Computational Statistics and Data Analysis 54 (2010) 3080–3094

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A cross-validation deletion–substitution–addition model selection algorithm: Application to marginal structural models

Thaddeus J. Haight [a,*], Yue Wang [b], Mark J. van der Laan [b,c], Ira B. Tager [a]

[a] Division of Epidemiology, School of Public Health, University of California-Berkeley, United States
[b] Division of Biostatistics, School of Public Health, University of California-Berkeley, United States
[c] Department of Statistics, University of California-Berkeley, United States

ARTICLE INFO

Article history:
Received 28 January 2009
Received in revised form 30 January 2010
Accepted 2 February 2010
Available online 11 February 2010

Keywords:
Cross-validation
Machine learning
Marginal structural models
Lung function
Cardiovascular mortality

ABSTRACT

The cross-validation deletion–substitution–addition (cvDSA) algorithm is based on data-adaptive estimation methodology to select and estimate marginal structural models (MSMs) for point treatment studies as well as models for conditional means where the outcome is continuous or binary. The algorithm builds and selects models based on user-defined criteria for model selection, and utilizes a loss function-based estimation procedure to distinguish between different model fits. In addition, the algorithm selects models based on cross-validation methodology to avoid "over-fitting" data. The cvDSA routine is an R software package available for download. An alternative R-package (DSA) based on the same principles as the cvDSA routine (i.e., cross-validation, loss function), but one that is faster and with additional refinements for selection and estimation of conditional means, is also available for download. Analyses of real and simulated data were conducted to demonstrate the use of these algorithms, and to compare MSMs where the causal effects were assumed (i.e., investigator-defined), with MSMs selected by the cvDSA. The package was used also to select models for the nuisance parameter (treatment) model to estimate the MSM parameters with inverse-probability of treatment weight (IPTW) estimation. Other estimation procedures (i.e., *G*-computation and double robust IPTW) are available also with the package.

1. Introduction

In recent years, epidemiologists' knowledge about the theory and application of marginal structural models (MSMs) to examine causal effects in observational studies has grown substantially. MSMs provide unbiased estimates of marginal effects in the presence of both causal intermediates in point treatment (exposure) studies and time-dependent confounding in longitudinal studies (Robins et al., 2000). Conventional (conditional) association models provide stratum-specific effects which are typically biased in these situations. MSMs eliminate the need to adjust for confounding in the models themselves. Instead, nuisance parameter models (e.g. treatment models) are used to address confounding, so that with MSMs one obtains a direct, unconditional assessment of the exposure on the response. While model selection procedures for nuisance parameters have been addressed in the published literature (Mortimer et al., 2005; Brookhart and van der Laan, 2006), procedures for the selection of MSMs have not. The recent development of a general cross-validated data-adaptive model selection procedure represents an important methodological advancement to better characterize the causal effects of interest through MSM selection and a more flexible examination of the exposure–response causal curve.

* Corresponding address: Division of Epidemiology, School of Public Health, University of California-Berkeley, 519 University Hall, MC:7360, Berkeley, CA 94720, United States. Tel.: +1 510 643 5716; fax: +1 510 643 7316.
E-mail address: tad@stat.berkeley.edu (T.J. Haight).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.02.002

The cross-validation deletion–substitution–addition (DSA) algorithm selects models adaptively for MSMs and nuisance parameter models for point treatment studies (Wang et al., 2004). The approach is derived from a general methodology that provides data-adaptive machine learning type algorithms based on user-supplied criteria (e.g., maximum model size) (van der Laan and Dudoit, 2003; Sinisi and van der Laan, 2004). Specifically, the algorithm builds a model space of candidate models based on so-called deletion, substitution and addition moves and utilizes a loss function-based estimation procedure to distinguish between different models with respect to model fit (van der Laan and Dudoit, 2003). The goal is to select a model that results in the best estimate of a given data distribution. Moreover, the algorithm selects models based partly on V-fold cross-validation (Efron and Tibshirani, 1993; Wang et al., 2004) and, thus, avoids the problem of "over-fitting" data that can occur with other data-adaptive model selection algorithms (e.g., StepAIC function, R-Software, current version, R Foundation for Statistical Computing).

This paper discusses methodological aspects of the algorithm and compares it with other model selection criteria. An illustrative analysis demonstrates how the algorithm works. Two R-packages are available which implement the algorithm: one is a well-developed package (DSA) for the selection of conditional models (e.g., nuisance parameter models); the second is for MSM selection for point treatment studies (cvDSA), and includes components for the selection of nuisance parameter models (cvGLM) and selection of MSMs (cvMSM). The second package (cvDSA) is less developed than the first in terms of ease of use and speed. We advise selection of the treatment model with the DSA package, and submission of this model to the cvMSM procedure for MSM selection. The discussion of the algorithm is in the context of its selection of MSMs, but it provides an overall view of the DSA algorithm as a general tool for model selection. Both packages are available for download from http://stat-www.berkeley.edu/~laan/Software/index.html. Additional background and technical details about the algorithm are available (Dudoit et al., 2003; van der Laan and Dudoit, 2003; Sinisi and van der Laan, 2004; Wang et al., 2004).

2. Background on MSMs

MSMs are used to define causal parameters of interest for exposure–response relations based on the concept of counterfactuals (Robins et al., 2000). This concept permits assessment of observational data in a hypothetical framework in which, contrary to fact, subjects were exposed to all possible levels of an exposure and had outcomes associated with those exposures. With counterfactual data, one can evaluate whether differences in the outcome are attributable to causal differences in the level of the exposure. To recreate the conditions under which observed data can be evaluated as counterfactual data requires several assumptions.

First, the observed data for any given subject represent one realization of his/her counterfactual data that correspond with the exposure actually received (consistency assumption) (Robins, 1999). In a point treatment study, the observed data can be represented as $O = (W, A, Y = Y(A))$, where W represents the baseline covariates, A the treatment (exposure) assignment, and $Y(A)$ the outcome under observed treatment A. The observed data $O = (W, A, Y)$ on a randomly sampled subject represent one realization/component of the counterfactual "full" data $X = ((Y(a), a \in \mathbb{A}), W)$ when exposure $a = A$.

A second assumption is the no "unmeasured confounders", or "randomization assumption": $Y(a) \perp A|W$ — i.e., the treatment of interest is "randomized" with respect to the outcome within strata of the measured covariates, W (Robins, 1999). To satisfy this assumption, one conditions on all the measurable confounders of the exposure and outcome through a nuisance parameter model. Estimation of nuisance parameters can occur either by a model of a regression of the outcome on treatment (exposure) and all potential confounders (W) (G-computation estimation, double robust inverse probability of treatment weight (DR-IPTW) estimation), or a model of the conditional probability of treatment given W (inverse probability of treatment weight (IPTW) estimation). Correct characterization of one of these nuisance parameter models is required to assess properly the effect of treatment on outcome without regard to potential extraneous factors.

Lastly, an additional assumption (experimental treatment assignment, or ETA) is required to provide unbiased estimates with IPTW estimation. This assumption states that all exposures have a positive probability of occurrence, given baseline covariates.

The parameter of interest in an MSM is the treatment-specific mean $E(Y(a)|V)$, possibly conditional on some baseline covariates V that are a subset of $W(V \subset W)$. When $V = W$, the MSM represents a traditional multiple regression model, where the effect of a is a fully adjusted causal parameter. Classical MSMs define a model for $E(Y(a)|V)$ such as a linear model $m(a, V|\beta)$, so that the parameter of interest is the regression parameter β in this assumed model. The goal of the cross-validation DSA algorithm is to achieve a correct characterization (i.e., fit) of the nuisance models and MSMs to evaluate causal effects for point treatment studies.

Additional details of the theory and application of MSMs are available (Robins, 1999; Hernán et al., 2000; van der Laan and Robins, 2002; Yu and van der Laan, 2002; Haight et al., 2003; Neugebauer and van der Laan, 2003; Bryan et al., 2004; Mortimer et al., 2005).

3. Overview of the cross-validation deletion–substitution–addition algorithm

A possible estimator of the treatment-specific mean (MSM) minimizes the empirical risk – a statistical criterion of model fit defined below – over all candidate treatment-specific means. However, since the model space of possible treatment-specific means is infinite dimensional, given the different parameterizations of the treatment variable and the baseline