University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2003 *Paper* 129

A Semiparametric Model Selection Criterion with Applications to the Marginal Structural Model

M. Alan Brookhart* Mark J. van der Laan†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, alanb@stat.berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley

http://biostats.bepress.com/ucbbiostat/paper129

A Semiparametric Model Selection Criterion with Applications to the Marginal Structural Model

M. Alan Brookhart and Mark J. van der Laan

Abstract

Estimators for the parameter of interest in semiparametric models often depend on a guessed model for the nuisance parameter. The choice of the model for the nuisance parameter can affect both the finite sample bias and efficiency of the resulting estimator of the parameter of interest. In this paper we propose a finite sample criterion based on cross validation that can be used to select a nuisance parameter model from a list of candidate models. We show that expected value of this criterion is minimized by the nuisance parameter model that yields the estimator of the parameter of interest with the smallest mean-squared error relative to the expected value of an initial consistent reference estimator. In a simulation study, we examine the performance of this criterion for selecting a model for a treatment mechanism in a marginal structural model (MSM) of point treatment data. For situations where all possible models cannot be evaluated, we outline a forward/backward model selection algorithm based on the cross validation criterion proposed in this paper and show how it can be used to select models for multiple nuisance parameters. We evaluate the performance of this algorithm in a simulation study of the one-step estimator of the parameter of interest in a MSM where models for both a treatment mechanism and a conditional expectation of the response need to be selected. Finally, we apply the forward model selection algorithm to a MSM analysis of the relationship between boiled water use and gastrointestinal illness in HIV positive men.

1 Introduction

The selection of the appropriate statistical model for the data and research question under investigation is a critical step in the practice of data analysis. The literature on model selection is extensive, covering such varied topics as selecting smoothing parameters in density estimation (Silverman, 1986), selecting the order of autoregressive processes in the analysis of time series (Akaike 1969; Parzen 1974) to the selection of covariates to be included in a multivariate linear regression model (Mallows, 1973; Allen, 1971). In this paper, we identify a new model selection problem relating to the estimation of nuisance parameters in semiparametric models and propose a model selection criterion appropriate to this problem based on cross-validation (Stone, 1974).

To understand the problem, consider a semiparametric model where the distribution of the data is modeled with $F_{\eta,\psi}$ where ψ is the parameter of interest and η is a high dimensional nuisance parameter. In this paper, we consider classes of estimating functions of ψ that depend on an estimated value for η. Additionally, we assume that the choice of ψ is completely determined by the specific research question, so that our model selection problem is concerned only with the selection of a model for η.

Due to the curse of dimensionality, a non-parametric estimate of η may not be possible; therefore some modelling assumption for η will be necessary. How η is modeled can affect both the finite sample bias and variance of the estimate of ψ. When the parameter η is orthogonal to ψ (in the sense that their scores are uncorrelated), the asymptotic variance of the estimator $\hat{\psi}(\hat{\eta})$ *decreases* as the dimension of the model for $\hat{\eta}$ gets larger (Robins and Rotnitsky 1992; van der Laan and Robins 2002). This result implies that asymptotically η should be estimated with as large a model as possible. However, in finite samples there is no established theory to guide the construction of the model for η. In general, larger models for η will have less bias, but may suffer from greater finite sample variability. This paper is concerned with the development of a finite sample criterion that can be used to select models of η.

Model selection criteria are often based on either prediction error, for example Allen's PRESS statistic and Mallows's C_p statistic, or on the Kullback-Liebler distance (Kullback and Liebler, 1951), for example Akaike's Information Criterion (Akaike, 1973). Due to the curse of dimensionality, in many semiparametric models the maximum likelihood estimate is either asymptotically inconsistent or has poor finite sample performance (Robins and Ritov,

1997; van der Laan and Robins 2002). For this reason, model selection criteria based directly on the Kullback-Liebler distance are not always feasible in semiparametric models. Additionally, since we are concerned with the estimation of a parameter rather than the prediction of an outcome, it seems natural to base our model selection criterion directly on the estimation error rather than on prediction error.

In this paper, we propose a model selection criterion based on cross-validation that aims to minimize the mean-squared error of $\hat{\psi}$. The development of this criterion assumes that we have available a consistent, but potentially highly variable estimate of the true value of ψ. The goal for our model selection criterion is to select a model from a given a set of candidate models of η, yielding estimators $\{\hat{\eta}_0, \hat{\eta}_1, \ldots, \hat{\eta}_K\}$, where the estimator of the parameter of interest $\hat{\psi}(\hat{\eta}_0)$ is assumed to be consistent and approximately unbiased for ψ although possibly highly variable. We show that the expected value of this criterion is minimized by the estimator $\hat{\psi}(\hat{\eta}_k) \in \{\hat{\psi}(\hat{\eta}_0), \hat{\psi}(\hat{\eta}_1), ..., \hat{\psi}(\hat{\eta}_K)\}$ with the smallest mean-squared error relative to $E[\hat{\psi}(\hat{\eta}_0)]$.

We demonstrate the use of the criterion in the context of the marginal structural model (MSM) for point treatment data of Robins (1998, 2000). The estimators that we consider are semiparametric and depend on the estimation of either one or two large nuisance parameters. In this setting, we examine the performance of model selection criterion and an associated iterative model selection algorithm through simulation studies and a data analysis.

This paper is organized as follows: in section 2, we review the marginal structural model for point treatment data and discuss two different estimators of its parameters, the inverse-probabilty of treatment weighted (IPTW) estimator and the one-step estimator. In section 3, we present our model selection criterion and discuss its characteristics. In section 4, we introduce a forward selection algorithm based on this criterion that can be used to select variables for multiple distinct nuisance parameter models. In section 5, we present a simulation study of the model selection criterion in an MSM for both a discrete and continuous outcome. In section 6, we present a simulation study of the forward selection algorithm for selecting variables to be included in a treatment mechanism and projection term model for the one-step estimator of an MSM parameter. Finally, in section 7 we demonstrate the use of our forward model selection algorithm in an MSM analysis of the causal relationship between boiled water use and gastrointestinal illness in HIV positive men.

2 Marginal structural model for point treatment data

We illustrate the use of the model selection methodology proposed in this paper in the context of the marginal structural model (MSM) of point treatment data (Robins, 1998). The data that we consider for this model are n iid realizations of (W, A, Y) where Y is the outcome, A is the treatment, and W is vector of potential confounders that may be related to both A and Y. The only assumption that we impose on the data is that there are no unmeasured confounders for treatment. This assumption is expressed formally with the following conditional independence statement:

$$A \perp (Y_a, a \in \mathcal{A})|W$$

where $\mathcal{A}$ is the set of all possible treatments and Y_a is the counterfactual outcome a randomly selected subject would have experienced if, possibly contrary to fact, he had been assigned treatment a.

A marginal structural model is a model for the mean of the counterfactual random variable Y_a, i.e.,

$$E[Y_a|V] = m_\psi(a, V)$$

where m is our model for the mean of Y_a parameterized with $\psi \in \mathcal{R}^M$, and $V \subset W$ are covariates on which we may want to condition.

Under the assumption of no unmeasured confounders for treatment, ψ can be consistently estimated by solving the so-called inverse probability of treatment weighted (IPTW) estimating equation:

$$\frac{1}{n}\sum_{i=1}^{n} U(Y_i, A_i, W_i; \psi, g) \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{h(A_i, V_i)\epsilon_i(\psi)}{g(A_i|W_i)} = 0$$

where $\epsilon_i(\psi) = (Y_i - m_\psi(A_i, V_i))$, g is the conditional distribution of A given W, and $h(A, V)$ is a vector function of A and V. This estimating equation follows directly from the identity:

$$E[\frac{h(A, V)\epsilon(\psi)}{g(A|W)}] = 0.$$

Note that for identifiability, we require that $g(a|W)$ is bounded away from zero for all $a \in \mathcal{A}$ or less restrictively that $\max_{a\in\mathcal{A}} |h(a, V)/g(a|W)| < \infty$

almost everywhere. For the purposes of this paper, we take

$$h(A,V) = \frac{\frac{d}{d\psi}m(A,V)g(A|V)}{E[\epsilon(\psi)^2|A,W]}.$$

Solving the estimating equation U for this choice of h is equivalent to performing a weighted regression of Y on A and V using as weights

$$w = \frac{g(A|V)}{g(A|W)}.$$

In either case, since g is typically not known, it is regarded as a nuisance parameter and estimated.

The efficiency and robustness of the IPTW estimator can be improved by subtracting from U its projection onto T_{RA}, where T_{RA} denotes the Hilbert space of scores for all one-dimensional sub-models of g satisfying the assumption of no unmeasured confounders for treatment (Robins 1998). The projection is given by

$$\Pi(U|T_{RA}) = E[\frac{h(A,V)\epsilon(\psi)}{g(A|W)}|A,W] - E[\frac{h(A,V)\epsilon(\psi)}{g(A|W)}|W]$$

where Π denotes the Hilbert space projection operator. This projection term can be written as:

$$\Pi(U|T_{RA}) = \frac{h(A,V)E[\epsilon(\psi)|A,W]}{g(A|W)} - \sum_{a\in\mathcal{A}} h(a,V)E[\epsilon(\psi)|A=a,W]$$

where $E[\epsilon(\psi)|A,W] = E[Y|A,W] - m_\psi(A,V)$. We denote this orthogonalized version of the IPTW estimating function as U^* and write it as:

$$0 = \frac{1}{n}\sum_{i=1}^{n} U^*(Y_i, A_i, W_i; \psi, Q, g) \equiv$$

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{h(A_i,V_i)\epsilon_i(\psi)}{g(A_i|W_i)} - \frac{h(A_i,V_i)E[\epsilon(\psi)|A_i,W_i]}{g(A_i|W_i)} - \sum_{a\in\mathcal{A}} h(a,V_i)E[\epsilon(\psi)|A_i=a,W_i]\right\}.$$

where $Q = E[\epsilon(\psi)|A,W]$. The estimator U^* now depends on two nuisance parameters: g, and Q.

In additional to being more efficient than U, U^* also has the property of being doubly robust (Scharfstein, Rotnitsky, Robins 1999; Robins 2000;

van der Laan and Robins 2002). So, even if the model for g is misspecifed, the estimator remains consistent provided that the model for Q is specified correctly.

It is possible to solve U^* directly, however, given an initial $\sqrt{n}-$consistent estimate of ψ, say $\hat{\psi}$ from the solution of U, we can use the principle of one-step estimation to derive an estimator that is asymptotically equivalent to the solution of U^*. To do this, we perform one iteration of the Newton-Raphson algorithm using $\hat{\psi}$ as our initial estimator. This yields the so-called one-step estimator:

$$\hat{\psi}^* = \hat{\psi} - C^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} U^*(Y_i, A_i, W_i; \hat{\psi}, Q, g) \right\}$$

where C is a p by p matrix of derivatives:

$$C = \frac{1}{n} \sum_{i=1}^{n} \frac{d}{d\psi} U^*(Y_i, A_i, W_i; \psi, Q, g) \Big|_{\psi=\hat{\psi}}.$$

Note that although the one-step estimator is computationally efficient, the double robust property is lost since the consistency of the one-step estimator depends on $\sqrt{n}-$consistency of the initial estimator.

Both the IPTW and the one-step estimators of parameters in the marginal structural model outlined here require the estimation of nuisance parameters. Misspecification of the nuisance parameter models can lead to inconsistent and/or highly variable estimators of the parameter of interest. For the IPTW estimator, the asymptotic variance of $\hat{\psi}$ decreases as the model for g gets larger. In the case of the one-step estimator $\hat{\psi}^*$, when Q is misspecified the asymptotic variance decreases as the model for g get larger. These asymptotic results have led to the suggestion of creating richly specified nuisance parameter models. While this is certainly correct asymptotically, in finite samples large nuisance parameters models may lead to estimators that are considerably more variable than more succinctly specified models. In the following section, we propose a semiparametric model selection criterion based on cross-validation that can be used to help select nuisance parameter models with the goal of balancing the potential variance of large nuisance parameter models with possible bias of smaller models.

3 Model selection methodology

The goal for our model selection criterion is to select a nuisance parameter model from a set of candidate models that minimizes the mean-squared error of the estimator of $\psi \in \mathcal{R}^M$. The development of our criterion is based on the assumption that the estimator $\hat{\psi}(\hat{\eta}_0)$ is a consistent and approximately unbiased, although potentially highly variable estimator of ψ. To denote the dependance of the estimator on both the data and the choice of model, we denote $\hat{\psi}(\hat{\eta}_k)$ as $\hat{\psi}_k(X)$. We assume that our estimators $\hat{\psi}_k(X)$, $k \in \mathcal{K} = \{0, 1, ..., K\}$ are all regular asymptotically linear (RAL) estimators.

For simplicity of exposition, we first consider univariate ψ (i.e., $M = 1$). Let the optimal estimator be referenced by k^*, i.e.,

$$k^* = \text{argmin}_{k \in \mathcal{K}} E_F[(\hat{\psi}_k(X) - \psi)^2],$$

where F is the true distribution of the data.

Our model selection criterion is based on a cross-validation procedure where the data are partitioned into 2 groups V times. Let X_v^0 be data in the training sample and X_v^1 be the data in the validation sample for the v^{th} partition of the data.

For univariate ψ, our criterion function is given by

$$C_V(k) = \frac{1}{V} \sum_{v=1}^{V} (\hat{\psi}_k(X_v^0) - \hat{\psi}_0(X_v^1))^2. \tag{1}$$

Our estimate of the optimal model is found by minimizing $C_V(k)$, i.e.,

$$\hat{k}^* = \text{argmin}_{k \in \mathcal{K}} C_V(k)$$

To see why this is a reasonable criterion, note that

$$E[C_V(k)] = \frac{1}{V} \sum_{v=1}^{V} \Big\{ E[(\hat{\psi}_k(X_v^0) - \psi)^2]$$

$$-2E[(\hat{\psi}_k(X_v^0) - \psi)(\hat{\psi}_0(X_v^1) - \psi)] + E[(\hat{\psi}_0(X_v^1) - \psi)^2] \Big\}.$$

and since X_v^0 is independent of X_v^1 by cross-validation,

$$E[C_V(k)] = \frac{1}{V} \sum_{v=1}^{V} \Big\{ E[(\hat{\psi}_k(X_v^0) - \psi)^2]$$

$$-2E[(\hat{\psi}_k(X_v^0)-\psi)]E[(\hat{\psi}_0(X_v^1)-\psi)]+E[(\hat{\psi}_0(X_v^1)-\psi)^2]\Big\}.$$

If $\hat{\psi}_0(X_v^1)$ is unbiased, then

$$E[C_V(k)]=\frac{1}{V}\sum_{v=1}^{V}\Big\{\text{MSE}(\hat{\psi}_k(X_v^0))+\text{VAR}[\hat{\psi}_0(X_v^1)]\Big\}.$$

Since $\text{VAR}[\hat{\psi}_0(X_v^1)]$ is constant across k, the estimator that yields the smallest mean squared error for the training data minimizes $E[C_V(k)]$.

If $\hat{\psi}_0$ is biased, the cross term $2E[(\hat{\psi}_k(X_v^0)-\psi)]E[(\hat{\psi}(X_v^1,0)-\psi)]$ in $E[C_V(k)]$ is not zero. However, since by assumption $\hat{\psi}_0$ is a consistent RAL estimate of ψ, the bias term $E[\hat{\psi}_0(X_v^1)-\psi]$ is $O(1/n)$. Therefore the cross-term $2E[(\hat{\psi}_k(X_v^0)-\psi)]E[(\hat{\psi}_0(X_v^1)-\psi)]$ will be asymptotically negligible because it will be either $O(1/n^2)$ if $\hat{\psi}$ is consistent, or if $\hat{\psi}_k$ is inconsistent the bias squared term $E[(\hat{\psi}_k-\psi)]^2$ will dominate $E[C_V(k)]$. However, it would be possible to estimate $2E[(\hat{\psi}_k(X_v^0)-\psi)]E[(\hat{\psi}_0(X_v^1)-\psi)]$ via the bootstrap and adjust the criterion function by minimizing instead $C_V(k)+2E_{\hat{F}}[(\hat{\psi}_k(X_v^0)-\psi)]E_{\hat{F}}[(\hat{\psi}_0(X_v^1)-\psi)]$.

For multidimensional ψ, our aim is to minimize the sum of the mean-squared error of ψ with respect to a M by M matrix B:

$$k^*=\text{argmin}_{k\in\mathcal{K}}\Big\{(\hat{\psi}_k(X)-\psi)^TB(\hat{\psi}_k(X)-\psi)\Big\}.$$

The matrix B is arbitrary and could reflect the relative importance of the elements of ψ. One natural choice of B would be, $\hat{\text{VAR}}[\hat{\psi}_k(X)]^{-1}$. For such a choice of B our criterion would correspond to minimizing the mean-squared error of ψ with respect to it's variance-covariance matrix.

Given this aim, we propose to use the following analogous criterion function

$$C_V(k)=\frac{1}{V}\sum_{v=1}^{V}\Big\{(\hat{\psi}_k(X_v^0)-\hat{\psi}_0(X_v^1))^TB(\hat{\psi}_k(X_v^0)-\hat{\psi}_0(X_v^1))\Big\}. \qquad (2)$$

For a fixed nuisance parameter model, asymptotic inference for ψ can be based on the variance-covariance matrix of the influence curve of ψ. However, when the nuisance parameter model is selected in a data adaptive way, as we propose here, this confidence interval may be inaccurate. In order for the variability of the model selection procedure to be reflected in the confidence interval, it would be possible to employ a bootstrap estimation

procedure where for each bootstrap resample, the nuisance parameter model is re-selected before ψ estimated.

An additional issue facing the analyst is how to optimally partition the data. For Monte-Carlo cross-validation, the data are randomly divided in a training and validation data set for each of the V divisions. For V-fold cross-validation, the data are split into V approximately equal sized partitions where for the v^{th} iteration, the validation sample consists of the v^{th} division of the data and the training sample consists of the remaining $v-1$ divisions. Recent theoretical and empirical results have suggested that V-fold cross-validation, while computationally simpler than Monte-Carlo cross-validation, is asymptotically equivalent (van der Laan, Dudoit, Keles 2003).

3.1 Forward/backward model selection of the nuisance parameter model based on cross-validation w.r.t. parameter of interest

In situations where it is not possible to evaluate all candidate nuisance parameter models, it will be necessary to conduct a search in the space of possible models to find the best model selection possible. For this task, we propose a forward/backwards model selection algorithm that can be used to build the nuisance model in a stepwise manner. Consider the situation where we need to select variables for a single nuisance parameter model, for example in the IPTW estimator where $\eta = g(A|W)$. Suppose that the models for g that we consider are regression models defined by the set of covariates we enter in the model. Let the set of p covariates be indexed with $\{1, \ldots, p\}$. For a subset $R \subset \{1, \ldots, p\}$, let β_R be the p-dimensional regression parameter with j-th component set equal to zero for all $j \notin R$. Each subset R defines a lower dimensional multivariate regression model for g. Let $\hat{\psi}_R$ be the corresponding estimator of the parameter of interest ψ. As before, let $\hat{\psi}_0$ be an approximately unbiased estimator of ψ. For example, in the context of the IPTW estimator, $\hat{\psi}_0$ could be given by estimating $g(A|W)$ with a regression of A onto all covariates with a significant marginal association with Y. As before, let X_v^0 be data in the training sample and X_v^1 be the data in the validation sample for the v^{th} partition of the data. We would like to find the model selection R^* that minimizes either the univariate or multivariate

criterion function defined previously, e.g.,

$$C_V(R) = \frac{1}{V}\sum_{v=1}^{V}(\hat{\psi}_R(X_v^0) - \hat{\psi}_0(X_v^1))^2.$$

However, because there are 2^p possible subsets of $\{1, \ldots, p\}$, it is computationally infeasible to evaluate all of possible nuisance parameter models for large p. In this situation, we propose to use a forward/backward selection algorithm for obtaining a good model selection $\hat{R}^*$. This algorithm works like a standard stepwise model selection algorithm, however instead of adding or subtracting variables that improve the predictive accuracy of the model, as measured by a criterion such as AIC, our model selection algorithm selects variable with the aim of improving the criterion with the ultimate goal of minimizing the MSE of the parameter of interest. The algorithm is formally defined as follows:

Initialization of forward selection. Let $k = 0$, $R(0) =$, $R^+(j,k) = R(k) \cup \{j\}$.

Forward selection. We have $R(k)$ is given. Let $j^* = \text{argmin}_{j \notin R(k)} C_V(R^+(j,k))$. If $C_V(R^+(j^*,k)) < C_V(R(k))$, then $k = k+1$, $R(k) = R^+(j^*,k)$ and repeat the previous step. Otherwise we stop with this forward selection and proceed to backward selection with this set $R(k)$ as start.

Backward selection. From the forward selection algorithm we obtain a set $R(k)$. Let $R^-(j,k) = R(k)/\{j\}$. Let $j^* = \text{argmin}_{j \in R(k)} C_V(R^-(j,k))$. If $C_V(R^-(j^*,k)) < C_V(R(k))$, then $k = k-1$, $R(k) = R^-(j^*,k)$ and repeat the previous step. Otherwise, we stop with this backward selection and stop or proceed with forward selection with this set $R(k)$.

Iterate forward/backward selection until convergence. Let $\hat{R}^* = R(k)$ be the final subset resulting from this algorithm.

The algorithm presented here can be generalized to situations where we need to select regression models for two or more nuisance parameters.

4 Simulation study of model selection in an MSM

Through a simulation study, we examine the performance of our proposed criterion for selecting a treatment mechanism model in the IPTW estimator of the parameters in a marginal structural model of point treatment data. In this experiment, we take W to be a 3-dimensional covariate, A to be a dichotomous treatment, and Y to be either a continuous or dichotomous outcome. We consider only simple linear combination of the W, so in this simulation we can examine all eight possible combinations of covariates in the treatment mechanism model.

We generate data using the following laws for the observed data:

- W is multivariate normal with mean 0 and $\Sigma = I$.
- A is conditionally Bernoulli with mean

$$E[A|W] = (1 + \exp\{-(\beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3)\})^{-1}.$$

- The conditional distribution of Y is either Gaussian with mean

$$E[Y|A, W] = \alpha_0 + \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 A$$

and variance σ_y or Bernoulli with

$$E[Y|A, W] = (1 + \exp\{-(\alpha_0 + \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 A)\})^{-1}.$$

Since treatment only depends on W, the simulated data satisfy the assumption of no unmeasured confounders for treatment.

For the simulations where we take Y to be continuous, the parameter of interest is $\psi_{(1)} = E[Y_1] - E[Y_0]$, so we fit the following marginal structural model:

$$m(a, V) = \psi_{(0)} + \psi_{(1)} a.$$

For the simulations where Y is dichotomous, we are interested in the causal log odds-ratio, i.e.,

$$\psi_{(1)} = \log \frac{E[Y_1](1 - E[Y_0])}{E[Y_0](1 - E[Y_1])},$$

so we fit the marginal structural model, where

$$E[Y_a] = m(a, V) = (1 + \exp\{-(\psi_{(0)} + \psi_{(1)} a)\})^{-1}$$

To compute the mean-squared error for the simulations, we need to determine the true value of ψ given the true value of the association parameters α and β. This can be done using the G-computation formula of Robins (1986).

Given this set-up described, we consider three scenarios based on different settings from the parameters of the data generating distribution. We evaluate each of these scenarios for both the continuous and dichotomous outcomes. In all three scenarios, the covariates W_1 and W_2 are related to the outcome, while W_3 is not ($\alpha_1 = \alpha_2 = 1$, $\alpha_3 = 0$). For all scenarios, the treatment has the same effect on the outcome ($\alpha_4 = 1$). The three scenarios differ in how the covariates are related to treatment:

- **Scenario 1** Treatment is completely randomized as it would be in a clinical trial ($\beta_1 = \beta_2 = \beta_3 = 0$). This scenario is depicted graphically in figure 1.
- **Scenario 2** Treatment depends on W_3 ($\beta_3 = 1$), but since this covariate is unrelated to the outcome there is no confounding. This scenario is depicted graphically in figure 2.
- **Scenario 3** Treatment is confounded through W_1 while W_3 still predicts treatment ($\beta_1 = \beta_3 = 1$). This scenario is depicted graphically in figure 3.

Additionally, the marginal variance of W_1, W_2, W_3 is one and the conditional variance of Y is one when Y is continuous, and is the nominal binomial variance when Y is dichotomous.

Simulation Experiment 1. To evaluate the performance of cross-validation under each scenario and for each type of outcome, we simulate 200 data sets for sample sizes $N = 200$ and $N = 1000$. For each data set, we evaluate the estimators corresponding to the eight possible combinations of W_1, W_2, W_3 in the treatment mechanism model and the estimator corresponding to an unweighted estimating equation. The cross-validation procedure evaluates all of these estimators for each data set and selects the one that minimizes 1 where $m = 25$. For these simulations, we take $\hat{\psi}_0$ to be the estimator derived from using a fully specified treatment mechanism model; i.e.,

$$\hat{g}(A|W) = \hat{\beta}_0 + \hat{\beta}_1 W_1 + \hat{\beta}_2 W_2 + \hat{\beta}_3 W_3.$$

The MSE of all estimators and the one selected by the cross-validation procedure are reported in tables 1 and 2. Additionally, we report the percentage of times a particular estimator was selected by cross-validation.

Simulation Experiment 2. To gain insight into how the criterion performs under different specifications of the null model, we repeat the previous experiment simulating data under each of the scenarios specified above for a single sample size ($n = 500$); however we consider two different specifications of the null model. In the Scenarios 1 and 2 where there is no confounding, any specification of the nuisance parameter model in $\hat{\psi}_0$ will yield a consistent estimate of ψ. For Scenario 3, only specifications of $g(A|W)$ that include the confounder W_3 will lead to a consistent estimate of ψ. For Scenarios 1, we compare cross-validation using as $\hat{\psi}_0$ an unweighted regression of Y on A, as might be done in a randomized clinical trial, to a null model based on a treatment mechanism model using both W_1 and W_2. For both Scenarios 2 and 3, we compare cross-validation based on a $\hat{\psi}_0$ using an estimate of g based on to a model using all covariates W_1, W_2, W_3 to one using only W_1, W_2. The results for continuous and dichotomous outcomes, respectively, are displayed in tables 3 and 4.

Under all three scenarios, the most efficient choice for the model of $g(A|W)$ is the one using both W_1 and W_2, the covariates related to the outcome. In all cases, adding W_3, the covariates related only to treatment, increases the finite sample variability of our estimator. These results are much more pronounced with the continuous outcome than with the dichotomous outcome. For simulations involving the continuous outcome, the estimator generated by cross-validation is as efficient, or nearly so, as the estimator derived from a treatment mechanism model using only W_1 and W_2. The performance of the model selection criterion for a dichotomous outcome is nearly as efficient as the most efficient choice of the model for g, given the efficient choice for the null model.

Increasing the sample size does not improve the performance of this procedure under either Scenario 1 or 2. However, increasing the sample size increases the relative MSE of the estimator chosen by cross-validation to any inconsistent estimator (i.e., any estimator in Scenario 3 that does not include the confounder in the model for $g(A|W)$). This suggests that increasing sample size will help rule out all inconsistent estimators, but will not necessarily help the criterion select between the consistent estimators.

The second simulation experiment suggests that the performance of the procedure is strongly dependent on the choice for the null model. The criterion is able to select models that are more efficient than the null (if they exist); however these models are selected more frequently when more efficient choices for $\hat{\psi}_0$ are used. This observation suggests that the data analyst

should include in η_0 not just confounders, in order to insure the approximate unbiasedness of $\hat{\psi}_0$, but also covariates related to the outcome in order to increase the efficiency of $\hat{\psi}_0$ and thus the overall performance of the algorithm.

5 Simulation study of forward model selection algorithm

We examine the performance of the forward component of our proposed forward/backward model selection algorithm to select the nuisance parameter models in a one-step estimator of an MSM through a simulation study. In these simulations, we use the same basic data generating set-up outlined in the previous section, although here we only consider continuous outcomes for a single sample size ($n = 500$). We use slightly different settings for the parameters of the observed data generating distributions.

In all three scenarios, the treatment has always the same effect on the outcome ($\alpha_4 = 1$), $\text{VAR}[W_2] = \text{VAR}[W_3] = \text{VAR}[Y|A, W] = 1$ an $\text{VAR}[W_1] = 2$.

- **Scenario 1: (One Confounder, strong confounding)** Treatment depends on W_1 and W_3 ($\beta_1 = \beta_3 = 1, \beta_2 = 0$) while both W_1 and W_2 are related to the outcome ($\alpha_1 = 1$, $\alpha_2 = 1.5$, $\alpha_3 = 0$).

- **Scenario 2: (Two Confounders)** Treatment depends on W_1 and W_3 ($\beta_1 = 1, \beta_3 = 1.2, \beta_2 = 0$) while all covariates are related to the outcome ($\alpha_1 = 1$, $\alpha_2 = 1.5$, $\alpha_3 = 0.5$).

Our model selection simulation considers each of the three components of W as possible covariates in both the model for g and the model for Q. For each data generating scenario, we run the forward model selection cross-validation algorithm based on 25 sample partitions for two different null models: an IPTW estimator and a one-step estimator. The IPTW used as $\hat{\psi}_0$ includes only covariates related to the outcome in the model for g. The one-step estimator used as $\hat{\psi}_0$ includes only the confounders in the treatment mechanism model, but includes all variables related to the outcome in the model for Q. The results for 200 simulated data sets are presented in table 5. We report the mean-squared error of the two estimators chosen by cross-validation as well as five other sample estimators:

- One Step 1: model for g includes W_1, W_2, W_3 model for Q include W_1, W_2, W_3, A.

- One Step 2: model for g includes W_1, model for Q include W_1, W_2,A.
- One Step 3: model for g includes W_1, W_2, model for Q include W_1, W_2,A.
- IPTW 1: model for g includes W_1, W_2, W_3.
- IPTW 2: model for g includes only W_1, W_2.

Consistent with Section 4, these results suggest that, for both scenarios, the performance of the model selection algorithm is strongly dependent on the efficiency of the null model used. In Scenario 1, cross-validation based on the IPTW has approximately three times the MSE of the estimator selected using as a null model a more efficient one-step estimator. For Scenario 2, the relative MSE is slightly larger than three.

6 Data analysis example: Estimating the causal relationship between boiled water use and acute gastrointestinal illness in HIV positive men.

In this section, we employ the forward component of our forward/backward model selection algorithm outlined in Section 3 to fit an MSM to an epidemiological survey data set. We use this algorithm to select both a model for g in an IPTW estimator and a model for g and Q in the one-step estimator.

The data that we use were gathered to estimate the causal effect of boiled water use on the incidence of diarrhea among HIV positive men in San Francisco. The data ($n = 499$) consist of a treatment A (boiled water use), an outcome Y (diarrhea during the past seven days), and 26 additional covariates that are all potential confounders. As potential confounders, these variables may both predict the use of boiled water and be independently related to diarrhea incidence. The potential confounders include factors such as risky sexual activity, ethnicity, presence of pets in the home, and consumption of high risk foods (e.g., shellfish). The data and sample are described in greater detail elsewhere (Eisenberg, et. al., 2002).

We assume that the 26 covariates contain all potential confounders, so that the assumption of no unmeasured confounders for treatment holds. The

parameter of interest in this analysis is the causal odds-ratio due to "boiled water use," so we estimate the following MSM

$$E[Y_a] = \frac{1}{1 - \exp\{\psi_{(0)} + \psi_{(1)}a\}}.$$

With many confounders, we might choose to select only those with marginal relationship with the outcome to constitute the treatment mechanism model in $\hat{\psi}_0$. However, with only 26 total covariates, we take $\hat{\psi}_0$ to be an IPTW estimator based a $\hat{g}$ derived from a logistic regression of A on all 26 covariates. Using this $\hat{\psi}_0$, we performed a forward selection procedure to select a model for g in an IPTW estimator and a separate forward selection procedure was used to select a model for g and Q in a one-step estimator of ψ.

The forward selection procedure arrived at the model given in table 7, for $\hat{g}$ in an IPTW. To understand the type of confounding that might be present, we also included these same variables in a logistic regression model of the outcome. These estimates are given in table 8. Interestingly, the indicator variables, "use of anti-diarrheal medication" and "use of bottled water," both predict the treatment, "use of boiled water," as well as diarrhea incidence itself. We speculated that these variables must be controlling for severity of illness; i.e., the sicker the patients are, the more likely they will employ other measures to try to limit gastrointestinal illness. Additionally, all variables except for "years of education" affect the treatment and the outcome in the same direction; i.e., increases in the levels of the confounders are associated with an increased likelihood of both treatment and diarrhea or a decreased likelihood of both. Therefore, we expected that an IPTW that adjusts for these confounders to yield an estimated treatment effect that is shifted to the left of an unweighted estimate.

Parameter estimates as well as estimated variances derived from 500 bootstrap re-samples of the data are reported in table 6. Kernel density estimates of the bootstrap distribution of $\hat{\psi}_{(1)}$ are depicted in figure 4 for the forward selected one-step estimator, the null model, and an unweighted estimator. Although there does not appear to be strong measured confounding within the data, the unweighted estimator appears to be slightly biased to the right, as expected. The one-step estimator is clearly less variable than $\hat{\psi}_0$ without appearing to have introduced much bias.

The estimated standard errors for the estimators considered are presented in table 6. There is little difference between the estimated standard error of the estimator chosen by forward model selection of g compared with the

forward selected one-step estimator. However, both of these estimators have a variance that is approximately 30% less than the variance $\hat{\psi}_0$, the estimator we might have naively used in an attempt to make the dimension of the model of $\hat{g}$ large and thereby minimize the asymptotic variance of $\hat{\psi}$.

7 Discussion

This paper has presented a general model selection methodology that can be used to select nuisance parameter models for semiparametric estimators of a parameter of interest. In this paper, we have looked at the MSM where the nuisance parameters were the conditional distribution of treatment and the conditional expectation of the response given all available covariates. However, the general approach outlined in the paper has wide applicability beyond the MSM to a broad of class of censored data and causal inference models (van der Laan and Robins, 2002). To illustrate this generality of this method, we present here two other applications where a model for a nuisance parameter needs to be selected: the inverse probability of censoring weighted estimator (IPCW) and a semiparametric missing data model.

The inverse probability of censoring weighted estimator (IPCW) of Robins and Rotnitsky (1992) allows for the estimation a survival function of a right censored random variable when the data satisfy the assumption of coarsening at random (CAR) (Gill, et. al. 1995). The observed data structure in this problem takes the form of n iid copies of $(\bar{W}(\tilde{T}), \tilde{T} = \min(T, C), \Delta = I(T = \tilde{T}))$, where T is a possibly censored outcome, Δ is indicator that T was not censored, $\bar{W}(t)$ is the history of a vector of time varying covariates through time t, and C is a random variable denoting the time of censoring.

The CAR assumption states that the hazard of censoring at time t only depends on the observed covariate process up until time t. Formally,

$$\lambda_C(t|\bar{W}(T), T) = \lambda_C(t|\bar{W}(t^-)).$$

If CAR holds and the survival function of censoring is bounded away from zero, i.e., $\bar{G}(T|\bar{W}(T)) > 0$ almost everywhere, then distribution function of T at time t can be estimated by solving the IPCW estimating function:

$$0 = \hat{F}(t) - \frac{1}{n}\sum_{i=1}^{n} \frac{I(T_i < t)\Delta_i}{\bar{G}(T_i|\bar{W}(T_i))}.$$

for $\hat{F}(t)$.

Since the conditional survival function of the time of censoring is not known in practice, it must be estimated. How this model is selected can affect both the consistency and efficiency of the estimate of $F(t)$. The model selection procedure outlined in this paper can be used to select a model for $\bar{G}(T|\bar{W}(T))$ in an attempt to minimize the mean-squared error of $\hat{F}(t)$.

A similar model selection problem occurs in the semiparametric missing data model of Robins, Rotnitzky, Zhao (1994). In this model, we observe n iid copies of $(Y, W, \Delta, E\Delta)$ where E is an exposure, W is a vector of covariates including surrogates for E, Y is an outcome, and Δ is a random variable indicating whether or not E was observed (i.e., if $\Delta = 0$ then E is missing).

We are interested in estimating the parameters of a model for the conditional mean of Y given E and perhaps some additional covariates V extracted from W. We denote the parameterized model with m_β, i.e.,

$$Y = m_\beta(V, E) + \epsilon,$$

where $E[\epsilon|V, E] = 0$.

Under the assumption that E is missing at random (Rubin, 1976), i.e.,

$$\Pr[\Delta = 1|E, W, Y] = \Pr[\Delta = 1|W, Y]$$

and that $\Pr[\Delta = 1|W, Y]$ is bounded away from 0 almost everywhere, β can be consistently estimated by solving

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{h(E_i, V_i)(Y_i - m_\beta(E_i, V_i))\Delta_i}{Pr(\Delta_i = 1|W_i, Y_i)}.$$

where h is an arbitrary vector function. Again, since $Pr(\Delta = 1|W, Y)$ will not be know in practice, the inverse probability weighted estimating function depends on the nuisance parameter $\hat{Pr}(\Delta = 1|W, Y)$. The efficiency and the consistency of this estimating function depends on how the model for $\hat{Pr}(\Delta = 1|W, Y)$ is chosen. Like the model for the treatment mechanism in an inverse probability of treatment weighted estimator and the model for the hazard of censoring in an inverse probability of censoring weighted estimator, the model for the missing data mechanism can be selected using the cross-validation approach detailed in this paper.

We have shown through simulation studies that the model selection criterion proposed in this paper is frequently able to identify the optimal finite sample estimators of a parameter of interest in a MSM in a mean-squared

error sense. In the simulated data examples considered in this research, *ad hoc* choices of the nuisance parameter model, for example a model including variables marginally related to the outcome, would have resulted in highly efficient estimators. However, for realistic data sets with many covariates exhibiting a complex multivariate relationship with both the outcome and treatment, it is no longer clear that such a simple strategy will continue to work well. In the data analysis presented in this paper, the estimators selected by cross-validation were considerably less variable than ones selected by a naive method, without introducing significant bias.

Since the criterion minimizes the mean-squared error of the estimator based on the training data set, rather than the larger, complete data set, one potential criticism of this approach is that it may tend to select models that are smaller than optimal. A potential solution approach to this problem would be to use the entire sample to fit the nuisance parameter model for both the null and candidate models. These models are then held constant across iterations of the cross-validation procedure which still respects the sample splits for the evaluation of the estimating functions. In future work, we intend to investigate the performance of such a procedure both theoretically and empirically through simulation studies and data analysis.

References

[1] Akaike, H. (1969) Fitting autoregressive models for prediction. Ann. Inst. Statist. Math. 21: 243-247.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Proceedings of the 2^{nd} International Symposium on Information Theory, 267-281.

[3] Allen, D.M. (1971) The prediction sum of squares as a criterion for selecting predictor variables. Technical Report No. 23, Department of Statistics, University of Kentucky.

[4] Bickel, P.J., Klaassen, C.A., J., Ritov, Y., Wellner, J.A. (1993) Efficient and adaptive estimation for semiparametric models. Johns Hopkins University Press.

[5] Eisenberg, J.N., Wade, T.J., Hubbard, A.E., Abrams, D.I., Leiser, R.J., Charles, S., Vu, M., Saha, S., Wright, C.C., Levy, D.A., Jensen, P., Colford Jr., J.M., (2002) Association between water treatment methods and diarrhea in HIV positive individuals. Epidemiology and Infection, 129,315-323.

[6] Gill, R.D.,van der Laan, M.J., Robins, J.M. (1995) Coarsening at Random: Characterizations, conjectures, and counterexamples. Springer Lecture Notes in Statistics.

[7] Hernn, M., Brumback, B., and Robins, J.M. (2000) Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 11, 561-570.

[8] Kullback, S., Liebler, R. A. (1951) On information and sufficiency. Ann. Math. Statist. 22, 79-86.

[9] Mallows, C.L. (1973) Some comments on C_p. Technometrics 15, 661-675.

[10] Robins, J.M. (1986) A new approach to causal inference in mortality studies with sustained exposure periods. Mathematical Modelling 7, 1393-1512.

[11] Robins, J.M. and Rotnitzky, A. (1992) Recovery of information and adjustment for dependent censoring using surrogate markers. AIDS Epidemiology - Methodological Issues, Eds: Jewell N., Dietz K., Farewell V. Boston, MA: Birkhuser, 297-331.

[12] Robins, J.M., Rotnitzky, A., Zhao L.P. (1994) Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89, 846-866.

[13] Robins, J.M. (1994) Correcting for non-compliance in randomized trials using structural nested mean models. Communications in Statistics 23, 2379-2412.

[14] Robins, J.M., Ritov, Y. (1997) Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. Statistics in Medicine 16, 285-319

[15] Robins, J.M., (1997) Marginal structural models. 1997 Proceedings of the American Statistical Association. Section on Bayesian Statistical Science, 1-10.

[16] Robins, J.M. (1999) Marginal structural models versus structural nested models as tools for causal inference. Statistical Models in Epidemiology: The Environment and Clinical Trials. M.E. Halloran and D. Berry, Editors, IMA Volume 116, NY: Springer-Verlag, 95-134.

[17] Robins, J.M., Hernan, M. and, Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11, 550-560.

[18] Robins, J.M. (2000) Robust estimation in sequentially ignorable missing data and causal inference models. Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999, 6-10.

[19] Rubin, D.B. (1976) Inference and missing data. Biometrika 63, 581-592.

[20] Scharfstein, D.O., Rotnitzky, A., and Robins, J.M. (1999) Adjusting for non-ignorable drop-out using semiparametric non-response models. Journal of the American Statistical Association 94, 1096-1120.

[21] Silverman, B. (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall.

[22] Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. J. Roy. Statist. Soc. Ser. B 36, 111-147.

[23] van der Laan, M.J., Robins, J.M. (2002) Unified methods for censored longitudinal data and causality. Springer-Verlag, New York.

[24] van der Laan, M.J., Dudoit, S., Keles, S (2003) Asymptotic optimality of likelihood based cross-validation. Tecnical Report 125, Division of Biostatistics, University of California at Berkeley.

Figure 1: The causal diagram for Scenario 1.

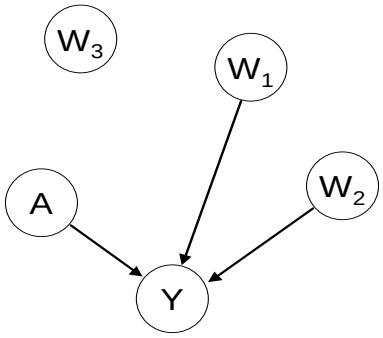

Figure 2: The causal diagram for Scenario 2.

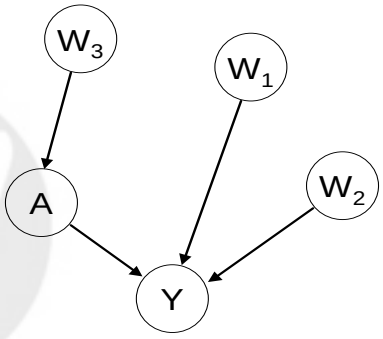

Figure 3: The causal diagram for Scenario 3.

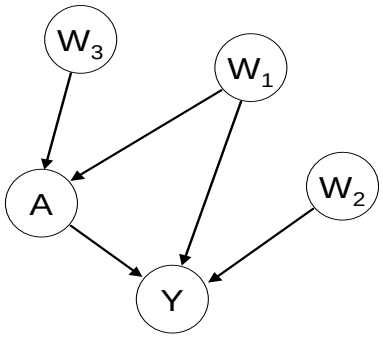

Figure 4: Kernel density estimates of the bootstrap distribution of $\hat{\psi}_0$ (solid line), $\hat{\psi}_k^*$ from a one-step estimator with $\hat{g}$ and $\hat{Q}$ selected by forward selection (dotted line), and an unweighted estimator (dash-dot line). Density estimates based on 500 bootstrap samples.

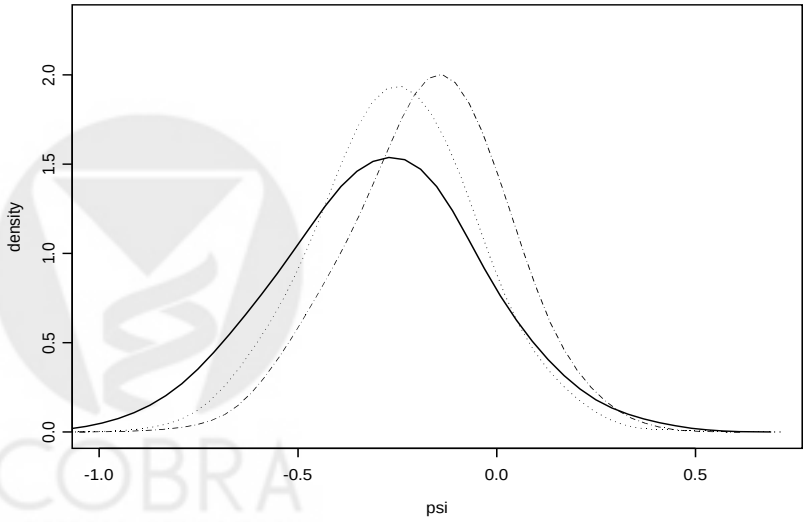

Table 1: Results from 200 simulations of data from a marginal structural model of a continuous outcome for different sample sizes. We compare the MSE of the estimator derived from cross-validation (CV) to all other possible estimators using the three available covariates and an estimator based on an unweighted estimating equation. Also reported is the percentage of times that cross-validation selects the estimator.

Scenario 1: Treatment is completely randomized, but W_1 and W_2 are related to outcome.

	CV	W_1	W_2	W_3	W_1, W_2	W_1, W_3	W_2, W_3	W_1, W_2, W_3	No Wghts.
N=250	0.021	0.046	0.043	0.076	0.020	0.047	0.043	0.020	0.075
		1.5%	5.0%	0.5%	55.5%	1.5%	3.5%	32.5%	0.0%
N=1000	0.0039	0.0092	0.0082	0.0136	0.0039	0.0092	0.0082	0.0039	0.0136
		4.0%	5.5%	1.5%	52.0%	2.5%	1.5%	33.0%	0.0%

Scenario 2: Treatment depends on W_3. Only W_1 and W_2 are related to outcome.

	CV	W_1	W_2	W_3	W_1, W_2	W_1, W_3	W_2, W_3	W_1, W_2, W_3	No Wghts.
N=250	0.025	0.042	0.046	0.083	0.022	0.059	0.054	0.032	0.067
		8.0%	3.5%	0.0%	74.5%	2.0%	0.0%	9.5%	2.5%
N=1000	0.0053	0.0098	0.0067	0.0168	0.0037	0.0137	0.0097	0.0067	0.0139
		8.0%	4.5%	0.0%	71.5%	0.0%	1.0%	13.5%	1.5%

Scenario 3: Treatment depends on W_3 and W_1. W_1 and W_2 are related to outcome.

	CV	W_1	W_2	W_3	W_1, W_2	W_1, W_3	W_2, W_3	W_1, W_2, W_3	No Wghts.
N=250	0.038	0.054	0.588	0.794	0.031	0.089	0.780	0.064	0.594
		7.0%	0.0%	0.0%	71.0%	2.5%	0.0%	19.0%	0.5%
N=1000	0.009	0.011	0.534	0.716	0.006	0.023	0.708	0.017	0.543
		17.0%	0.0%	0.0%	69.5%	0.5%	0.0%	13.0%	0.0%

Table 2: Results from 200 simulations of data from a marginal structural model of a dichotomous outcome for different sample sizes. We compare the MSE of the estimator derived from cross-validation (CV) to all other possible estimators using the three available covariates and an estimator based on an unweighted estimating equation. Also reported is the percentage of times that cross-validation selects the estimator.

Scenario 1: Treatment is completely randomized, but W_1 and W_2 are related to outcome.

	CV	W_1	W_2	W_3	W_1, W_2	W_1, W_3	W_2, W_3	W_1, W_2, W_3	No Wghts.
N=250	0.021	0.046	0.043	0.076	0.020	0.047	0.043	0.020	0.075
		8.5%	7.5%	2.5%	40.0%	7.0%	4.5%	26.0%	4.0%
N=1000	0.012	0.015	0.015	0.019	0.011	0.015	0.015	0.011	0.019
		5.0%	8.5%	3.5%	36.5%	5.5%	10.5%	25.0%	5.5%

Scenario 2: Treatment depends on W_3. Only W_1 and W_2 are related to outcome.

	CV	W_1	W_2	W_3	W_1, W_2	W_1, W_3	W_2, W_3	W_1, W_2, W_3	No Wghts.
N=250	0.079	0.073	0.068	0.118	0.063	0.112	0.099	0.093	0.078
		15.0%	19.5%	2.0%	45.0%	1.5%	2.0%	8.5%	6.5%
N=1000	0.014	0.014	0.014	0.023	0.011	0.020	0.020	0.017	0.016
		19.0%	16.5%	1.0%	47.0%	0.0%	2.0%	7.0%	7.5%

Scenario 3: Treatment depends on W_3 and W_1. W_1 and W_2 are related to outcome.

	CV	W_1	W_2	W_3	W_1, W_2	W_1, W_3	W_2, W_3	W_1, W_2, W_3	No Wghts.
N=250	0.130	0.095	0.401	0.563	0.084	0.143	0.538	0.122	0.417
		23.5%	9.0%	0.0%	52.0%	2.0%	0.5%	6.5%	6.5%
N=1000	0.020	0.017	0.320	0.430	0.015	0.034	0.430	0.031	0.320
		23.5%	0.0%	0.0%	65.5%	0.5%	0.0%	10.0%	0.5%

Table 3: Results from 200 simulations of data from a marginal structural model of a continuous outcome (n=500) for different null models. We compare the MSE of the estimator derived from cross-validation (CV) to all other possible estimators using the three available covariates and an estimator based on an unweighted estimating equation. Also reported is the percentage of times that cross-validation selects the estimator.

Scenario 1: Treatment is completely randomized, but W_1 and W_2 are related to outcome.

Null Model	CV	W_1	W_2	W_3	W_1, W_2	W_1, W_3	W_2, W_3	W_1, W_2, W_3	No Wghts.
no covariates	0.015	0.017	0.016	0.026	0.007	0.017	0.016	0.007	0.026
		9.0%	7.0%	0.0%	43.0%	4.5%	6.5%	30.0%	0.0%
W_1, W_2	0.007	.	.	.	.	.	.	.	.
		2.5%	3.5%	0.0%	50.5%	3.0%	1.0%	39.0%	0.5%

Scenario 2: Treatment depends on W_3. Only W_1 and W_2 are related to outcome.

Null Model	CV	W_1	W_2	W_3	W_1, W_2	W_1, W_3	W_2, W_3	W_1, W_2, W_3	No Wghts.
W_1, W_2, W_3	0.011	0.014	0.016	0.033	0.008	0.023	0.026	0.015	0.022
		8.0%	8.0%	0.0%	68.5%	0.5%	0.5%	12.5%	2.0%
W_1, W_2	0.009	.	.	.	.	.	.	.	.
		2.0%	5.0%	0.0%	82.0%	1.0%	0.5%	7.5%	2.0%

Scenario 3: Treatment depends on W_3 and W_1. W_1 and W_2 are related to outcome.

Null Model	CV	W_1	W_2	W_3	W_1, W_2	W_1, W_3	W_2, W_3	W_1, W_2, W_3	No Wghts.
W_1, W_2, W_3	0.017	0.022	0.531	0.709	0.012	0.038	0.714	0.029	0.526
		14.0%	0.0%	0.0%	68.5%	2.0%	0.0%	15.5%	0.0%
W_1, W_2	0.012	.	.	.	.	.	.	.	.
		4.0%	0.0%	0.0%	85.5%	2.0%	0.0%	8.5%	0.0%

Table 4: Results from 200 simulations of data from a marginal structural model of a dichotomous outcome (n=500) for different null models. We compare the MSE of the estimator derived from cross-validation (CV) to all other possible estimators using the three available covariates and an estimator based on an unweighted estimating equation. Also reported is the percentage of times that cross-validation selects the estimator.

Scenario 1: Treatment is completely randomized, but W_1 and W_2 are related to outcome.

Null Model	CV	W_1	W_2	W_3	W_1, W_2	W_1, W_3	W_2, W_3	W_1, W_2, W_3	No Wghts.
no covariates	0.027	0.028	0.027	0.031	0.023	0.028	0.027	0.023	0.031
		8.5%	9.0%	1.0%	37.5%	5.0%	6.5%	29.5%	3.0%
W_1, W_2	0.024	.	.	.	.	.	.	.	.
		7.5%	9.0%	1.5%	35.5%	7.0%	11.0%	25.5%	3.0%

Scenario 2: Treatment depends on W_3. Only W_1 and W_2 are related to outcome.

Null Model	CV	W_1	W_2	W_3	W_1, W_2	W_1, W_3	W_2, W_3	W_1, W_2, W_3	No Wghts.
W_1, W_2, W_3	0.029	0.030	0.026	0.045	0.024	0.044	0.038	0.038	0.032
		14.5%	20.0%	1.0%	44.5%	1.5%	4.0%	7.0%	7.5%
W_1, W_2	0.024	.	.	.	.	.	.	.	.
		14.5%	8.5%	2.5%	52.0%	5.0%	2.5%	10.5%	4.5%

Scenario 3: Treatment depends on W_3 and W_1. W_1 and W_2 are related to outcome.

Null Model	CV	W_1	W_2	W_3	W_1, W_2	W_1, W_3	W_2, W_3	W_1, W_2, W_3	No Wghts.
W_1, W_2, W_3	0.054	0.046	0.350	0.487	0.035	0.070	0.472	0.055	0.364
		28.0%	2.0%	0.0%	58.0%	3.0%	0.0%	7.0%	2.0%
W_1, W_2	0.036	.	.	.	.	.	.	.	.
		17.0%	0.0%	0.0%	72.0%	3.5%	0.0%	7.5%	0.0%

Table 5: Results from 200 simulated data sets ($N = 500$) from a marginal structural model of a continuous outcome. We report the MSE of the estimator derived from forward selection based on cross-validation (CV) of both the treatment mechanism and the regression model used to estimate the projection term to all five other possible estimators (see text for details). We repeat the simulation for different data generating distribution, one with two confounders and one with three confounders.

	CV (IPTW)	CV (One Step)	One Step 1	One Step 2	One Step 3	IPTW 1	IPTW 2
Scenario 1	0.096	0.028	0.041	0.021	0.027	0.100	0.190
Scenario 2	0.205	0.055	0.046	0.296	0.294	0.431	0.246

Table 6: Results from analysis of effect of boiled water use on diarrhea incidence. Estimates reported from four estimators: 1) an unweighted estimating equation; 2) an IPTW using all covariates in model for treatment mechanism (Full IPTW), 3) an IPTW estimator using $\hat{g}$ selected by forward model selection, 4) an one-step estimator using $\hat{g}$, and $\hat{Q}$ selected by forward model selection. Estimated variances are derived from 500 bootstrap resamples of data.

Estimator	$\hat{\psi}_{(1)}$	$\hat{\text{VAR}}_{\hat{F}}[\hat{\psi}_{(1)}]$
Unweighted	-0.163	0.035
Full IPTW	-0.280	0.060
Forward Selected IPTW	-0.234	0.039
Forward Selected One-Step	-0.254	0.037

Table 7: Treatment mechanism in an IPTW selected by forward model selection.

	Value	Std. Error	t value
(Intercept)	0.337	0.140	2.41
use of anti-diarrheal meds	0.106	0.050	2.11
income	-0.031	0.022	-1.38
currently employed	-0.097	0.055	-1.75
glasses/day water consumed	0.002	0.003	0.78
use of bottled water	0.140	0.039	3.61
years of education	-0.007	0.009	-0.83
age in years	-0.001	0.002	-0.55

Table 8: Variables from forward selected treatment mechanism model used in logistic regression model of outcome.

	Value	Std. Error	t value
(Intercept)	0.226	0.149	1.52
use of anti-diarrheal meds	0.167	0.054	3.10
ordinal measure of income	-0.018	0.024	-0.77
currently employed	-0.027	0.059	-0.46
glasses/day water consumed	0.003	0.003	0.95
use of bottled water	0.046	0.041	1.12
years of education	0.022	0.009	2.42
age in years	-0.003	0.002	-1.41