University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2006 *Paper* 212

Extending Marginal Structural Models through Local, Penalized, and Additive Learning

Daniel Rubin* Mark J. van der Laan†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, daniel.rubin@fda.hhs.gov

†Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

Extending Marginal Structural Models through Local, Penalized, and Additive Learning

Daniel Rubin and Mark J. van der Laan

Abstract

Marginal structural models (MSMs) allow one to form causal inferences from data, by specifying a relationship between a treatment and the marginal distribution of a corresponding counterfactual outcome. Following their introduction in Robins (1997), MSMs have typically been fit after assuming a semiparametric model, and then estimating a finite dimensional parameter. van der Laan and Dudoit (2003) proposed to instead view MSM fitting not as a task of semiparametric parameter estimation, but of nonparametric function approximation. They introduced a class of causal effect estimators based on mapping loss functions suitable for the unavailable counterfactual data to those suitable for the data actually observed, and then applying what has been known in nonparametric statistics as empirical risk minimization, or global learning.

However, it has long been recognized in the statistical learning community that global learning is only one of several paradigms for estimator construction. Building upon van der Laan and Dudoit's work, we show how marginal structural models for causal effects can be extended through the alternative techniques of local, penalized, and additive learning. We discuss how these new methods can often be implemented by simply adding observation weights to existing algorithms, demonstrate the gains made possible by these extended MSMs through simulation results, and conclude that nonparametric function estimation methods can be fruitfully applied for making causal inferences.

1 Introduction

Marginal structural models (MSMs) were introduced by Robins (1997) as tools for drawing causal inferences from data. Let $\mathcal{A} \subset \mathbb{R}^d$ denote a set of possible treatments that can be given to subjects. Let Y_a denote the outcome or response for a subject that would have occured if, possibly contrary to fact, treatment $a \in \mathcal{A}$ had been administered. When studying how treatment $a \in \mathcal{A}$ affects the outcome Y_a, we would ideally observe for every subject,

$$X = (W, \{Y_a : a \in \mathcal{A}\}), \tag{1}$$

for W a vector of baseline covariates. However, suppose that only a single random treatment $A \in \mathcal{A}$ is actually given to each subject. Hence, consider the scenario in which the observed data on a subject is,

$$O = (W, A, Y_A), \tag{2}$$

and we have collected an i.i.d. sample $\{O_i\}_{i=1}^n$ representing data on n subjects. An MSM can be used to estimate features of the marginal distributions of the counterfactual responses $\{Y_a : a \in \mathcal{A}\}$, and consequently to analyze the causal effect of a subject's treatment on their outcome. The necessity for MSMs arises because when the covariates W influence both the treatment and outcome, the distribution of Y_a does not necessarily equal the conditional distribution of $\{Y_A|A = a\}$. Knowledge of this more traditional object of study would merely provide information concerning the association of treatment $a \in \mathcal{A}$ with the subject responses.

Estimation of causal effects cannot be done with observational data unless a vital assumption is made. The key requirement is that enough baseline covariates W are collected so that there is no unmeasured confounding. This is taken to mean that the treatment and set of counterfactual responses are conditionally independent given the covariates, written formally as,

$$\{A \perp \{Y_a : a \in \mathcal{A}\}|W\}. \tag{3}$$

The assumption (3) will be satisfied in a randomized trial, or any situation where treatment is randomly assigned to subjects, with the randomizing mechanism depending only on the baseline covariates W. From a temporal standpoint, we can safely assume

that there is no unmeasured confounding if the baseline covariates are first measured, and treatment is then assigned before any additional knowledge concerning the subject's potential outcome is available. In general, there is no way to empirically verify from the observed data $\{O_i\}_{i=1}^n$ whether (3) holds. This difficulty has made causal inference somewhat controversial with observational data, where there may be no way to tell if enough covariates have been collected to ensure that there is no unmeasured confounding.

Assuming that (3) holds, an MSM places restrictions on a function ψ mapping the treatment a to a feature of the marginal distribution of Y_a. Examples of such possible functions are,

$$\begin{aligned}
\psi_1(a) &= E[Y_a], \\
\psi_2(a) &= F_{Y_a}(y) = P(Y_a \leq y), \\
\psi_3(a) &= F_{Y_a}^{-1}(\tau) = \sup\{y : F_{Y_a}(y) \leq \tau\}, \\
\psi_4(a) &= P(Y_a = 1), \\
\psi_5(a) &= I(P(Y_a = 1) \geq \frac{1}{2}), \\
\psi_6(a)(\cdot) &= F_{Y_a}(\cdot), \\
\psi_7(a)(\cdot) &= f_{Y_a}(\cdot) = F'_{Y_a}(\cdot).
\end{aligned} \tag{4}$$

Note that $\psi_1, ..., \psi_5$ map treatment a to a real number summarizing the distribution of Y_a. ψ_1, ψ_2, and ψ_3 respectively map a to the mean of Y_a, the CDF of Y_a at a fixed point y, and the τ-quantile of Y_a. ψ_4 is of interest with dichotomous responses $Y_a \in \{0, 1\}$ indicating success or failure, and maps $a \in \mathcal{A}$ to the success probability corresponding to this treatment. ψ_5 is also of use with dichotomous responses, when there is interest in classifying which treatments will lead to success or failure, as it maps treatment a to the Bayes classification of Y_a. ψ_6 and ψ_7 are more complicated because they map $a \in \mathcal{A}$ to entire functions related to the distribution of Y_a, instead of a single real-valued functional of this distribution. These two choices of ψ respectively relate treatment a to the CDF and density function of the counterfactual response Y_a.

In fact, MSMs can be used to model features of the conditional distribution $\{Y_a|V\}$, for V a subset of the covariates W. They can also be utilized to examine causal effects in longitudinal studies, where the treatment A is adjusted according to past treatment and subject history. For expositional purposes, we will not focus on these more complex

marginal structural models.

In this paper we will further restrict attention to the case where the treatment A is a continuous random variable, with $g(A|W)$ denoting the conditional density of A given W. This has not been the usual setting when applying MSMs for causal inference, where treatment has typically been taken to be categorical or ordinal, and warrants explanation. Our restriction to continuous A in this work is meant to elucidate the gains made possible by viewing estimation of ψ as an exercise in nonparametric function approximation, rather than in semiparametric parameter estimation. There are myriad situations with randomized or observational data where learning the causal effect of a continuous treatment A could be of interest, when intervention on A is possible, such as treatment representing a continuous drug dosage. The examination of continuous treatments could also be worthwhile in many previously studied cases where treatment was discretized prior to analysis, but in fact treatment measurements with greater precision were available.

As originally conceived, an MSM will specify a functional form for $\psi(\cdot)$ such as $\psi(a) = \psi(a|\beta)$, parameterized by an unknown $\beta \in \mathbb{R}^p$. Determination of β from the observed data $\{O_i\}_{i=1}^n$ then becomes a problem of semiparametric estimation, and procedures have been developed for this purpose for a wide class of MSMs, as summarized in van der Laan and Robins (2002). Once β is estimated with $\hat{\beta}$, the estimate of the function ψ becomes $\hat{\psi}(a) = \psi(a|\hat{\beta})$. Although MSMs have been associated with semiparametric models, as a matter of notation we will refer to any estimates of such causal parameters $\psi(\cdot)$ as fits to marginal structural models.

van der Laan and Dudoit (2003) proposed to nonparametrically estimate the types of functions $\psi(\cdot)$ previously described, through what they termed *loss based estimation*, to be described in section 2. Instead of parameterizing $\psi(\cdot)$ by a Euclidean $\beta \in \mathbb{R}^p$, van der Laan and Dudoit's estimation techniques relied on mapping loss functions that would have been used with the complete data $\{X_i\}_{i=1}^n$ to those suitable with only the observed data $\{O_i\}_{i=1}^n$. Using these mapped loss functions, the approach essentially reduces to what has been termed empirical risk minimization in the statistical learning literature. The present paper can be viewed as an extension of this work.

It is well known in the statistical learning community that empirical risk minimization is only one of several broad classes of techniques for estimator construction. In an overview of nonparametric regression, Györfi et al. (2002) make the distinction be-

tween what they term the "four related paradigms" of local averaging, local modeling, global modeling, and penalized modeling. While van der Laan and Dudoit's empirical risk minimization approach to function approximation falls into the framework of global modeling, their idea of replacing a complete data loss function with an observed data loss function can be applied to alternative learning paradigms. In section 3 we propose methods to fit general $\psi(\cdot)$ through local learning, penalized learning, and additive learning. We will present simulation results demonstrating the gains available from using these new procedures for nonparametric function estimation, and discuss how these methods can often be implemented by adding observation weights to existing algorithms. We conclude in section 4 by briefly noting how the local, penalized, and additive learning techniques based on the loss function replacement ideas described in section 3 are not limited to the causal inference setting, but can be applied to estimation problems with general types of incomplete data.

2 MSMs, Global Learning, and Cross-Validation

For many causal parameters $\psi(\cdot)$, the quality of an approximation $\hat{\psi}(\cdot)$ can be quantified by the magnitude of a risk function,

$$R(\hat{\psi}) = E[L(X, \hat{\psi})] = E[\int_{\mathcal{A}} \mu(a) l(Y_a, \hat{\psi}(a)) da].$$

Here $l(Y_a, \hat{\psi}(a)) \in \mathbb{R}$ defines the loss incurred by $\hat{\psi}(a)$ in predicting some feature of the counterfactual response Y_a. The *loss function* $L(X, \hat{\psi})$ simply integrates these counterfactual losses across the set of treatments, and the risk of $\hat{\psi}(\cdot)$ is defined by the expected value of the loss function. In this formulation, $\mu : \mathcal{A} \rightarrow \mathbb{R}$ is a user-supplied weight function, meant to specify regions of $\mathcal{A}$ where the precision of $\hat{\psi}(\cdot)$ is given increased or decreased importance.

For the causal parameters $\psi_1, ..., \psi_7$ as in (4), natural values for the losses $l(Y_a, \hat{\psi}(a))$

can be given by,

$$\begin{aligned}
l_1(Y_a, \hat{\psi}(a)) &= |Y_a - \hat{\psi}(a)|^2, \\
l_2(Y_a, \hat{\psi}(a)) &= |I(Y_a \leq y) - \hat{\psi}(a)|^2, \\
l_3(Y_a, \hat{\psi}(a)) &= |Y_a - \hat{\psi}(a)| + (2\tau - 1)|Y_a - \hat{\psi}(a)|, \\
l_4(Y_a, \hat{\psi}(a)) &= -Y_a \log(\hat{\psi}_a) - (1 - Y_a)\log(1 - \hat{\psi}(a)), \\
l_{5,1}(Y_a, \hat{\psi}(a)) &= I(Y_a \neq \hat{\psi}(a)),\ l_{5,2}(Y_a, \hat{\psi}(a)) = \max(1 - (2Y_a - 1)\hat{\psi}(a), 0), \\
l_6(Y_a, \hat{\psi}_a(\cdot)) &= \int |I(Y_a \leq y) - \hat{\psi}_a(y)|^2 dy, \\
l_{7,1}(Y_a, \hat{\psi}_a(\cdot)) &= -\log \hat{\psi}_a(Y_a),\ l_{7,2}(Y_a, \hat{\psi}_a(\cdot)) = \int \hat{\psi}_a^2(y)dy - 2\hat{\psi}_a(Y_a). \qquad (5)
\end{aligned}$$

Most of these losses should be recognized immediately. Here l_1 is the usual squared error loss function for prediction of Y_a, l_2 measures the squared error for prediction of the indicator $I(Y_a \leq y)$, l_3 gives the standard loss function for predicting the τ-quantile for Y_a, and l_4 gives the well known cross-entropy loss for prediction of a binary response. $l_{5,1}$ is the misclassification loss function, returning a loss of one if the prediction of binary Y_a is misclassified, and zero loss otherwise. The support vector machine loss function $l_{5,2}$ has recently generated a great deal of interest for use in classification problems, and both classification losses $l_{5,1}$ and $l_{5,2}$ have as their risk minimizers the Bayes classifier $\psi_5(a)$. The loss l_6 simply gives an integrated version of l_2. $l_{7,1}$ and $l_{7,2}$ provide the negative log-likelihood and least squares loss for estimators $\hat{\psi}(a)(\cdot)$ of the density of Y_a, the latter introduced in Rudemo (1982). Choosing candidate density estimators $\hat{\psi}(a)(\cdot)$ to minimize risks $E[l_{7,1}(Y_a, \hat{\psi}(a)(\cdot))]$ and $E[l_{7,2}(Y_a, \hat{\psi}(a)(\cdot))]$ will respectively result in the estimators minimizing the Kullbach-Leibler divergence and integrated squared distance from the true density function of the counterfactual response Y_a.

Before considering estimation of a causal parameter $\psi(\cdot)$ from the observed data $\{O_i\}_{i=1}^n$, we will first mention how the problem could be solved if the counterfactual data $\{X_i\}_{i=1}^n$ were available. Given a (possibly infinite) collection Ψ of candidate estimators for $\psi(\cdot)$, we would ideally want to choose the estimator ψ_n having the smallest risk, or

$$\psi_n = \text{argmin}_{\{\hat{\psi} \in \Psi\}} R(\hat{\psi}).$$

Unfortunately, such a ψ_n could never be used in practice. Even with the counterfactual data $\{X_i\}_{i=1}^n$, the risk function would depend on the unknown data generating distribu-

tion, and could not be evaluated. *Empirical risk minimization* attacks this problem by noting that the risk of $\hat{\psi}$ is the expected value of $L(X, \hat{\psi})$, which can be approximated by the empirical mean $\frac{1}{n}\sum_{i=1}^{n} L(X_i, \hat{\psi})$. The resulting estimator then becomes,

$$\psi_n = \text{argmin}_{\{\hat{\psi}\in\Psi\}} \frac{1}{n}\sum_{i=1}^{n} L(X_i, \hat{\psi}). \tag{6}$$

For example, suppose that we are interested in estimating $\psi(\cdot) : a \to E[Y_a]$. With the set of candidates given by,

$$\Psi_p = \{\hat{\psi}(\cdot) : \hat{\psi}(a) = \sum_{j=0}^{p} \beta_j a^j\}, \tag{7}$$

we could use the loss $l(Y_a, \hat{\psi}(a)) = |Y_a - \hat{\psi}(a)|^2$ when attempting to make the best polynomial fit to $\psi(\cdot)$ of degree p. Whenever the causal parameter $\psi(\cdot)$ can be represented by a basis expansion, a common technique is to consider candidate sets Ψ_p consisting of linear combinations of the first p functions in this expansion.

Clearly, there is a bias-variance type tradeoff as the size of the candidate set Ψ grows. When introducing more candidates to Ψ, we decrease the risk of the risk minimizer, but curtail our ability to uniformly control the differences across Ψ between the true and empirical risks. The size of the candidate set Ψ is frequently increased as the sample size n grows. Because this empirical risk minimization approach selects among candidates that attempt to approximate the causal parameter $\psi(\cdot)$ as a function of the entire treatment set $\mathcal{A}$, this estimator ψ_n is occasionally also said to have been built from *global modeling* or *global learning* (as opposed to local learning, to be described in the subsequent section).

When only the observed data $\{O_i\}_{i=1}^{n}$ is available, the empirical risk minimizer (6) cannot be used. If we could somehow find an observed data loss function $L^\star(O, \hat{\psi})$ having the same expected value as $L(X, \hat{\psi})$, then its expected value would equal the risk. Just as $\frac{1}{n}\sum_{i=1}^{n} L(X_i, \hat{\psi})$ would be an empirical estimate of the risk, so would $\frac{1}{n}\sum_{i=1}^{n} L^\star(O_i, \hat{\psi})$, and we could use the estimator,

$$\psi_n = \text{argmin}_{\hat{\psi}\in\Psi} \frac{1}{n}\sum_{i=1}^{n} L^\star(O_i, \hat{\psi}). \tag{8}$$

Unfortunately, such an observed data loss function $L^\star(O, \hat{\psi})$ will generally not be an explicit function of the observed data O and the candidate $\hat{\psi}$, but will depend on

nuisance parameters that themselves must be estimated from the data. For instance, we will consider the observed data loss function,

$$L^\star(O, \psi|g) = \frac{\mu(A)}{g(A|W)} l(Y_A, \psi(A)). \tag{9}$$

The nuisance parameter involved in this observed data loss function is $g(\cdot|W)$, the conditional density function of the treatment A given the baseline covariates W. Hence, to evaluate the loss function and estimate ψ_n as in (8), we would first have to perform conditional density estimation. If $E[L(X, \hat{\psi})]$ is finite and there exists an $\epsilon > 0$ such that the identifiability condition

$$\mu(a) > 0 \text{ implies } g(a|W) > \epsilon \text{ with probability one} \tag{10}$$

holds, we see that indeed,

$$\begin{aligned} E[L^\star(O, \hat{\psi}|g)] &= E[E[L^\star(O, \hat{\psi}|g)|X]] \\ &= E[E[L^\star(O, \hat{\psi}|g)|W, \{Y_a : a \in \mathcal{A}\}]] \\ &= E[\int_{\mathcal{A}} \frac{\mu(a)}{g(a|W)} l(Y_a, \hat{\psi}(a)) g(a|W) da] \\ &= E[\int_{\mathcal{A}} \mu(a) l(Y_a, \hat{\psi}(a)) da] \\ &= E[L(X, \hat{\psi})]. \end{aligned} \tag{11}$$

Note that the loss function $L^\star(O, \hat{\psi}|g)$ is attractive from a computational standpoint, because it is simply a weighted version of $l(Y_A, \hat{\psi}(A))$. $\mu(A)$ weighs $l(Y_a, \hat{\psi}(A))$ to account for the importance of $\hat{\psi}(\cdot)$ being accurate when evaluated at the observed treatment A, while weighing by the inverse density of treatment $\frac{1}{g(A|W)}$ adjusts for the confounding of A and Y_A present in the baseline covariates W. Whenever software is available to minimize $\frac{1}{n} \sum_{i=1}^n l(Y_A, \hat{\psi}(A))$, and can be made to take observation weights, the estimator in (8) can be immediately implemented once the conditional treatment density g is fit with $\hat{g}$.

Rather than constructing a single empirical risk minimizer based on a candidate set Ψ, van der Laan and Dudoit (2003) considered the sieve-based approach of nesting candidate sets $\Psi_1 \subset \Psi_2 \subset ...$, constructing an empirical risk estimator $\psi_{n,p}$ for many values of p based on the observed data loss function and candidate set Ψ_p, and then selecting among these estimators with a certain form of cross-validation. For instance,

if using the Ψ_p giving degree p polynomials as in (7), then cross-validation would be used to select the degree p of the desired polynomial fit.

Just as empirical risk minimization depends on an observed data loss function $L^\star(O, \hat{\psi})$ having the same expectation as $L(X, \hat{\psi})$, so does van der Laan and Dudoit's version of cross-validation, and in fact the two procedures are closely related. Their cross-validation technique amounts to splitting the data into two groups (possibly repeated over several folds), using a training set for constructing a set of candidates Ψ, and using a validation set to choose among this Ψ with empirical risk minimization based on $L^\star(O, \hat{\psi})$. Formally, if $\hat{P}$ is an empirical probability mass function putting mass $\frac{1}{n}$ on $\{O_i\}_{i=1}^n$, $\psi_k(\hat{P})$ denotes the estimator produced by the kth estimation procedure when fed data $\{O_i\}_{i=1}^n$, and $L^\star(O, \hat{\psi}|\hat{g})$ is the estimated observed data loss function as in (9), then the cross-validation procedure would work as follows.

$$
\begin{aligned}
\hat{P}(o) &= \frac{1}{n}\sum_{i=1}^{n} I(O_i = o) \text{ is the empirical probability mass function (PMF),} \\
\{B_i\}_{i=1}^n &\in \{0,1\}^n \text{ is a random vector indicating the training and validation samples,} \\
\hat{P}_{B,0}(o) &= \frac{1}{\sum_{i=1}^n (1-B_i)}\sum_{i=1}^{n} I(O_i = o, B_i = 0) \text{ is the training sample empirical PMF,} \\
\hat{P}_{B,1}(o) &= \frac{1}{\sum_{i=1}^n B_i}\sum_{i=1}^{n} I(O_i = o, B_i = 1) \text{ is the validation sample empirical PMF,} \\
\hat{k} &= \text{argmin}_{1\leq k\leq K} E_B \int L^\star(o, \psi_k(\hat{P}_{B,0})|\hat{g}) d\hat{P}_{B,1}(o), \\
\hat{\psi} &= \psi_{\hat{k}}(\hat{P}) \text{ is the estimator selected by cross-validation.}
\end{aligned} \tag{12}
$$

In a method they termed their *loss based estimation* approach, van der Laan and Dudoit introduced general ways of mapping full data loss functions $L(X, \hat{\psi})$ into observed data loss functions $L^\star(O, \hat{\psi})$ having the same expectation, of which (9) is a special case. We should note that (9) is not the optimal observed data loss function in terms of either efficiency or robustness. However, it will suffice for our purposes of showing how such an observed data loss function $L^\star(O, \hat{\psi})$ can be used for causal inference with local, penalized, and additive modeling. The fact that the observed data loss function $L^\star(O, \hat{\psi}|g)$ in (9) is simply a weighted version of $l(Y_A, \hat{\psi}(A))$ will also come in handy, when implementing the estimation procedures to be described in the sequel.

3 MSMs Using Alternative Learning Paradigms

While the observed data loss function $L^\star(O, \hat{\psi}|g)$ can be used for empirical risk minimization as in (8), it can also be used for other types of estimator construction. Hastie et al. (2001) mentions that,

> The variety of nonparametric regression techniques or learning methods fall into a number of different classes depending on the nature of the restrictions imposed. These classes are not distinct, and indeed some methods fall into several classes.

Empirical risk minimization, or global learning, is simply one of these different classes of learning methods. In this section we describe local, penalized, and additive learning, which are motivated by different considerations, and how the observed data loss function $L^\star(O, \hat{\psi}|g)$ can also be exploited when using these classes of learning tools to estimate causal parameters.

3.1 Local Learning

One popular technique for forming estimators is based on trying to locally approximate a function of interest. That is, rather than the global learning approach of attempting to select the closest $\hat{\psi}(\cdot)$ to $\psi(\cdot)$ from a candidate set Ψ, we could estimate $\psi(a_0)$ by only considering candidates in Ψ that well approximate $\psi(\cdot)$ in a neighborhood of a_0. Local estimators generally start with a candidate set Ψ that is somewhat smaller than would be used in the global learning approach. For example, when estimating the counterfactual mean process $\psi : a \rightarrow \text{Median}(Y_a)$ we could consider the candidate set,

$$\Psi = \{\psi : \psi(a) = \beta_0 + \beta_1 a\}, \tag{13}$$

consisting of linear functions of the treatment. Although we might not expect $\psi(\cdot)$ to be globally well approximated by a linear function, the first order Taylor expansion

$$\psi(a) \simeq \psi(a_0) + (a - a_0)\psi'(a_0) \tag{14}$$

suggests that the causal parameter $\psi(\cdot)$ might behave like a member of Ψ in a neighborhood of treatment $a_0 \in \mathcal{A}$. An overview of local modeling is provided in Fan and Gijbels (1996).

Local learning often depends on a kernel function $K_h : \mathbb{R} \to \mathbb{R}^+$, with bandwidth a h, chosen so that $K_h(\|a - a_0\|)$ becomes large if treatment a is far from a_0. For example,

$$\begin{aligned} K_{1,h}(x) &= \frac{1}{\sqrt{2\pi}h}\exp(-\frac{1}{2h^2}x^2), \\ K_{2,h}(x) &= I(|x| \le h), \\ K_{3,h}(x) &= \frac{3}{4}(1 - \frac{x^2}{h^2})I(|x| \le h), \\ K_{4,h}(x) &= (1 - \frac{|x|^3}{h^3})I(|x| \le h), \end{aligned}$$

define the Gaussian, box, Epanechnikov, and tri-cube kernels commonly used in smoothing applications. If the counterfactual data $\{X_i\}_{i=1}^n$ were available, we could image locally approximating a causal parameter $\psi(\cdot)$ in a neighborhood of $a_0 \in \mathcal{A}$ with,

$$\psi_{n,a_0} = \text{argmin}_{\{\hat{\psi}\in\Psi\}}\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{A}}\mu(a)l(Y_{a,i},\hat{\psi}(a))K_h(\|a - a_0\|)da,$$

resulting in the estimator,

$$\psi_n(a) = \psi_{n,a}(a). \tag{15}$$

Such an estimator attempts to find the candidate $\hat{\psi} \in \Psi$ minimizing the locally weighted risk,

$$R_{a_0}(\hat{\psi}) = E[\int_{\mathcal{A}}\mu(a)l(Y_a,\hat{\psi}(a))K_h(\|a - a_0\|)da], \tag{16}$$

and for each $a_0 \in \mathcal{A}$ selects the candidate ψ_{n,a_0} via empirical risk minimization. By adding the kernel weight $K_h(\|a - a_0\|)$ to the integrand in the loss function, we more heavily weigh losses $l(Y_a, \hat{\psi}(a))$ for treatments a close to a_0. The bandwidth h calibrates the size of the neighborhood over which we hope to locally approximate the causal parameter $\psi(\cdot)$. As is well known, changing this bandwidth results in a bias-variance tradeoff, and h would typically be chosen with cross-validation.

When only having access to the observed data $\{O_i\}_{i=1}^n$, computing the estimator (15) is impossible. But just as we can construct empirical risk estimators with the surrogate loss function $L^\star(O, \hat{\psi}|g)$ defined in (9), we can form local empirical risk estimators at $a_0 \in \mathcal{A}$ through weighing this loss function by $K_h(\|A - a_0\|)$. That is, it can be shown as in (11) that $L^\star(O, \hat{\psi}|g)K_h(\|A - a_0\|)$ is unbiased for the locally weighted risk defined $R_{a_0}(\hat{\psi})$ defined in (16). The observed data analog to the complete data estimator of (15) is to first estimate the nuisance parameter g with $\hat{g}$, and then

choose the candidate $\psi_{n,a_0} \in \mathcal{A}$ minimizing the locally weighted empirical risk. This gives,

$$\begin{aligned}\psi_{n,a_0}(\cdot) &= \text{argmin}_{\hat{\psi}\in\Psi}\frac{1}{n}\sum_{i=1}^{n} K_h(\|A_i - a_0\|)L^\star(O_i, \hat{\psi}|\hat{g}) \\ &= \text{argmin}_{\hat{\psi}\in\Psi}\frac{1}{n}\sum_{i=1}^{n} \frac{\mu(A_i)}{\hat{g}(A_i|W_i)} l(Y_{Ai}, \hat{\psi}(A_i))K_h(\|A_i - a_0\|),\end{aligned}$$

resulting in the estimator $\psi_n(a) = \psi_{n,a}(a)$. If a software routine exists to minimize the local empirical risk $\frac{1}{n}\sum_{i=1}^{n} l(Y_{A,i}, \hat{\psi}(A_i))K_h(\|A_i - a_0\|)$, and accepts observation weights, this estimator ψ_n can be conveniently computed by adding observation weights $\frac{\mu(A_i)}{\hat{g}(A_i|W_i)}$ to the existing algorithm. The bandwidth h indexes a class of estimators, and could be selected with van der Laan and Dudoit's observed data cross-validation method (12), described in the previous section.

To illustrate the local learning of a causal parameter, we ran a simulation based on $n = 200$ observations. We considered estimating the function $\psi(\cdot) : \mathcal{A} = [0, 1] \rightarrow$ Median(Y_a). We chose $\mu(a) = 1$ as our weight function, which gave no extra weight to any region of counterfactuals in the fitting process. The observed data was generated according to,

$$\begin{aligned} W &\sim U(0,1) \text{ for } U(0,1) \text{ the uniform distribution on } [0,1], \\ A &\sim \beta(\frac{1}{2} + W,\ \frac{3}{2} - W) \text{ for } \beta(a,b) \text{ the Beta distribution with shapes a and b}, \\ Y_a &\sim \psi(a) + W - 1 + \beta(2,2). \end{aligned} \tag{17}$$

Indeed, one can verify that $\psi(a) = E[Y_a] = \text{Median}(Y_a)$. We considered the four choices of the causal parameter,

$$\begin{aligned}\psi(a) &= \sin(15a), \\ \psi(a) &= a, \\ \psi(a) &= 4(a - \frac{1}{2})^2, \\ \psi(a) &= I(a > \frac{1}{2}),\end{aligned}$$

giving an oscillating, linear, quadratic, and step function. We estimated these $\psi(\cdot)$ with the median regression loss function $l(Y_a, \hat{\psi}(a)) = |Y_a - \hat{\psi}(a)|$ corresponding to l_3 in (5) with $\tau = \frac{1}{2}$. We then attempted to form locally linear approximations to the median

function $\psi(\cdot)$. We estimated the conditional density function $g(\cdot|W)$ with $\hat{g}(\cdot|W)$ based on the **hare()** function in the polspline R package, which fit a hazard regression model using linear splines. For implementing this median regression, we merely had to add observation weights $\frac{\mu(A_i)}{\hat{g}(A_i|W_i)}$ to the **rq()** function in the R quantreg package, which could be used to find the linear function minimizing the empirical absolute deviation $\frac{1}{n}\sum_{i=1}^{n}|Y_A - \beta_0 - \beta_1 A|$. Our fits were based on using a Gaussian kernel, and fixed bandwidth $h = \frac{1}{10}$. The results in Figure 1 suggest that the local procedure is indeed able to smoothly fit causal median curves $\psi(\cdot)$ of various shapes.

3.2 Penalized Learning

If minimizing empirical risk over a set of candidates Ψ that is in some sense too large, one cannot ensure that the risks $E[L(X, \hat{\psi})]$ will be well approximated by their empirical versions $\frac{1}{n}\sum_{i=1}^{n} L(X_i, \hat{\psi})$ uniformly over $\hat{\psi} \in \Psi$. We will then possess no guarantee that the empirical risk minimizer will be a decent estimate of the parameter of interest. A tremendous amount of work in empirical process theory has been directed toward specifying exactly which function classes Ψ are "too large" in learning problems, and the "size" of Ψ can often be controlled through restricting its Vapnik-Chernovenkis dimension. An alternative approach is to continue using a large candidate set Ψ, such as

$$\Psi = \{\psi : \psi \text{ maps } \mathcal{A} \text{ to } \mathbb{R}, \text{ and has two continuous derivatives}\}, \tag{18}$$

but also penalize the empirical risk by the complexity of $\hat{\psi} \in \Psi$.

Such penalization depends on a penalty functional $J : \Psi \to \mathbb{R}^+$. When using the candidate set Ψ defined in (18), a common approach in many smoothing problems is to penalize the complexity of $\hat{\psi}$ by the curvature of the function, which can be quantified through,

$$J(\hat{\psi}) = \int_{\mathcal{A}} \{\hat{\psi}''(a)\}^2 da. \tag{19}$$

If the counterfactual data $\{X_i\}_{i=1}^{n}$ were available, we could then estimate the causal parameter $\psi(\cdot)$ with,

$$\psi_n = \text{argmin}_{\hat{\psi}\in\Psi} \frac{1}{n}\sum_{i=1}^{n} L(X_i, \hat{\psi}) + \lambda J(\hat{\psi}). \tag{20}$$

Here $\lambda \geq 0$ is a smoothing parameter, used to specify the tradeoff between empirical risk and complexity penalization for a candidate $\hat{\psi}$. When $\lambda = 0$, $\psi_n(\cdot)$ can be any

function minimizing empirical risk. When $\lambda = \infty$, $\psi_n(\cdot)$ must be a linear function of the treatment, because a nonzero second derivative will imply an infinite penalty. As noted by Hastie et al. (2001), these choices of λ lead to estimators ψ_n that "vary from very rough to very smooth, and the hope is that $\lambda \in (0, \infty)$ indexes an interesting class of functions in between." This smoothing parameter λ would generally be chosen with cross-validation. Greater detail on penalized modeling can be found in Wahba (1990).

With only the observed data $\{O_i\}_{i=1}^n$, we could not hope to directly evaluate the penalized empirical risk $\frac{1}{n}\sum_{i=1}^n L(X_i, \hat{\psi}) + \lambda J(\hat{\psi})$. But as with the global and local learning approaches to function approximation previously discussed, we can tackle the problem by using the observed data loss function $L^\star(O, \hat{\psi}|g)$. Because $\frac{1}{n}\sum_{i=1}^n L^\star(O_i, \hat{\psi}|g) + \lambda J(\hat{\psi})$ is unbiased for the penalized empirical risk, it is a natural estimate of this quantity. After fitting the nuisance parameter g with $\hat{g}$ based on conditional density estimation, we could then perform penalized learning with the observed data by forming the estimator,

$$\begin{aligned} \psi_n &= \text{argmin}_{\hat{\psi}\in\Psi} \frac{1}{n}\sum_{i=1}^n L^\star(O_i, \hat{\psi}|\hat{g}) + \lambda J(\hat{\psi}) \\ &= \text{argmin}_{\hat{\psi}\in\Psi} \frac{1}{n}\sum_{i=1}^n \frac{\mu(A_i)}{\hat{g}(A_i|W_i)} l(Y_{A,i}, \hat{\psi}(A_i)) + \lambda J(\hat{\psi}). \end{aligned} \tag{21}$$

As with local learning, the smoothing parameter λ can be selected with van der Laan and Dudoit's observed data cross-validation scheme, given in (12).

Even if Ψ is an infinite dimensional function space as in (18), an elegant feature of penalized learning is that the estimator ψ_n can often be easily computed. When using the squared error loss $l(Y_A, \hat{\psi}(A)) = |Y_A - \hat{\psi}(A)|^2$, the estimator ψ_n becomes a natural cubic spline with knots at the observed treatments $(A_1, ..., A_n)$. As with the global and local learning estimators, the penalized learning estimator of ψ can often be trivially implemented if existing software routines can minimize $\frac{1}{n}\sum_{i=1}^n l(Y_A, \hat{\psi}(A)) + \lambda J(\hat{\psi})$ over $\hat{\psi} \in \Psi$, and can take observation weights $\frac{\mu(A_i)}{\hat{g}(A_i|W_i)}$.

To demonstrate the potential benefits of penalized learning in causal inference problems, we again generated $n = 200$ observations according to (17), this time attempting to estimate the parameter $\psi(a) = E[Y_a] = \sin(15a)$. As in the previous subsection, we estimated the nuisance parameter $g(\cdot|W)$ with the **hare()** R function. The estimator ψ_n defined in (21) was based on the squared error loss $l(Y_A, \hat{\psi}(A)) = |Y_A - \hat{\psi}(A)|^2$. We were able implement ψ_n by adding the relevant obser-

vation weights to the **smooth.spline()** function in R, which also chose the smoothing parameter λ.

From Figure 2, we see that the penalized learning procedure accurately fit the causal parameter $\psi(\cdot)$. However, in this same simulation we also estimated $\psi(\cdot)$ with a global learning approach, based on minimizing empirical risk over candidate sets Ψ containing polynomials of degree up to four. Even though penalized learning led to accurate curve fitting, these attempts at globally approximating $\psi(\cdot)$ failed, as shown in Figure 3. While van der Laan and Dudoit considered using $L^\star(O, \hat{\psi}|g)$ for sieve-based empirical risk minimization, these results seem to demonstrate the potential benefits of also considering estimators built from this observed data loss function in a different way.

3.3 Additive Learning

While our previous simulations have focused on a univariate $a \in \mathcal{A} \subset \mathbb{R}$, in many studies there will be an interest in a multivariate treatment,

$$\mathbf{a} = (a_1, ..., a_d) \in \mathcal{A} = \mathcal{A}_1 \times ... \times \mathcal{A}_d \subset \mathbb{R}^d. \tag{22}$$

For even moderate treatment dimension d, we may be plagued by the "curse of dimensionality" when trying to learn a causal function such as $\psi : \mathbf{a} \rightarrow E[Y_{\mathbf{a}}]$. Often the only way out is to make assumptions concerning the structure of the multivariate function of interest. While the most traditional approach to approximating multivariate functions has been to make a linear fit, a more flexible technique is to assume an additive model, as discussed in Hastie and Tibshirani (1990). Such a model would imply that,

$$\psi(\mathbf{a}) = \alpha + \sum_{j=1}^{d} \psi_j(a_j). \tag{23}$$

The constant α will not be identifiable unless the additive components $\psi_j(\cdot)$ are somehow centered. A convenient choice is to center such that $E[\psi_j(A_j)] = 0$. The set of candidates can then be written as,

$$\Psi = \{\hat{\psi} : \hat{\psi}(\mathbf{a}) = \alpha + \sum_{j=1}^{d} \hat{\psi}_j(a_j),\ E[\hat{\psi}_j(A_j)] = 0\}.$$

A popular approach to fitting additive models is known as *backfitting*, in which each of the additive components $\psi_j(\cdot)$ are iteratively fit with univariate smoothing.

The essential idea is that if fits to the intercept α and $\psi_k(\cdot)$, $k \neq j$ have already been made, then estimation of $\psi(\cdot)$ has been reduced to estimation of the univariate $\psi_j(\cdot)$. This component could then be fit with the natural cubic spline resulting from penalized estimation, as discussed in section 3.2. For $J(\cdot)$ the penalty functional of (19), and $\hat{g}$ an estimate of the nuisance parameter g involved in the observed data loss function (9), the backfitting algorithm can be written as follows, adapted from Algorithm 9.1 of Hastie et al. (2001).

1. Initialize $\psi_{n,j}(\cdot) = 0$, $\alpha_n = \text{argmin}_{\alpha \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} L^{\star}(O_i, \alpha | \hat{g})$
2. Cycle: j=1,2,... ,d,1,... until convergence of α_n and the functions $\psi_{n,1}, ..., \psi_{n,d}$

$$\psi_{n,j} \leftarrow \text{argmin}_{\{\hat{\psi}_j : \mathcal{A}_j \rightarrow \mathbb{R}\}} \frac{1}{n} \sum_{i=1}^{n} L^{\star}(O_i, \alpha_n + \sum_{k \neq j} \psi_{n,k} + \hat{\psi}_j | \hat{g}) + \lambda_j J(\hat{\psi}_j)$$

$$\psi_{n,j} \leftarrow \psi_{n,j} - \frac{1}{n} \sum_{i=1}^{n} \psi_{n,j}(A_{i,j})$$

$$\alpha_n \leftarrow \text{argmin}_{\alpha \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} L^{\star}(O_i, \alpha + \sum_{j=1}^{d} \psi_{n,j} | \hat{g}) \tag{24}$$

Here $\lambda_1, ...\lambda_d$ are smoothing parameters, and could be chosen with the observed data cross-validation method as in (12). As with local and penalized estimation of causal parameters, this procedure could often be implemented by adding observation weights to existing procedures. Suppose that $\psi(\cdot)$ was not necessarily a causal parameter, but simply predicted some feature of the observed response Y_A from the multivariate treatment A. For $\psi(\cdot)$ minimizing the risk $E[l(Y_A, \psi(A))]$, we could imagine studying the association between the treatment and outcome by fitting an additive model for $\psi(\cdot)$. If centering the additive components so that $\psi_j(A_j)$ has mean zero, then (24) could be implemented by adding observation weights $\frac{\mu(A_i)}{\hat{g}(A_i|W_i)}$ to an existing backfitting routine.

4 Learning in General Incomplete Data Structures

While we have focused on fitting marginal structural models, van der Laan and Dudoit's approach to empirical risk minimization with mapped loss functions was originally developed in greater generality. They in fact discussed how their approach could apply to global learning with right censored data, and other types of incomplete data structures.

Abstracting from the causal inference setting thus far described, consider a situation where O represents the observed data, while X represents the unobserved complete data that we would have preferred to measure. Whenever a function $L(X, \hat{\psi})$ could have been used with the complete data X to measure the loss incurred by a candidate estimator $\hat{\psi}$, van der Laan and Dudoit considered performing empirical risk minimization using a surrogate loss function $L^\star(O, \hat{\psi})$, having the property that $E[L^\star(O, \hat{\psi})] = E[L(X, \hat{\psi})]$. They described an explicit construction of a class of such surrogate loss functions suitable for the observed data O, applicable in incomplete data structures satisfying what has been known as *coarsening at random* following Heitjan and Rubin (1991) and Gill et al. (1997). The construction was based on the *doubly robust mapping* defined in van der Laan and Robins (2002), and the mapping of the full data loss function $L(X, \hat{\psi})$ to the observed data loss function $L^\star(O, \hat{\psi}|g)$ in (9) is a special case of this approach. An observed data loss function $L^\star(O, \hat{\psi})$ of this form will generally depend on nuisance parameters that must themselves be estimated from the data, just as (9) depends on the unknown conditional density g.

The alternative learning paradigms described in this section can also apply to general incomplete data structures, when based on a loss function $L^\star(O, \hat{\psi})$ having the same expectation as $L(X, \hat{\psi})$. Note that the penalized estimator (20) is essentially defined by the class Ψ of candidates and the full data loss function $L(X, \hat{\psi})$. For constructing the observed data analog as in (21), the modus operandi is to simply replace $L(X, \hat{\psi})$ everywhere in the algorithm with $L^\star(O, \hat{\psi}|\hat{g})$. Clearly, the local and additive observed data estimators of the previous section are also essentially based on using full data procedures with weighted versions of the losses $l(Y_A, \hat{\psi})$. After forming $L^\star(O, \hat{\psi}|\hat{g})$ for general incomplete data structures based on van der Laan and Dudoit's mappings, we expect that using this observed data loss function for local, penalized, and additive learning should often be a fairly straightforward task.

5 Discussion

In many areas of statistics, there are well known tradeoffs involved in moving from semiparametric modeling to nonparametric function approximation. When attempting to perform causal inference, nonparametric estimators generally require a larger number of observations for reliable results, can be more burdensome from a computational

standpoint, and cannot always be used in a simple manner to test the null hypothesis of no treatment effect. However, the function approximation approach can compensate for these deficiencies by relying on fewer assumptions, and capturing finer structure of the causal function of interest. Suppose that one hopes to use the fit of the causal parameter $\psi : \mathcal{A} \rightarrow \mathbb{R}$ to intervene on a continuous univariate treatment $A \in \mathcal{A}$. The simulation results of section 3 seem to suggest plotting a smooth fit of $\psi(\cdot)$ may sometimes transmit more information about the causal effect of treatment than could an estimated $\hat{\beta} \in \mathbb{R}^d$ parameterizing a statistical model.

If taking the nonparametric route to causal inference, the present work demonstrates how the loss function replacement methodology of van der Laan and Dudoit can be combined with function approximation paradigms developed in the statistical learning community, to greatly expand the toolbox of available estimators.

References

Fan, J. and Gijbels, I. (1996). Local Polynomial Modeling and Its Applications. Chapman & Hall.

Gill, R.D., van der Laan, M.J., and Robins, J.M. (1997). Coarsening at random: characterizations, conjectures and counter-examples. *Proceedings of the First Seattle Symposium in Biostatistics*, 1995. D.Y. Lin and T.R Fleming (editors), Springer Lecture Notes in Statistics, 255-294. math.uu.nl/people/gill/Preprints/car0.pdf

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall, London.

Hastie, T., Tibshirani, R., and Friedman, J.H. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.

Heitjan, D.F. and Rubin, D.B. (1991). Ignorability and coarse data. *Annals of Statis-*

tics, **19**, 2244-2253.

Robins, J.M. (1997). Marginal Structural Models. Proceedings of the American Statistical Association, Section on Bayesian Statistical Science, 1998, 1-10.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, 65-78.

van der Laan, M.J. and Dudoit, S. (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 130. bepress.com/ucbbiostat/paper130.

van der Laan, M.J. and Robins, J.M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York.

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia.

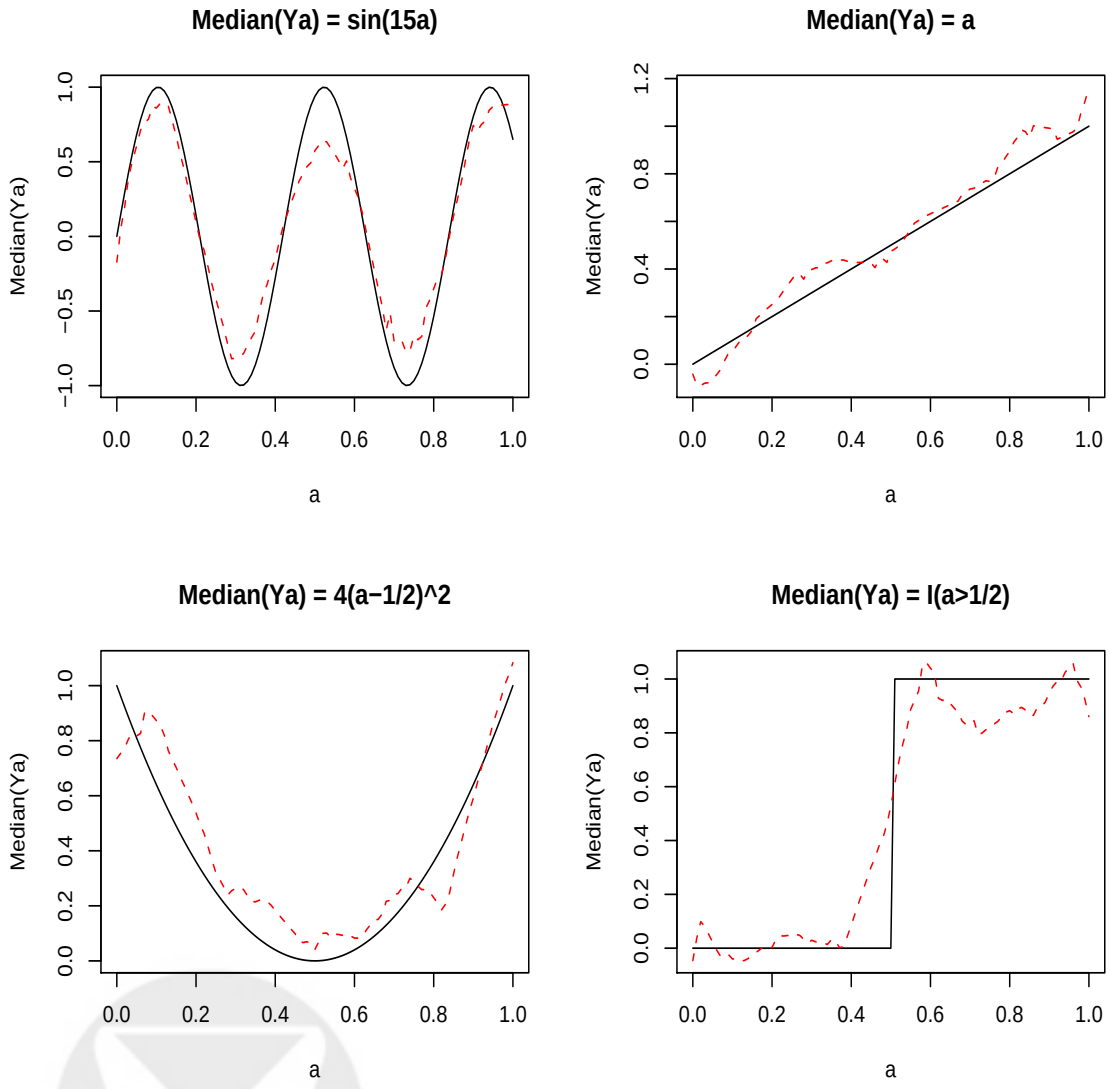

Figure 1: These plots show the results of locally linear median regression using the observed data loss function $L^\star(O, \hat{\psi}|\hat{g})$. The fits are based on $n = 200$ observations, and a Gaussian kernel with a bandwidth of $h = 1/10$. The solid black lines represent the true median functions, and the dashed red lines represent the fitted functions. The local procedure is able to fit curves of varying shapes, as the four plots correspond to the median of counterfactual response Y_a behaving as a linear, oscillating, quadratic, and step function of the treatment $a \in \mathcal{A} = [0, 1]$.

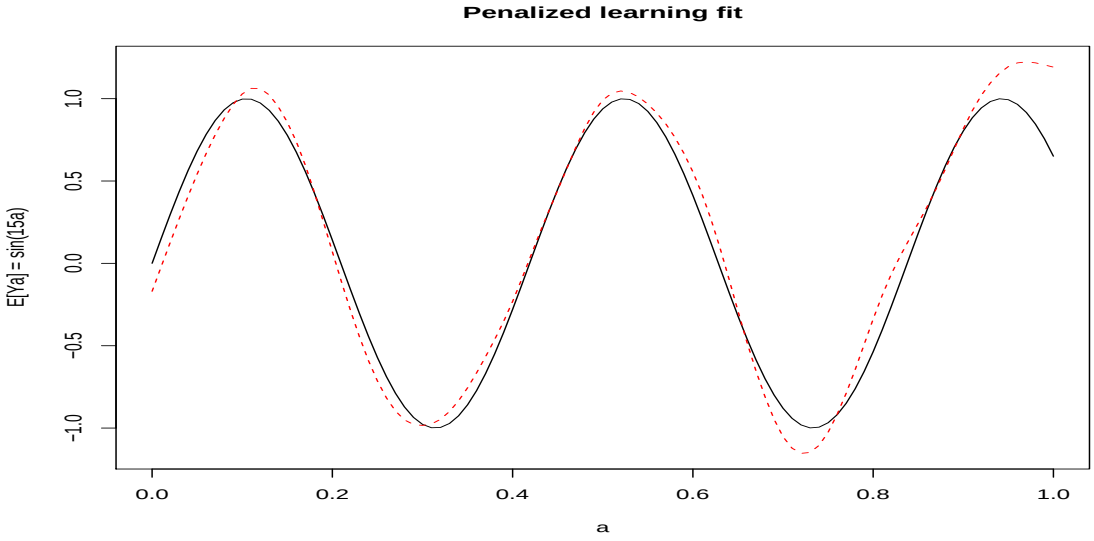

Figure 2: The penalized learning procedure, based on $n = 200$ observations and the observed data loss function $L^\star(O, \hat{\psi}|\hat{g})$, accurately fits the oscillating function $\psi : a \to E[Y_a]$. The solid black line represents the true counterfactual mean function, while the dashed red line represents the fitted function.

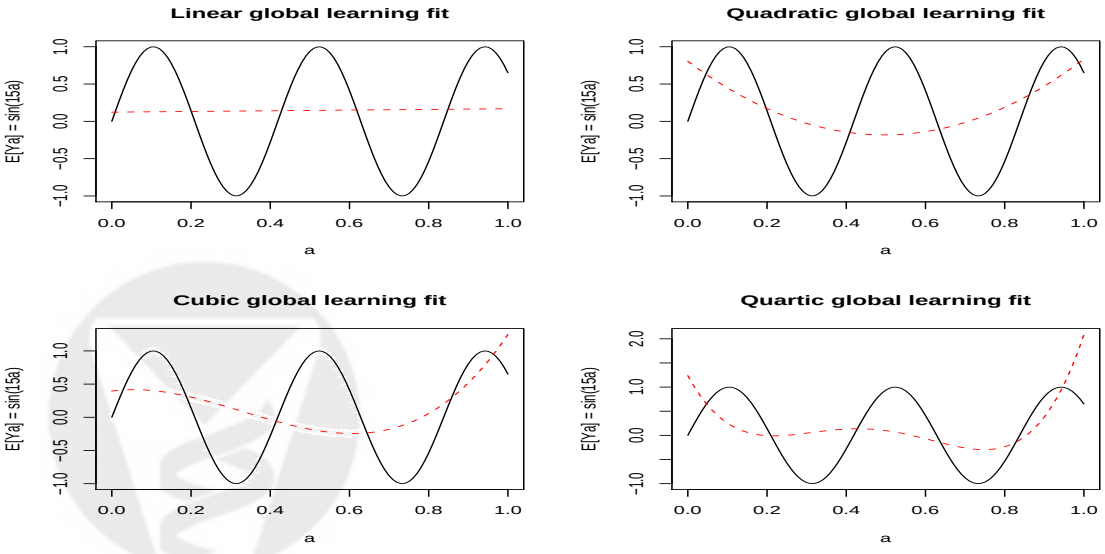

Figure 3: The global modeling procedures, based on $n = 200$ observations and the observed data loss function $L^\star(O, \hat{\psi}|\hat{g})$, fail to accurately fit the oscillating function $\psi : a \to E[Y_a]$. The solid black lines represent the true counterfactual mean functions, while the dashed red lines represent the fitted functions. The four plots represent attempts to globally model the counterfactual mean function with polynomials of higher degree, extending up to degree four.