014) address the first question by discussing model selection in the context of estimating the treatment effect conditional on covariates. They found that within a given candidate set of models, the best model for treatment effect estimation may be different than the best model for response estimation or prediction. This phenomenon also was discussed in Qian and Murphy (2011) in the context of optimizing treatment decisions. While targeted model selection tools are a step in the right direction for accurate estimation of treatment effects, post-model selection estimators from finite samples often have large variability because of model selection instability.

This article addresses the second question by introducing a model combination method for estimation of treatment effects. This method is (to the best of

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

our knowledge) the first combination algorithm specifically targeted at estimating the effect of a treatment conditional on covariates. It has been well-established (e.g., Yang, 2003) that for estimating the full regression function or forecasting, model combination algorithms often lead to more accurate estimates and predictions than model selection procedures when selection instability is high. The same is true for conditional treatment effect estimation, which motivates the work in this article.

In this article, the terms "model" and "procedure" have particular meanings that are helpful to understand. We consider a model to mean a parametric specification of possible distributions of the observations. Given such a model, the familiar estimation methods, e.g., MLE, MME and Bayes estimators, can be used to estimate the unknown parameters. A procedure, on the other hand, is any algorithm that provides an estimate of the conditional treatment effect (for our interest), given data. A formal definition of a treatment effect estimation procedure is given in Section 2. A parametric model, together with a chosen parameter estimation method, leads to an estimate of the treatment effect. This process of using the parametric model is a procedure for estimating the treatment effect. However, there are nonparametric procedures or machine learning algorithms that do not have any parametric family of distributions behind them. Even in the parametric framework, as soon as there are multiple candidate models and one uses a model selection method to choose one for final estimation, the final estimate is not from a model but based on a procedure (that selects one model from a list and then performs estimation).

The distinction between models and procedures is important because the method proposed in this article applies generally to procedure combination, despite "model combination" being the more common phrase. Our method assigns combination weights to members of a general collection of candidate procedures. By contrast, some other popular combination methods, such as Bayesian Model Averaging or model selection criterion based weighting, are not applicable to many nonparametric or machine learning procedures.

Furthermore, our proposed method is targeted to estimate *the conditional effect of a treatment* rather than the full regression function. Since models that are good for response estimation or prediction may not be good for treatment effect estimation (and vice versa), an algorithm targeted to the specific goal of treatment effect estimation is needed to ensure that models doing a good job of estimating the treatment effect receive higher weights. We call our method Treatment Effect Estimation by Mixing, abbreviated TEEM. (Incidentally, TEEM is a homophone for "team"; this is fitting because the candidate estimators work together to produce an accurate estimate.) TEEM relies on data splitting to evaluate the candidate procedures and therefore can combine multiple types of regression procedures and estimates; any procedure that, given data, produces an estimate of the treatment effect conditional on covariates can be used as a candidate in TEEM. Furthermore, the theoretical results we present for TEEM do not assume that any of the candidate models are correct. These features give the method flexibility

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

to be used in a wide variety of settings. However, it should be noted that in order to interpret the estimates produced by TEEM as conditional causal effects, the assumption of unconfounded assignment (sometimes called selection on observables) that is common in the program evaluation literature (e.g., Imbens and Wooldridge, 2009) must hold for the treatment, outcome, and covariates under study. This assumption is formalized in the next section.

The rest of the article is organized as follows. A mathematical formulation of the conditional treatment effect estimation problem is presented in Section 2. Section 3 introduces the TEEM algorithm and suggests how dimension reduction techniques may be used with TEEM in situations where there are many covariates. The method is applied to the benchmark National Supported Work Demonstration dataset (LaLonde, 1986) in Section 4, and in Section 5, we perform a simulation based on the same data to compare our method to other model selection and combination techniques. Section 6 contains some concluding remarks. An online supplement to this article, available at Cambridge Journals Online (journals.cambridge.org/ect), presents an alternative version of the TEEM algorithm and provides its theoretical justification in the form of a risk bound for estimation of the conditional treatment effect. The supplement also contains additional simulation studies.

2. FRAMEWORK

We consider a general regression framework in which the distribution of the response Y may depend on a binary treatment variable $T \in \{t, c\}$ and one or more baseline covariates $\mathbf{U} \in \mathbb{R}^p$. In order to isolate the treatment difference of primary interest, we express the observations as follows:

$$Y_i = [f_t(\mathbf{U}_i) + \sigma_t(\mathbf{U}_i)\varepsilon_i]I(T_i = t) + [f_c(\mathbf{U}_i) + \sigma_c(\mathbf{U}_i)\varepsilon_i]I(T_i = c), \qquad 1 \le i \le n. \tag{1}$$

In the above, $f_t(\mathbf{u}) \equiv E(Y_i|T_i = t, \mathbf{U}_i = \mathbf{u})$ and $f_c(\mathbf{u}) \equiv E(Y_i|T_i = c, \mathbf{U}_i = \mathbf{u})$ denote the regression mean functions for the treatment and control groups, respectively. The covariates $\mathbf{U}_i$ are assumed to be i.i.d. from some unknown distribution $P_{\mathbf{U}}$ with support $\mathcal{U}$, and realized values of $\mathbf{U}$ are denoted by $\mathbf{u}$. The ε_i are i.i.d. noise variables independent from $(T_i, \mathbf{U}_i)_{i=1}^n$ with a known probability density function $\phi(t), t \in \mathbb{R}$, having zero mean and unit variance (often ϕ will be taken to be the standard Gaussian pdf). The error variance functions are allowed to differ under treatment and control, and these are denoted by σ_t^2 and σ_c^2, respectively. As suggested by the notation in (1), these variance functions may be heteroscedastic with respect to $\mathbf{U}$. The treatment variable T_i may be deterministic or random.

The object of interest in our work is $\Delta(\mathbf{u}) := f_t(\mathbf{u}) - f_c(\mathbf{u})$, the difference between the regression functions under treatment and control. As we discuss next, some conditions are needed to properly interpret $\Delta(\mathbf{u})$ as a causal effect. We define causal effects using the potential outcomes framework of the Rubin Causal

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

Model (Holland, 1986). That is, let $Y_{i,(t)}$ denote the response that would have been observed had $T_i = t$, and let $Y_{i,(c)}$ denote the corresponding potential outcome if $T_i = c$. Then the causal effect of the treatment T on unit i is the unobserved random variable $Y_{i,(t)} - Y_{i,(c)}$. Following Imbens and Wooldridge (2009), we define the Conditional Average Treatment Effect (CATE) as the expectation of this random variable conditional on the observed value of the covariate vector $\mathbf{U}_i$:

$$\text{CATE}(\mathbf{u}) := \mathbb{E}[(Y_{i,(t)} - Y_{i,(c)})|\mathbf{U}_i = \mathbf{u}]. \tag{2}$$

Note that $\Delta(\mathbf{u}) = \mathbb{E}(Y_i|T_i = t, \mathbf{U}_i = \mathbf{u}) - \mathbb{E}(Y_i|T_i = c, \mathbf{U}_i = \mathbf{u})$. We call Δ the Conditional Average Treatment Difference (CATD) to distinguish it from the CATE because Δ may be influenced by unobserved confounding variables and therefore not accurately represent the conditional effect of the treatment.

If the pair of potential outcomes for each i is conditionally independent of the treatment given covariates; that is,

$$(Y_{i,(t)}, Y_{i,(c)}) \perp\!\!\!\perp T_i|\mathbf{U}_i, \tag{3}$$

then $\Delta(\mathbf{u}) = \text{CATE}(\mathbf{u})$ and therefore $\Delta(\mathbf{u})$ represents the (mean) causal effect of the treatment variable given the covariate value $\mathbf{u}$. The assumption (3) is known to hold in randomized experiments, but it only holds in observational studies if all variables confounding the treatment and response are observed in the covariate vector $\mathbf{U}$. In this work, we focus on estimation of Δ because it is identifiable in most experiments and observational studies. However, it is important to keep in mind that in order for $\Delta(\mathbf{u})$ to represent a causal effect of the treatment variable T, additional assumptions such as (3) are needed.

We define the L_2 norm with respect to the probability distribution of the covariates,

$$\|f\|_2 := \left[\int_{\mathcal{U}} |f(\mathbf{u})|^2 P_{\mathbf{U}}(d\mathbf{u})\right]^{1/2}, \tag{4}$$

where $P_{\mathbf{U}}$ denotes the probability distribution of $\mathbf{U}_i$ for $1 \le i \le n$. This norm will be used to measure the average discrepancy between Δ and various estimates $\widehat{\Delta}$ over $\mathcal{U}$.

The TEEM method involves combining a finite collection of procedures proposed for estimating the function Δ. A treatment effect estimation procedure, denoted by ψ, refers to a method of estimating the functions Δ, σ_t^2, and σ_c^2 based on $\mathbf{Z}^l = (Y_i, T_i, \mathbf{U}_i)_{i=1}^l$ at each sample size l. (Estimators of σ_t^2 and/or σ_c^2 can be shared between the procedures if desired.) Let ψ_j, $1 \le j \le J$, denote the candidate treatment effect estimation procedures, and let $\widehat{\Delta}_{l,j}$, $\hat{\sigma}^2_{t,l,j}$, and $\hat{\sigma}^2_{c,l,j}$ denote the estimators of the functions Δ, σ_t^2, and σ_c^2, respectively, resulting from the application of procedure ψ_j to the data $\mathbf{Z}^l$.

For the TEEM method, ψ may be any sort of procedure or algorithm that, given data, produces estimates of Δ, σ_t^2, and σ_c^2. A given candidate ψ_j could represent

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

a parametric, nonparametric, or semiparametric statistical model, a nonstatistical machine learning procedure, or subjective expert judgment. Some candidate procedures may utilize only a subset of the covariates or none of them at all (as in the application described in Section 4); nevertheless, they are considered by our method to have the same estimands (Δ, σ_t^2, and σ_c^2) because they provide estimates of these functions at any value of $\mathbf{u}$. The TEEM method of procedure combination is very general and has the ability to combine different types of estimators. This flexibility is an important advantage for solving the problem of treatment effect estimation because, as discussed in Section 1, there are a variety of approaches proposed in the literature for estimation of conditional treatment effects.

The proposed TEEM algorithm provides combining weights for any finite collection of candidate procedures that estimate the function Δ. While additional assumptions such as (3) may be necessary to properly interpret $\Delta(\mathbf{u})$ as a causal effect conditional on $\mathbf{u}$, the purpose of TEEM is to accurately estimate Δ, not to ensure that (3) holds for the resulting estimator. Of course, (3) often can be ensured by randomization or other considerations at the study design stage.

3. TREATMENT EFFECT ESTIMATION BY MIXING (TEEM)

The TEEM algorithm for combining estimates of conditional treatment effects is based on data splitting, as in cross-validation. The candidate procedures are fit using a subset of the data (called the training set) and evaluated on the remaining subset (the evaluation set). Traditional cross-validation for procedure selection compares the individual responses in the evaluation set with their predicted values from the candidate procedures. However, for treatment effect estimation, since each observational unit is in either the treatment or control group, individual treatment effects are not available to compare with the estimates. Our solution to this problem is to approximate individual treatment effects in the evaluation data by using pairs of nearby observations, one from each treatment group.

Suppose we have a pair of observations (i, j) such that $T_i = t$ and $T_j = c$. If individuals i and j have the same baseline covariates ($\mathbf{U}_i = \mathbf{U}_j$), then within the framework of the previous section, $Y_i - Y_j$ is an observation from $N(\Delta(\mathbf{U}_i), \sigma_t^2(\mathbf{U}_i) + \sigma_c^2(\mathbf{U}_i))$ and this difference can be used to evaluate the accuracy of estimates $\widehat{\Delta}(\mathbf{U}_i)$. If the covariates of i and j do not match, then $Y_i - Y_j \sim N(\Delta(\mathbf{U}_i) + (f_c(\mathbf{U}_i) - f_c(\mathbf{U}_j)), \sigma_t^2(\mathbf{U}_i) + \sigma_c^2(\mathbf{U}_j))$, and the bias for $Y_i - Y_j$ as an estimate of $\Delta(\mathbf{U}_i)$ is represented by $f_c(\mathbf{U}_i) - f_c(\mathbf{U}_j)$. If the distance between $\mathbf{U}_i$ and $\mathbf{U}_j$ is small and the control regression function f_c is smooth, this bias will be small and the paired difference $Y_i - Y_j$ will be a nearly unbiased estimate of $\Delta(\mathbf{U}_i)$. The TEEM algorithm uses differences between such nearby treatment/control pairs to evaluate the candidate estimates of treatment effects and assign to them appropriate weights. Section 3.1 below describes the TEEM algorithm in detail.

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

3.1. The TEEM Algorithm

Step 0. Select a fraction $\rho \in (0, 1)$ of the n observations that will be used to fit the models. Denote $\lfloor \rho n + 0.5 \rfloor$ by n_1; n_1 is the number of observations used to fit the models. Similarly denote the size of the evaluation set, $n - n_1$, by n_2.

Step 1. Randomly permute the order of the n observations; call this permutation π. Split the resulting ordered data into two parts: the training part $\mathbf{Z}^{(1)} = (Y_i, T_i, \mathbf{U}_i)_{i=1}^{n_1}$ and the evaluation part $\mathbf{Z}^{(2)} = (Y_i, T_i, \mathbf{U}_i)_{i=n_1+1}^{n}$.

Step 2. For each unit i in $\mathbf{Z}^{(2)}$ (regardless of its treatment status), let i^* denote its nearest neighbor, in terms of normalized Euclidean distance, from the other treatment group in $\mathbf{Z}^{(2)}$. Specifically, let

$$\widetilde{\mathbb{U}} = \begin{bmatrix} \widetilde{\mathbf{U}}_1^T \\ \vdots \\ \widetilde{\mathbf{U}}_n^T \end{bmatrix}$$

denote the standardized version of the covariate design matrix $\mathbb{U}$, so that each column of $\widetilde{\mathbb{U}}$ has unit variance, and let $d(\cdot)$ denote Euclidean distance. Then for each unit i in $\mathbf{Z}^{(2)}$, define i^* as an index such that $d(\widetilde{\mathbf{U}}_i, \widetilde{\mathbf{U}}_{i^*})$ is the smallest among $d(\widetilde{\mathbf{U}}_i, \widetilde{\mathbf{U}}_k)$ with $T_i \neq T_k$ and $n_1 + 1 \leq k \leq n$. (In the case of discrete covariates, there may be multiple observations in the other treatment group that are equidistant from i. Any tiebreaking method may be used to choose one i^* in this case.)

In some situations, it may be useful to apply a caliper to bound the matching discrepancy so that none of the matched pairs are too far apart. Althauser and Rubin (1970) and others have argued that caliper matching can remove a large percentage of the total bias induced by matching while only removing a small percentage of the matched pairs. For TEEM with nearest neighbors, if a caliper is applied, observations without a "close enough" match in the other treatment group are not used to evaluate the treatment effect estimation procedures.

Step 3. For each of the n_2 treatment-control pairs produced by Step 2 (or the $\widetilde{n}_2$ pairs for some $\widetilde{n}_2 \leq n_2$ if a caliper is applied), create approximate treatment effects

$$\widetilde{\delta}_i = [2I(T_i = t) - 1](Y_i - Y_{i^*}). \tag{5}$$

In other words, for each pair (i, i^*), $\widetilde{\delta}_i$ is the response of the treated unit minus the response of the control unit.

Step 4. For each procedure $j = 1, \ldots, J$, fit the procedure ψ_j to the data $\mathbf{Z}^{(1)}$ to obtain $\widehat{\Delta}_{n_1,j}$ and

$$\hat{\sigma}_{\widetilde{\delta}_i, n_1, j} = \begin{cases} \sqrt{\hat{\sigma}^2_{t,n_1,j}(\mathbf{U}_i) + \hat{\sigma}^2_{c,n_1,j}(\mathbf{U}_{i^*})} & \text{if } T_i = t \\ \sqrt{\hat{\sigma}^2_{t,n_1,j}(\mathbf{U}_{i^*}) + \hat{\sigma}^2_{c,n_1,j}(\mathbf{U}_i)} & \text{if } T_i = c. \end{cases}$$

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

Step 5. Use the performance of the estimates $\widehat{\Delta}_{n_1,j}$ on the evaluation data to create the weights $W_{\pi,j}$, where the subscript π represents the permutation applied in Step 1A.

$$W_{\pi,j} = \frac{\prod_{i=n_1+1}^{n} \phi\left\{\left[\widetilde{\delta}_i - \widehat{\Delta}_{n_1,j}(\mathbf{U}_i)\right]/\hat{\sigma}_{\widetilde{\delta}_i,n_1,j}\right\}/\hat{\sigma}_{\widetilde{\delta}_i,n_1,j}}{\sum_{k=1}^{J}\prod_{i=n_1+1}^{n} \phi\left\{\left[\widetilde{\delta}_i - \widehat{\Delta}_{n_1,k}(\mathbf{U}_i)\right]/\hat{\sigma}_{\widetilde{\delta}_i,n_1,k}\right\}/\hat{\sigma}_{\widetilde{\delta}_i,n_1,k}}, \tag{6}$$

where ϕ is the error density.

Step 6. Repeat Steps 1A–5A a total of P times for some $P \geq 1$. For each j average the weights $W_{\pi,j}$ over the permutations to create averaged weights $\overline{W}_j$.

$$\overline{W}_j = \frac{1}{P}\sum_{p=1}^{P} W_{\pi_p,j}, \tag{7}$$

where for each iteration $1 \leq p \leq P$, π_p denotes the permutation applied in Step 1A of the iteration. Note that $\sum_{j=1}^{J}\overline{W}_j = 1$.

Step 7. Fit each candidate procedure j to the entire sample of n observations to obtain estimates $\widehat{\Delta}_{n,j}(\mathbf{u})$ for any $\mathbf{u} \in \mathcal{U}$.

Step 8. Create the final TEEM estimator

$$\overline{\widehat{\Delta}}_A(\mathbf{u}) = \sum_{j=1}^{J}\overline{W}_j\widehat{\Delta}_{n,j}(\mathbf{u}). \tag{8}$$

3.2. The Two Versions of TEEM

We have developed two related versions of the TEEM algorithm for this article. The version of TEEM outlined in Section 3.1 is intended for practical use and is demonstrated in the numerical results presented in Sections 4 and 5. In the online supplement to this article, we present an alternate version of the algorithm, called TEEM_A, to facilitate theoretical development. In TEEM_A, the treatment-control pairs used to evaluate the candidate procedures are produced by partitioning the covariate space instead of by searching for nearest neighbors. The partitioning in TEEM_A ensures that the paired differences are independent of each other and that the two units within each pair are sufficiently close in $\mathbf{U}$. In the supplement, it is shown that under some regularity conditions, a partition can be constructed that enables the resulting TEEM_A estimator to achieve the desired risk bound.

Practically, if the data results from a matched study design and the sample size is large relative to the number of covariates, possession of many independent nearby pairs may be realistic. In the absence of these conditions, however, requiring the treatment-control pairs to be independent (as in TEEM_A) seems too restrictive. In particular, much economic analysis features unbalanced data and large numbers of covariates. In a related context for estimating average treatment effects, Abadie and Imbens (2006) showed that matching with replacement (as in

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

the TEEM algorithm of Section 3.1) produces higher-quality matches than does matching without replacement. For similar reasons, we recommend the version of TEEM described in Section 3.1 for general practical use. The numerical analysis in Sections 4 and 5 uses the version of TEEM described in Section 3.1.

3.3. Sufficient Dimension Reduction for TEEM

Both versions of the TEEM algorithm are subject to the so-called "curse of dimensionality" in the pairing step. If p is moderate or large, for a given observation i it may be difficult to find an observation in the other treatment group that possesses covariates similar to $\mathbf{U}_i$. The pairing task becomes much easier if the dimension of the vector used for pairings can be reduced. The framework of sufficient dimension reduction (see Cook, 1998 for an overview) can be helpful for identifying ways to reduce the dimension of the vector used in pairings without any loss of information relevant to our target Δ.

In the framework of sufficient dimension reduction, a mean dimension reduction subspace S for the regression of Y on $\mathbf{U}$ is a subspace such that $Y \perp\!\!\!\perp E(Y|\mathbf{U})|\eta^T\mathbf{U}$, where $\perp\!\!\!\perp$ denotes independence and η is a matrix whose columns form a basis in S. When the intersection of all mean dimension reduction subspaces is itself a mean dimension reduction subspace, it is called the central mean subspace and denoted by $S_{E(Y|\mathbf{U})}$. Under mild conditions on the regression, the central mean subspace exists and is unique (Cook and Li, 2002). In many cases, $\dim(S_{E(Y|\mathbf{U})})$ is much lower than the dimension of $\mathbf{U}$. For example, $\dim(S_{E(Y|\mathbf{U})}) = 1$ in linear models and generalized linear models.

In Section 2, f_t and f_c are used to denote the regression (mean) functions of Y on $\mathbf{U}$ for the treatment and control groups, respectively. If we let S_t and S_c denote the central mean subspaces for these regressions, then $\Delta(\mathbf{u}) = \Delta(\eta_{tc}^T\mathbf{u})$, where η_{tc} is a matrix whose columns span both S_t and S_c. Therefore, the central mean subspaces for the treatment and control regressions can be combined to produce a projection of $\mathbf{U}$ that does not result in any loss of information about Δ. If we let $S_{tc} = \{s_t + s_c | s_t \in S_t, s_c \in S_c\}$, then S_{tc} contains all of the information that $\mathbf{U}$ has about Δ and has dimension at most $\dim(S_t) + \dim(S_c)$.

Let d_{tc} denote the dimension of S_{tc}. If $d_{tc} \ll p$, then finding a treatment-control pair (i, i^*) such that $d(\eta_{tc}^T\mathbf{U}_i, \eta_{tc}^T\mathbf{U}_{i^*})$ is small is much easier than finding a pair for which $d(\mathbf{U}_i, \mathbf{U}_{i^*})$ is small. Because S_{tc} is sufficient for Δ, Theorem 1 in the online supplement will hold if the m cells are created from a partition of $S_{tc}\mathcal{U}$ and p is replaced by d_{tc} throughout the algorithm, theorem, and proof. If $d_{tc} \ll p$, such a dimension reduction will reduce the risk bound and increase the convergence rate substantially.

The benefits of sufficient dimension reduction in our setting can be realized in practice. Methods available to estimate vectors in the central mean subspace (or the more general central subspace; see Cook and Li, 2002) without assuming a model for the data include sliced inverse regression (SIR) (Li, 1991), iterative Hessian transformation (IHT) (Cook and Li, 2002), and large-p methods

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

proposed in Li (2007) and Li and Yin (2008). For linear models, estimation of the central mean subspace is equivalent to estimation of the regression coefficients. The simulation study in Section S2.2 of the online supplement demonstrates the use of sufficient dimension reduction with the TEEM algorithm and illustrates the potential benefit of this approach.

4. APPLICATION: LABOR TRAINING PROGRAM

4.1. The LaLonde Data

In this section we apply the TEEM method to the well-known LaLonde (1986) National Supported Work (NSW) Demonstration data set. The NSW Demonstration was a federally and privately funded program in the 1970s that provided work experience to individuals who were struggling financially. Eligible participants were randomly assigned to the treatment or control group, and follow-up interviews were conducted with both groups to obtain information about post-intervention earnings. LaLonde (1986) analyzed the male and female participants separately, and we will focus on the study's male participants. The male participants from this experiment were previously analyzed by Dehejia and Wahba (1999) in a study of propensity scores and by Imai and Ratkovic (2013), who used a penalized regression method to estimate heterogeneity of the treatment effect.

There were $n = 722$ male participants in the experiment; 297 were treated and 425 were in the control group. The outcome variable Y in our analysis is the change in the square root of income from 1975 (pretreatment) to 1978 (post-treatment); square root transformations on income are done to reduce skewness. The treatment variable T equals 1 if the person was treated in the NSW demonstration and 0 otherwise. Four baseline covariates, measured before the treatment was applied, are used to identify heterogeneity of the treatment effect in some of the candidate procedures. These covariates are $\sqrt{1975\text{ income}}$, age, years of education, and marital status, with variable names Inc75, Age, Educ, and Married, respectively. Racial indicators for black and Hispanic individuals also are available in the LaLonde dataset, but these variables were not used because a preliminary analysis provided no evidence that race moderated the treatment effect. Descriptions, means, and standard deviations for each of the variables used in our analysis are given in Table 1.

TABLE 1. Variables used in LaLonde NSW data analysis

Name	Type	Description	Mean	SD
Y	Outcome	$\sqrt{1978\text{ income}} - \sqrt{1975\text{ income}}$	20.4	56.4
T	Treatment	$T = 1$ if enrolled in training; otherwise, $T = 0$	0.411	0.492
Inc75	Covariate	$\sqrt{1975\text{ income}}$	37.5	40.5
Educ	Covariate	Years of education	10.3	1.7
Age	Covariate	Age in years	24.5	6.6
Married	Covariate	Married = 1 if married; otherwise; married = 0	0.162	0.369

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

4.2. Candidate Procedures

A feature of this analysis that seems typical of many studies of treatment effect heterogeneity is the plausibility of nonlinear treatment effects with respect to some of the covariates. For example, intuitively it seems quite possible that the job training program may be most beneficial for those in the middle of the income, education, or age distributions. Therefore, our set of candidate procedures includes linear models and additive models that are possibly nonlinear.

Preliminary analysis shows that baseline income (Inc75) is clearly related to the response, so this covariate is included in every candidate model. The linear candidate models are those containing different subsets of the variables {T, Educ, Age, Married, T*Inc75, T*Educ, T*Age, T*Married}. Model hierarchy is enforced, meaning that if a treatment-covariate interaction is included in the model, both corresponding main effects must be included as well. This constraint applied to the above set of variables allows for 62 possible linear models.

The additive candidate models are estimated with the `gam` function from the R `mgcv` package (Wood, 2006). Each term in the additive model is a smooth, possibly nonlinear function of a single variable. Treatment-covariate interactions are estimated by allowing these functions to differ for each covariate depending on the value of the treatment variable. The default choice of smoothing parameter (based on generalized cross-validation) is used to fit each model. Terms involving categorical variables (T, Married, and T*Married) in the additive models are linear. Model hierarchy is similarly enforced as in the linear model consideration set, generating 62 possible additive models and 124 candidate procedures in all.

Sixteen of the models have no treatment effects (i.e., $\widehat{\Delta} = 0$). Since these models produce different estimates of the full regression function, they are considered as separate procedures in the TEEM combination algorithm so that all selection and combination methods utilize the same set of candidates. Note that each of these models will have the same weight $\overline{W}_j$ in the TEEM algorithm.

4.3. Model Selection and Combination Methods

Nine model selection or combination methods were applied to select from (or combine) the candidate procedures described in the previous section.

Model Selection Methods. Five methods of model selection are used to choose one of the 124 candidate procedures. Two of these are the familiar criteria of AIC (Akaike, 1974) and BIC (Schwarz, 1978). These likelihood-based criteria can be used to compare additive models to linear models because the additive model can be represented as a penalized likelihood, where the penalty is a measure of the roughness of the function. The third model selection method used is the traditional cross-validation (CV) that uses individual prediction errors of the response to select a model. To implement CV, we use half of the data to estimate the candidate models, then compute the average squared prediction error of each candidate

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

estimate on the other half. We repeat the process 100 times to average out the variability in data splitting, and the model chosen by CV is the one with the lowest average squared error over the 100 splits.

The weighted FIC (wFIC; Claeskens and Hjort, 2008) with Δ as the focus parameter is the fourth model selection method used. See Rolling and Yang (2014) for more details on the derivation of wFIC for targeting Δ. Note that the wFIC is restricted to compare and choose among the linear candidate models because (to the best of our knowledge) wFIC is not suitable for comparing linear and nonlinear models of the focus parameter. Finally, we apply treatment effect cross-validation (TECV; Rolling and Yang, 2014), a form of cross-validation targeted to select a procedure that accurately estimates the treatment effect. As with traditional CV, we randomly split the data into equal-sized training and evaluation samples 100 times. The model with the lowest average TECV statistic over the 100 splits is chosen by treatment effect cross-validation.

Model Combination Methods. Four model combination methods are analyzed in this study. Each of them forms an estimate of the treatment effect Δ from a convex combination of the candidate procedures. That is, each combination method produces a $\widehat{\Delta}$ function by

$$\widehat{\Delta}(\mathbf{u}) = \sum_{j=1}^{J} w_j \widehat{\Delta}_{n,j}(\mathbf{u}), \tag{9}$$

where $J = 124$ in this case and $\widehat{\Delta}_{n,j}$ is the estimate of Δ produced by applying procedure j to the entire sample of size n. The combination methods differ in their choices of the weights w_j.

One benchmark method of model combination is Bayesian Model Averaging (BMA), in which each w_j represents the posterior probability of model j. Raftery (1995) derived the approximation

$$w_j = \frac{\exp(-\frac{1}{2}\mathrm{BIC}_j)}{\sum_{j=1}^{J}\exp(-\frac{1}{2}\mathrm{BIC}_j)} \tag{10}$$

to the posterior probability of model j when the models have equal prior probabilities. Buckland, Burnham, and Augustin (1997) suggested combining models based on AIC by replacing BIC_j with AIC_j in (10). We refer to these two weighting schemes as BMA and cAIC, respectively. Adaptive Regression by Mixing (ARM; Yang, 2001) was discussed briefly in Section 1. Essentially, the ARM weights are based on each model's relative ability to predict the response on evaluation data. We construct the ARM weights w_j by assuming normal errors, using each procedure separately to estimate the error variance, and averaging the weights over 100 different 50/50 data splittings. Finally, the version of TEEM described in Section 3.1 is used with no caliper, assuming normal errors, and averaging over 100 different 50/50 data splittings. To compare the methods most

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

accurately, the same 100 random data splittings are used for CV, TECV, ARM and TEEM.

4.4. Results

Table 2 summarizes the results of the model selection and combination methods when applied to the LaLonde data. For each model selection method, the table lists the type of model (additive or linear) that was selected, along with the active variables in the model. For each selection and combination method ν, the mean and standard deviation (SD) of the resulting estimator $\widehat{\Delta}_\nu(\mathbf{U}_i)$ over the 722 data values of $\mathbf{U}_i$ are reported in Table 2. The standard deviation of $\widehat{\Delta}_\nu(\mathbf{U}_i)$ measures the amount of treatment effect heterogeneity indicated by the estimator resulting from the method ν.

Interestingly, each of the five model selection methods considered in our study chooses a different model, with each implying something different about the treatment's effect on the outcome. The additive model selected by AIC implies the treatment effect varies nonlinearly with baseline income and age, as well as being different for married vs. single people. The model selected by BIC implies the NSW treatment has no effect at all on the outcome, while the model chosen by traditional CV implies a homogeneous positive treatment effect. Among the set of linear candidate models, wFIC chooses one containing treatment interactions with marital status and baseline income. Finally, TECV selects a linear model implying the treatment effect differs only by marital status. The coefficient of the *T**Married interaction is positive in the model selected by TECV, indicating the treatment provides a greater benefit to married men.

TABLE 2. $\widehat{\Delta}_\nu(\mathbf{u})$ estimates from model selection and combination methods applied to the LaLonde NSW data

			$\widehat{\Delta}_\nu(\mathbf{U}_i)$ values[b]	
Method	Model Type	Active variables[a]	Mean	SD
AIC	Additive	*T**s(Inc75), *T**s(Educ), *T**married, s(age)	5.7	12.5
BIC	Linear	Inc75	0	0
CV	Linear	T, Inc75	6.6	0
wFIC	Linear	*T**Inc75, *T**married	6.6	8.6
TECV	Linear	Inc75, *T**married	6.7	6.4
cAIC			5.6	6.0
BMA			1.4	0.1
ARM			5.3	2.9
TEEM			5.6	3.9

[a] The presence of interaction terms implies the presence of both main effects.
[b] The mean and standard deviation of $\widehat{\Delta}_\nu(\mathbf{U}_i)$ over the $n = 722$ sample values of $\mathbf{U}_i$, where ν denotes a model selection or combination method that produces an estimate $\widehat{\Delta}(\mathbf{u})$.

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

The top half of Table 2 suggests substantial model selection uncertainty in this analysis. In such situations, model combination often provides a good compromise between similar-performing models that give quite different estimates. In Figure 1, the $\widehat{\Delta}_v(\mathbf{U}_i)$ values from the LaLonde data resulting from AIC, CV, TECV, and TEEM are plotted against pretreatment income. There is substantial treatment effect heterogeneity suggested by $\widehat{\Delta}_{\text{AIC}}$, with an overall negative association between $\widehat{\Delta}_{\text{AIC}}$ and pretreatment income. For the models selected by CV and TECV, there is no interaction between the treatment and baseline income. The TEEM model combination lies somewhere in between these extremes, exhibiting some negative association between $\widehat{\Delta}_{\text{TEEM}}$ and pretreatment income but with much less variability than $\widehat{\Delta}_{\text{AIC}}$. Although we do not know how the treatment effect truly varies with pretreatment income, it seems reasonable to believe that those who entered the program with a higher income may have benefited less from the program overall. At the same time, it seems unlikely that the treatment effect is as heterogeneous as $\widehat{\Delta}_{\text{AIC}}$ suggests.

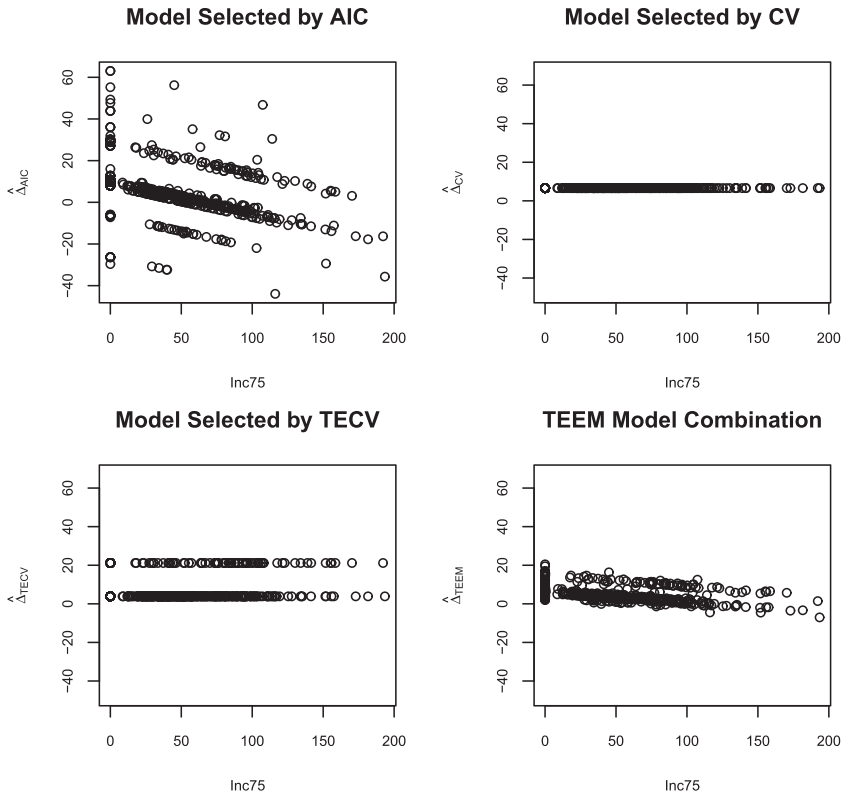


FIGURE 1. $\widehat{\Delta}_v(\mathbf{U}_i)$ from methods AIC, CV, TECV and TEEM plotted against the Inc75 variable for the $n = 722$ observations in the LaLonde NSW data.

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

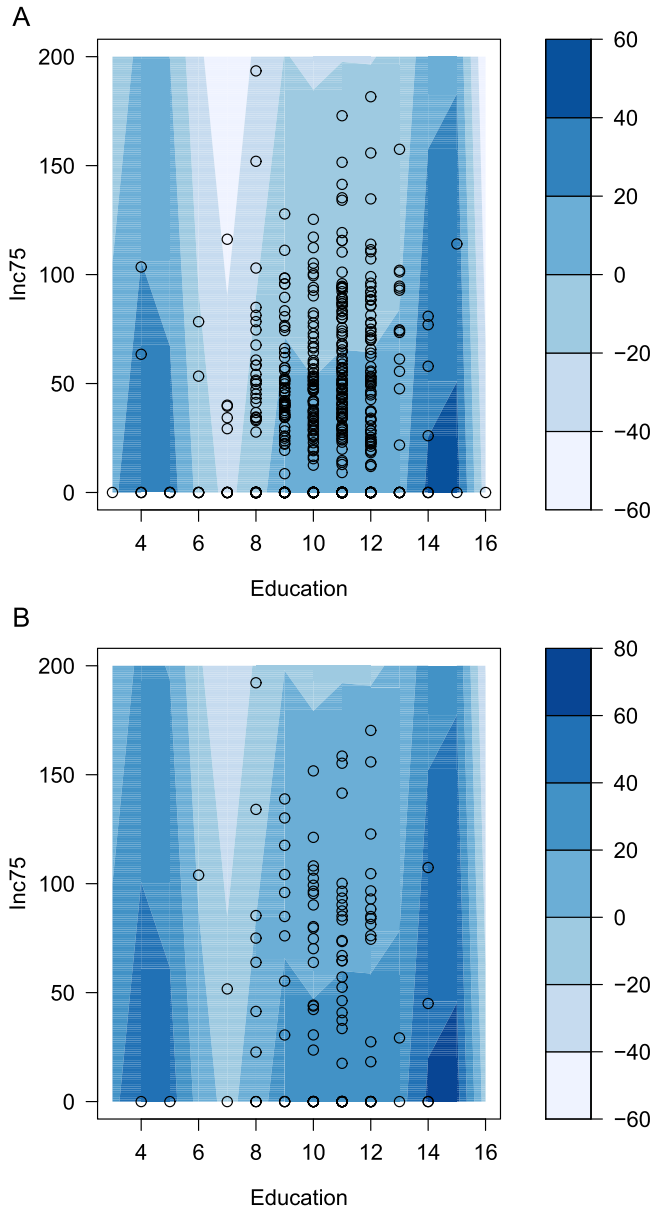


FIGURE 2. Contour plots for $\widehat{\Delta}_{\mathrm{AIC}}$ for the LaLonde NSW data. The circles indicate the locations of one or more original data points.

The decreased variability in $\widehat{\Delta}_{\mathrm{TEEM}}$ compared with $\widehat{\Delta}_{\mathrm{AIC}}$ can also be seen in Figures 2 and 3. These are contour plots of $\widehat{\Delta}_{\nu}$ from AIC and TEEM, respectively,

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

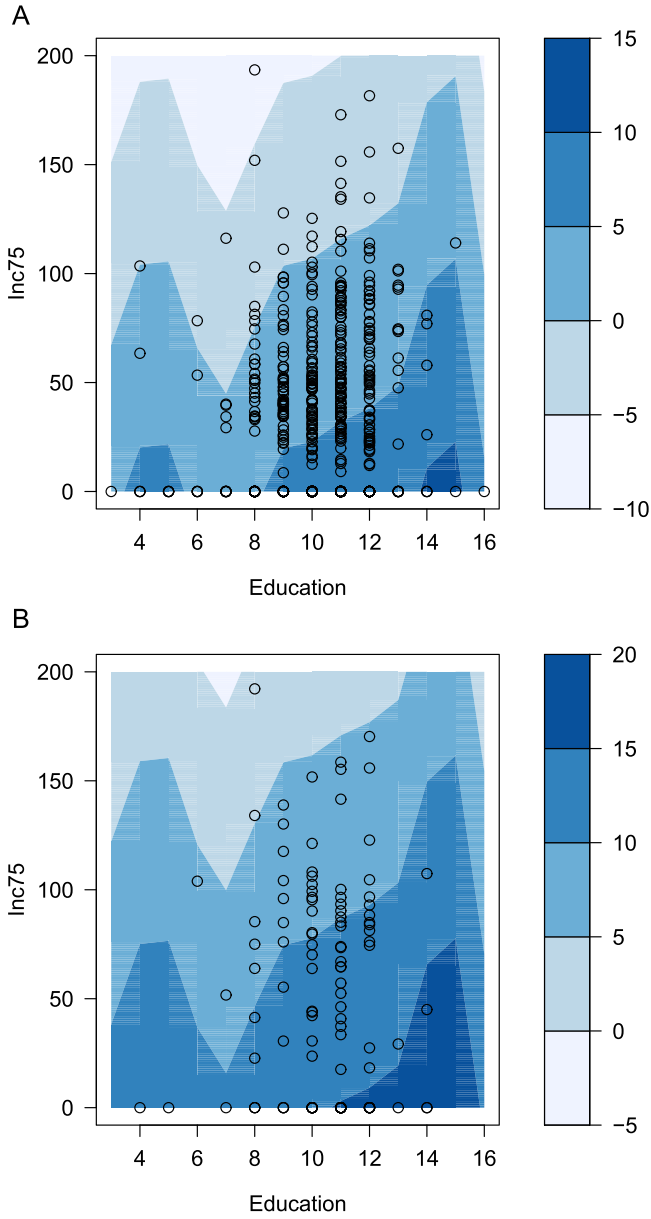


FIGURE 3. Contour plots for $\widehat{\Delta}_{\text{TEEM}}$ for the LaLonde NSW data. The circles indicate the locations of one or more original data points.

over the range of Inc75 and Education in the LaLonde data. Age is set to 25 (the sample mean), and subfigures (a) and (b) show results for single and married

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

males, respectively. The points on the contour plots represent values of Inc75 and Education realized in the LaLonde sample for single and married subgroups.

Figure 2 reveals substantial heterogeneity in $\widehat{\Delta}_{\mathrm{AIC}}$ with respect to both the Inc75 and Education variables. In particular, the heterogeneity and nonlinearity of the estimated treatment effect with respect to education in Figure 2 seem implausible. The plots in Figure 3 show a much more reasonable degree of heterogeneity (note the differing scales in Figures 2 and 3), with the program estimated to be more beneficial for those with little or no income, those with more education, and married participants.

5. SIMULATION STUDY

5.1. Setup

To gain further insight into the performance of model selection and combination methods on the LaLonde data analyzed in the previous section, we perform a guided simulation experiment that evaluates the methods under different simulation scenarios consistent with the data. Each simulation scenario is based on a model selected by one of the model selection methods. Li, Lue, and Chen (2000) called this type of simulation a "cross-examination".

Specifically, for each model selection method τ considered in our study, we generate a response vector Y_τ^* by adding i.i.d. noise to the estimated regression functions from the model selected by τ at the $n = 722$ sample values of $(T, \mathbf{U})$. The noise is generated by a mean-zero Gaussian distribution with variance equal to the error variance estimate from the model selected by τ. For each τ, all nine model selection and combination methods are then applied to $(Y_\tau^*, T, \mathbf{U})$ to produce an estimate of Δ. Since the true Δ_τ is known for each choice of τ, we can compare the performances of the estimators from the selection and combination methods under each version of the truth put forth by the five model selection methods, respectively. For each selection or combination method ν, we estimate the risk of each $\widehat{\Delta}_{\tau,\nu}$ under squared error loss by averaging $\left[\Delta_\tau(\mathbf{U}_i) - \widehat{\Delta}_{\tau,\nu}(\mathbf{U}_i)\right]^2$ over the $n = 722$ sample values of $\mathbf{U}_i$. To average out the variability in the random errors, the results are aggregated over 100 different realizations of the error vector.

The active variables involved in generating Y_τ^* under each of the five scenarios can be found by looking at the model selected by each method in Table 2. While Δ_{AIC} varies nonlinearly with baseline income and education level, the BIC scenario has $\Delta_{\mathrm{BIC}} = 0$. The scenario under CV has $\Delta_{\mathrm{CV}} = 6.6$. Under wFIC, Δ_{wFIC} varies linearly with baseline income and marital status. Finally, Δ_{TECV} differs by marital status but is otherwise constant; specifically, Δ_{TECV} is 20.5 for married males and 3.8 for unmarried males.

5.2. Results

Table 3 shows the estimated risk (average mean squared error) of $\widehat{\Delta}_{\tau,\nu}$ as an estimator of Δ_τ for each of the nine model selection and combination methods

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

TABLE 3. Guided simulation results: Estimated risk[a] (SE) of $\widehat{\Delta}_{\tau,\nu}$ under five scenarios consistent with the LaLonde data

	Model selection method determining $\mathbb{E}(Y\|T,\mathbf{U})$				
Method	AIC	BIC	CV	wFIC	TECV
AIC	157.3 (6.0)	41.3 (6.2)	54.9 (5.5)	86.9 (6.3)	74.4 (6.0)
BIC	191.6 (2.8)	**1.4** (1.4)	37.0 (2.1)	109.8 (1.6)	72.3 (1.5)
CV	169.0 (4.4)	12.6 (3.7)	31.1 (3.4)	81.0 (4.3)	59.6 (3.7)
wFIC	149.3 (4.0)	21.0 (4.5)	39.7 (3.9)	69.9 (4.9)	53.2 (4.2)
TECV	160.2 (3.9)	21.4 (4.2)	36.4 (3.6)	75.8 (4.2)	**48.7** (4.0)
cAIC	131.7 (5.2)	28.0 (4.9)	35.8 (4.8)	65.5 (5.1)	51.4 (5.0)
BMA	166.5 (2.1)	**1.3** (1.1)	**24.1** (1.7)	88.4 (2.2)	57.4 (1.8)
ARM	**111.5** (2.7)	11.6 (1.7)	**16.6** (1.8)	**47.1** (2.5)	**32.5** (2.1)
TEEM	**107.9** (2.9)	15.1 (1.9)	**20.4** (2.0)	**46.4** (2.7)	**32.2** (2.3)

[a] Numbers in bold represent the methods with the lowest estimated risks for each scenario and those not significantly different (using Tukey's HSD method of multiple comparisons) from the lowest-risk method.

under each of the five scenarios. For the scenario with a nonlinear Δ (the model selected by AIC), the model combination methods of ARM and TEEM significantly outperform all other methods, including AIC. AIC does not perform very well even on its own "home field" because even if the true additive model is selected, there is substantial variability involved in estimating the nonlinear Δ_{AIC} with a nonlinear estimate. When $\Delta = 0$ (the BIC scenario), BIC usually chooses a model with no treatment effect and thus has the lowest risk. In the more moderate scenarios based on the models chosen by wFIC, CV and TECV, the model combination methods of ARM and TEEM perform the best. Overall, the performances of ARM and TEEM are similar, with TEEM performing perhaps slightly better when there is more actual heterogeneity in the treatment effect.

A summary of each method's estimated risk across all five scenarios can be found in Figure 4. Each data point in the boxplot represents an average of $\left[\Delta_\tau(\mathbf{U}_i) - \widehat{\Delta}_{\tau,\nu}(\mathbf{U}_i)\right]^2$ over the $n = 722$ realized $\mathbf{U}_i$ values. There are 100 realizations generated from each of the five τ scenarios; therefore, 500 data points are summarized in each boxplot. Both Table 3 and Figure 4 show that model combination methods possessed lower risk than model selection methods on average, with the methods of ARM and TEEM performing the best overall.

Additional simulation studies are presented in the online supplement to this article. These include settings, commonly encountered in practice, where the treatment effect mean function (and in one setting, the error distribution) is misspecified by all of the candidate models. Since our method of TEEM does not assume that any of the candidate models are correct, it provides good performance in situations where all models are wrong, but some may be useful. The supplement

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

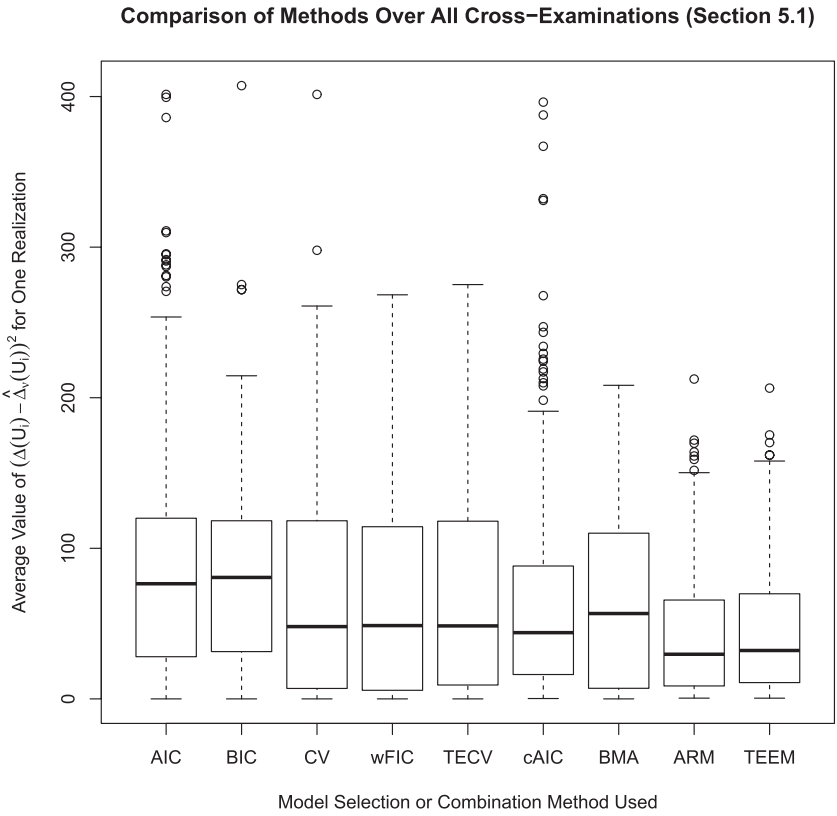


FIGURE 4. Results of the LaLonde NSW cross-examination (Section 5) over the five scenarios combined.

also presents a simulation where the TEEM algorithm is paired with dimension reduction techniques, as described in Section 3.3.

6. CONCLUSION

There is much theoretical and empirical evidence in the literature to support the practice of model combination rather than model selection when the goal is accurate estimation or prediction of a response. Our work indicates that model combination also is useful for the purpose of estimating a possibly heterogeneous treatment effect. The problem of treatment effect estimation differs from traditional regression in some ways and may result in different candidate models needing to receive higher weights; within a given candidate set, models that are the best for prediction may not be the best for treatment effect estimation. Therefore, there is a need for model combination methods targeted to the treatment effect. The TEEM algorithm proposed in this work is one such method.

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

TEEM has the flexibility to simultaneously combine traditional regression models, machine learning procedures, and any other current or future method proposed in the growing area of treatment effect estimation. We provide an oracle inequality for the TEEM estimator under squared error loss that guarantees the TEEM estimator will converge to the true Δ as long as at least one of the candidate estimators converges to Δ; the convergence will be at nearly the best rate offered by the candidates if the covariate vector is one-dimensional. An analysis of the benchmark LaLonde labor training data shows that TEEM provides a sensible data-driven weighting of linear and nonlinear treatment effect estimates, and simulation studies provide evidence that TEEM compares favorably with other selection and combination methods in providing accurate estimates of the treatment effect.

Combining treatment effect estimators is a rich topic with abundant practical applications and a number of future research directions. The method we propose in this article measures accuracy under the traditional squared error loss; however, in practice other loss functions and characteristics of the estimator may be of interest. An example with particular relevance to econometrics is optimal treatment targeting; see Bhattacharya and Dupas (2012), Hirano and Porter (2009), Kitagawa and Tetenov (2015), and Stoye (2009) for recent studies on optimally assigning treatments to units. The goal of accurately estimating the treatment effect function is related to, but distinct from, the goal of optimal treatment targeting. For model combination, different objective functions will lead to different formulas for combining the candidate estimators. This is an interesting area for future research.

We do not claim that our method of combining estimates of treatment effects assigns the optimal weights to the procedures. Calculation of theoretically optimal combining weights requires estimation of the covariances between the candidate estimates, a notoriously difficult task that can introduce substantial estimation error. Smith and Wallis (2009), Qian, Rolling, Cheng, and Yang (2017), and Claeskens, Magnus, Vasnev, and Wang (2016) discuss this issue for the problem of forecast combination. Yang (2004), also in a forecast combination context, terms the search for the optimal linear combination of procedures "combining for improvement" and distinguishes it from the more modest goal of "combining for adaptation", which targets the performance of the best candidate. Combining for improvement is shown to pay an unavoidable price in performance as a result of searching for the optimal weights.

The TEEM method falls into the category of combining for adaptation, because its aim is to achieve the best performance (for estimating the treatment effect) among the candidates. In this sense, the objective of TEEM is similar to the objective of a model selection method that targets the treatment effect; however, the TEEM estimator may outperform post-model selection estimators because of the latter's high variability. Indeed, our results indicate that when there is a large degree of model selection uncertainty, the TEEM method of model combination often results in an estimator of the treatment effect that is more accurate than an estimator from a single selected candidate. A properly targeted method of model

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

combination can offer large advantages over model selection in the important setting of treatment effect estimation.

REFERENCES

Abadie, A. & G.W. Imbens (2006) Large sample properties of matching estimators for average treatment effects. *Econometrica* 74, 235–267.

Abadie, A. & G.W. Imbens (2011) Bias-corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics* 29, 1–11.

Abrevaya, J., Y.-C. Hsu, & R.P. Lieli (2015) Estimating conditional average treatment effects. *Journal of Business & Economic Statistics* 33, 485–505.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19, 716–723.

Althauser, R.P. & D. Rubin (1970) The computerized construction of a matched sample. *American Journal of Sociology* 76, 325–346.

Bhattacharya, D. & P. Dupas (2012) Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics* 167, 168–196.

Buckland, S., K. Burnham, & N. Augustin (1997) Model selection: An integral part of inference. *Biometrics* 53, 603–618.

Cai, T., L. Tian, P.H. Wong, & L. Wei (2011) Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* 12, 270–282.

Chang, M., S. Lee, & Y.-J. Whang (2015) Nonparametric tests of conditional treatment effects with an application to single-sex schooling on academic achievements. *The Econometrics Journal* 18, 307–346.

Claeskens, G. & N.L. Hjort (2008) Minimizing average risk in regression models. *Econometric Theory* 24, 493–527.

Claeskens, G., J.R. Magnus, A.L. Vasnev, & W. Wang (2016) The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting* 32, 754–762.

Cook, R.D. (1998) *Regression Graphics*. Wiley.

Cook, R.D. & B. Li (2002) Dimension reduction for conditional mean in regression. *The Annals of Statistics* 30, 455–474.

Dehejia, R.H. & S. Wahba (1999) Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94, 1053–1062.

Green, D.P. & H.L. Kern (2012) Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly* 76, 491–511.

Hirano, K. & J.R. Porter (2009) Asymptotics for statistical treatment rules. *Econometrica* 77, 1683–1701.

Holland, P.W. (1986) Statistics and causal inference. *Journal of the American Statistical Association* 81, 945–960.

Hsu, Y.-C. (2017) Consistent tests for conditional treatment effects. *The Econometrics Journal* 20, 1–22.

Imai, K. & M. Ratkovic (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7, 443–470.

Imbens, G.W. (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86, 4–29.

Imbens, G.W. & J.M. Wooldridge (2009) Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47, 5–86.

Kitagawa, T. & A. Tetenov (2015) Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice. Cemmap Working paper, CWP10/15.

LaLonde, R.J. (1986) Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review* 76, 604–620.

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press

Li, K.-C. (1991) Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* 86, 316–327.
Li, K.-C., H.-H. Lue, & C.-H. Chen (2000) Interactive tree-structured regression via principal Hessian directions. *Journal of the American Statistical Association* 95, 547–560.
Li, L. (2007) Sparse sufficient dimension reduction. *Biometrika* 94, 603–613.
Li, L. & X. Yin (2008) Sliced inverse regression with regularizations. *Biometrics* 64, 124–131.
Qian, M. & S.A. Murphy (2011) Performance guarantees for individualized treatment rules. *The Annals of Statistics* 39, 1180–1210.
Qian, W., C.A. Rolling, G. Cheng, & Y. Yang (2017) On the forecast combination puzzle. Preprint. Available at https://arxiv.org/abs/1505.00475v1.
Raftery, A.E. (1995) Bayesian model selection in social research. *Sociological Methodology* 25, 111–163.
Rolling, C.A. & Y. Yang (2014) Model selection for estimating treatment effects. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*) 76, 749–769.
Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics* 6, 461–464.
Smith, J. & K. Wallis (2009) A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics* 71, 331–355.
Stoye, J. (2009) Minimax regret treatment choice with finite samples. *Journal of Econometrics* 151, 70–81.
Taddy, M., M. Gardner, L. Chen, & D. Draper (2016) A nonparametric Bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics* 34, 661–672.
Wood, S.N. (2006) *Generalized Additive Models*: *An Introduction with R*. Chapman and Hall/CRC.
Yang, Y. (2001) Adaptive regression by mixing. *Journal of the American Statistical Association* 96, 574–588.
Yang, Y. (2003) Regression with multiple candidate models: Selecting or mixing? *Statistica Sinica* 13, 783–809.
Yang, Y. (2004) Combining forecasting procedures: Some theoretical results. *Econometric Theory* 20, 176–222.
Zhang, B. (2016) Empirical likelihood in causal inference. *Econometric Reviews* 35, 201–231.

https://doi.org/10.1017/S0266466618000397 Published online by Cambridge University Press