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

Year 2006 *Paper* 208

Doubly Robust Censoring Unbiased Transformations

Daniel Rubin* Mark J. van der Laan†

*Division of Biostatisitcs, 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

Doubly Robust Censoring Unbiased Transformations

Daniel Rubin and Mark J. van der Laan

Abstract

We consider random design nonparametric regression when the response variable is subject to right censoring. Following the work of Fan and Gijbels (1994), a common approach to this problem is to apply what has been termed a censoring unbiased transformation to the data to obtain surrogate responses, and then enter these surrogate responses with covariate data into standard smoothing algorithms. Existing censoring unbiased transformations generally depend on either the conditional survival function of the response of interest, or that of the censoring variable. We show that a mapping introduced in another statistical context is in fact a censoring unbiased transformation with a beneficial double robustness property, in that it can be used for nonparametric regression if either of these two conditional distributions are estimated accurately. Advantages of using this transformation for smoothing are illustrated in simulations and on the Stanford heart transplant data. Additionally, we discuss how doubly robust censoring unbiased transformations can be utilized for regression with missing data, in causal inference problems, or with current status data

1 Introduction

Random design nonparametric regression is a popular subject of study in the statistical literature, because informally the regression function provides the best prediction of a response given covariates. Nonparametric methods are often required because modern datasets are complicated enough so that any assumed parametric or semiparametric model would almost certainly be misspecified.

When the responses are subject to right censoring, additional complications arise in most smoothing problems. Let $X = (W, Y)$ denote the possibly unavailable covariate and response pair, and suppose our interest is in estimating the regression function $m(w) = E[Y|W = w]$, for the purpose of being able to predict response values at different vector-valued covariate levels. But instead of observing an i.i.d. sample $\{X_i\}_{i=1}^n$, assume that we only can observe each survival time Y_i up to a random censoring time C_i. Formally, consider observing an i.i.d. sample $\{O_i\}_{i=1}^n$, where

$$O = (W,\ \Delta = 1(Y \leq C),\ \tilde{Y} = Y \wedge C),$$

and let $\bar{F}(\cdot|W)$ and $\bar{G}(\cdot|W)$ denote the conditional survival functions of the desired response Y and censoring time C given the covariates W. For convenience, we will assume that the survival and censoring times are continuous random variables, although this is unnecessary.

Throughout this work we will also make the standard assumption that the response and censoring time are conditionally independent given the covariates W, or that

$$\{Y \perp C|W\}. \tag{1}$$

In fact, the regression function $m(W) = E[Y|W]$ is often unidentifiable from such observed data. Consider the case of a censoring time corresponding to a study endpoint, which the true response time Y may sometimes exceed. Nothing can be known about the survival time distribution beyond this endpoint, and hence the regression function will be unidentifiable. For regression to remain a worthwhile endeavor with right censored data, the response must often first be transformed. We can consider truncating the responses at some value τ and estimating the regression function $w \to E[Y \wedge \tau|W = w]$. The response is also often transformed to the log scale, and in this case we would consider the parameter of interest to be the function

$w \to E[\log(Y)|W]$. To simplify notation, we will assume that such transformations have already been incorporated into the response and censoring times Y and C, and continue to let $m(w) = E[Y|W = w]$ denote the desired regression function.

A popular approach to prediction with right censored data is to replace the possibly unavailable responses $\{Y_i\}_{i=1}^n$ with surrogate values $\{Y^\star(O_i)\}_{i=1}^n$ using an appropriate mapping $Y^\star(\cdot)$ of the observed data, and then enter the imputed data $\{W_i, Y^\star(O_i)\}_{i=1}^n$ into standard smoothing algorithms. The key requirement is that imputation mapping $Y^\star(\cdot)$ is approximately what Fan and Gijbels (1996) term a "censoring unbiased transformation," meaning that

$$E[Y^\star(O)|W] = E[Y|W] = m(W). \tag{2}$$

The motivation behind such a requirement is that adaptive smoothing techniques would ideally still be able to estimate the regression function with imputed response data under (2), due to $Y^\star(O)$ having the correct conditional mean structure.

Unfortunately, censoring unbiased transformations generally depend on nuisance parameters, and cannot be directly applied to the observed data. As will be discussed in the following section, existing censoring unbiased transformations that have been proposed for right censored data fall into two categories.

1. Transformations depending on the conditional survival function $\bar{F}(\cdot|W)$, or a functional of this conditional distribution. To apply such transformations, one would first have to construct a preliminary estimate for the conditional distribution of the response Y given covariates W.

2. Transformations depending on the censoring mechanism, or the function $\bar{G}(\cdot|W)$. Applying such transformations thus necessitates estimating this censoring mechanism, which is not directly related to the parameter of interest $m(W) = E[Y|W]$.

In this paper we propose using a *doubly robust* censoring unbiased transformation, to be given in section 2.3. While this mapping has been introduced in statistical contexts unrelated to nonparametric regression, we give the new result that it will have the correct conditional mean structure as in (2) if at least one of the two nuisance parameters $\bar{F}(\cdot|W)$ or $\bar{G}(\cdot|W)$ is correctly specified. Hence, the doubly robust mapping gives the data analyst two chances to form a valid censoring unbiased transformation, and this property can be utilized to form enhanced smoothing procedures.

We give an overview of existing censoring unbiased transformations and present the doubly robust transformation in section 2. Advantages of the doubly robust procedure are highlighted in simulations and with the Stanford heart transplant data in sections 3 and 4. Doubly robust censoring unbiased transformations can be utilized for more general types of censored responses than arise in the right censored data structure, and in section 5 we discuss the applicability of such imputation schemes to regression problems with missing responses, in causal inference problems, and with current status data. Proofs of formal statements concerning double robustness are given in Appendix 1, while double robustness in abstract censored data structures satisfying "coarsening at random" is treated in Appendix II.

2 Censoring Unbiased Transformations

Our overview of censoring unbiased transformations in this section is partially adapted from the discussion of Fan and Gijbels (1996).

2.1 The Buckley-James Transformation

One of the earliest censoring unbiased transformations was the Buckley-James (1979) mapping, given by,

$$Y^{\star}(O) = \Delta Y + (1-\Delta) Q_{\bar{F}}(W, C), \tag{3}$$

for

$$Q_{\bar{F}}(w, y) = E[Y|W=w, Y>y] = \frac{1}{\bar{F}(y|W=w)} \int_{y}^{\infty} y dF(y|W=w). \tag{4}$$

This transformation is the best predictor of the original response, in that it minimizes $E|Y^{\star}(O) - Y|^2$ among all censoring unbiased transformations $Y^{\star}(\cdot)$, leading Fan and Gijbels to note that it can be regarded as the "best restoration." The nuisance parameter required to evaluate (3) is the function $Q_{\bar{F}}(W, \cdot)$, which is in turn a functional of the conditional survival function $\bar{F}(\cdot|W)$ associated with the response Y. The original proposal by Buckley and James for estimation of this nuisance parameter depended on strong assumptions, such as the linearity of the true regression function $m(w) = E[Y|W=w]$. Fan and Gijbels (1994, 1996) considered more adaptive estimates of $Q_{\bar{F}}$, that corresponded to local average estimators and locally linear estimators.

2.2 Transformations Depending on the Censoring Mechanism

While the Buckley-James mapping given in (3) depends on $\bar{F}(\cdot|W)$, other censoring unbiased transformations have been proposed that instead depend only the censoring mechanism. For example, Koul et al. (1981) considered the mapping

$$Y^{\star}(O) = \frac{Y\Delta}{\bar{G}(Y|W)}, \tag{5}$$

which has also been termed the *inverse probability of censoring weighted* (IPCW) mapping. To evaluate this transformation, one would first have to estimate the conditional distribution of the censoring time given the covariates. It is frequently the case the censoring time is completely independent of both the response time Y and covariates W, as might be the case if censoring is caused by the end of a study. In this setting, the conditional survival function $\bar{G}(\cdot|W) = \bar{G}(\cdot)$ could be estimated efficiently with the Kaplan-Meier estimator

$$\hat{\bar{G}}(c) = \prod_{\{i:\, \tilde{Y}_i \le c\}} \left(1 - \frac{1}{\#\{j : \tilde{Y}_j \ge \tilde{Y}_i\}}\right)^{1-\Delta_i}. \tag{6}$$

Zheng (1987) studied more general censoring unbiased transformations, which for non-negative and continuous response and monitoring times took the form,

$$Y^{\star}(O) = \int_0^{\tilde{Y}} \frac{1}{\bar{G}(c|W)} dc + \int_0^{\tilde{Y}} \frac{d(W,c)}{\bar{G}(c|W)} d\bar{G}(c|W) + (1-\Delta) d(W,C).$$

Fan and Gijbels (1994) considered this mapping with

$$d(w,c) = \frac{\alpha c}{\bar{G}(c|W=w)},$$

for different choices of α. They noted that the IPCW transformation given in (5) corresponded to $\alpha = -1$, while a transformation given by Leurgans (1987) corresponded to $\alpha = 0$. Fan and Gijbels proposed to instead apply the mapping with the data-dependent choice of

$$\hat{\alpha} = \min_{\{i:\, \Delta_i = 1\}} \frac{\int_0^{Y_i} \{\bar{G}(c|W_i)\}^{-1} dc - Y_i}{Y_i \{\bar{G}(Y_i|W_i)\}^{-1} - \int_0^{Y_i} \{\bar{G}(c|W_i)\}^{-1} dc},$$

after constructing an estimator of $\bar{G}(\cdot|W)$.

2.3 A Doubly Robust Censoring Unbiased Transformation

The censoring unbiased transformations considered in sections 2.1 and 2.2 respectively depend on the nuisance parameters $\bar{F}(\cdot|W)$ and $\bar{G}(\cdot|W)$. Simulations in the following section will show that a poor preliminary estimator for $\bar{F}(\cdot|W)$ can degrade the performance of regression based on the Buckley-James transformation, while a poor preliminary estimator for $\bar{G}(\cdot|W)$ can degrade the performance of regression based on the transformations given in section 2.2.

In fact, it is possible to construct a censoring unbiased transformation $Y^\star(O)$ that will have the correct conditional mean structure if *either* $\bar{F}(\cdot|W)$ and $\bar{G}(\cdot|W)$ is well approximated. This "doubly robust" transformation provides a clear advantage over the existing procedures described in sections 2.1 and 2.2, because with such a transformation one only needs to solve at least one of two function approximation problems. For $Q_{\bar{F}}(\cdot,\cdot)$ defined as in (4), we propose using the censoring unbiased transformation given by,

$$\begin{aligned} Y^\star(O) &= Y^\star_{\bar{F},\bar{G}}(O) \\ &= \frac{Y\Delta}{\bar{G}(Y|W)} + \frac{Q_{\bar{F}}(W,C)(1-\Delta)}{\bar{G}(C|W)} - \int_{-\infty}^{\tilde{Y}} \frac{Q_{\bar{F}}(W,c)}{\bar{G}^2(c|W)} dG(c|W), \end{aligned} \tag{7}$$

possessing the double robustness property formalized in the following theorem. The theorem is proven in Appendix I.

Theorem 1. *Suppose the conditional independence assumption (1) holds, that Y and C are continuous random variables, and that the conditional distribution of $\{C|W\}$ has a conditional density $g(\cdot|W)$. Assume that $Y \leq \tau < \infty$ for some τ and that $\bar{F}_1(\tau|W) = 0$ with probability one for some conditional survival function $\bar{F}_1(\cdot|W)$. Suppose further that $\bar{G}_1(\tau|W) \geq \epsilon > 0$ for some ϵ and conditional survival function $\bar{G}_1(\cdot|W)$, with corresponding conditional density $g_1(\cdot|W)$. Assume that $g_1(\cdot|W = w)$ is absolutely continuous with respect to $g(\cdot|W = w)$ for all w. We will use the convention that $Q_{\bar{F}_1}(w, y)$ is set to zero if $\bar{F}_1(y|W = w) = 0$. Then,*

$$E[Y^\star_{\bar{F}_1,\bar{G}_1}(O)|W] = E[Y|W] \text{ if either } \bar{F}(\cdot|W) = \bar{F}_1(\cdot|W) \text{ or } \bar{G}(\cdot|W) = \bar{G}_1(\cdot|W). \tag{8}$$

The statistical literature concerning double robustness is primarily related to estimation in semiparametric models, and is discussed in great detail in van der Laan and

Robins (2003). In fact, the doubly robust mapping (7) can be seen as a special case of the doubly robust mappings in chapter 3 of this work, where doubly robust mappings are used to construct estimating equations for regular parameters with censored data. Theorem 2.1 of van der Laan and Robins implies the weaker form of (8) that

$$E[Y^{\star}_{\bar{F}_1,\bar{G}_1}(O)] = E[Y] \text{ if } \bar{F}(\cdot|W) = \bar{F}_1(\cdot|W) \text{ or } \bar{G}(\cdot|W) = \bar{G}_1(\cdot|W). \tag{9}$$

Later work by van der Laan and Dudoit (2003) used the property (9) for doubly robust model selection with censored data, and for constructing M-estimates of irregular parameters. The novelty in our work lies in the result that the doubly robust mapping $Y^{\star}_{\bar{F}_1,\bar{G}_1}(O)$ not only has the correct mean if one of $\bar{F}_1(\cdot|W)$ or $\bar{G}_1(\cdot|W)$ is correctly specified as in (9), but also the correct conditional mean given observed covariates as in (2), along with the realization that this property of being a censoring unbiased transformation is precisely what is needed for nonparametric regression.

One can verify that the Buckley James transformation of (3) corresponds to using the function $\bar{G}(\cdot|W) = 1$ in the doubly robust mapping (7). Such a $\bar{G}(\cdot|W)$ gives the interpretation of the censoring time being a point mass at $+\infty$, but the transformation will remain a censoring unbiased transformation if $\bar{F}(\cdot|W)$ is correctly specified. The IPCW mapping of (5) corresponds to using the function $Q_{\bar{F}}(w, c) = 0$ in (7). Such an $\bar{F}(\cdot|W)$ gives the interpretation of the response time being a point mass at $-\infty$, but the mapping will again remain a censoring unbiased transformation if $\bar{G}(\cdot|W)$ is correctly specified.

3 Simulations

We assessed the quality of the doubly robust transformation through simulations, and compared its performance to that of the Buckley-James transformation (3) and the IPCW transformation (5).

Implementing the regression procedures based on these transformations required estimates of the function $Q(w, y) = E[Y|W = w, Y > y]$, the censoring mechanism $\bar{G}(\cdot|W)$, and choosing a smoothing procedure to use with the imputed data $\{W_i, Y^{\star}(O_i)\}_{i=1}^{n}$. For the smoothing procedure, we used the **smooth.spline()** function in the R language, which fit a cubic smoothing spline to the imputed responses. In all simulations we fit the censoring mechanism through the Kaplan-Meier estimator (6), which has been the standard recommendation in the statistical literature related to censoring unbiased transformations.

We fit $Q(\cdot,\cdot)$ through a nearest neighbor estimate that was similar to that proposed by Fan and Gijbels (1994, 1996). We estimated $Q(w, y)$ by taking the mean of the k uncensored responses greater than y, whose covariate value W was closest to w. If less than k such responses were available, we took the average of these responses. If no such responses were available, we estimated $Q(w, y)$ by y itself. Like Fan and Gijbels, we chose the number of nearest neighbors k by implementing the Buckley-James transformation for each $k \leq \frac{n-1}{2}$ to form imputed data $\{W_i, Y_k^\star(O_i)\}_{i=1}^n$, evaluated the squared error leave-one-out cross-validation criterion for the **smooth.spline()** regression fit to this data, and selected the k minimizing this criterion.

Our first set of simulations demonstrated that regression based on the doubly robust censoring unbiased transformation could indeed adapt to the shape of a regression curve, if given sufficient data. For univariate covariates W, errors ϵ, and censoring times C generated independently, we generated $n = 200$ observations O through the following mechanism.

$$\begin{aligned} W &\sim U(0,1) \\ \epsilon &\sim 2(\text{Beta}(4,4) - \frac{1}{2}) \\ C &\sim \text{exponential}(\frac{1}{2}) - 1 \\ Y &= m(W) + \epsilon \\ O &= (W,\ \Delta = 1(Y \leq C),\ \tilde{Y} = Y \wedge C) \end{aligned} \tag{10}$$

We generated such data using four choices for the regression function m, corresponding to linear, quadratic, sigmoidal, and oscillating functions. For such data, the censoring times were indeed independent of the covariates and response times, so we expected the Kaplan-Meier estimator to be a good fit. We also expected no problems with the fit for Q, as nearest neighbor methods typically do not break down with univariate data. For the four choices of regression function $m(W)$, 52%, 44%, 51%, and 52% of the responses were censored. The results are displayed in Figure 1, and show the doubly robust procedure could accurately approximate these four smooth curves.

In a second set of simulations, we compared the doubly robust transformation with the Buckley-James and IPCW transformations. We generated $n = 200$ replicates of O

according to the following mechanism, where the true regression function was simply the identity function $m(W) = W$.

$$\begin{aligned} W &\sim U(0,1) \\ \epsilon &\sim 2(\text{Beta}(4,4) - \frac{1}{2}) \\ \{C|W \le \frac{1}{2}\} &= +\infty \text{ (meaning no censoring)} \\ \{C|W > \frac{1}{2}\} &\sim \text{exponential}(1) - 1 \\ Y &= W + \epsilon \\ O &= (W,\ \Delta = 1(Y \le C),\ \tilde{Y} = Y \wedge C). \end{aligned} \tag{11}$$

The censoring mechanism here depended on the covariates, as censoring never occured if the covariate W did not exceed $\frac{1}{2}$. Hence, the assumptions for the Kaplan-Meier estimate of $\bar{G}(\cdot|W)$ were violated. In this simulation, 38% of the responses were censored. From the results in the top row of Figure 2, we see that the regressions using the Buckley-James and doubly robust transformations accurately fit the regression line, while the IPCW estimator behaved erratically.

In a final set of simulations, we considered covariates not only consisting of the univariate W, but also of a $\{0,1\}$ random variable V. We generated $n = 400$ replicates as follows, where again the regression function $E[Y|W] = m(W) = W$ was simply the identity function.

$$\begin{aligned} W &\sim U(0,1) \\ V &\sim \text{Bernoulli}(\frac{1}{2}) \\ \epsilon &\sim 2(\text{Beta}(4,4) - \frac{1}{2}) \\ \{C|V = 0\} &= +\infty \text{ (meaning no censoring)} \\ \{C|V = 1\} &\sim \text{exponential}(\frac{1}{3}) - 2 \\ Y &= W + 2(V - \frac{1}{2}) + \epsilon \\ O &= (W,\ V,\ \Delta = 1(Y \le C),\ \tilde{Y} = Y \wedge C) \end{aligned} \tag{12}$$

We considered correctly modeling the censoring mechanism, so that we set $\bar{G} = 1$ for all observations with $V = 0$ in the IPCW and doubly robust transformations, while using the Kaplan-Meier estimator for observations with $V = 1$. In practice, one would expect to notice with $n = 400$ data points if censoring never occured at a certain level of a binary covariate, so such a fit might be fairly realistic. However, we did not correctly model the nuisance parameter $Q(w, y) = E[Y|W, Y > y]$, because we fit the function while ignoring the covariate V. We imagine that such an estimate could also be fairly common in practice, because if a univariate smoother was desired for a specific covariate, one might be reluctant to adjust for additional covariates. The problem with ignoring V in the fit of Q was that the conditional independence assumption $\{Y \perp C|W\}$ did not hold, but rather the conditional independence $\{Y \perp C|W, V\}$. This was due to the event $\{V = 1\}$ being associated with both large Y values and small censoring times. Under this censoring mechanism, 32% of the responses were censored. The results displayed in the bottom row of Figure 2 show that the IPCW and doubly robust mappings led to fairly accurate fits of the regression line, while the Buckley-James transformation led to a severe underestimate of this line.

Therefore, simulated data from the mechanisms in (11) and (12) show that a misspecified censoring mechanism $\bar{G}(\cdot|W)$ can degrade the performance of the IPCW transformation, while a misspecified $Q_{\bar{F}}(W, \cdot))$ can degrade the Buckley-James transformation. In colloquial jargon, the Buckley-James and IPCW transformations put all of their eggs in one basket. The doubly robust transformation can be applied whenever either of the Buckley-James or IPCW function approximation problems has been solved, even if the data analyst is not sure which of $\bar{G}(\cdot|W)$ or $Q_{\bar{F}}(W, \cdot)$ has been well-approximated, and is in this sense a superior censoring unbiased transformation.

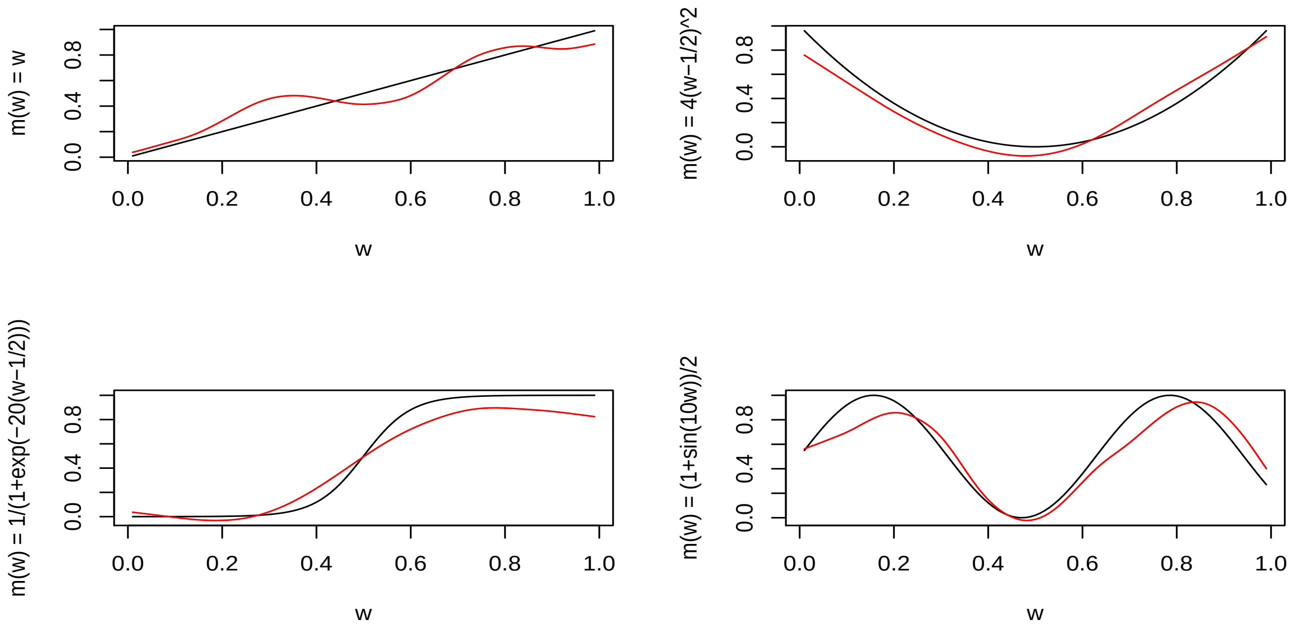

Figure 1: Doubly robust fits of four regression functions, for data generated as in (10). Black lines indicate the regression function, and red lines indicate the fit.

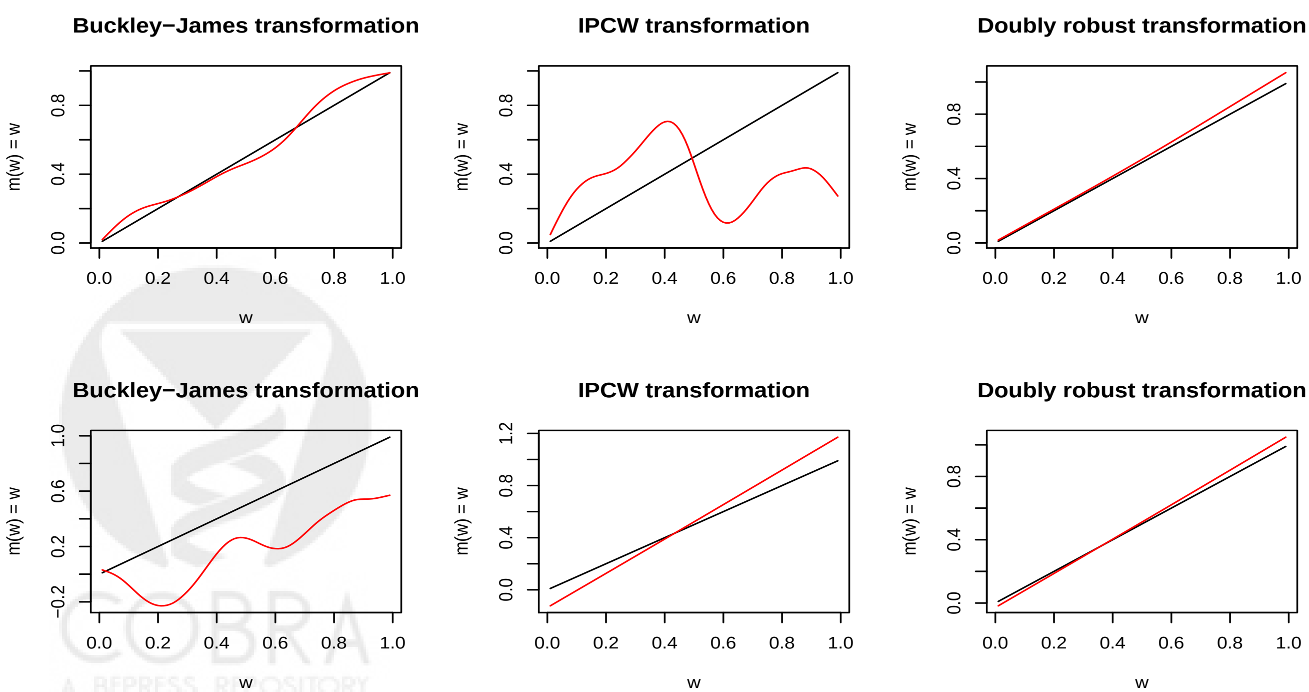

Figure 2: The first row gives the fits for data generated as in (11), and the second row for the data generated as in (12). Black lines indicate the regression function, and red lines indicate the fits.

4 Stanford Heart Transplant Data

We applied censoring unbiased transformations to the Stanford heart transplant data, which has been studied by Miller and Halpern (1982), Doksum and Yandell (1982), and Fan and Gijbels (1994), and is somewhat of a benchmark dataset for right censored regression methods. In this study, patients receiving a heart transplant were followed until either death or a single study endpoint. Among other covariates, the age of each patient at transplantation was recorded, and there appears to have been medical interest in determining how heart transplantation risk was associated with age. For comparison with previous analyses, we considered the dataset to consist of only the 157 patients for which there was information on the tissue type, 55 of whom had censored survival times, and we used the $\log_{10}$(days) time scale for the survival and censoring times.

While Miller and Halpern and Doksum and Yandell considered linear and quadratic fits of $E[\log_{10}(\text{Days of Survival})|\text{Age}]$ based on various regression models, Fan and Gijbels attempted to fit this function through adaptive smoothing. After fitting the Buckley-James nuisance parameter $Q_{\bar{F}}(\text{Age}, \cdot)$ with a local averaging estimator, and then applying a local linear smoother to estimate the regression function from the transformed data, Fan and Gijbels concluded that their fitted curve

> ...reflects the fact that for earlier age, the log-survival time is nearly independent of age, but at later age it decreases linearly with aging.

From the smoothed curve, they suggested the relationship

$$E[\log_{10}(\text{Days of survival})|\text{Age in years}] = 2.74 - 0.078(\text{Age} - 48)_+, \qquad (13)$$

which is shown in the top left panel of Figure 3. Commenting on the utility of smoothing methods for censored data, in comparison to the linear and quadratic fits that had been implemented previously for the Stanford heart data, Fan and Gijbels concluded about (13) that

> ...such a relation appears to be new. The result supports our intuition and moreover, gives a deeper insight into the heart transplantation risk at various ages. In comparison with previous studies by, for example, Miller and Halpern (1982) and Doksum and Yandell (1982), our analysis gives a more precise description of the data structure.

In analyzing this heart transplantation data, we considered slightly modifying the parameter of interest to the function

$$E[\log_{10}(\text{Days of survival}) \wedge \tau | \text{Age in years}], \tag{14}$$

for $\tau = 3.26$. As a practical matter, we simply truncated the 21 values of $\log_{10}(\tilde{Y})$ exceeding τ to τ, set the censoring indicator Δ to one for these observations, and then attempted to estimate the regression function as if the original observations had been this transformed data. Our motivation was the fact that a regression function with right censored data can only be estimated when the response time is sufficiently small so that given any covariate values, the response has a nontrivial chance of being uncensored. The Kaplan-Meier fit for the censoring time distribution suggested this did not hold, as the fit gave extremely small values of $\hat{\bar{G}}(\tilde{Y}_i)$ for some observed data points, and zero for one data point. The level $\tau = 3.26$ in (14) corresponded to truncating survival at the five year mark, and the refit Kaplan-Meier curve gave values of $\hat{\bar{G}}(\tilde{Y}_i)$ no smaller than 0.40 for all observed data points. Truncation as in (14) is a useful tactic in many applied regression problems with right censored data, because it allows us to handle identifiability problems, while retaining an interpretable parameter of interest.

We first estimated the regression function (14) with the Buckley-James transformation, estimating the nuisance parameter $Q_{\bar{F}}(\text{Age}, \cdot)$ with the nearest neighbor method described in the previous section. The cross-validation method previously described selected $k = 6$ nearest neighbors to use for this nuisance parameter estimate. After obtaining the resulting imputed response values, we again used the **smooth.spline()** procedure to estimate the regression function. The resulting curve fit is displayed in the top right panel of Figure 3. Notice that the fit closely resembles the suggested relationship of Fan and Gijbels, in that the curve is roughly constant (very slightly increasing) until between the ages of 40 and 50, when it begins to decrease linearly. Our Buckley-James fit appears slightly smaller than the Fan and Gijbels piecewise linear function, possibly due to our truncation scheme.

We next applied the IPCW censoring unbiased transformation to the Stanford heart transplant data. We used the Kaplan-Meier estimator (6) to fit the censoring mechanism $\bar{G}(\cdot|\text{Age})$, which ignored the age values and fit a marginal survival function. Using the **smooth.spline()** once more with the transformed responses, we obtained an estimate of the regression function (14). The resulting fit is presented in the bottom left panel of Figure 3, and appears very different from Fan and Gijbels' suggested relationship, or our regression fit based on the Buckley-James transformation. In fact, using the IPCW transformation would have led us to the counterintuitive conclusion that a patient's expected log survival time actually slightly increases with age.

Thus, two popular censoring unbiased transformations led to contradictory interpretations of how heart transplantation risk was associated with age. The two transformations respectively depended on accurate estimation of the conditional distributions of the survival and censoring times, given age. On the surface, it does not appear either of these function approximation problems should have been difficult to solve. Nearest neighbor methods are generally reliable in low dimensions, and we had no particular reason to distrust our estimate of $Q(\text{Age}, \cdot)$. Because censoring was caused by the end of the study, domain knowledge also suggested that the censoring time distribution did not depend on the age of the subject, and hence that the Kaplan-Meier estimator of the censoring mechanism should have been reliable. Indeed, one can verify that a Cox model for the censoring distribution does not show any significance for age.

Because the doubly robust mapping was immune to misspecification of one of $Q_{\bar{F}}(\text{Age}, \cdot)$ or $\bar{G}(\cdot|\text{Age})$, it served as a data analytic tool to resolve the inconsistencies stemming from the currently used censoring unbiased transformations. That is, if either the Buckley-James or IPCW fits were accurate, we would have expected the doubly robust fit to also be accurate. Again using the nearest neighbor fit for $Q_{\bar{F}}(\text{Age}, \cdot)$, the Kaplan-Meier estimator $\bar{G}(\cdot|\text{Age})$, and the **smooth.spline()** function with the transformed responses, we fit the doubly robust estimator to the heart transplant data. The resulting curve in the bottom right panel of Figure 3 in fact looks almost identical to the curve based on the Buckley-James transformation. This seems to support the conclusion that $Q_{\bar{F}}(\text{Age}, \cdot)$ and not $\bar{G}(\cdot|\text{Age})$ has been well approximated, and give further credence to the relationship between transplantation risk and age suggested by Fan and Gijbels.

Of course, both $Q_{\bar{F}}(\text{Age}, \cdot)$ and $\bar{G}(\cdot|\text{Age})$ could have been misspecified, and factors other than misspecification of the nuisance parameters in a censoring unbiased transformation can also contribute to inaccurate regression fits. Such factors might include violations of the i.i.d. assumption or the conditional independence assumption (1), or the regression function $m(\text{Age})$ being complex and difficult to estimate even with full and uncensored covariate and response data $\{\text{Age}_i, \text{Survival Time}_i\}_{i=1}^n$.

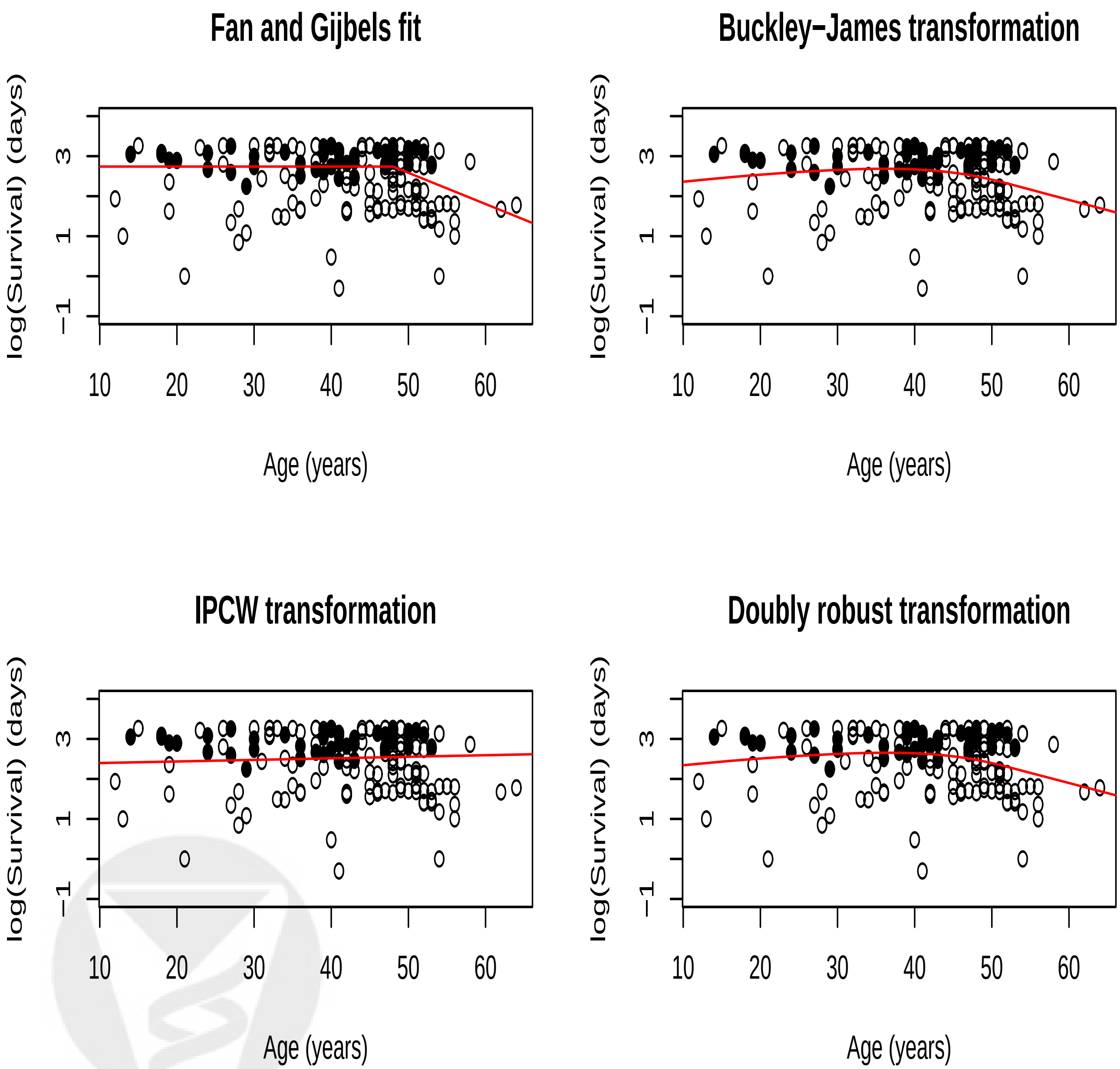

Figure 3: Smoothed Stanford heart transplant data using different censoring unbiased transformations. Open circles represent uncensored observations, while solid circles represent censored observations.

5 Additional Censored Data Structures

Doubly robust procedures can be implemented when the censoring mechanism does not necessarily correspond to right censoring. In this section, we discuss how to form doubly robust mappings when attempting to perform regression with a missing response, in causal inference problems, or with current status data, and highlight the advantages over existing approaches. An abstract treatment of how to construct doubly robust mappings for general censored data structures is differed to Appendix II.

5.1 Regression with a Missing Response

Consider the situation where the possibly unavailable full covariate and response data is given by $X = (W, Y)$ with an interest in prediction of Y from W, but now where the response values can be missing. In this case, the observed data is given by

$$O = (W, \Delta, \Delta Y), \tag{15}$$

for $\Delta \in \{0, 1\}$ an indicator of whether the response is available. Further assume that,

$$\pi(W) \equiv P(\Delta = 1|W) = P(C = 1|X) \tag{16}$$

$$\pi(W) \geq \epsilon \text{ with probability one, for some } \epsilon > 0, \tag{17}$$

so that $\{Y \perp \Delta|W\}$ and the probability of missingness given the covariates W is bounded away from one. Two common approaches to handling missing responses are as follows.

1. Performing a complete case analysis, or fitting the regression function by ignoring the observations with a missing response. This is justified because by (16), $Q(W) \equiv E[Y|W, \Delta = 1] = E[Y|W] = m(W)$. The only loss relative to full data methods is a reduced sample size.

2. Methods based on the propensity scores $\{\pi(W_i)\}_{i=1}^n$. Specifically, we can impute the *inverse probability of missingness weighted* (IPMW) responses $Y^\star(O) = \frac{Y\Delta}{\pi(W)}$ to form a new set of responses $\{Y^\star(O_i)\}_{i=1}^n$ and then apply standard smoothing algorithms. In fact, one can verify that $E[Y^\star(O)|W] = E[Y|W] = m(W)$, making the IPMW mapping a valid imputation target. While the resulting regression fit will be based on all n observations, $\pi(\cdot)$ will have to be estimated. Often the

assumption of missingness *completely at random* is made, meaning that $\pi(W)$ is a constant function of W, which can be estimated by the proportion of $\{\Delta_i\}_{i=1}^n$ equal to one. More generally, π can estimated from binary regression applied to $\{W_i, \Delta_i\}_{i=1}^n$.

Methods 1 and 2 therefore respectively depend on estimating the full data parameter Q from a reduced sample size, or the missingness mechanism π. In fact, we can instead target the doubly robust censoring unbiased transformation

$$\begin{aligned} Y^\star(O) &= Y^\star_{Q,\,\pi}(O) \\ &= \frac{Y\Delta}{\pi(W)} - \frac{\Delta}{\pi(W)}Q(W) + Q(W), \end{aligned} \tag{18}$$

with estimates for both Q and π, having the correct conditional mean if either Q or π is correctly specified. The result is formalized in the following theorem, proven in Appendix I.

Theorem 2. *Let Q_1 be an alternative function of W and π_1 be an alternative conditional distribution of Δ given W. Suppose (16) holds, that (17) holds for π_1, and that Y and $Q_1(W)$ are integrable. Then $E[Y^\star_{Q_1,\,\pi_1}(O)|W] = E[Y|W] = m(W)$ if either $Q(\cdot) = Q_1(\cdot)$ or $\pi(\cdot) = \pi_1(\cdot)$.*

Consequently, the doubly robust imputation scheme can be thought of as a way to combine two natural approaches to handling missing data, so that the overall smoothing procedure should perform well if at least one of the two original schemes was successful.

5.2 Causal Inference in a Point Treatment Study

For $\mathcal{C}$ a finite set, let $\{Y_c : c \in \mathcal{C}\}$ denote a set of responses for a subject, and as before let W denote the subject's covariates. Suppose the interest is in estimating the regression function $m(V) = E[Y|V]$ for

$$Y \equiv \sum_{c\,\in\,\mathcal{C}} b_c Y_c \tag{19}$$

a known linear combination of the responses, and V a subset of the covariates W. Such a formulation could allow us to estimate the regression function associated with each of the responses, or with contrasts of these responses. However, suppose that instead of observing the full covariate and multiple response data $X = (W, \{Y_c : c \in \mathcal{C}\})$, we only observe the covariates W and a single random response in $\{Y_c : c \in \mathcal{C}\}$. That is, let C denote a random variable taking values in $\mathcal{C}$, and consider observing i.i.d. replicates of

$$O = (W, C, Y_C).$$

Such a scenario might arise in a medical study where W are baseline covariates collected about each subject, $\mathcal{C}$ represents the finite set of treatment options, C denotes the treatment administered to the subject, and Y_c corresponds to the *counterfactual* outcome that would have been recorded had the subject's treatment been set to level $c \in \mathcal{C}$. If a new patient arrives after the study is completed, and only a subset V of the covariates represented by W can be measured for that patient, one would want to know which treatment to give the patient. Interest might then lie in predicting each potential treatment outcome Y_c, or more generally a linear combination as in (19).

For regression to be possible, we will suppose that $\{Y_c : c \in \mathcal{C}\}$ is conditionally independent of C given W. In the causal inference literature, this is often termed the assumption of *no unmeasured confounding*. Such an assumption cannot be checked from the data, and is often controversial for observational studies. We feel obligated to stress that the regression methodology we will introduce depends on this assumption of no unmeasured confounding, which must be verified in a case by case basis from domain knowledge before blindly applying the imputation procedures.

Following the introduction of *marginal structural models* by Robins (1997), a typical approach to estimation of $m(V) = E[Y|V]$ would rely on a semiparametric model. That is, one would specify a functional form $m(V) = m(V|\beta)$ parameterized by some unknown vector-valued β, and then attempt to estimate β from the observed data. As reviewed in the manuscript of van der Laan and Robins (2003), estimation approaches typically fall into one of the following three categories.

1. Methods dependent on estimating the function

$$Q : (w, c) \to E[Y_C|W = w, C = c]. \tag{20}$$

   We can fit Q from the data because Y_C, W and C are observed for each subject, so we can simply regress Y_C on the (W, C) pair.

2. Methods dependent on correctly specifying the function

$$g : (c, w) \to P(C = c|W = w). \tag{21}$$

This function gives the conditional probabilities of treatment given the baseline covariates, and can be fit through a polychotomous regression on $\{W_i, C_i\}_{i=1}^n$. As in the previous example of performing regression with a missing response, the values $g(C|W)$ are termed *propensity scores* for each subject.

3. Doubly robust methods allowing one to estimate either the full data parameter Q or the propensity scores $g(C|W)$.

Indeed, we can form a doubly robust censoring unbiased transformation,

$$Y^{\star}_{Q,g}(O) = \sum_{c \in \mathcal{C}} b_c \{ \frac{Y_c 1(C=c)}{g(c|W)} - \frac{1(C=c)}{g(c|W)} Q(W,c) + Q(W,c) \}, \tag{22}$$

which is a valid imputation target if either Q or g is correctly specified, as formalized in the following theorem.

Theorem 3. *Let Q_1 be an alternative function of (W,C) and $g_1(\cdot|W)$ an alternative conditional probability mass function for $\{C|W\}$. Suppose that for $b_c \neq 0$, Y_c and $Q_1(W,c)$ are integrable and that $g(c|W)$ is bounded away from zero with probability one. Then under no unmeasured confounding, as discussed previously,*

$$E[Y^{\star}_{Q_1,g_1}(O)|V] = E[Y|V] = m(V)$$

if either $Q(W,\cdot) = Q_1(W,\cdot)$ or $g(\cdot|W) = g_1(\cdot|W)$.

Thus, the nuisance parameters Q and g necessary for performing doubly robust semiparametric estimation of the regression function $m(V)$ are exactly those needed for doubly robust imputation in the nonparametric regression problem. For a sufficiently large sample size n, the estimated $\{Y^{\star}_{Q,g}(O_i)\}_{i=1}^n$ with covariates $\{V_i\}_{i=1}^n$ could be entered into a vast array of available software or "black boxes" designed for the nonparametric regression problem, such as decision trees, neural networks, MARS, etc. Such a procedure would ameliorate the complication of finding a clever semiparametric parameterization for $E[Y|V] = m(V|\beta)$, and could lead to more adaptive methods for drawing causal inferences.

5.3 Prediction with Current Status Data

Once more, let $X = (W, Y)$ denote the possibly unavailable covariate and response data, and consider estimating the regression function $m(W) = E[Y|W]$. Current status data can arise in cross sectional studies, where it is only known whether the survival time Y exceeds a single random monitoring time C. Formally, let $\bar{F}(\cdot|W)$ denote the conditional survival function of Y given covariates W, let $\bar{G}(\cdot|W)$ denote the conditional survival function of the monitoring time C given W, and consider observing

$$O = (W, C, \Delta = 1(Y \leq C),$$

assuming that $\{Y \perp C|W\}$. It is easy to see that without additional assumptions the regression function could be unidentifiable from such a data structure. If C were always a constant value, there would be no method to estimate the regression function from knowledge of whether the survival time exceeded this constant value. Indeed, we will often have to work with an interval truncated survival time for regression to be possible, and consider the parameter of interest to be

$$E[Y'|W] = E[a1(Y < a) + Y1(a \leq Y \leq b) + b1(Y > b)|W].$$

It is our experience that the regression function $w \rightarrow E[Y'|W = w]$ remains a worthwhile object of study in applied problems, so long as $[a, b]$ is a moderately wide interval. To simplify notation, will we assume that $a \leq Y \leq b$ with probability one, which can always be achieved through truncation.

The conditional survival function $\bar{F}(\cdot|W)$ can then be identified by the data generating distribution for O through the relationship,

$$\bar{F}(y|W = w) = P(\Delta = 0|W = w, C = c).$$

This suggests that one could estimate $\bar{F}(\cdot|W)$, and hence the regression function $m(W) = \int_a^b y dF(y|W)$, by fitting a binary regression of $\{\Delta_i\}_{i=1}^n$ on $\{W_i, C_i\}_{i=1}^n$. Special care must be taken to ensure that the resulting conditional survival function corresponds to a proper conditional distribution. In fact, a variety of parametric, semiparametric, and nonparametric estimators have been proposed for current status that operate in this manner. In particular, Shiboski (1998) has suggested an elegant regression model for current status data using a combination of generalized additive modeling and isotonic regression.

However, such direct estimators of $\bar{F}(\cdot|W)$ "ignore" the censoring mechanism $g(\cdot|W)$, as do all estimators based on maximizing a likelihood for the observed data. If this conditional density were known, we could simply impute responses

$$Y^\star(O) = 1(a \le C \le b)\frac{1-\Delta}{g(C|W)} + a, \tag{23}$$

and observe that $E[Y^\star(O)|W] = E[Y|W] = m(W)$, making for a valid imputation target. With current status data, we might expect the study designers to have quite a bit of distributional information concerning the monitoring time C, because presumably they would have a say in when to monitor their subjects. Partial knowledge toward this end might enable us to fit the conditional density $g(\cdot|Z)$ from the $\{Z_i, C_i\}_{i=1}^n$ data. For example, if we guessed that the monitoring time C was independent of the covariates Z, we could estimate g through univariate density estimation.

In fact, a preliminary estimator of the conditional survival function $\bar{F}(\cdot|W)$ can be combined with a preliminary estimator of the monitoring time conditional density $g(\cdot|W)$ to form a doubly robust censoring unbiased transformation, which will be a valid imputation target if at least one of $\bar{F}$ or g is correctly specified. The formal result is given as follows, and proven in Appendix I.

Theorem 4. *Suppose the conditional independence $\{Y \perp C|W\}$, and that $a \le Y \le b$ with probability one. Let $\bar{F}_1$ denote a conditional survival function such that $\bar{F}_1(a|W) = 1$ and $\bar{F}_1(b|W) = 0$ with probability one. Assume that C is a continuous monitoring time with conditional density $g(\cdot|W)$, and let $g_1(\cdot|W)$ be another conditional density bounded away from zero on $[a,b]$. Then for,*

$$\begin{aligned} Y^\star(O) &= Y^\star_{\bar{F}_1,g}(O) \\ &= 1(a \le C \le b)\frac{1-\Delta}{g(C|W)} - 1(a \le C \le b)\frac{\bar{F}_1(C|W)}{g(C|W)} + (\int_a^b \bar{F}_1(y|W)dy + a). \end{aligned}$$

we have that $E[Y^\star_{\bar{F}_1,g}(O)|W] = E[Y|W] = m(W)$ if either $\bar{F}(\cdot|W) = \bar{F}_1(\cdot|W)$ or $g(\cdot|W) = g_1(\cdot|W)$ on $[a,b]$.

6 Concluding Remarks

We have introduced a general strategy for performing nonparametric regression from censored data, with the appealing property of double robustness. We conclude with several pieces of advice for anyone contemplating implementing our method in an actual data analysis.

- The doubly robust mappings given in this work generally depend on inverse weighting by quantities related to the censoring mechanism, such as the factors $\bar{G}(\cdot|W)$, $\pi(W)$, $g(\cdot|W)$, and $g(C|W)$ in the survival analysis, missing data, causal inference, and current status data examples given previously. Anytime we divide by an estimated probability or quantity between zero and one, we must make sure that the estimate is bounded away from zero for the procedure to retain stability.

- For regression to remain a worthwhile task with censored data, the parameter of interest must sometimes be transformed, as discussed in the survival analysis and current status examples.

- The imputation technique given for prediction with right censored data (as well as for the other data structures discussed in section 5) should not necessarily be thought of as an "off the shelf" method when performing regression. Rather, we have presented an attractive censoring unbiased transformation $Y^{\star}_{\bar{F},\bar{G}}(O)$, but one must accurately approximate at least one of the $\bar{F}(\cdot|W)$ and $\bar{G}(\cdot|W)$ components of this doubly robust mapping. Of course, the performance of the imputation method will depend on the performance of the statistician in estimating these two nuisance parameters.

Appendix I: Proofs of Theorems 1-4

We here prove Theorems 1-4, which state the double robustness of the proposed mappings for the right censored, missing response, point treatment, and current status data structures. Although these results can also be derived as corrollaries of Theorem 5 (to be given in Appendix II), where doubly robust mappings are defined for general censored data structures, it is perhaps easier and more illuminating to furnish direct proofs.

Proof of Theorem 1. Recall that we use the convention that $Q(W,c) = \frac{1}{\bar{F}_1(c|W)}\int_c^\infty y dF_1(y|W)$ is zero if $\bar{F}_1(c|W) = 0$. This and the theorem assumptions ensure that the resulting conditional expectations below are well defined and finite. We write,

$$\begin{aligned} Y^\star(O) &= Y^\star_{\bar{F},\bar{G}}(O) \\ &= \frac{Y\Delta}{\bar{G}(Y|W)} + \frac{Q_{\bar{F}}(W,C)(1-\Delta)}{\bar{G}(C|W)} - \int_{-\infty}^{\tilde{Y}} \frac{Q_{\bar{F}}(W,c)}{\bar{G}^2(c|W)} dG(c|W) \\ &= T_1 + T_2 - T_3. \end{aligned}$$

First observe that,

$$\begin{aligned} E[T_1|W] &= E[\frac{Y\Delta}{\bar{G}_1(Y|W)}|W] \\ &= E[E[\frac{Y\Delta}{\bar{G}_1(Y|W)}|W,Y]|W] \\ &= E[\frac{Y}{\bar{G}_1(Y|W)}P(\Delta=1|W,Y)|W] \\ &= E[\frac{Y}{\bar{G}_1(Y|W)}\bar{G}(Y|W)|W] \\ &= \int_{-\infty}^{\tau} y \frac{\bar{G}(y|W)}{\bar{G}_1(y|W)} dF(y|W). \end{aligned} \quad (24)$$

Next note that,

$$
\begin{aligned}
E[T_2|W] &= E[\frac{Q_1(W,C)(1-\Delta)}{\bar{G}_1(C|W)}|W] \\
&= E[E[\frac{Q_1(W,C)(1-\Delta)}{\bar{G}_1(C|W)}|W,C]|W] \\
&= E[\frac{Q_1(W,C)}{\bar{G}_1(C|W)}P(\Delta=0|W,C)|W] \\
&= E[\frac{Q_1(W,C)}{\bar{G}_1(C|W)}\bar{F}(C|W)|W] \\
&= E[\frac{\bar{F}(C|W)}{\bar{F}_1(C|W)}\int_C^\tau ydF_1(y|W)\bar{G}_1^{-1}(C|W)|W] \\
&= \int_{-\infty}^{\infty}\frac{\bar{F}(c|W)}{\bar{F}_1(c|W)}\bar{G}_1^{-1}(c|W)\{\int_c^\tau ydF_1(y|W)\}dG(c|W) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{\frac{\bar{F}(c|W)}{\bar{F}_1(c|W)}\bar{G}_1^{-1}(c|W)1(y<\tau)1(y>c)y\}dF_1(y|W)dG(c|W) \\
&= \int_{-\infty}^{\infty}1(y<\tau)y\{\int_{-\infty}^{\infty}1(c<y)\frac{\bar{F}(c|W)}{\bar{F}_1(c|W)}\bar{G}_1^{-1}(c|W)dG(c|W)\}dF_1(y|W) \\
&= \int_{-\infty}^{\tau}y\{\int_{-\infty}^{y}\frac{\bar{F}(c|W)}{\bar{F}_1(c|W)}\bar{G}_1^{-1}(c|W)dG(c|W)\}dF_1(y|W). \qquad (25)
\end{aligned}
$$

Finally, observe that,

$$\begin{aligned}
E[T_3|W] &= E[\int_{-\infty}^{\min(Y,C)} \frac{Q_1(W,c)}{\bar{G}_1^2(c|W)} dG_1(c|W)|W] \\
&= E[\int_{-\infty}^{\infty} 1(Y>c)1(C>c)\frac{Q_1(W,c)}{\bar{G}_1^2(c|W)} dG_1(c|W)|W] \\
&= \int_{-\infty}^{\infty} P(Y>c, C>c|W)\frac{Q_1(W,c)}{\bar{G}_1^2(c|W)} dG_1(c|W) \\
&= \int_{-\infty}^{\infty} P(Y>c|W)P(C>c|W)\frac{Q_1(W,c)}{\bar{G}_1^2(c|W)} dG_1(c|W) \text{ as } \{Y \perp C|W\} \\
&= \int_{-\infty}^{\infty} \bar{F}(c|W)\bar{G}(c|W)\frac{Q_1(W,c)}{\bar{G}_1^2(c|W)} dG_1(c|W) \\
&= \int_{-\infty}^{\infty} \frac{\bar{G}(c|W)}{\bar{G}_1^2(c|W)}\{\frac{\bar{F}(c|W)}{\bar{F}_1(c|W)}\int_c^{\tau} y dF_1(y|W)\} dG_1(c|W) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{\frac{\bar{G}(c|W)}{\bar{G}_1^2(c|W)}\{\frac{\bar{F}(c|W)}{\bar{F}_1(c|W)} 1(y<\tau)1(y>c)y\} dF_1(y|W) dG_1(c|W) \\
&= \int_{-\infty}^{\infty} 1(y<\tau)y\{\int_{-\infty}^{\infty} 1(y>c)\frac{\bar{G}(c|W)}{\bar{G}_1^2(c|W)}\frac{\bar{F}(c|W)}{\bar{F}_1(c|W)} dG_1(c|W)\} dF_1(y|W) \\
&= \int_{-\infty}^{\tau} y\{\int_{-\infty}^{y} \frac{\bar{F}(c|W)}{\bar{F}_1(c|W)}\frac{\bar{G}(c|W)}{\bar{G}_1^2(c|W)} dG_1(c|W)\} dF_1(y|W) \\
&= \int_{-\infty}^{\tau} y\{\int_{-\infty}^{y} \frac{\bar{F}(c|W)}{\bar{F}_1(c|W)}\frac{\bar{G}(c|W)}{\bar{G}_1^2(c|W)}\frac{dG_1}{dG}(c|W) dG(c|W)\} dF_1(y|W). \qquad (26)
\end{aligned}$$

Further, note from elementary calculus that for $c < \tau$ (so the demoninator is nonzero),

$$\frac{d}{dc}\{\frac{\bar{G}}{\bar{G}_1}(c|W)\} = -\{\frac{1}{\bar{G}_1(c|W)} - \frac{\bar{G}(c|W)}{\bar{G}_1^2(c|W)}\frac{g_1(c|W)}{g(c|W)}\} g(c|W). \qquad (27)$$

Thus, combining (24), (25), (26), and (27), we see that,

$$\begin{aligned}
E[Y^\star(O)|W] &= E[T_1 + T_2 - T_3|W] \\
&= E[T_1|W] + E[T_2|W] - E[T_3|W] \\
&= \int_{-\infty}^{\tau} y \frac{\bar{G}(y|W)}{\bar{G}_1(y|W)} dF(y|W) \\
&+ \int_{-\infty}^{\tau} y\{\int_{-\infty}^{y} \frac{\bar{F}(c|W)}{\bar{F}_1(c|W)} \bar{G}_1^{-1}(c|W) dG(c|W)\} dF_1(y|W) \\
&- \int_{-\infty}^{\tau} y\{\int_{-\infty}^{y} \frac{\bar{F}(c|W)}{\bar{F}_1(c|W)} \frac{\bar{G}(c|W)}{\bar{G}_1^2(c|W)} \frac{dG_1}{dG}(c|W) dG(c|W)\} dF_1(y|W) \\
&= \int_{-\infty}^{\tau} y \frac{\bar{G}(y|W)}{\bar{G}_1(y|W)} dF(y|W) \\
&+ \int_{-\infty}^{\tau} y\{\int_{-\infty}^{y} \frac{\bar{F}(c|W)}{\bar{F}_1(c|W)} [\frac{1}{\bar{G}_1(c|W)} - \frac{\bar{G}(c|W)}{\bar{G}_1^2(c|W)} \frac{dG_1}{dG}(c|W)] dG(c|W)\} dF_1(y|W) \\
&= \int_{-\infty}^{\tau} y \frac{\bar{G}(y|W)}{\bar{G}_1(y|W)} dF(y|W) \\
&- \int_{-\infty}^{\tau} y\{\int_{-\infty}^{y} \frac{\bar{F}(c|W)}{\bar{F}_1(c|W)} [\frac{d}{dc} \frac{\bar{G}(c|W)}{\bar{G}_1(c|W)}] dc\} dF_1(y|W). \qquad (28)
\end{aligned}$$

If $G = G_1$, then $\frac{d}{dc}\frac{\bar{G}(c|W)}{\bar{G}_1(c|W)} = 0$, so the second term in (28) vanishes, and we are left with

$$E[Y^\star(O)|W] = \int_{-\infty}^{\tau} y \frac{\bar{G}(y|W)}{\bar{G}(y|W)} dF(y|W) = \int_{-\infty}^{\tau} y dF(y|W) = E[Y|W] = m(W).$$

If $F = F_1$, then (28) becomes

$$\begin{aligned}
E[Y^\star(O)|W] &= \int_{-\infty}^{\tau} y \frac{\bar{G}(y|W)}{\bar{G}_1(y|W)} dF(y|W) - \int_{-\infty}^{\infty} y\{\int_{-\infty}^{y} \frac{d}{dc} \frac{\bar{G}(c|W)}{\bar{G}_1(c|W)} dc\} dF(y|W) \\
&= \int_{-\infty}^{\tau} y\{\frac{\bar{G}(y|W)}{\bar{G}_1(y|W)} - \int_{-\infty}^{y} \frac{d}{dc} \frac{\bar{G}(c|W)}{\bar{G}_1(c|W)} dc\} dF(y|W) \\
&= \int_{-\infty}^{\tau} y\{\frac{\bar{G}(y|W)}{\bar{G}_1(y|W)} - [\frac{\bar{G}(y|W)}{\bar{G}_1(y|W)} - \frac{\bar{G}(-\infty|W)}{\bar{G}_1(-\infty|W)}]\} dF(y|W) \\
&= \int_{-\infty}^{\tau} y\{\frac{\bar{G}(y|W)}{\bar{G}_1(y|W)} - \frac{\bar{G}(y|W)}{\bar{G}_1(y|W)} + \frac{1}{1}\} dF(y|W) \\
&= \int_{-\infty}^{\tau} y dF(y|W) \\
&= E[Y|W] \\
&= m(W).
\end{aligned}$$

This proves the desired result. □

Proof of Theorem 2. The theorem assumptions on Y, π_1 and Q_1 ensure that the conditional expectations given below are well defined and finite. We write,

$$\begin{aligned} Y^\star(O) &= Y^\star_{Q,\,\pi}(O) \\ &= \frac{Y\Delta}{\pi(W)} - \frac{\Delta}{\pi(W)}Q(W) + Q(W) \\ &= T_1 - T_2 + T_3. \end{aligned} \tag{29}$$

Note that,

$$\begin{aligned} E[T_1|W] &= E[\frac{Y\Delta}{\pi_1(W)}|W] = E[E[\frac{Y\Delta}{\pi_1(W)}|W,Y]|W] = E[\frac{Y}{\pi_1(W)}P(\Delta=1|W,Y)|W] \\ &= E[\frac{Y}{\pi_1(W)}P(\Delta=1|W)|W] \text{ as } \{Y \perp \Delta|W\} \\ &= E[Y\frac{\pi}{\pi_1}(W)|W] = \frac{\pi}{\pi_1}(W)E[Y|W] \\ &= \frac{\pi}{\pi_1}(W)E[Y|W,\Delta=1] \text{ as } \{Y \perp \Delta|W\} \\ &= \frac{\pi}{\pi_1}(W)Q(W). \end{aligned}$$

Additionally,

$$E[T_2|W] = E[\frac{\Delta}{\pi_1(W)}Q_1(W)|W] = \frac{Q_1(W)}{\pi_1(W)}P(\Delta=1|W) = \frac{\pi}{\pi_1}(W)Q_1(W), \tag{30}$$

and,

$$E[T_3|W] = E[Q_1(W)|W] = Q_1(W).$$

Therefore,

$$\begin{aligned} E[Y^\star(O)|W] &= E[T_1 - T_2 + T_3|W] = E[T_1|W] - E[T_2|W] + E[T_3|W] \\ &= \frac{\pi}{\pi_1}(W)(Q(W) - Q_1(W)) + Q_1(W). \end{aligned}$$

From this it is immediate that $E[Y^\star(O)|W] = Q(W) = E[Y|W] = m(W)$ if either $Q = Q_1$ or $\pi = \pi_1$. This complete the proof. □

Proof of Theorem 3. We will sketch the proof, because the double robustness can easily be seen to follow from the double robustness in the missing response problem given in Theorem 2. Note that for any fixed $c \in \mathcal{C}$ such that $b_c \neq 0$, we could have further censored the observed data $O = (W, C, Y_C)$ into $O = (W, 1(C=c), 1(C=c)Y_c)$. If the goal were to estimate $E[Y_c|W]$, this would be exactly the missing response problem considered previously, with $1(C=c)$ and $g(c|W)$ playing the roles of Δ and $\pi(W)$. Note that the doubly robust mapping would then be $\frac{Y_c 1(C=c)}{g(c|W)} - \frac{1(C=c)}{g(c|W)} Q_1(W,c) + Q_1(W,c)$, and it follows from Theorem 2 for missing data problem that this would have conditional mean equal to $E[Y_c|W]$ if either $Q=Q_1$ or $g=g_1$, under the theorem assumptions. As each term of the linear combination comprising $Y^\star(O)$ thus has the "correct" conditional mean given W if $Q=Q_1$ or $g=g_1$, it follows that $E[Y^\star(O)|W]=E[Y|W]$, and that as $\sigma(V)\subset\sigma(W)$, $E[Y^\star(O)|V]=E[E[Y^\star(O)|W]|V]=E[E[Y|W]|V]=E[Y|V]$. This proves the desired result. $\Box$

Proof of Theorem 4. The theorem assumptions on Y and g_1 on $[a,b]$ ensure that the conditional expectations given below are well defined and finite. We write,

$$\begin{aligned} Y^\star(O) &= Y^\star_{\bar{F}_1,g}(O) \\ &= 1(a\le C\le b)\frac{1-\Delta}{g(C|W)} - 1(a\le C\le b)\frac{\bar{F}_1(C|W)}{g(C|W)} + (\int_a^b \bar{F}_1(y|W)dy + a) \\ &= T_1 - T_2 + T_3. \end{aligned}$$

First observe that,

$$\begin{aligned} E[T_1|W] &= E[1(a\le C\le b)\frac{1-\Delta}{g_1(C|W)}|W] = E[1(a\le C\le b)\frac{1(Y>C)}{g_1(C|W)}|W] \\ &= E[E[1(a\le C\le b)\frac{1(Y>C)}{g_1(C|W)}|W,Y]|W] = E[\int_a^b 1(Y>c)\frac{g}{g_1}(c|W)dc|W] \\ &= \int_a^b P(T>c|W)\frac{g}{g_1}(c|W)dc = \int_a^b \bar{F}\frac{g}{g_1}(c|W)dc. \end{aligned}$$

Also,

$$E[T_2|W] = E[1(a\le C\le b)\frac{\bar{F}_1(C|W)}{g_1(C|W)}|W] = \int_a^b \bar{F}_1\frac{g}{g_1}(c|W)dc.$$

Therefore, recalling that $T_3 \equiv \int_a^b \bar{F}_1(c|W)dc + a$, it follows that

$$E[Y^\star(O)|W] = E[T_1|W] - E[T_2|W] + E[T_3|W] = \int_a^b \{\frac{g}{g_1}(\bar{F}-\bar{F}_1)+\bar{F}_1\}(c|W)dc + a.$$

It is immediate that the integrand of the first term is $\bar{F}(c|W)$ if either $\bar{F}=\bar{F}_1$ or $g=g_1$, and hence that $E[Y^\star(O)|W]=\int_a^b \bar{F}(y|W)dy + a$. Recalling from elementary probability that this is equal to $\int_a^b y dF(y|W)$ as $Y\ge a$, we thus have proved the desired double robustness. $\Box$

Appendix II: General Censored Data Structures

We now describe doubly robust mappings for general censored data structures, having the appropriate conditional mean if at least one of the full data distribution or censoring mechanism is correctly specified. There is a specific construction for the double robust mapping based on the data generating distribution provided in van der Laan and Robins (2003), so that no cleverness is required when confronted with forms of censored data not considered in this paper. For instance, regression could be desired for responses that are both right censored and subject to missingness. Once more, what is new here is the robustness result for the conditional mean instead of unconditional mean, and the application of this fact to regression problems. The development in this appendix will be purposefully abstract, with the view that the technique could be applied when wanting to perform regression under virtually any type of censoring.

Consider the triplet of random variables (X, C, O) defined on a probability space $(\Omega, \mathcal{F}, \mu)$, and taking values in $\mathcal{X} \times \mathcal{C} \times \mathcal{O}$. Here X will denote the *full data*, or random variable we would have observed had there been no censoring. In the regression context, this will typically mean that $\sigma(W, Y) \subset \sigma(X)$, for Y a real-valued response, W a vector-valued set of covariates, and $\sigma(\cdot)$ denoting the sigma field generated by a random variable. Our interest will be in forming rules to predict the response Y from the covariates W, and hence in the regression function

$$m : w \to E[Y|W = w].$$

Here C represents a *censoring variable* that determines how much of the full data we can actually observe. The *observed data* is defined by $O \equiv \Phi(X, C)$, for Φ a known measurable mapping of the full data and censoring variable. It is this data structure O that is assumed available to the statistician, based on i.i.d. copies $\{O_i\}_{i=1}^n$. We will assume that the covariates are uncensored and available from the observed data when estimating the desired regression function, meaning that $\sigma(W) \subset \sigma(O)$. Let F denote the distribution of the full data X, and P denote the distribution of the observed data O. Further, we will assume a regular conditional distribution G for the distribution of O given X, and recall that the regular conditional distribution will always exist when (X, O) is defined on a nice measurable space. It is clear that the distribution P is determined by the pair (F, G), so we will write $O \sim P = P_{F,G}$ to denote the distribution of the observed data.

For the regression function to even be identifiable from the distribution P of the observed data O, we will generally need the assumption of *coarsening at random*. This notion was introduced for discrete random variables in Heitjan and Rubin (1991) and generalized in Gill et al. (1997). Our definition in this section is based on the latter reference, to which we refer for a more detailed discussion. For $o \in \mathcal{O}$, we let $\alpha(o)$ denote the restricted support of X implied by O being a coarsening of X. That is, we define

$$\alpha(o) \equiv \{x \in \mathcal{X} : o = \Phi(x, c) \text{ for some } c \in \mathcal{C}\}$$

and assume

$$(x, o) \rightarrow I(x \in \alpha(o)) \text{ is jointly measurable in } (x, o).$$

We then say that the regular conditional distribution $G_{O|X}$ of O given X satisfies coarsening at random if a version of G can be chosen so that for F-almost all $x, x' \in \mathcal{X}$

$$G_{O|X=x}(do) = G_{O|X=x'}(do) \text{ on } \{o : x \in \alpha(o)\} \cap \{o : x' \in \alpha(o)\}.$$

In words, this means that the conditional distribution of the observed data O given the full data value $X = x$ does not depend on the specific $x \in \mathcal{X}$, other than the requirement imposed by O being a coarsening of X. Unfortunately, there is generally never a way to examine the validity of the coarsening at random assumption in any practical problem.

Several further definitions are needed before presenting our imputation method. Consider the Hilbert spaces $L^2(F)$ and $L^2(P_{F,G})$ consisting of all measurable real-valued square integrable functions of X and O respectively, endowed with the inner products

$$< s_1(X), s_2(X) >_{L^2(F)} \quad = \quad E_F[s_1(X)s_2(X)] \tag{31}$$

$$< h_1(O), h_2(O) >_{L^2(P_{F,G})} \quad = \quad E_{F,G}[h_1(O)h_2(O)]. \tag{32}$$

We define the *score operator* $l_{F,G} : L^2(F) \rightarrow L^2(P_{F,G})$ as

$$l_{F,G}(s(X)) = E_{F,G}[s(X)|O].$$

Its adjoint $l_G^T : L^2(P_{F,G}) \rightarrow L^2(F)$ is given by

$$l_G^T(h(O)) = E_G[h(O)|X].$$

Finally, we define the *information operator* $I_{F,G} : L^2(F) \to L^2(F)$ as the composition

$$I_{F,G} = l_G^T \circ l_{F,G}.$$

Recalling that Y is the full data response variable, we will make the *experimental censoring assumption* that there exists a unique (up to null sets) element $I^{-1}(Y) \in L^2(F)$ such that $I \circ I^{-1}(Y) = Y$ almost surely.

Sufficient conditions for this experimental censoring assumption are shown through the proof of Lemma 3.3 in van der Laan (1998). Specifically, if $\|h\|_{L^2(F)} > 0$ implies $\|l_{F,G}(h)\|_{L^2(P_{F,G})} > 0$ then the information operator is one-to-one. If there exists an $\epsilon > 0$ such that $\|l_{F,G}(h)\|_{L^2(P_{F,G})} \geq \epsilon \|h\|_{L^2(F)}$ then the information operator is onto. Hence, these two conditions together imply the experimental censoring assumption, and the inverse of the information operator is given by the Neumann series

$$I_{F,G}^{-1} = \sum_{i=0}^{\infty} (J - I_{F,G})^i,$$

for J the identity mapping. Whenever $\sigma(\Delta X) \subset \sigma(O)$ for $\Delta \in \{0, 1\}$, it follows immediately from this result that the experimental censoring assumption holds if there is an $\epsilon > 0$ such that

$$P_{F,G}(\Delta = 1|X) \geq \epsilon > 0 \text{ a.s.}$$

In words, this is a simple condition to check when the coarsening mechanism allows for the entire full data structure X to be part of the observed data, as was the case for several important examples described previously (such as regression with a right censored or missing response, but not the point treatment or current status data problems). The doubly robust mapping can now be defined as,

$$Y_{F,G}^{\star}(O) \equiv l_{F,G} \circ I_{F,G}^{-1}(Y) \in L^2(P_{F,G}). \tag{33}$$

It generally holds that,

$$E_{F,G}[Y_{F_1,G_1}^{\star}(O)|W] = E_F[Y|W] = m(W) \text{ if either } F = F_1 \text{ or } G = G_1.$$

The result is stated formally in the following theorem.

Theorem 5. *Let $P_{F,G}$ and P_{F_1,G_1} denote two distributions for the observed data O, so that each satisfy coarsening at random and P_{F_1,G_1} satisfies the experimental censoring* *assumption. Suppose that $G_1(\cdot|X)$ satisfies $G(\cdot|X=x) << G_1(\cdot|X=x)$ for F_1-almost all $x \in \mathcal{X}$, so we can define the Radon-Nikodym derivative $\frac{dG}{dG_1}(\cdot|X=x) \in L^2(P_{F_1,G_1})$ for F_1-almost all $x \in \mathcal{X}$.*

Then $E_{F,G}[Y^\star_{F_1,G_1}|W] = E_F[Y|W] \equiv m(W)$ a.s. if either $F = F_1$ or $G = G_1$.

The proof is given below. A weaker unconditional version of this result, showing that $E_{F,G}[Y^\star_{F_1,G_1}] = E_F[Y]$, is given in Theorem 2.1 of van der Laan and Robins (2003).

Proof of Theorem 5. Note that the experimental censoring assumption is only needed to ensure that the doubly robust mapping $Y^\star_{F_1,G_1}(O)$ is well defined as in (33). We first prove the theorem for $G = G_1$. We use the definition that $l_G^T(s(O)) \equiv E_G[s(O)|X]$ for $s(O) \in L^2(P_{F_1,G})$, the definition of the information operator as the adjoint l_G^T composed with the score operator $l_{F_1,G}$, and the definition of $Y^\star_{F_1,G}$ in (33). Recalling that $Y^\star_{F_1,G}$ is a function $Y^\star_{F_1,G_1}(O)$ of the observed data O, we first notice that

$$\begin{aligned} E_{F_1,G}[Y^\star_{F_1,G}|X] &= E_G[Y^\star_{F_1,G}(O)|X] \\ &= l_G^T(Y^\star_{F_1,G}(O)) \\ &= l_G^T \circ l_{F_1,G} \circ I^{-1}_{F_1,G}(Y) \\ &= I_{F_1,G} \circ I^{-1}_{F_1,G}(Y) \\ &= Y. \end{aligned}$$

As $\sigma(W) \subset \sigma(X)$, we conclude by conditioning on the full data X that,

$$\begin{aligned} E_{F,G}[Y^\star_{F_1,G}|W] &= E_{F,G}[E_G[Y^\star_{F_1,G}|X]|W] \\ &= E_{F,G}[Y|W] \\ &= E_F[Y|W] \\ &= m(W). \end{aligned}$$

This completes the proof for the case of $G = G_1$. We now consider the case of $F = F_1$. Define conditional inner products on $L^2(F)$ and $L^2(P_{F,G_1})$ by

$$\begin{aligned} < s_1, s_2 >_{\mathcal{X},\mathcal{W}} &\equiv E_F[s_1(X)s_2(X)|W] \\ < h_1, h_2 >_{\mathcal{O},\mathcal{W}} &\equiv E_{F,G_1}[h_1(O)h_2(O)|W] \end{aligned}$$

Because of the inclusions $\sigma(W) \subset \sigma(X)$ and $\sigma(W) \subset \sigma(O)$, elementary manipulations yield that,

$$\begin{aligned}
< h(O), l_{F,G_1}(s) >_{\mathcal{O},\mathcal{W}} &= E_{F,G_1}[h(O)E_{F,G_1}[s(X)|O]|W] \\
&= E_{F,G_1}[E_{F,G_1}[h(O)s(X)|O,W]|W] \\
&= E_{F,G_1}[h(O)s(X)|W] \\
&= E_{F,G_1}[E_{F,G_1}[s(X)h(O)|X,W]|W] \\
&= E_{F,G_1}[E_{F,G_1}[s(X)h(O)|X]|W] \\
&= E_{F,G_1}[s(X)E_{F,G_1}[h(O)|X]|W] \\
&= E_{F,G_1}[s(X)E_{G_1}[h(O)|X]|W] \\
&= E_{F,G_1}[s(X)l^T_{G_1}(h)|W] \\
&= E_F[s(X)l^T_{G_1}(h)|W] \\
&= < l^T_{G_1}(h), s(X) >_{\mathcal{X},\mathcal{W}} \qquad (34)
\end{aligned}$$

Consequently, we note that if

$$h(O) \in T_{\text{CAR}} \equiv \{h : E_{G_1}[h(O)|X] = 0\} \subset L^2(P_{F,G_1}),$$

then (34) implies that,

$$\begin{aligned}
E_{F,G_1}[h(O)Y^\star_{F,G_1}(O)|W] &= < h(O), Y^\star_{F,G_1}(O) >_{\mathcal{O},\mathcal{W}} \\
&= < h(O), l_{F,G_1} \circ I^{-1}_{F,G_1}(Y) >_{\mathcal{O},\mathcal{W}} \\
&= < l^T_{G_1} \circ h(O), I^{-1}_{F,G_1}(Y) >_{\mathcal{X},\mathcal{W}} \\
&= < E_{G_1}[h(O)|X], I^{-1}_{F,G_1}(Y) >_{\mathcal{X},\mathcal{W}} \\
&= < 0, I^{-1}_{F,G_1}(Y) >_{\mathcal{X},\mathcal{W}} \\
&= 0. \qquad (35)
\end{aligned}$$

In fact $\frac{dG}{dG_1}(O|X) - 1$ belongs to T_{CAR} because formula (8) in Gill et al. (1997) shows that the Radon-Nikodym derivative $\frac{dG}{dG_1}(O|X)$ can be written as a function of O when both $P_{F,G}$ and P_{F,G_1} satisfy coarsening at random as is assumed in the theorem statement, and for F-almost all $x \in \mathcal{X}$,

$$E_{F,G_1}[\frac{dG}{dG_1}(O|X)|X = x] = \int_{\mathcal{O}} \frac{dG}{dG_1}(o|x)dG_1(o|x) = \int_{\mathcal{O}} dG(o|x) = 1.$$

Thus, (35) implies that,

$$E_{F,G_1}[(\frac{dG_1}{dG}(O|X)-1)Y^\star_{F,G_1}(O)|W]=0,$$

which together with the implication of (34) that $E_{F,G_1}[Y^\star_{F,G_1}|W]=m(W)$ finally gives

$$\begin{aligned}
m(W) &= E_{F,G_1}[Y^\star_{F,G_1}(O)|W] \\
&= E_{F,G_1}[Y^\star_{F,G_1}(O)\frac{dG}{dG_1}(O|X)|W] \\
&= E_{F,G_1}[E_{G_1}[Y^\star_{F,G_1}(O)\frac{dG}{dG_1}(O|X)|X]|W] \\
&= \int_{\mathcal{X}}\{\int_{\mathcal{O}} Y^\star_{F,G_1}(O)\frac{dG}{dG_1}(o|x)dG_1(o|x)\}dF(x|W) \\
&= \int_{\mathcal{X}}\{\int_{\mathcal{O}} Y^\star_{F,G_1}(o)dG(o|x)\}dF(x|W) \\
&= E_{F,G}[Y^\star_{F,G_1}(O)|W]
\end{aligned}$$

This completes the proof for the case of $F=F_1$, and hence of the desired result. □

References

Buckley, J. and James, I.R. (1979). Linear regression with censored data. *Biometrika*, **66**, 429-436.

Doksum, K.A. and Yandell, B.S. (1982). Properties of regression estimates based on censored survival data. In *A Festschrift for Erich Lehmann* (P.J. Bickel, K.A. Doksum and J.L. Hodges, Jr., eds), 140-156. Wadsworth & Brooks/Cole, California.

Fan, J. and Gijbels, I. (1994). Censored regression: local linear approximations and their applications. *Journal of the American Statistical Association*, **89**, 560-570.

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

Heitjan, D.F. and Rubin, D.B. (1991). Ignorability and coarse data. *Annals of Statistics*, **19**, 2244-2253.

Koul, H., Susarla, V. and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *Annals of Statistics*, **9**, 1276-1288.

Leurgans, S. (1987). Linear models, random censoring and synthetic data. *Biometrika*, **74**, 301-309.

Miller, R.G. and Halpern, J. (1982). Regression with censored data. *Biometrika*, **69**, 521-539.

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

Shiboski, S.C. (1998). Generalized Additive Models for Current Status Data. *Lifetime Data Analysis 4*, 29-50.

van der Laan, M.J. (1998). Identity for NPMLE in censored data models. *Lifetime Data Models*, **4**, 83-102.

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.

Zheng, Z. (1987). A class of estimators of the parameters in linear regression with censored data. *Acta Mathem. Applic. Sinica*, **3**, 231-241.