2015

Estimating Controlled Direct Effects of Restrictive Feeding Practices in the `Early Dieting in Girls' Study

Yeying Zhu, *University of Waterloo*
Debashis Ghosh, *university of colorado denver*
Donna L Coffman, *Pennsylvania State University*
Jennifer S Williams, *Pennsylvania State University*

Estimating Controlled Direct Effects of Restrictive Feeding Practices: An Application to Early Dieting in Girls Study

Yeying Zhu

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada.

Debashis Ghosh

Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, USA.

Donna L. Coffman

The Methodology Center, Pennsylvania State University, University Park, USA.

Jennifer Savage Williams

The Center for Childhood Obesity Research, Pennsylvania State University, University Park, USA

Summary. In this article, we examine the causal effect of parental restrictive feeding practices on children's weight status. An important mediator we are interested in is children's self-regulation status. Traditional mediation analysis (Baron and Kenny, 1986) applies a structural equation modelling (SEM) approach and decomposes the intent-to-treat (ITT) effect into direct and indirect effects. More recent approaches interpret the mediation effects based on the potential outcomes framework. In practice, there often exist confounders that jointly influence the mediator and the outcome. Inverse probability weighting based on propensity scores are used to adjust for confounding and reduce the dimensionality of confounders simultaneously. We show that combining machine learning algorithms (such as a generalized boosted model) and logistic regression to estimate the propensity scores can be more accurate and efficient in estimating the controlled

direct effects than using logistic regression alone. The proposed methods are general in the sense that we can combine multiple candidate models and use the cross-validation criterion to select the optimal subset of the candidate models for combining. The criterion achieves a balance between the number of models we combine and the variability of the resulting estimator. A data application to the Early Dieting in Girls Study shows that the causal effect of mother's restrictive feeding differs according to whether the daughter eats in the absence of hunger.

Keywords: Causal inference, Generalized boosted model, Logistic regression, Mediator, Model combining, Random forests

1. Introduction

Mediation analysis is often used to understand how a treatment, intervention or prevention program may affect an outcome through an intermediate variable. For example, the JOBS II study investigates whether a particular five-day training session has a positive impact on the mental health of unemployed workers by improving their sense of mastery (Imai et al., 2010a). In this example, the sense of mastery mediates the hypothesized causal relationship between the treatment and the outcome and is therefore called a mediator.

Mediation analysis can also play an important role in studies with imperfect compliance. For instance, in many clinical trials, the treatment is randomly assigned, but the patient may not fully comply with his/her assigned treatment. Therefore, the intent-to-treat (ITT) estimate may be biased for the actual effect of the treatment. One solution is to treat patients' compliance status as a mediator and decompose the ITT effect into direct and indirect effects or look at the effect of the treatment among those who comply with their assigned treatment.

In the literature, the traditional approach for mediation analysis uses a linear structural equation modelling (LSEM) approach (Baron and Kenny, 1986; Judd and Kenny, 1981). Based on Baron and Kenny's four-step procedure, the total effect of the treatment can be decomposed into the direct and indirect ef-

fects, where the latter implies the amount of mediation. However, this method assumes there is no interaction between the treatment and the mediator. This no-interaction assumption is later relaxed by Jo (2008) and Imai et al. (2010b). Another criticism about LSEM is that the outcome variable is modelled by conditioning on the observed treatment and mediator. However, the mediator is assumed to be affected by the treatment and therefore is a post-treatment variable. Consequently, the direct and indirect effects can not necessarily be interpreted causally. Based on the framework of counterfactuals, more recent approaches interpret the mediation effect as natural effects, controlled effects and principal stratification effects, all of which can be interpreted causally because they are based on the difference among the potential outcomes within the same subject. Such approaches include Imai et al. (2010b), who propose nonparametric identification of natural direct and indirect effects; Angrist et al. (1996), who apply two-stage least squares to estimate principal stratification effects among compliers; Ten Have et al. (2007), who propose rank preserving models (RPM) for controlled effects, and Gallop et al. (2009), who focus on Bayesian approaches for principal stratification effects.

In this article, we focus on the inverse probability weighting (IPW) method to estimate controlled direct effects using marginal structural models (Robins, 1999; VanderWeele, 2009; Coffman and Zhong, 2012). Controlled direct effects are of particular interest to researchers who would like to estimate the causal effect of an exposure on an outcome. However, the exposure may change the status of another risk factor that mediates the relationship between the exposure and the outcome. Consequently, unbiased estimates of the controlled direct effects can only be obtained by controlling for the intermediate variable.

In the IPW method, an important issue is how to estimate the weights based on propensity scores and how to avoid extreme weights. Large weights may increase the variance of the estimated controlled effects. The traditional approach trims extremely large weights to a fixed value, say w_0 (Potter, 1990; Cole and Hernán, 2008). More recent approaches include Bayesian "weight pooling" (El-

liott, 2008) and normalization (Xiao et al., 2010). In this paper, we aim to deal with this problem from a model averaging perspective. We propose combining a parametric estimator with a nonparametric estimator of propensity scores for calculating the weights. This is an extension of Zhu et al. (2014) in the context of mediation analysis . However, the innovative part of this study lies in Section 4.2: when there are more than two candidate models, we employ a cross-validation criterion to select the best subset of the models for combining. The criterion eventually minimizes the mean squared error of the estimated controlled direct effects.

The layout of this paper is as follows. In Section 2, we describe the motivating example and the research question. In Section 3, we introduce controlled direct effects based on the potential outcomes framework and describe the IPW method. We point out that an important issue in the IPW approach is the estimation of the propensity scores and describe two different types of estimation methods: parametric and nonparametric algorithms. In Section 4, we propose a class of estimators that combines both parametric and nonparametric algorithms and propose a Monte-Carlo cross-validation criterion to determine the best subset of candidate models for combining. In Section 5, we apply the proposed methodology to the Early Dieting in Girls study and simulation studies are conducted to show the performance of the proposed methods. Some discussion concludes Section 6.

2. Early Diet in Girls Study

It has been shown that almost 17% of children and adolescents aged 2-19 are obese leading to both immediate and long-term effects on health and well-being (Ogden et al., 2012). The current obesogenic food environment is characterized by large amounts of inexpensive, readily available, and palatable foods. It promotes unhealthy diets, weight gain, and obesity (Hill et al., 2008). In response to this environment, parents who are concerned about their children's diet or weight may try to control what and when their child eats (i.e., restrictive feeding

practices) with the intent of preventing their child from eating too much and gaining excessive weight. However, restrictive feeding practices can have unintended consequences. For example, experimental studies reveal that restricting children's access to highly palatable foods can promote liking, requests, and the overconsumption of the restricted food when that food becomes readily available (Faith et al., 2012). Further, excessive control of children's eating may result in the development of poorer self-regulation of energy intake, which is associated with greater weight gain across childhood (Birch et al., 2003). Thus, it is plausible that susceptibility to eating in the absence of hunger (i.e., impaired self-regulation) may mediate the causal influence of restrictive feeding on subsequent child weight status. Therefore, to investigate the causal effect of restrictive feeding, it is necessary to control for the child's self-regulation condition. In this article, we will apply the proposed methodology to answer the above-mentioned research question using data collected from the Early Dieting in Girls study, which will be introduced in Section 5. Although studies suggest that restrictive feeding practices are associated with childrens self-regulation and weight status, many of these studies are correlational and none of them consider the children's self-regulation condition as an important mediator. Our study shows that the causal effect of mother's restrictive feeding on daughter's subsequent weight status significantly differs according to whether the child eats in the absence of hunger.

3. Mediation Analysis Based on Propensity Scores

3.1. Controlled Direct Effects Based on Marginal Structural Models

Let Y denote the response of interest, M a mediator, T an indicator of the treatment, and let $\mathbf{X}$ be a p-dimensional vector of baseline covariates. The observed data can be represented as $(Y_i, M_i, T_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, a random sample from $(Y, M, T, \mathbf{X})$. In addition to the observed quantities, we further define $Y_i(t, m)$ as the potential outcome if subject i is assigned to treatment t and the mediator is set to the value of m. Generally speaking, the treatment

and the mediator could be of any type: categorical, ordinal or continuous. In this paper, we focus on the case when both the treatment and the mediator are binary. As a result, there are four potential outcomes for each subject: $Y_i(1,1)$, $Y_i(1,0)$, $Y_i(0,1)$ and $Y_i(0,0)$.

Following Robins and Greenland (1992), the controlled direct effect (CDE) is defined as the causal effect of the treatment on the outcome, setting the mediator to a specific value, say m. It can be denoted as $CDE_m = E[Y(1,m) - Y(0,m)]$. When the mediator is binary, we have two controlled direct effects: $CDE_1 = E[Y(1,1) - Y(0,1)]$ and $CDE_0 = E[Y(1,0) - Y(0,0)]$. Mediation analysis should normally be applied to longitudinal studies as the mediator should occur after the treatment and the outcome variable should happen after both the treatment and the mediator. Consequently, IPW technique for time-varying treatments proposed by Robins (1999) can be applied to estimate the controlled effects (VanderWeele, 2009). The method works by fitting a marginal structural model for the potential outcomes:

$$E[Y(t,m)] = \alpha_0 + \alpha_1 t + \alpha_2 m + \alpha_3 tm. \tag{1}$$

Model (1) is marginal in that it is defined for the expected value of potential outcomes without conditioning on any covariates (which is different from regression models). In the case of binary mediator, $CDE_1 = \alpha_1 + \alpha_3$ and $CDE_0 = \alpha_1$. To consistently estimate the parameters in (1), the following weights are calculated for $i = 1, \ldots, n$:

$$w_i^T = \frac{1}{P(T = T_i | \mathbf{X} = \mathbf{X}_i)} \tag{2}$$

$$w_i^M = \frac{1}{P(M = M_i | T = T_i, \mathbf{X} = \mathbf{X}_i)}. \tag{3}$$

For simplicity, we assume $\mathbf{X}$ includes all covariates that confound the relationship among T, M and Y. In the case there are post-treatment confounders $\mathbf{W}$ that jointly affect M and Y, we only need to re-write the denominator of w_i^M as $P(M = M_i | T = T_i, \mathbf{X} = \mathbf{X}_i, \mathbf{W} = \mathbf{W}_i)$ (VanderWeele, 2009). Under the assumption that there are no unmeasured confounders among T, M and Y, the parameters in (1) can be estimated by a weighted regression of Y on T, M and

$T \times M$. The weight for subject i is calculated by $w_i^T \times w_i^M$. More often, the following stabilized weights are used:

$$w_i^T = \frac{P(T = T_i)}{P(T = T_i|\mathbf{X} = \mathbf{X}_i)}$$

$$w_i^M = \frac{P(M = M_i|T = T_i)}{P(M = M_i|T = T_i, \mathbf{X} = \mathbf{X}_i)}.$$

Robins (1999) proves that using stabilized weights, IPW estimators of the parameters in (1) are consistent and efficient given that the denominators in w_i^T and w_i^M are correctly specified. In practice, the true models for the denominators in w_i^T and w_i^M are unknown and estimated from the observed data. However, estimates close to zero will result in enormous weights given that the numerators are bounded away from zero. Consequently, the variance of the estimated coefficients could be very large. In addition, estimates close to zero may indicate misspecification in the fitted models.

In the rest of this paper, we focus on the estimation of the weights in (2) and (3) for controlled direct effects, especially the denominators of the weights: $P(T = T_i|\mathbf{X} = \mathbf{X}_i)$ and $P(M = M_i|T = T_i, \mathbf{X} = \mathbf{X}_i)$.

3.2. Estimation of the Inverse Probability Weights

In practice, researchers usually collect a large number of baseline covariates or potential confounders, which means the dimension of $\mathbf{X}$ is large. Denote $\pi(\mathbf{X}_i) = P(T = 1|\mathbf{X} = \mathbf{X}_i)$ and $e(\mathbf{X}_i, T_i) = P(M = 1|T = T_i, \mathbf{X} = \mathbf{X}_i)$. These can be regarded as two different versions of propensity scores. In the literature, the most popular way to estimate $\pi(\mathbf{X}_i)$ and $e(\mathbf{X}_i, T_i)$ is logistic regression. However, when $\mathbf{X}$ is high-dimensional and some of the covariates are correlated with each other, it is challenging to specify a logistic regression model. Kang and Schafer (2007) show through simulation studies that when there is a misspecification in the logistic regression model, the IPW estimates are highly biased and inconsistent in the sense that the variance of the estimates gets larger as the sample size gets larger. This is due to the fact that IPW methods are

sensitive to extreme weights and logistic regression fails to estimate correctly when the denominators in (2) or (3) are close to zero.

An alternative class of methods for estimating $\pi(\mathbf{X}_i)$ and $e(\mathbf{X}_i, T_i)$ is nonparametric machine learning algorithms. These methods include classification and regression trees (CART) and its various extensions, such as pruned CART, bagged CART, Bayesian CART, random forests and boosting algorithms. Other nonparametric algorithms include support vector machines and K-nearest neighbors. See Hastie et al. (2009) for a detailed discussion of each algorithm. Compared to logistic regression, nonparametric machine learning algorithms are more flexible in the form of the model. In particular, the tree classifiers (CART and its various extensions) can deal with a large number of covariates, even when some of the covariates are highly correlated with each other. The algorithms can automatically select important covariates, interaction terms, and nonlinear terms by splitting nodes based on different predictors. Among the various tree classifiers, ensemble methods, such as random forests and boosting, have repeatedly shown superior performance for many classification problems in the literature (e.g., Svetnik et al. (2003)). They are resistant to outliers and avoid over-fitting. Next, we will briefly introduce random forests and a special case of boosting: the generalized boosted model.

Random forests (Breiman, 2001) is an algorithm in which multiple tree models are fit to the data and then predictions are averaged over the multiple models. Each tree is built on a bootstrap sample of the original data. At each node of a tree, instead of determining the best split among all available covariates, the best split is calculated on a random subset of the covariates. The size of the subset is usually taken to be $\sqrt{p}$ for classification problems and $p/3$ for regression problems, where p is the total number of covariates.

Generalized boosted models (GBMs, McCaffrey et al. (2004)) are one type of boosting algorithm that can directly produce estimates of the probabilities. The algorithm aims to maximize the expected Bernoulli log-likelihood by fitting additive models. Each model is a regression tree fitting the residuals from

the previous step. The algorithm stops when the number of trees reaches the minimum of average standardized absolute mean difference (ASAM) of the covariates. The calculation of ASAM will be introduced in Section 5.3.

In the following section, we propose a model averaging technique to estimate the inverse probability weights in (2) and (3). The proposed methods combine parametric and nonparametric estimators of propensity scores. We consider random forests or GBMs as the nonparametric component because of the superior performance and wide application of these two algorithms. By averaging over several plausible models, we can reduce the variance of the estimated causal effects because extreme weights are shrunk to more reasonable values without *ad-hoc* adjustments.

4. The Proposed Methods

4.1. Combining Two Models

First, we suggest combining a parametric model with a nonparametric model while estimating the inverse probability weights. Let $\pi_1(\mathbf{X}_i)$ be the estimate of $\pi(\mathbf{X}_i) = P(T = 1|\mathbf{X} = \mathbf{X}_i)$ from a logistic regression model and $\pi_2(\mathbf{X}_i)$ be the estimate from a nonparametric model, such as a random forests model or a GBM. We propose the following estimator:

$$\hat{\pi}_i = \frac{a_i}{a_i + b_i}\pi_1(\mathbf{X}_i) + \frac{b_i}{a_i + b_i}\pi_2(\mathbf{X}_i), \tag{4}$$

where

$$a_i = \pi_1(\mathbf{X}_i)^{T_i}[1 - \pi_1(\mathbf{X}_i)]^{1-T_i}$$

and

$$b_i = \pi_2(\mathbf{X}_i)^{T_i}[1 - \pi_2(\mathbf{X}_i)]^{1-T_i}.$$

Then, w_i^T in (2) is estimated by $1/\hat{\pi}_i$ if $T_i = 1$ and $1/(1 - \hat{\pi}_i)$ if $T_i = 0$. We can further replace the numerator of w_i^T by $\hat{P}(T = T_i)$ to stabilize the weights. Strictly speaking, $\hat{\pi}_i$ is not the estimated propensity score but the weighted average of two models, because its value depends on both T_i and $\mathbf{X}_i$.

Denote $e_1(\mathbf{X}_i, T_i)$ as the estimate of $P(M = 1|T = T_i, \mathbf{X} = \mathbf{X}_i)$ from a logistic regression model and $e_2(\mathbf{X}_i, T_i)$ as the estimate of the propensity score from a nonparametric algorithm. Then, we let

$$\hat{e}_i = \frac{c_i}{c_i + d_i} e_1(\mathbf{X}_i, T_i) + \frac{d_i}{c_i + d_i} e_2(\mathbf{X}_i, T_i), \tag{5}$$

where

$$c_i = e_1(\mathbf{X}_i, T_i)^{M_i}[1 - e_1(\mathbf{X}_i, T_i)]^{1-M_i} \tag{6}$$

and

$$d_i = e_2(\mathbf{X}_i, T_i)^{M_i}[1 - e_2(\mathbf{X}_i, T_i)]^{1-M_i}. \tag{7}$$

Then, w_i^M in (3) is estimated by $1/\hat{e}_i$ if $M_i = 1$ and $1/(1 - \hat{e}_i)$ if $M_i = 0$. We can further replace the numerator of w_i^M by $\hat{P}(M = M_i|T = T_i)$ to stabilize the weights.

As can be seen, the proposed estimator is a weighted average of the two models. The weight placed on each model is proportional to the Bernoulli likelihood. In Appendix A of the supplementary material, we provide a justification for the choice of the mixing weights. In summary, weights are chosen in (6) and (7) by the following principle: if $M_i = 1$, more weight is placed on the higher value of $e_1(\mathbf{X}_i, T_i)$ and $e_2(\mathbf{X}_i, T_i)$; if $M_i = 0$, more weight is given to the lower value of $e_1(\mathbf{X}_i, T_i)$ and $e_2(\mathbf{X}_i, T_i)$. In extreme cases when $M_i = 1$ and $e_1(\mathbf{X}_i, T_i) \approx 0$, the inverse probability weight will be very large if we rely on logistic regression alone to estimate propensity scores. However, in our proposed estimator, given that $e_2(\mathbf{X}_i, T_i)$ is bounded away from zero, $\hat{e}_i \approx e_2(\mathbf{X}_i, T_i)$. Hence, extremely large weights while calculating w_i^M can be avoided in a systematic way. The same principle applies to the calculation of w_i^T. Since the final weights used to estimate the controlled direct effects are calculated as $w_i^T \times w_i^M$, our proposed method simultaneously prevents extreme values of w_i^M and w_i^T, and therefore improves the estimation of controlled direct effects by stabilizing the weights.

4.2. Combining Multiple Models

Similarly, we can combine more than two candidate models/algorithms for the inverse probability weights using the same kind of principles as in the previous section. Assume we want to combine K different models. We let

$$\hat{\pi}_i = \sum_{j=1}^{K} w_{i,j}\pi_j(\mathbf{X}_i),$$

where

$$w_{i,j} = \frac{\pi_j(\mathbf{X}_i)^{T_i}[1-\pi_j(\mathbf{X}_i)]^{1-T_i}}{\sum_{k=1}^{K}\pi_k(\mathbf{X}_i)^{T_i}[1-\pi_k(\mathbf{X}_i)]^{1-T_i}}, \tag{8}$$

and $\pi_k(\mathbf{X}_i)$ is the estimator of $\pi(\mathbf{X}_i)$ from the k-th model. Similarly,

$$\hat{e}_i = \sum_{j=1}^{K} \tilde{w}_{i,j}e_j(\mathbf{X}_i, T_i),$$

where

$$\tilde{w}_{i,j} = \frac{e_j(\mathbf{X}_i, T_i)^{M_i}[1-e_j(\mathbf{X}_i, T_i)]^{1-M_i}}{\sum_{k=1}^{K}e_k(\mathbf{X}_i, T_i)^{M_i}[1-e_k(\mathbf{X}_i, T_i)]^{1-M_i}}, \tag{9}$$

and $e_k(\mathbf{X}_i, T_i)$ is the estimator of $e(\mathbf{X}_i, T_i)$ from the k-th model. It is natural to think that if logistic regression is insufficient to fit the propensity score model, adding a nonparametric fit will improve the accuracy. However, when more and more models are combined, the variability of the estimated controlled direct effect will also increase. Therefore, we need to have a balance between the number of models we want to combine and the variability of the resulting estimator.

4.3. Selecting the Optimal Subset of Candidate Models for Combining

If the total number of candidate models K is not too large, we can make an exhaustive search for the optimal one among $2^K - 1$ possible combinations. In other words, we can rewrite the weights in (8) and (9) in the following way. Denote $\lambda_j \equiv \mathrm{I}\{\text{the } j\text{-th candidate model is selected for combining}\}$ where $I()$ is the indicator function. Then w_{ij} in (8) and $\tilde{w}_{ij}$ in (9) are replaced by

$$w_{i,j} = \frac{\lambda_j\pi_j(\mathbf{X}_i)^{T_i}[1-\pi_j(\mathbf{X}_i)]^{1-T_i}}{\sum_{k=1}^{K}\lambda_k\pi_k(\mathbf{X}_i)^{T_i}[1-\pi_k(\mathbf{X}_i)]^{1-T_i}},$$

and

$$\tilde{w}_{i,j} = \frac{\lambda_j e_j(\mathbf{X}_i, T_i)^{M_i}[1 - e_j(\mathbf{X}_i, T_i)]^{1-M_i}}{\sum_{k=1}^{K} \lambda_k e_k(\mathbf{X}_i, T_i)^{M_i}[1 - e_k(\mathbf{X}_i, T_i)]^{1-M_i}}.$$

How should we determine the value of λ_j (0 or 1) for $j = 1, \ldots, K$? The criterion we will rely on should take into account both model accuracy and variability. Generally speaking, the estimation of controlled direct effects can be regarded as a two-stage procedure. In the first stage, the weights $w^T \times w^M$ are estimated based on the observed data. A marginal structural model shown in (1) is built at the second stage and a weighted regression is fitted using the weights calculated in the first stage. The controlled direct effects are simply a function of the estimated coefficients. The two-stage procedure fits the model structure discussed by Brookhart and van der Laan (2006). Using their notation, if we denote a two-dimensional vector $\psi \equiv (CDE_0, CDE_1)^T$ and the propensity scores as $\eta \equiv (\pi(\mathbf{X}), e(\mathbf{X}, T))^T$, ψ hence represents the parameters of interest and η the nuisance parameters. Note that we can also denote the parameters of interest as $\psi \equiv (\alpha_1, \alpha_3)^T$ where α_1 and α_3 are the coefficients in the marginal structural model in (1). The issue of choosing an optimal combination of candidate models can be restated as follows: assuming we have $2^K - 1$ different approaches to estimating η in stage one, which approach is optimal for stage two? Denote the resulting estimates of ψ as $\hat{\psi}_1(\mathbf{X})$,...,$\hat{\psi}_{2^K-1}(\mathbf{X})$.To account for the fact that there is a trade-off between bias and variance while estimating ψ, we aim to find a combination of models that minimizes $E\{(\hat{\psi}_k(\mathbf{X}) - \psi)^T B(\hat{\psi}_k(\mathbf{X}) - \psi)\}$, where B is a matrix that shows the relative importance of the components of ψ. Since the quadratic loss function is unknown, Brookhart and van der Laan (2006) propose a Monte-Carlo cross-validation criterion for selecting the optimal estimator of the candidate models. First, let $\hat{\psi}_0(\mathbf{X})$ be an approximately unbiased but highly variable estimate for ψ. The model used to estimate η in $\hat{\psi}_0(\mathbf{X})$ is regarded as the reference model. In practice, the reference model could be an over-fitted logistic regression model. In the v-th iteration of the Monte-Carlo cross-validation, the data set is divided into two parts: the training data

(X_v^0) and the testing data (X_v^1). The criterion function is defined as follows:

$$C_v(k) = \frac{1}{V}\sum_{v=1}^{V}\{\hat{\psi}_k(X_v^0) - \hat{\psi}_0(X_v^1)\}^T B\{\hat{\psi}_k(X_v^0) - \hat{\psi}_0(X_v^1)\}, \tag{10}$$

where $B = [\widehat{Var}\{\hat{\psi}_k(X_v^0)\}]^{-1}$ is the estimated variance-covariance matrix of $\hat{\psi}_k(X_v^0)$ and V is the total number of iterations.

The best subset of models for combining is then chosen to be the one that leads to the smallest C_v among the $2^K - 1$ possible combinations. If $\hat{\psi}_0$ is an unbiased estimator of ψ, the optimal combination selected by the Monte Carlo cross-validation criteria leads to the smallest mean square error of the parameter(s) of interest.

When K is large, it is computationally expensive to compare all possible combinations of K models. Instead, we can employ a stepwise selection method based on Monte-Carlo cross-validation. The detailed algorithm can be found in Section 3.1 of Brookhart and van der Laan (2006).

5. Data Application

5.1. Application to the Early Dieting in Girls Study

The Early Dieting in Girls study is a longitudinal study that aims to examine parental influences on daughters' growth and development from ages 5 to 15. The study involves 197 daughters and their mothers, who are from non-Hispanic, White families living in central Pennsylvania. The participants were assessed at daughters' age five (Wave 1), seven (Wave 2), nine (Wave 3), eleven (Wave 4), thirteen (Wave 5) and fifteen (Wave 6). At each wave, both daughters and their mothers were interviewed during a scheduled visit to the laboratory. The motivating question in this mediation analysis is whether mother's restriction in the daughter's dietary intake has a significant causal effect on subsequent daughter's weight status. The mediator we are interested in is whether the daughter eats in the absence of hunger (i.e., daughter's self-regulation).

Based on the research question, the outcome variable selected is the daughter's BMI at Wave 4 and the treatment variable is whether the mother restricts

the daughter's dietary intake, which is measured at Wave 3. The mediator variable is the total calories consumed by the daughter in the Eating in the Absence experiment at Wave 3 (See Figure 1). Before the experiment, a subjective measure of hunger was obtained after eating a standard ad-libitum lunch and compulsory preload. None of the participants reported being hungry. In the experiment, the girls were offered toys and generous portions of 10 sweet and savory snacks for 10 minutes. Then, the total number of calories each girl consumed were recorded. Since all girls took at least one taste of each food at the beginning of this procedure to assess preference, we dichotomize the mediator at its median value (202.88 calories) to measure the daughter's self-regulation. $M = 1$ means the daughter eats in the absence of hunger while $M = 0$ means the daughter does not eat in the absence of hunger. Although both the treatment and the mediator are measured at Wave 3, it is guaranteed that the treatment takes place before the mediator because the treatment is a summary of the mother's restriction behaviour prior to the lab visit. Because neither the treatment nor the mediator are randomized, we need to adjust for confounders in order to get unbiased estimates of the desired causal treatment effects. The potential confounders in this study include: family history of diabetes, income; daughter's weight concerns, body esteem, dieting status, satisfaction with body, restrained eating, emotional and external disinhibited eating behavior, peer influence on eating, daughter's perception of own current shape and negative emotions about eating snack foods; mother's BMI, restrained, disinhibited and hunger eating behavior, maternal perception of being overweight, perceived girl overweight subscale score, restriction subscale score, division of feeding responsibility subscale score, and weight concerns subscale score.

We only include those observations with complete data for T, M and Y, which results in a total of 159 observations in the analysis.

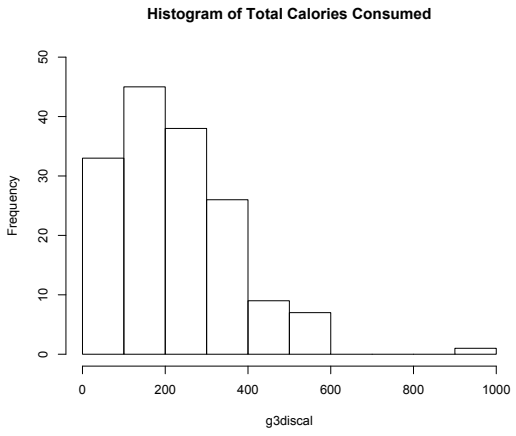

Fig. 1. Histogram of the continuous mediator (g3discal): total calories consumed.

5.2. Methodology

A crude regression analysis shows that the daughter's self-regulation is significantly related to the treatment (p-value= 0.033) and the outcome (p-value= 0.009). Therefore, the daughter's self-regulation is intermediate on the causal pathway from the treatment to the outcome and thus should be controlled for. In other words, we are interested in estimating the controlled direct effect of the treatment on the outcome when the mediator is set to a fixed level. To draw a causal inference, we rely on the potential outcomes framework to perform the mediation analysis. We fit the marginal structural model defined in (1). Based on the model, $CDE_0 = \alpha_1$ and $CDE_1 = \alpha_1 + \alpha_3$. We employ IPW to estimate α_1 and α_3. In particular, the weights are calculated as $w^T \times w^M$, where

$$w^T = \frac{P(T = T_i)}{P(T = T_i|\mathbf{X}_2 = \mathbf{X}_{2i})}, \tag{11}$$

$$w^M = \frac{P(M = M_i|T = T_i)}{P(M = M_i|T = T_i, \mathbf{X}_2 = \mathbf{X}_{2i}, \mathbf{X}_3 = \mathbf{X}_{3i})}. \tag{12}$$

$\mathbf{X}_2$ denotes the potential confounders measured at Wave 2 and $\mathbf{X}_3$ denotes the potential confounders measured at Wave 3. We apply both parametric and nonparametric machine learning algorithms to estimate the denominators in w^T and w^M: logistic regression, random forests and GBM. When applying logistic

regression to estimate $P(T = T_i|\mathbf{X}_2 = \mathbf{X}_{2i})$, we only include the covariates that are marginally related to the treatment at a significance level of 0.25. That is, we first fit a logistic regression model of the treatment on each covariate one at a time and select the covariates whose p-values are smaller than 0.25. The selected covariates are then treated as predictors for estimating $P(T = T_i|\mathbf{X}_2 = \mathbf{X}_{2i})$. This approach is similar to that of Hirano and Imbens (2001). When estimating $P(M = M_i|T = T_i, \mathbf{X}_2 = \mathbf{X}_{2i}, \mathbf{X}_3 = \mathbf{X}_{3i})$, the logistic regression model is fit as a function of the treatment, as well as the covariates that are marginally related to the outcome while the treatment variable is in the model. When fitting the random forests and GBM models, the algorithms will automatically select important variables and interaction terms without specifying a parametric model.

To estimate the denominators in w^T and w^M, we also apply the proposed methods: LR+RF, LR+GBM, RF+GBM and LR+RF+GBM. The performance metrics of the estimated controlled direct effects are shown in Table 1. The standard errors of the estimates are obtained by the robust sandwich formula using *survey* package in R.

5.3. Results

Table 1 shows that, among the propensity score adjusted models, the parametric algorithm (LR) leads to the larger standard error compared to the other two nonparametric algorithms (RF and GBM). In the literature, it has been discussed that in inverse weighted estimation, the standard error of the estimator is greatly influenced by large weights (Kang and Schafer, 2007). Figure 2 displays the inverse probability weights produced by each method. As shown, logistic regression yields the largest range of weights and many weights are larger than 2 (since the stabilized weights are used, the average of "well-behaved" weights should be around one (Cole and Hernán, 2008)). Compared to logistic regression, combining parametric and nonparametric algorithms can generally reduce the standard error. Take CDE_0 for example. Table 1 shows combining logistic

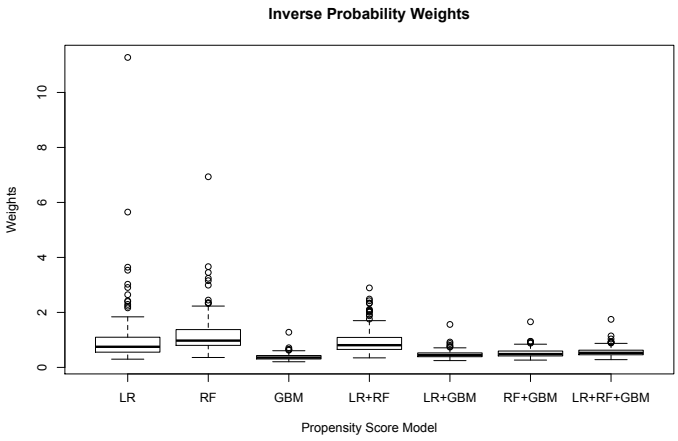

Fig. 2. Boxplots of inverse probability weights by different propensity score models.

regression with random forests can reduce the standard error by 20.1% while combining logistic regression and GBM can reduce the standard error by 19.1%.

Next, we employ the Monte-Carlo cross-validation criterion defined in (10) to determine the optimal subset of candidate models for combining. The cross-validation values with $V = 100$ are displayed in the last column of Table 1. The reference model for the nuisance parameters is a logistic regression model with all available covariates as main effects. The C_v values are the same for CDE_0 and CDE_1 because we consider the parameters of interest as a vector (i.e., $(\alpha_1, \alpha_3)^T$). It can be seen that combining the parametric model with the nonparametric model (LR+RF and LR+GBM) produces the smallest C_v values among all the methods we tested.

The main idea of IPW is to adjust for confounding and essentially achieve covariate balance among groups so that the subjects in different groups can be treated as if they were randomly assigned. Therefore, we calculate the ASAM values between $T = 1$ and $T = 0$ (See ASAM(T) in Table 2). A smaller ASAM means the covariates are more balanced among groups. It is calculated in the following way: for the j-th covariate, calculate the absolute value of the difference (d_j) in the weighted means between $T = 1$ and $T = 0$; then, divide d_j

Table 1. Estimated controlled effects

	CDE_0	CDE_1	
Method	Estimate (SE)	Estimate (SE)	C_v
LR	0.002 (1.034)	2.555 (0.842)	8.717
RF	-0.121 (1.023)	2.046 (0.836)	7.776
GBM	0.390 (0.901)	2.383 (0.778)	7.344
LR+RF	0.176 (0.826)	2.498 (0.762)	7.165
LR+GBM	0.275 (0.837)	2.441 (0.752)	7.280
RF+GBM	0.323 (0.927)	2.306 (0.799)	7.334
LR+RF+GBM	0.305 (0.851)	2.422 (0.768)	7.288

by the standard deviation of the covariate for the group of $T = 1$ and denote it as $d_j^{'}$; finally, average $d_j^{'}$ over all the covariates (McCaffrey et al., 2004). Here, the weights we use should be w^T in (11). Meanwhile, the covariates should be balanced in the mediator groups after adjusting by the weights w^M. We also calculate the ASAM values between $M = 1$ and $M = 0$ (See ASAM(M) in Table 2). As shown, the ASAM values without weighting are 0.281 for T and 0.208 for M, which means there is a slight imbalance in the distribution of the covariates. However, after applying the weights, differences between groups decrease significantly. It is shown that LR+RF achieves the smallest ASAMs among all the methods we tested. In other words, LR+RF performs best in terms of removing selection bias. This is consistent with our conclusion based on the Monte-Carlo cross-validation criterion.

We further look at each covariate individually. The circle in Figure 3 represents the standardized mean difference in each covariate without weighting and the triangle is the standardized weighted mean difference after adjusting for the weights, which is $d_j^{'}$ in the above definition. In some literature, $d_j^{'}$ is called the absolute effect size. The weights are calculated by the proposed method: LR+RF. As displayed in Figure 3, the absolute effect sizes are generally reduced by the proposed method.

Table 2. Average standardized absolute mean distance (ASAM)

Method	ASAM (T)	ASAM (M)
Without Weighting	0.281	0.208
LR	0.149	0.144
RF	0.150	0.149
GBM	0.176	0.174
LR+RF	0.140	0.135
LR+GBM	0.157	0.162
RF+GBM	0.167	0.163
LR+RF+GBM	0.158	0.156

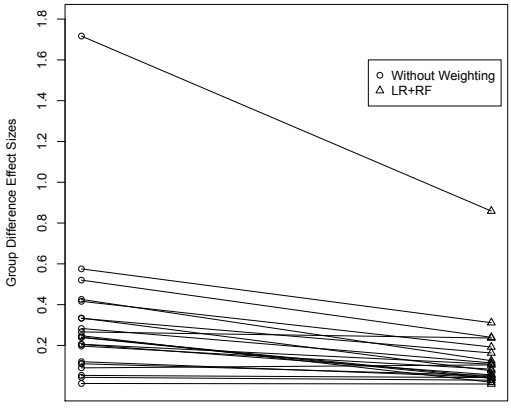

Fig. 3. Standardized mean difference (effect size) in each covariate between $T = 1$ and $T = 0$.

5.4. Conclusion

Since the combination of LR and RF yields the best finite performance, we draw conclusion based on this approach. The 95% confidence interval for CDE_0 is (-1.443, 1.795) and the 95% confidence interval for CDE_1 is (1.003, 3.992). In other words, CDE_0 is not significantly different from zero while CDE_1 is significantly larger than zero. It means for those girls who eat in the absence of hunger, mother's restriction behaviour during the daughter's childhood will lead to an increase in the daughter's BMI at age 11. On the other hand, for those who do not eat in the absence of hunger, mother's restriction behaviour does not have a significant effect on the daughter's BMI at age 11. This implies that mother should not restrict daughters' food intake during childhood.

5.5. Simulation Studies

In the Early Dieting in Girls study, the true value of the controlled effects are unknown. To further test the performance of our proposed methods, we conducted simulation studies. The simulation setup is based on the analysis in Section 5.3. Appendix B in the supplementary material presents the details of the simulation setup and tables for the results.

With the existence of post-treatment covariates in the simulation setup, the traditional multiple regression method fails to work properly. Therefore, we focus on the IPW method to estimate the controlled direct effects. Compared to logistic regression, combining parametric and nonparametric algorithms to estimate propensity scores can generally reduce the bias, variance and MSE of the estimated controlled direct effects. Meanwhile, the proposed methods produce confidence interval coverages closer to 95%.

6. Discussion

In this paper, we focused on the estimation of controlled direct effects in mediation analysis. IPW is often used to adjust for measured confounders. However, IPW methods are sensitive to extreme weights and a logistic regression model

with main effects is sometimes insufficient to estimate propensity scores, especially when the number of confounders is large. We proposed a model averaging strategy that combines parametric and nonparametric estimators of propensity scores for calculating the inverse probability weights. Unlike the existing model combining/averaging techniques, the weight placed on each model is data-adaptive and varies at every data point. By averaging over several plausible models, extreme weights are shrunk to more reasonable values without *ad-hoc* adjustments. Consequently, the bias and variance of the estimated effects can be reduced. The proposed method is general in the sense that we can combine more than two candidate models and use the cross-validation criterion to select the optimal subset of the candidate models for combining. On the other hand, based on our simulation studies and data analysis results, we find that in most cases, combining logistic regression with random forests or combining logistic regression with boosting is sufficient to achieve good results.

In the data analysis example, we applied the proposed methodology to a longitudinal study: the Early Dieting in Girls study. Among all the methods we tested, the combined estimator of logistic regression and random forests (LR+RF) gives the smallest standard errors of the estimated controlled effects and the smallest cross-validation value. Meanwhile, it also achieves the best balance in the covariates after they are adjusted by the weights.

Although this article focuses on mediation analysis, the proposed methodology can be applied to the general case with time-varying treatments and time-varying confounders. It provides a new perspective on how to deal with extreme weights in the IPW procedure in a systematic way. Besides binary mediators, continuous mediators are very common in practice. Robins et al. (2000) discussed IPW for estimating causal effects of continuous time-varying treatments. In this case, the propensity score is defined as the conditional density of observing the level of the treatment given covariates. When the vector of covariates is high-dimensional, the traditional approach for conditional density estimation suffers from the curse of dimensionality. An alternative is to employ machine

learning algorithms to estimate propensity scores, such as boosting algorithms. It is also desirable to extend the proposed approaches in this paper to continuous mediators.

Acknowledgments

This work was supported by awards P50DA010075-16 from the National Institute on Drug Abuse (NIDA), 5R21DK082858-02 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), and HD32973-09 from the National Institute of Child Health and Human Development (NICHD). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIDA, NIDDK or NICHD or the National Institutes of Health. We would also like to thank Leann Birch for permission to use the Early Dieting in Girls Study data, which was funded by National Institute of Child Health and Human Development R01 HD32973. The content is solely the responsibility of the authors and does not necessarily represent the official views of NICHD.

References

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*(434), 444–455.

Baron, R. M. and D. A. Kenny (1986). The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology 51*(6), 1173–1182.

Birch, L. L., J. O. Fisher, and K. K. Davison (2003). Learning to overeat: maternal use of restrictive feeding practices promotes girls eating in the absence of hunger. *The American Journal of Clinical Nutrition 78*(2), 215–220.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Brookhart, M. A. and M. J. van der Laan (2006). A semiparametric model selection criterion with applications to the marginal structural model. *Computational Statistics and Data Analysis 50*(2), 475–498.

Coffman, D. L. and W. Zhong (2012). Assessing mediation using marginal structural models in the presence of confounding and moderation. *Psychological Methods 17*(4), 642–664.

Cole, S. R. and M. A. Hernán (2008). Constructing inverse probability weights for marginal structural models. *American journal of epidemiology 168*(6), 656–664.

Elliott, M. R. (2008). Model averaging methods for weight trimming. *Journal of official statistics 24*(4), 517.

Faith, M. S., K. S. Scanlon, L. L. Birch, L. A. Francis, and B. Sherry (2012). Parent-child feeding strategies and their relationships to child eating and weight status. *Obesity Research 12*(11), 1711–1722.

Gallop, R., D. S. Small, J. Y. Lin, M. R. Elliott, M. Joffe, and T. R. Ten Have (2009). Mediation analysis with principal stratification. *Statistics in Medicine 28*(7), 1108–1130.

Hastie, T., R. Tibshirani, and J. H. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Verlag.

Hill, J. O., J. C. Peters, V. A. Catenacci, and H. R. Wyatt (2008). International strategies to address obesity. *Obesity Reviews 9*(s1), 41–47.

Hirano, K. and G. W. Imbens (2001). Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services and Outcomes Research Methodology 2*(3), 259–278.

Imai, K., L. Keele, and D. Tingley (2010a). A general approach to causal mediation analysis. *Psychological Methods 15*(4), 309–334.

Imai, K., L. Keele, and T. Yamamoto (2010b). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science 25*(1), 51–71.

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods 13*(4), 314–336.

Judd, C. M. and D. A. Kenny (1981). Process analysis. *Evaluation Review 5*(5), 602–619.

Kang, J. D. Y. and J. L. Schafer (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods 9*(4), 403–425.

Ogden, C. L., M. D. Carroll, B. K. Kit, and K. M. Flegal (2012). *Prevalence of obesity in the United States, 2009-2010*. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Potter, F. J. (1990). A study of procedures to identify and trim extreme sampling weights. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 225–230.

Robins, J. M. (1999). Association, causation, and marginal structural models. *Synthese 121*(1), 151–179.

Robins, J. M. and S. Greenland (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology 3*(2), 143–155.

Robins, J. M., M. Hernán, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology 11*(5), 550–560.

Svetnik, V., A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston (2003). Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences 43*(6), 1947–1958.

Ten Have, T. R., M. M. Joffe, K. G. Lynch, G. K. Brown, S. A. Maisto, and A. T. Beck (2007). Causal mediation analyses with rank preserving models. *Biometrics 63*(3), 926–934.

VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology 20*(1), 18–26.

Xiao, Y., M. Abrahamowicz, and E. E. Moodie (2010). Accuracy of conventional and marginal structural cox model estimators: A simulation study. *The International Journal of Biostatistics 6*(2).

Zhu, Y., D. Ghosh, N. Mitra, and B. Mukherjee (2014). A data-adaptive strategy for inverse weighted estimation of causal effect. *Health Services and Outcomes Research Methodology 14*(3), 69–91.