causalglm

Semiparametric and nonparametric generalized linear models for conditional causal inference using Targeted Maximum Likelihood Estimation

Lars van der Laan

September 2021

1 Introduction to causalglm

1.1 Semiparametric and nonparametric generalized linear models for conditional causal inference using Targeted Maximum Likelihood Estimation

It is possible to get robust and efficient inference for causal quantities using machine-learning. In the search for answers to causal questions, assuming parametric models can be dangerous. With even a seemingly small amount of confounding and misspecificaton, they can give biased answers. One way of mitigating this challenge is to instead assume a parametric model for only the feature of the data-generating distribution that you care about. That is, assume a semiparametric model! Let the data speak for itself and use machine-learning to model the nuisance features of the data that are not directly related to your causal question. Why worry about things that don't matter for your question? It is not worth the risk of being wrong.

In this package, we utilize targeted machine-learning to generalize the parametric generalized linear models commonly used for treatment effect estimation (e.g. the R package glm) to the world of semi and nonparametric models. There is little-to-no loss in precision/p-values/confidence-interval-widths with these semiparametric methods relative to parametric generalized linear models, but the bias reduction from these methods can be substantial! Simulations suggest that these methods can work well with small sample sizes. We employ ensemble machine-learning (Super-Learning) that adapts the aggressiveness of the ML algorithms with sample size, thereby allowing for robust and correct inference in a diverse range of settings. All methods utilize targeted maximum likelihood estimation (TMLE) (van der Laan, Rose, 2011).

Each estimand considered in this package can be modeled with a user-specified parametric model that is either assumed correct ('spglm' and 'causalglmnet') or as an approximation, i.e. working model, of the nonparametric true estimand ('npglm'). The former approach provides interpretable estimates and correct inference only when the parametric model is correct, and the latter approach provides interpretable estimates and nonparametrically correct inference even when the parametric model is incorrect.

By default, the Highly Adaptive Lasso (HAL) and its semiparametric variants are used for estimation. See Benkeser et al. (2016) for an overview of HAL and its statistical performance. For the theoretical analysis of HAL including its fast rate of convergence in high dimensions, see Bibaut et al. (2019) and van der Laan (2017) . We implement the HAL estimators using the R package hal9001 (Coyle et al., 2021).

Noticable features supported:

- Efficent semiparametric and nonparametric inference for user-specified parametric working-models of conditional treatment-effect functions with 'spglm' and 'npglm'.
- Efficient nonparametric inference for marginal structural models for the CATE, CATT, TSM and RR with 'npglm'.
- General machine-learning tools with the tlverse/sl3 generalized machine-learning ecosystem (Coyle et al., 2021).
- High dimensional covariates and variable selection for confounders with the wrapper function 'causalglmnet'.

- Interpretable semiparametric and nonparametric estimates and efficient inference even with adaptive estimation and variable selection.
- Designed-easy-to-use interface and built-in machine-learning routines for diverse settings and immediate use.

1.2 spglm: semiparametric causal inference for generalized linear models

This function supports semiparametric efficient estimation of following point-treatment estimands:

- Conditional average treatment effect (CATE). (Causal semiparametric linear regression)
- Conditional odds ratio (OR) between two binary variables. (Causal semiparametric logistic regression)
- Conditional relative risk (RR) for nonnegative outcomes and a binary treatment. (Causal semiparametric log-linear relative-risk regression)

1.3 npglm: nonparametric causal inference for generalized linear models

The function npglm supports nonparametric efficient working-model-based estimation of following point-treatment estimands:

- Conditional average treatment effect (CATE).
- Conditional average treatment effect among the treated (CATT)
- Conditional treatment-specific mean (TSM)
- Conditional odds ratio (OR) between two binary variables.
- Conditional relative risk (RR) for nonnegative outcomes and a binary treatment.

This function also automatically supports marginal structural models by specifying lower dimensional formulas/working-models for the following estimands

- Marginal structural models for the conditional average treatment effect (CATE).
- Marginal structural models for the conditional average treatment effect among the treated (CATT).
- Marginal structural models for the conditional treatment-specific mean (TSM).
- Marginal structural models for the conditional relative risk (RR).

1.4 causalglmnet: high dimensional causal inference for generalized linear models

The function causalglmnet supports causal inference with high dimensional confounders for the following estimands:

- Conditional average treatment effect (CATE).
- Conditional odds ratio (OR) between two binary variables.
- Conditional relative risk (RR) for nonnegative outcomes and a binary treatment.

Note that the function causalglmnet is a custom wrapper function around the function spglm.

1.5 npcoxph: assumption-lean inference for the conditional hazard ratio

The function npcoxph supports the following survival estimands:

- Nonparametric inference for a user-specified working-model for the conditional hazard ratio between two treatments with ‘npcoxph‘.
- The estimands supported by ‘npcoxph‘ based on lower dimensional formulas can immediately be interpreted as marginal structural models for the hazard ratio.

2 Data-Structure and treatment-effect estimands

We will mainly consider the point-treatment data-structure $O = (W, A, Y)$ where W represents a vector of baseline covariates (i.e. possible confounders), A is a binary treatment assignment, and Y is some outcome variable. As an example, for a given observation O, W could be measurements: age, sex, a risk-score, location, income; A could take the value 1 if the individual receives the treatment and 0 if they do not receive the treatment; and Y is a binary or continuous variable that captures the effect of the treatment. For the goal of assessing heterogeneity of the treatment effect, there are a number of popular estimands:
The conditional average treatment effect (CATE):

$$CATE(w) := E[Y|A = 1, W = w] - E[Y|A = 0, W = w], \tag{1}$$

which is an additive measure of the effect of the treatment $(A = 1)$ relative to no treatment $(A = 0)$.

The conditional odds ratio (OR) for when Y is binary:

$$OR(w) := \frac{P(Y = 1|A = 1, W = w)/P(Y = 0|A = 1, W = w)}{P(Y = 1|A = 0, W = w)/P(Y = 0|A = 0, W = w)} \tag{2}$$

The conditional relative risk (RR) for when Y is nonnegative (e.g. a binary or count variable):

$$RR(w) := \frac{E[Y|A = 1, W = w]}{E[Y|A = 0, W = w]}, \tag{3}$$

which is a relative measure of the effect of the treatment $(A = 1)$ relative to no treatment $(A = 0)$.

In some application, non-contrast measures like the conditional treatment-specific mean (TSM) may be of interest:

$$TSM_a(w) := E[Y|A = a, W = w]. \tag{4}$$

2.1 Conventional estimators using parametric generalized linear models

In order to estimate the estimands of the previous section, parametric generalized linear models are often employed (e.g. the R package *glm*). For the CATE,the following linear regression model is often used:

$$E[Y|A, W = w] = \beta_0 A + \beta_1^T w \cdot A + \beta_2^T w.$$

This model is equivalent to assuming both the nuisance linear model

$$E[Y|A = 0, W = w] = \beta_2^T w$$

and target linear model

$$CATE(w) = \beta_0 + \beta_1^T w.$$

Thus, the coefficient in front of the treatment interactions can be directly interpreted as a measure of the conditional average treatment effect when the parametric model is correct. However, we see that very strong parametric assumptions are made on the orthogonal nuisance function $w \mapsto E[Y|A = 0, W = w]$, which has little to do with the CATE.

Next, for the conditional odds ratio, the following logistic regression model is often used

$$\text{logit}\{P(Y = 1|A, W = w)\} = \beta_0 A + \beta_1^T w \cdot A + \beta_2^T w.$$

This model is equivalent to assuming both the nuisance logistic model

$$\text{logit}\{P(Y = 0|A = 0, W = w)\} = \beta_2^T w$$

and the target logistic model

$$\log OR(w) = \beta_0 + \beta_1^T w.$$

Once again, we see that the conventional logistic regression model makes strong parametric assumptions on the orthogonal nuisance parameter $P(Y = 0|A = 0, W = w)$.

Finally, for the conditional relative-risk, the poisson or log-linear regression model is a well-known approach:

$$\log \{E[Y|A, W = w]\} = \beta_0 A + \beta_1^T w \cdot A + \beta_2^T w.$$

This is equivalent to assuming the nuisance log-linear model

$$\log \{E[Y|A = 0, W = w]\} = \beta_2^T w$$

and the target log-linear model

$$\log \{RR(w)\} = \beta_0 + \beta_1^T w.$$

Besides the strong parametric assumptions on the nuisance parameter $E[Y|A = 0, W = w]$, another issue with this approach is that conventional methods only provide inference when the outcome is Poisson distributed, which is not useful if the outcome is binary. Log-link binomial generalized-linear-models are one way to overcome this limitation.

Under the parametric assumptions, standard generalize linear model software can be used to obtain estimates and inference for the coefficients in the above models for the treatment-effect estimands. However, these methods make much stronger assumptions than are needed. In particular, the parametric assumptions on $E[Y|A = 0, W]$ provides little-to-no benefit in interpretability since we are interested in the coefficients for the treatment interaction terms, and it comes at a possibly substantial cost in bias due to model misspecification. Additionally, these methods do not allow for any adaptive estimation of $E[Y|A = 0, W]$ (e.g. using the LASSO, variable selection, or machine-learning) and therefore do not perform well in both estimation and inference in high dimensions. In the next section, we consider a partial relaxation of the parametric models through so-called partially-linear generalized linear models.

3 Semiparametric generalized linear models for treatment effect estimation with spglm

In this section, we give an overview of semiparametric treatment-effect estimation in partially-linear generalized linear models which allows for adaptive estimation of nuisance parameters of the data-generating distribution that are not directly relevant for the problem at end. Semiparametric models are statistical models in which one component of the data-generating distribution is parametric and the remaining components are nonparametric. For background on semiparametric models and estimators in causal inference, we refer to Bickel et al. (1993) and van der Laan, Robins (2003).

These methods allow for:

1. Interpretable (coefficient-based) inference for user-specified parametric models for conditional treatment effect estimands.
2. Adaptive machine-learning and variable selection methods including generalized additive models, LASSO, MARS and gradient-boosting can be used to estimate nuisance parameters nonparametrically, thereby substantially relaxing assumptions for valid inference although still assuming a parametric model for the conditional estimand of interest.

3.1 Conditional average treatment effect (CATE) and partially-linear least-squares regression

Let $\underline{f}(w)$ be an arbitrary known vector-valued function of the covariates and consider the linear parametric model $\beta^T \underline{f}(w)$ for $CATE(w)$. The partially-linear least-squares model is of the form:

$$E[Y|A, W = w] = \beta^T \underline{f}(w) \cdot A + h_0(w),$$

where $h_0(w) := E[Y|A = 0, W = w]$ is an unspecified nuisance function that is to be learned from the data nonparametrically. The parametric component of the model is the coefficient vector β. This model is equivalent to *only* assuming:

$$CATE(w) = \beta^T \underline{f}(w).$$

Thus, this semiparametric model only makes assumptions that directly relate the estimand of interest!
A concrete model is choosing $\underline{f}(W) = (1, W)$ which gives the linear model

$$CATE(w) = \beta_0 + \beta_1^T w.$$

Estimates and inference for the coefficient vector in this semiparametric model can be obtained by applying the R function *spglm* with the option 'estimand = "CATE"'. We employ machine-learning for initial estimation of the relevant components of the data-generating distribution and then use targeted maximum likelihood estimation for bias-correction, thereby allowing for valid efficient inference. The estimand is estimated using targeted maximum likelihood estimation and the theory and pseudo-code for the method can be found in the working paper, van der Laan (2009).

3.2 Conditional odds ratio (OR) and partially-linear logistic regression

Let $\underline{f}(w)$ be an arbitrary known vector-valued function of the covariates and consider the parametric model $\beta^T \underline{f}(w)$ for $\log OR(w)$. The partially-linear logistic model is given by:

$$\text{logit}\{E[Y|A, W = w]\} = \beta^T \underline{f}(w) \cdot A + h_0(w),$$

where $h_0(w) := \text{logit}\{E[Y|A = 0, W = w]\}$ is an unspecified nuisance function that is to be learned from the data nonparametrically. This model is equivalent to *only* assuming:

$$\log OR(w) = \beta^T \underline{f}(w).$$

Estimates and inference for the coefficient vector in this semiparametric model can be obtained by applying the R function *spglm* with the option 'estimand = "OR"'. The estimand is estimated using targeted maximum likelihood estimation and the theory and pseudo-code for the method can be found in the working paper, van der Laan (2009). For a similar estimator based on estimating equations and additional background on the semiparametric logistic regression model, see Tchetgen Tchetgen (2010).

3.3 Conditional relative-risk (RR) and partially-linear log-linear regression

Let $\underline{f}(w)$ be an arbitrary known vector-valued function of the covariates and consider the parametric model $\beta^T \underline{f}(w)$ for $\log RR(w)$. The partially-linear log-linear model is of the form:

$$\log\{E[Y|A, W = w]\} = \beta^T \underline{f}(w) \cdot A + h_0(w),$$

where $h_0(w) := \log\{E[Y|A = 0, W = w]\}$ is an unspecified nuisance function that is to be learned from the data nonparametrically. This model is equivalent to *only* assuming:

$$\log RR(w) = \beta^T \underline{f}(w).$$

Estimates and inference for the coefficient vector in this semiparametric model can be obtained by applying the R function *spglm* with the option 'estimand = "RR"'. The estimand is estimated using targeted maximum likelihood estimation and the theory and pseudo-code for the method can be found in the working paper, Tuglus et al. (2011).

3.4 spglm in practice

Let us generate a mock dataset that has a constant CATE of value 1.

```
> library(causalglm)
> n <- 250
> W <- runif(n, min = -1,  max = 1)
> A <- rbinom(n, size = 1, prob = plogis(W))
> Y <- rnorm(n, mean = A + W, sd = 0.3)
> data <- data.frame(W,A,Y)
```

We specify the parametric form of the CATE through the formula argument. In this case, we will use the intercept-only formula which is equivalent to assuming the CATE is constant. The output consists of: a coefficient estimate for the intercept, lower and upper confidence intervals, an asymptotic standard error estimate for the estimator, a Z-score and p-value. The argument 'W' should be a character vector of variable names in data for which to adjust, 'A' should be the name of a treatment variable, and 'Y' should be the name of an outcome variable. We set the argument 'estimand = "CATE"' to estimate the conditional average treatment effect.

```
> formula <- ~ 1
> output <-
+   spglm(
+     formula,
+     data,
+     W = "W", A = "A", Y = "Y",
+     estimand = "CATE",
+     verbose = FALSE
+   )
> summary(output)

   type       param  tmle_est         se     lower    upper  Z_score p_value
1: CATE (Intercept) 0.9726337 0.04082418 0.8926197 1.052648 376.7054       0
```

We can also do a much more complex model with higher dimensional treatment effect interactions. The following data distribution has $CATE(w) = 1 + w$.

```
> Y <-
+   rnorm(n,
+         mean = A * W + A + poly(W, degree = 3) + sin(4 * W),
+         sd = 0.4)
> data <- data.frame(W, A, Y)
> formula <- ~ 1 + W
> output <-
+   spglm(
+     formula,
+     data,
+     W = "W", A = "A", Y = "Y",
+     estimand = "CATE",
+     verbose = FALSE
+   )
> summary(output)

   type       param  tmle_est         se     lower    upper  Z_score p_value
1: CATE (Intercept) 0.9894983 0.05007017 0.8913625 1.087634 312.4683       0
2: CATE           W 0.9736947 0.09559303 0.7863358 1.161054 161.0522       0
```

Currently, the nonparametric learning is performed using the partially-linear first-order Highly Adaptive Lasso (HAL) implemented using the R package tlverse/hal9001. HAL is an adaptive piece-wise linear regression spline estimator and the custom implementation performs the risk minimization entirely within the semiparametric model. It is implemented using LASSO regression with the parametric treatment interactions (as specified by the formula argument) unpenalized and a rich penalized spline basis for the nonparametric component of the model. This method performs risk minimization entirely within the semiparametric model. There are a number of other built-in learning options: c("HAL", "SuperLearner", "glm", "glmnet", "gam", "mars", "ranger", "xgboost")

```
> Y <- rnorm(n, mean = A * W + A + W, sd = 0.4)
> data <- data.frame(W, A, Y)
> # generalized additive models:
> output <-
+   spglm(
+     formula,
+     data,
+     W = "W", A = "A", Y = "Y",
+     estimand = "CATE",
+     learning_method = "gam",
+     verbose = FALSE
+   )
> summary(output)

   type       param tmle_est         se     lower    upper  Z_score p_value
1: CATE (Intercept) 1.064734 0.05493813 0.9570570 1.172411 306.4342       0
2: CATE           W 1.120836 0.11262147 0.9001016 1.341570 157.3587       0

> # multivariate adaptive regression splines:
> output <-
+   spglm(
+     formula,
+     data,
+     W = "W", A = "A", Y = "Y",
+     estimand = "CATE",
+     learning_method = "mars",
+     verbose = FALSE
+   )
> summary(output)

   type       param tmle_est         se     lower    upper  Z_score p_value
1: CATE (Intercept) 1.051676 0.05642054 0.9410939 1.162258 294.7235       0
2: CATE           W 1.098311 0.10912470 0.8844307 1.312192 159.1374       0

> # gradient-boosting with xgboost: :
> output <-
+   spglm(
+     formula,
+     data,
+     W = "W", A = "A", Y = "Y",
+     estimand = "CATE",
+     learning_method = "xgboost",
+     verbose = FALSE
+   )
> summary(output)
```

```
   type       param tmle_est         se     lower    upper  Z_score p_value
1: CATE (Intercept) 1.094323 0.06419361 0.9685056 1.220140 269.5403       0
2: CATE           W 1.193905 0.13667646 0.9260241 1.461786 138.1167       0
```

The default learning algorithm "HAL" can be customized with the HAL_args_Y0W argument (see the arguments in hal9001 for more description). 'max_degree' $= 1$ corresponds with estimating the nuisance function $E[Y|A = 0, W]$ with an additive model. 'num_knots' specifies for each interaction degree how many variable knot points are used to generate the tensor product interaction basis functions.

```
> HAL_args_Y0W <-
+   list(
+     smoothness_orders = 1,
+     max_degree = 2,
+     num_knots = c(10, 5, 1)
+   )
> output <-
+   spglm(
+     formula,
+     data,
+     W = "W", A = "A", Y = "Y",
+     estimand = "CATE",
+     learning_method = "HAL",
+     HAL_args_Y0W = HAL_args_Y0W,
+     verbose = FALSE
+   )
> summary(output)
```

```
   type       param tmle_est         se     lower    upper  Z_score p_value
1: CATE (Intercept) 1.057640 0.05445743 0.9509052 1.164374 307.0794       0
2: CATE           W 1.104797 0.11170567 0.8858578 1.323736 156.3786       0
```

It is also possible to employ custom learners using the tlverse/sl3 framework and the sl3_Learner_Y (to estimate $E[Y|A = 1, W]$ and $E[Y|A = 0, W]$) and sl3_Learner_A (to estimate $P(A = 1|W)$) argument. Take a look at the argument append_interaction_matrix to understand the design matrix that is given as input to the learner sl3_Learner_Y. In particular, it is important to note that sl3_Learner_Y will be sent to Lrnr_glm_semiparametric.

```
> library(sl3)
> lrnr_glmnet <- Lrnr_glmnet$new()
> lrnr_xgboost <- Lrnr_xgboost$new(max_depth = 4)
> lrnr_earth <- Lrnr_earth$new()
> lrnr_stack <-
+   make_learner(Stack, lrnr_glmnet, lrnr_xgboost, lrnr_earth)
> lrnr_cv <- Lrnr_cv$new(lrnr_stack, full_fit = TRUE)
> # A custom superlearner
> lrnr_sl <- make_learner(Pipeline, lrnr_cv, Lrnr_cv_selector$new())
> output <-
+   spglm(
+     formula,
+     data,
+     W = "W", A = "A", Y = "Y",
+     estimand = "CATE",
+     sl3_Learner_A = lrnr_sl  ,
+     sl3_Learner_Y = lrnr_sl ,
+     verbose = FALSE
```

```
+   )
> summary(output)

   type       param tmle_est         se     lower    upper  Z_score p_value
1: CATE (Intercept) 1.054905 0.05693366 0.9433174 1.166493 292.9641       0
2: CATE           W 1.102858 0.10809111 0.8910031 1.314713 161.3242       0
```

That's all there is to it! spglm also supports the RR and OR estimands which are run in a completely analagous way. Use the option 'estimand = "OR"' to estimate the conditional odds ratio, and use the option 'estimand = "rR"' to estimate the conditional relative risk. Note that the parametric model specified by formula is actually for the log odds ratio and log relative risk (i.e. at the log scale). Thus, the coefficients returned are for the log-transformed OR and RR. We also provide the exponential-transformed coefficients and confidence intervals, which may be more interpretable as measures of the OR and RR.

```
> n <- 250
> W <- runif(n, min = -1,  max = 1)
> A <- rbinom(n, size = 1, prob = plogis(W))
> # OR
> Y <- rbinom(n, size =  1, prob = plogis(A + A * W + W + sin(5 * W)))
> data <- data.frame(W, A, Y)
> formula ~ 1 + W

formula ~ 1 + W

> output <-
+   spglm(
+     formula,
+     data,
+     W = "W", A = "A", Y = "Y",
+     estimand = "OR" ,
+     verbose = FALSE
+   )
> summary(output)

   type       param tmle_est        se     lower    upper  psi_exp lower_exp
1:   OR (Intercept) 1.130524 0.2928748 0.5564998 1.704548 3.097279  1.744555
2:   OR           W 1.394404 0.4856034 0.4426386 2.346169 4.032569  1.556810
   upper_exp  Z_score p_value
1:   5.49890 61.03342       0
2:  10.44547 45.40219       0

> # RR
> Y <- rpois(n, lambda = exp(A + A * W + sin(5 * W)))
> data <- data.frame(W, A, Y)
> formula ~ 1 + W

formula ~ 1 + W

> output <-
+   spglm(
+     formula,
+     data,
+     W = "W", A = "A", Y = "Y",
+     estimand = "RR" ,
+     verbose = FALSE
+   )
> summary(output)
```

```
   type       param  tmle_est        se     lower    upper  psi_exp lower_exp
1:   RR (Intercept) 0.9188326 0.1071738 0.7087759 1.128889 2.506363  2.031503
2:   RR           W 1.2294934 0.1870166 0.8629477 1.596039 3.419497  2.370137
   upper_exp  Z_score p_value
1:  3.092220 135.5557       0
2:  4.933453 103.9480       0

>
```

4 Robust nonparametric generalized linear models for treatment effect estimation with npglm

In the previous section, we considered semiparametric generalized linear models where the data distribution component of interest (the estimand) is modeled parametrically and nuisance components are modeled nonparametrically. While this is much more robust than typical parametric methods, the parametric assumption on the estimand of interest can still be quite strong. It is therefore of interest to develop fully nonparametric methods that provide correct and interpretable estimates and inference under no parametric assumptions. To do so, we will still utilize user-specified parametric models for the estimand of interest, however, we will not assume that these parametric models are correct. We will treat these parametric models as interpretable approximations of the true nonparametric estimand. That is, we utilize the parametric model as a "working-model" that is only used to derive an interesting nonparametric estimand.

It turns out that many of these working models have desirable properties. In particular, by specifying parametric models/formulas that depend on $V \subset W$ where V is a subvector of baseline covariates, we can actually learn marginal structural models. Notably, the intercept working model that approximates the true estimand by a constant often corresponds with a marginal causal estimand like the average treatment effect (ATE or ATT), the marginal treatment-specific mean, or the marginal relative risk. For these reasons, these nonparametrics method allow for the estimation of an even more rich class of parameters than the analagous semiparametric methods.

Nonparametric working-model based estimators for the estimands are implemented in the function 'npglm'. These methods have:

- Interpretable coefficient-based estimates and inference
- Nonparametric and causal estimates and inference even when the parametric model is incorrect
- Many of the estimands correspond with marginal structural models when lower dimensional working-models are used.

We refer to van der Laan, Robins (2003), Neugebauer, van der Laan (2008), and Robins et al. (1994) for background on marginal structural models. See also the working paper, van der Laan (2009), for more on marginal structural models in the context of TMLE.

4.1 Conditional average treatment effect (CATE) estimation with a linear working-model

To define a nonparametric approximation of the true CATE with a parametric linear-working model, we utilize the least-squares projection. Let $\underline{f}(w)$ be an arbitrary known vector-valued function of the covariates and consider the linear parametric working-model $\beta^T \underline{f}(w)$ for $CATE(w)$.
Define the risk function,

$$R_{CATE}(\beta) = E\left\{CATE(W) - \beta^T \underline{f}(W)\right\}^2$$

Our estimand of interest is given by β^* which is defined as the minimizer of R_{CATE}.

Consider the simple working model $\beta^T \underline{f}(W) := \beta_0 + \beta_1^T W$. The risk function then reduces to

$$R_{CATE}(\beta_0, \beta_1) = E\left\{CATE(W) - \beta_0 - \beta_1^T W\right\}^2$$

which can be viewed as the ordinary least-squares regression of the true estimand $CATE(W)$ onto W. By specifying a lower dimensional working model $\underline{f}(W) := V$ for some $V \subset W$, the risk function further reduces to

$$R_{CATE}(\beta_0, \beta_1) = E\left\{E[CATE(W)|V] - \beta_0 - \beta_1^T V\right\}^2.$$

The risk minimizer is now the least-squares projection of the true marginal structural CATE model $E[CATE(W)|V]$ onto the linear working model. Note if $E[CATE(W)|V] = \beta_0 + \beta_1^T V$, so that the working model is correct, then this estimand can be directly interpreted as a marginal structural CATE function.
If we use the intercept model then the risk function reduces to

$$R_{CATE}(\beta_0) = E\left\{E[CATE(W)] - \beta_0\right\}^2.$$

and the risk minimizer is exactly given by the nonparametric ATE $E[CATE(W)]$! Thus, the intercept model can be used for marginal ATE estimation. Quite remarkably, this estimands based on such least-squares working model projections automatically reduce to marginal structural model parameters when lower dimensional working models are used. No user or developer intervention is needed for this to happen!
These estimators and estimands are inspired by Chambaz et al. (2012) and is based on discussions with Prof. Mark van der Laan.

4.2 Conditional average treatment effect among the treated (CATT) estimation with a linear working-model

We now define an alternative working model for the CATE that focuses on the treatment effect among the treated. Again, let $\underline{f}(w)$ be an arbitrary known vector-valued function of the covariates and consider the linear parametric working-model $\beta^T \underline{f}(w)$ for $CATE(w)$.
Define the risk function,

$$R_{CATT}(\beta) = E\left\{E[Y|A,W] - A \cdot \beta^T \underline{f}(W) - E[Y|A=0,W]\right\}^2$$

Our estimand of interest is β^* which is defined as the minimizer of R_{CATT}. This working model can be viewed as the least-squares regression of the true conditional mean $E[Y|A,W]$ onto the interaction model $A \cdot \beta^T \underline{f}(W)$ using as offset the true placebo conditional mean $E[Y|A=0,W]$. It turns out that we can rewrite this risk function as

$$R_{CATT}(\beta) = E\left\{A\left[CATE(W) - \beta^T \underline{f}(W)\right]\right\}^2,$$

which is the least-squares projection of the true CATE onto the linear working model using only the observations with $A = 1$ (the treated). Because of this, we call estimands based on this risk function measures of the conditional average treatment effect among the treated (CATT).

This method can also be used to estimate marginal structural models for treatment effects among the treated. Specifically, by specifying a lower dimensional working model $\underline{f}(W) := V$ for some $V \subset W$, the risk function further reduces to

$$R_{CATT}(\beta_0, \beta_1) = E\left\{A\left[E[CATE(W)|V, A=1] - \beta_0 - \beta_1^T V\right]^2\right\}.$$

The risk minimizer is now the least-squares projection of the true marginal structural CATT model $E[CATE(W)|V, A = 1]$ onto the linear working model. Next, if we use the intercept model then the risk function reduces to

$$R_{CATT}(\beta_0) = E\left\{A \cdot [E[CATE(W)|A=1] - \beta_0]^2\right\}.$$

and the risk minimizer is exactly given by the nonparametric ATT $E[CATE(W)|A = 1]$! Thus, the intercept model can be used for marginal ATT estimation.

These estimators and estimands are directly due to Chambaz et al. (2012).

4.3 Conditional treatment-specific mean (TSM) estimation with a linear working-model

A similar working model-based estimand can be constructed for the conditional treatment specific mean. Let a be a level of a categorical treatment assignment A. Again, let $\underline{f}(w)$ be an arbitrary known vector-valued function of the covariates and consider the linear parametric working-model $\beta^T \underline{f}(w)$ for $CATE(w)$.
Define the risk function,

$$R_{TSM}(\beta) = E\left\{E[Y|A=a,W] - \beta^T \underline{f}(W)\right\}^2$$

Our estimand of interest is β^* which is defined as the minimizer of R_{TSM}. This estimand corresponds with the least-squares projection of $E[Y|A=a,W=w]$ onto the linear working model.

By specifying a lower dimensional working model $\underline{f}(W) := V$ for some $V \subset W$, the risk function further reduces to

$$R_{TSM}(\beta_0,\beta_1) = E\left\{E[E[Y|A=a,W]|V] - \beta_0 - \beta_1^T V\right\}^2.$$

The risk minimizer is now the least-squares projection of the true marginal structural TSM model E[E[Y|A=a,W]|V] onto the linear working model. If we use the intercept model then the risk function reduces to

$$R_{TSM}(\beta_0) = E\left\{E[E[Y|A=a,W]] - \beta_0\right\}^2.$$

and the risk minimizer is exactly given by the nonparametric TSM $E[E[Y|A=a,W]]$! Thus, the intercept model can be used for marginal TSM estimation.

These estimators and estimands are inspired by Chambaz et al. (2012) and is based on discussions with Prof. Mark van der Laan.

4.4 Conditional odds-ratio (OR) estimation with a logistic working-model

Define the working model

$$P_\beta(Y=1|A,W) := \text{expit}\left\{A \cdot \beta^T \underline{f}(W) + \text{logit}(P(Y=0|A,W))\right\},$$

which is not assumed correct.
Consider the log-likelihood projection risk function

$$R_{OR}(\beta) = E\left\{P(Y=1|A,W)\log(P_\beta(Y=1|A,W)) + P(Y=0|A,W)\log(P_\beta(Y=0|A,W))\right\}.$$

We define the nonparametric OR estimand as the risk minimizer β^* of R_{OR}. This estimand unfortnately does not reduce to a marginal structural model estimand when $\underline{f}(W)$ lower dimensional.

4.5 Conditional relative-risk regression (RR) with a log-linear working-model

Define the log-linear multiplicative working model

$$E_\beta[Y|A=1,W] := \exp\left\{\beta^T \underline{f}(W)\right\} E[Y|A=0,W],$$

which is not assumed correct and $E_\beta[Y|A=0,W] := E[Y|A=0,W]$ is left correctly specified. We define the projection using the log-linear generalized linear model,

$$R_{RR}(\beta) = E\left\{E[Y|A=0,W]\exp\left\{\beta^T \underline{f}(W)\right\} - E[Y|A=1,W]\beta^T \underline{f}(W)\right\}.$$

We define the nonparametric RR estimand as the risk minimizer β^* of R_{RR}.

By specifying a lower dimensional working model $\underline{f}(W) := V$ for some $V \subset W$, the risk function further reduces to

$$R_{RR}(\beta) = E\left\{E[E[Y|A=0,W]|V]\exp\left\{\beta^T V\right\} - E[E[Y|A=1,W]|V]\beta^T V\right\}.$$

The risk minimizer is now the projection of the true marginal structural RR model $\frac{E[E[Y|A=1,W]|V]}{E[E[Y|A=0,W]|V]}$ onto the log-linear working model. Thus, if the working model is correctly specified, the estimand is $\frac{E[E[Y|A=1,W]|V]}{E[E[Y|A=0,W]|V]}$. If we use the intercept model then the risk function reduces to

$$R_{RR}(\beta) = E\left\{E[E[Y|A=0,W]]\exp\{\beta\} - E[E[Y|A=1,W]]\beta\right\}.$$

and the risk minimizer is exactly given by the nonparametric marginal relative risk $\frac{E[E[Y|A=1,W]]}{E[E[Y|A=0,W]]}$! Thus, the intercept model can be used for marginal RR estimation.

4.6 npglm in action

npglm operates in almost exactly the same way as spglm (with a few less optional arguments). The main difference is that it supports new estimands like the CATT and TSM. All previous discussion on learner customization operates in the exact same way but it should be noted that nuisance functions are now estimated nonparametrically (and no longer semiparametrically) so it is important to use learners that include interactions. The following block of the code runs npglm for each supported estimand.

```
> n <- 250
> W <- runif(n, min = -1,  max = 1)
> A <- rbinom(n, size = 1, prob = plogis(W))
> # CATE
> Y <- rnorm(n, mean = A + A * W + W + sin(5 * W), sd = 0.5)
> data <- data.frame(W, A, Y)
> formula ~ 1 + W

formula ~ 1 + W

> # Use the formula_Y argument to specify the design matrix for glm, glmnet or mars (not supported for c
> output <-
+   npglm(
+     formula,
+     data,
+     W = "W", A = "A", Y = "Y",
+     estimand = "CATE" ,
+     learning_method = "glm",
+     formula_Y = ~ . + .*A,
+     verbose = FALSE
+   )
> summary(output)

   type       param init_est tmle_est        se     lower    upper
1: CATE (Intercept) 1.038896 1.032209 0.1066999 0.8230810 1.241337
2: CATE           W 1.230973 1.171545 0.1692565 0.8398083 1.503282
   psi_transformed lower_transformed upper_transformed
1:        1.032209         0.8230810          1.241337
2:        1.171545         0.8398083          1.503282

> # CATT
> Y <- rnorm(n, mean = A + A * W + W + sin(5 * W), sd = 0.5)
> data <- data.frame(W, A, Y)
> formula ~ 1 + W

formula ~ 1 + W
```

```
> output <-
+   npglm(
+     formula,
+     data,
+     W = "W", A = "A", Y = "Y",
+     estimand = "CATT" ,
+     verbose = FALSE
+   )
> summary(output)

   type       param init_est tmle_est         se     lower    upper
1: CATT (Intercept) 1.029281 1.030827 0.07147802 0.8907322 1.170921
2: CATT           W 1.134408 1.136782 0.13424711 0.8736625 1.399901
   psi_transformed lower_transformed upper_transformed
1:        1.030827         0.8907322          1.170921
2:        1.136782         0.8736625          1.399901

> # TSM
> Y <- rnorm(n, mean = A + A * W + W + sin(5 * W), sd = 0.5)
> data <- data.frame(W, A, Y)
> formula ~ 1 + W

formula ~ 1 + W

> output <-
+   npglm(
+     formula,
+     data,
+     W = "W", A = "A", Y = "Y",
+     estimand = "CATT" ,
+     learning_method = "mars",
+     formula_Y = ~ . + .*A,
+     verbose = FALSE
+   )
> summary(output)

   type       param  init_est  tmle_est         se     lower    upper
1: CATT (Intercept) 1.0387156 1.0385957 0.06647905 0.9082991 1.168892
2: CATT           W 0.9914426 0.9748958 0.13642220 0.7075132 1.242278
   psi_transformed lower_transformed upper_transformed
1:       1.0385957         0.9082991          1.168892
2:       0.9748958         0.7075132          1.242278

> # OR
> Y <- rbinom(n, size =  1, prob = plogis(A + A * W + W + sin(5 * W)))
> data <- data.frame(W, A, Y)
> formula ~ 1 + W

formula ~ 1 + W

> output <-
+   npglm(
+     formula,
+     data,
```

```
+     W = "W", A = "A", Y = "Y",
+     estimand = "OR" ,
+     verbose = FALSE
+   )
> summary(output)

   type       param init_est tmle_est        se     lower    upper
1:   OR (Intercept)  1.48634 1.645172 0.4507338 0.7617500 2.528594
2:   OR           W  1.87996 2.145839 0.9005735 0.3807473 3.910930
   psi_transformed lower_transformed upper_transformed
1:        5.181901          2.142021          12.53587
2:        8.549210          1.463378          49.94540

> # RR
> Y <- rpois(n, lambda = exp(A + A * W + sin(5 * W)))
> data <- data.frame(W, A, Y)
> formula ~ 1 + W

formula ~ 1 + W

> output <-
+   npglm(
+     formula,
+     data,
+     W = "W", A = "A", Y = "Y",
+     estimand = "RR" ,
+     verbose = FALSE
+   )
> summary(output)

   type       param init_est tmle_est        se     lower    upper
1:   RR (Intercept) 1.154087 1.154609 0.1009330 0.9567843 1.352434
2:   RR           W 1.356124 1.386268 0.2474941 0.9011882 1.871347
   psi_transformed lower_transformed upper_transformed
1:        3.172784          2.603311          3.866827
2:        3.999893          2.462527          6.497043

>
```

5 High dimensional semiparametric generalized linear models for treatment effect estimation using the LASSO with causalglmnet

In high dimensional settings (e.g. dim(W) >= 50-1000), conventional machine-learning algorithms may be computationally expensive or poorly behaved. In such scenarios, we can utilize lasso-penalized regression (Tibshirani, 1994) to estimate the nuisance parameters, allowing for adaptive variable selection and adjusting of confounders. The function causalglmnet is a specialized wrapper for spglm that uses the lasso implementation provided by the state-of-the-art R package glmnet (Friedman, 2010) for estimation of all nuisance parameters. Its use is exactly the same as spglm except learners no longer need to be specified.

```
> n <- 200
> W <- replicate(100,runif(n, min = -1,  max = 1))
> colnames(W) <- paste0("W", 1:100)
> beta <- runif(10, -1, 1)/20
```

```
> A <- rbinom(n, size = 1, prob = plogis(W[,10*(1:10)] %*% beta))
> # CATE
> Y <- rnorm(n, mean = A + W[,10*(1:10)] %*% beta, sd = 0.5)
> data <- data.frame(W, A, Y)
> formula = ~ 1
> output <-
+   causalglmnet(
+     formula,
+     data,
+     W = colnames(W), A = "A", Y = "Y",
+     estimand = "CATE" ,
+     verbose = FALSE
+   )
> summary(output)
> # OR
> Y <- rbinom(n, size =  1, prob = plogis( A + W[,10*(1:10)] %*% beta))
> data <- data.frame(W, A, Y)
> formula  = ~ 1
> output <-
+   causalglmnet(
+     formula,
+     data,
+     W = colnames(W), A = "A", Y = "Y",
+     estimand = "OR" ,
+     verbose = FALSE
+   )
> summary(output)
> # RR
> Y <- rpois(n, lambda = exp( A + W[,10*(1:10)] %*% beta))
> data <- data.frame(W, A, Y)
> formula  = ~ 1
> output <-
+   causalglmnet(
+     formula,
+     data,
+     W = colnames(W), A = "A", Y = "Y",
+     estimand = "RR" ,
+     verbose = FALSE
+   )
> summary(output)
>
```

6 Robust nonparametric inference for the hazard ratio with npcoxph

This is in development.

References

Benkeser, D. and van der Laan, M. (2016). The highly adaptive lasso estimator. *International Conference on Data Science and Advanced Analytics*, pages 689–696.

Bibaut, A. F. and van der Laan, M. J. (2019). Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm.

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

Chambaz, A., Neuvial, P., and van der Laan, M. J. (2012). Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic Journal of Statistics*, 6(none):1059 – 1099.

Coyle, J. R., Hejazi, N. S., Malenica, I., and Sofrygin, O. (2021a). *sl3: Modern Pipelines for Machine Learning and Super Learning*. R package version 1.4.2.

Coyle, J. R., Hejazi, N. S., and van der Laan, M. J. (2021b). *hal9001: The scalable highly adaptive lasso*. R package version 0.2.7.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Hejazi, N. S., Coyle, J. R., and van der Laan, M. J. (2020). hal9001: Scalable highly adaptive lasso regression in R. *Journal of Open Source Software*.

Laan, M. J. V. D. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

Neugebauer, R. and van der Laan, M. (2007). Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137(2):419–434.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Marginal structural models and causal inference in epidemiology. *Journal of the American statistical Association*, 89(427):846–866.

Tchetgen Tchetgen, E., Robins, J., and Rotnitzky, A. (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika*, 97:171–180.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Tuglus, C., Porter, K., and van der Laan, M. (2011). Targeted maximum likelihood estimation of conditional relative risk in a semi-parametric regression model. *U.C. BERKELEY DIVISION OF BIOSTATISTICS WORKING PAPER SERIES*, Working paper 283.

van der Laan, M. (2009). Readings in targeted maximum likelihood estimation. *Biostatistics Working Paper Series Working Paper 253*, pages 621–622,626–629.

van der Laan, M. (2017). A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *The international journal of biostatistics*, 13(2).

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6:number 1.

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

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York.