David Benkeser*, Andrew Mertens, John M. Colford, Jr., Alan Hubbard,
Benjamin F. Arnold, Aryeh D. Stein and Mark J. van der Laan

A machine learning-based approach for estimating and testing associations with multivariate outcomes

<https://doi.org/10.1515/ijb-2019-0061>

Received May 22, 2019; accepted June 18, 2020; published online August 13, 2020

Abstract: We propose a method for summarizing the strength of association between a set of variables and a multivariate outcome. Classical summary measures are appropriate when linear relationships exist between covariates and outcomes, while our approach provides an alternative that is useful in situations where complex relationships may be present. We utilize machine learning to detect nonlinear relationships and covariate interactions and propose a measure of association that captures these relationships. A hypothesis test about the proposed associative measure can be used to test the strong null hypothesis of no association between a set of variables and a multivariate outcome. Simulations demonstrate that this hypothesis test has greater power than existing methods against alternatives where covariates have nonlinear relationships with outcomes. We additionally propose measures of variable importance for groups of variables, which summarize each groups' association with the outcome. We demonstrate our methodology using data from a birth cohort study on childhood health and nutrition in the Philippines.

Keywords: canonical correlation; epidemiology; machine learning; multivariate outcomes; variable importance.

1 Introduction

In statistical analysis, we often encounter situations where a correlation between a multivariate set of covariates and a multivariate outcome is of scientific interest. A classical approach to this problem is canonical correlation analysis [1, 2]. Canonical correlation maximizes the correlation between a linear combination of the multivariate outcome and a linear combination of the covariates. Several test statistics have been proposed for significance testing of canonical correlation including Wilks' Λ [3], the Hotelling-Lawley trace [1, 2], the Pillai-Bartlett trace [4, 5], and Roy's largest root [6]. One potential drawback to canonical correlation analysis is that it may fail to identify associations when nonlinear relationships exist between covariates and outcomes. This has led to recent interest in nonlinear extension of canonical correlation analysis including kernel canonical correlation analysis [7–9].

Another analytic approach in settings with multivariate outcomes is latent variable analysis. Many definitions of latent variables appear in the literature, and we refer interested readers to more thorough discussions

*Corresponding author: David Benkeser, Department of Biostatistics and Bioinformatics, Emory University, School of Public Health, Atlanta, 30322, USA, E-mail: benkeser@emory.edu. <https://orcid.org/0000-0002-1019-8343>

Andrew Mertens and John M. Colford, Jr.: Department of Epidemiology, University of California, Berkeley, Berkeley, USA, E-mail: amertens@berkeley.edu (A. Mertens), jcolford@berkeley.edu (J.M. Colford)

Alan Hubbard and Mark J. van der Laan: Department of Biostatistics, University of California, Berkeley, Berkeley, USA, E-mail: hubbard@berkeley.edu (A. Hubbard), laan@berkeley.edu (M.J. van der Laan)

Benjamin F. Arnold: Francis I. Proctor Foundation, University of California, San Francisco, USA, E-mail: ben.arnold@ucsf.edu

Aryeh D. Stein: Hubert Department of Global Health, Emory University Rollins School of Public Health, Atlanta, USA, E-mail: aryeh.stein@emory.edu

in [10–14]. A commonly used definition is that a latent variable is a hypothetical construct that cannot be measured by the researcher and that describes some underlying characteristic of the participant. Others have strongly rejected the use of latent variables, instead opting for empirical explanations of observed phenomenon [15]. For our purposes it suffices to say that latent variable analysis tends to entail an unsupervised grouping of the observed outcomes into a set of lower-dimensional latent features. One technique commonly employed to this end is principal components analysis (PCA). Researchers often use PCA to reduce the observed multivariate outcome to a low-dimensional (e.g., univariate) outcome and may examine the factor loadings to ascribe a-posteriori meaning to the reduced outcomes. Researchers further may use these outcomes to test for associations of covariates. However, such tests may fail to identify associations between predictors and outcomes due to the unsupervised nature of the outcomes' construction.

In this work, we propose an alternative method for measuring association between a multivariate outcome and a set of covariates. Our approach is related to canonical correlation analysis, but rather than maximizing the correlation between linear functions of covariates and outcomes, we maximize the predictive accuracy of a flexible machine learning algorithm and a convex combination of the outcome. The method identifies the univariate outcome that is “most easily predicted” by the covariates and a summary measure of how well that “easiest-to-predict” outcome may be predicted. This approach adapts to complex relationships between covariates and outcomes and identifies associations in situations where traditional canonical correlation does not. However, in contrast to recent nonparametric canonical correlation proposals, our method provides a measure of association on a familiar scale, and also provides asymptotically justified inference, including confidence intervals and hypothesis tests. In certain situations, our method may further provide a novel means of constructing a clinically interpretable latent variable. However, we view our method as a latent variable approach only in the sense of Harman (1960) [16], who described latent variables as a convenient means of summarizing a number of variables in many fewer factors. Our approach fits in this definition by providing a single summary measure of the strength of association between covariates and multivariate outcomes.

2 Methodology

Suppose we observe n independent copies of $O := (X_1, \dots, X_D, Y_1, \dots, Y_J)$, where $X := (X_1, \dots, X_D)$ is a D -dimensional vector of covariates and $Y := (Y_1, \dots, Y_J)$ is a J -dimensional vector of real-valued outcomes. Without loss of generality, we assume that each outcome has mean zero and standard deviation one and otherwise make no assumptions about the true distribution P_0 of O . We define $Y_\omega := \sum_{j=1}^J \omega_j Y_j$ as a convex combination of the outcomes, where $\omega_j \geq 0$ for all j and $\sum_{j=1}^J \omega_j = 1$. We are interested in the outcome Y_ω that is “easiest-to-predict” based on X using a particular prediction function. To solidify some key ideas, for now assume that we are given weights ω and a prediction function ψ_ω that takes x as input and outputs $\psi_\omega(x)$, a real-valued prediction of Y_ω . We assume for now that such a prediction function is available at the outset, e.g., a machine learning algorithm trained on external data.

A measure of how well a given algorithm ψ_ω predicts Y_ω is mean squared-error,

$$\xi_{0,\omega}(\psi_\omega) := E_0 \{ [Y_\omega - \psi_\omega(X)]^2 \}, \quad (1)$$

which measures the average squared distance between the prediction made by ψ_ω and the outcome Y_ω . Here we use the notation $E_0 \{f(X, Y)\}$ to denote the average value of $f(X, Y)$ under the true distribution P_0 . Mean squared-error can be scaled to obtain a measure of accuracy that does not depend on the scale of the outcome. Specifically, let $\mu_\omega := E_0(Y_\omega)$ denote the marginal mean of Y_ω , which by assumption equals zero, and define

$$\rho_{0,\omega}^2(\psi_\omega) := 1 - \frac{\xi_{0,\omega}(\psi_\omega)}{\xi_{0,\omega}(\mu_\omega)},$$

as the proportional reduction in mean squared-error when using ψ_ω as opposed to μ_ω to predict Y_ω . We refer to this as a nonparametric R^2 , as this measure may be interpreted similarly as the R^2 measure in the familiar

context of linear models: values near to one indicate that nearly all of the variance in Y_ω is explained by ψ_ω . However, we note that the squared notation here is a misnomer in that $\rho_{0,\omega}^2(\psi_\omega)$ can be negative, which indicates that the marginal mean of Y_ω provides a better fit than $\psi_{0,\omega}$. Nevertheless, we maintain the colloquial squared-notation. If ψ_ω provides reasonably accurate predictions of Y_ω then $\rho_{0,\omega}^2(\psi_\omega)$ may serve as a summary measure of the strength of the association between X and Y . For example, rejecting the null hypothesis that $\rho_{0,\omega}^2(\psi_\omega) = 0$ may allow us to conclude that there is an association between X and at least one of the components of Y .

With these ideas in mind, we turn now to the setting where we must simultaneously (i) develop a function ψ_ω to predict the combined outcome, (ii) learn a reasonable choice of weights ω , and (iii) based on choices, develop a sensible estimate of an association between X and Y . If ψ_ω is restricted to linear functions of X , and when X and Y are of relatively low dimension, then canonical correlation analysis provides a straightforward means of accomplishing these tasks. However, when more aggressive data fitting approaches are employed to develop ψ_ω , over fitting becomes a concern and may prevent simple correlative statistics from being used to provide accurate summaries of associations. To address these challenges, we propose an approach centered around cross-validation and data-adaptive target parameters [17]. Here, we provide a high-level overview of the procedure, which allows us to introduce notation to describe the parameter that we are ultimately interested in estimating. In the following subsection, we describe in detail how the full estimation procedure unfolds in a given data set.

First, we note that for a given ω , the maximizer of $\rho_{0,\omega}^2(\psi_\omega)$ over all functions mapping an observation x to a real value is $\psi_{0,\omega} = E_0(Y_\omega|X)$. Due to linearity of expectation,

$$E_0(Y_\omega|X) = E_0\left(\sum_{j=1}^J \omega_j Y_j\right) = \sum_{j=1}^J \omega_j E_0(Y_j|X). \quad (2)$$

This implies that for any ω , a prediction function for Y_ω may be constructed by building a prediction function (or equivalently, an estimate of the conditional mean) for each Y_j and weighting those prediction functions by ω . An alternative approach is to use multivariate regression methods to simultaneously predict the whole vector Y [18]. However, we find the univariate approach appealing in that it allows a rich collection of univariate learning methods to be utilized. One approach that is particularly appealing is regression stacking or super learning [19–21], the motivation for and details of which we discuss in detail in the Appendix. Given the choice of algorithm for developing a prediction function, we proceed with a description of our approach.

To begin, we assume that the data are randomly partitioned into K splits of approximately equal size. For $k = 1, \dots, K$, we use $T_{n,k} \subset \{1, \dots, n\}$ to denote the indices of observations in the k -th training sample (consisting of all the splits excepting the k -th) and $V_{n,k}$ to denote the indices of observations in the k -th validation sample (i.e., the k -th split itself).

To develop a procedure for sensibly determining outcome weights, we propose an algorithm Ω that maps a given set of data indices into weights that are most promising for prediction using the selected algorithm. The weights based on the k -th training sample are selected by maximizing a (nested) cross-validated measure of accuracy of predictions made by the learner of the combined outcome. We use $\omega_{n,k} = \Omega(T_{n,k})$ denote the weights selected in the k -th training set.

Next, the k -th training set is used to develop a prediction function for $Y_{\omega_{n,k}}$. We denote by $\Psi_{\omega_{n,k}}$ a mapping between a given set of data indices and an algorithm for predicting $Y_{\omega_{n,k}}$, and let $\psi_{n,k} = \Psi_{\omega_{n,k}}(T_{n,k})$. Thus, $\psi_{n,k}$ is a prediction function developed in the k -th training set that maps x to a prediction of $Y_{\omega_{n,k}}$, the composite outcome based on weights also developed in the k -th training set.

Finally, using the respective validation samples, we estimate the following data-adaptive parameter to assess association between X and Y ,

$$\rho_{0n,\Omega}^2(\Psi) = 1 - \frac{\sum_{k=1}^K \xi_{0,\omega_{n,k}}(\psi_{n,k})}{\sum_{k=1}^K \xi_{0,\omega_{n,k}}(\bar{Y}_{n,k})},$$

where $\bar{Y}_{n,k}$ is the training sample average of $Y_{\omega_{n,k}}$ and we remind readers of the definition of $\xi_{0,\cdot}$ in equation (1). The quantity $\rho_{0n,\Omega}^2(\Psi)$ describes average performance of the training sample-specific learners for predicting each training sample-specific outcome. The notation for this parameter deliberately connotes that the parameter is dependent on P_0 (via the mean squared error portion), the observed data (through the random training splits), our specific procedure for developing weights Ω , and the chosen learner Ψ .

2.1 Detailed approach

We now provide a step-by-step approach of how our procedure unfolds in a given data set. We include accompanying illustrations for the two-fold cross-validation case (Figures 1 and 2).

Step 1. Partition the data. Randomly partition the full data set into K splits of approximately equal size. Let $T_{n,k} \subset \{1, \dots, n\}$ to denote the indices of observations in the k -th training sample and $V_{n,k}$ to denote the indices of observations in the k -th validation sample for $k = 1, \dots, K$.

Step 2. Train a learner for the multivariate outcome using the training data. For each split k , use the training sample to train the chosen learner of the multivariate outcome, resulting in learner $\Psi(T_{n,k})$ for $k = 1, \dots, K$.

Figure 1: A flowchart illustrating computation of two-fold cross-validated measure of association between X and Y . This procedure maps the full data set into $\rho_{n,\Omega}^2(\Psi)$, an estimate of $\rho_{0n,\Omega}^2(\Psi)$. Notation: $T_{n,k}$, the k -th training sample; $V_{n,k}$, the k -th validation sample; $\Psi_j(T_{n,k})$, the learner for the j -th outcome fit in the k -th training sample; $\Omega(T_{n,k})$, outcome weights computed in the k -th training sample; $\Psi_{\Omega(T_{n,k})}(T_{n,k})$, the learner fits from the k -th training sample combined using weights also computed in the k -th training sample; $\rho_{n,\Omega}^2(\Psi_{\Omega})$, the cross-validated nonparametric R^2 for the composite learner.

Figure 2: A flowchart illustrating two-fold outcome weighting procedure, which maps a set of observed data indices, $T_{n,k}$ into $\omega_{n,k}$, a vector of convex weights for the outcome Y . Notation: $T_{n,k}$, the k -th training sample; $V_{n,k}$, the k -th validation sample; $T_{n,k,k'}$, the k' -th nested training sample; $V_{n,k,k'}$, the k' -th nested validation sample; $\Psi_{\omega}(T_{n,k,k'})$, the prediction function for the multivariate outcome Y_{ω} trained using the k' -th nested training sample; $\rho_{n,k,\omega}^2(\Psi_{\omega})$, the cross-validated R^2 for the composite learner.

Step 3. *Fit outcome weights using the training data.* We now detail our procedure Ω for mapping a given training set into a vector of weights. The following procedure (illustrated in Figure 2 for $K' = 2$) is applied in each of the K training samples.

Step 3a. *Partition the given training data.* Given a training data set consisting of observations $\{O_i : i \in T_{n,k}\}$, randomly partition the data into K' nested splits of approximately equal size. Let $T_{n,k,k'} \subset T_{n,k}$ denote the indices of observations in the k' -th training sample nested in $T_{n,k}$ and similarly use $V_{n,k,k'}$ to denote the indices of observations in the k' -th validation sample nested in $T_{n,k}$ for $k' = 1, \dots, K'$.

Step 3b. *Fit learner for each outcome in training data.* For $k' = 1, \dots, K'$, develop a prediction function using observations $T_{n,k,k'}$ of the multivariate outcome, resulting in prediction function $\Psi_{\omega}(T_{n,k,k'})$ for a given set of weights ω . We note that if the procedure for developing the prediction function itself involves cross-validation, this step will involve further nested cross-validation of $T_{n,k,k'}$.

Step 3c. *Compute cross-validated R^2 .* For any ω , we can use the data in the k -th validation sample to compute an estimate of mean squared-error of $\Psi_{\omega}(T_{n,k,k'})$,

$$\xi_{n,\omega,k,k'}(\Psi_{\omega}(T_{n,k,k'})) = \frac{1}{|V_{n,k,k'}|} \sum_{i \in V_{n,k,k'}} \{Y_{\omega,i} - \Psi_{\omega}(T_{n,k,k'})(X_i)\}^2.$$

The cross-validated mean squared-error is the average over the validation folds, $\xi_{n,\omega,k}(\Psi_{\omega}) = \frac{1}{K'} \sum_{k'=1}^{K'} \xi_{n,\omega,k,k'}(\Psi_{\omega}(T_{n,k,k'}))$. Similarly, define

$$\bar{Y}_{\omega,n,k} = \frac{1}{|T_{n,k}|} \sum_{i \in T_{n,k}} Y_{\omega,i} \text{ and } \xi_{n,\omega}(\bar{Y}_{\omega,n,k}) = \frac{1}{|T_{n,k}|} \sum_{i \in T_{n,k}} \{Y_{\omega,i} - \bar{Y}_{\omega,n,k}\}^2,$$

as the empirical average weighted outcome among observations in $T_{n,k}$ and the empirical mean squared-error for predicting the composite outcome Y_{ω} using the training sample mean $\bar{Y}_{\omega,n,k}$. For any ω , we can compute the cross-validated nonparametric R^2 measure as

$$\rho_{n,k,\omega}^2(\Psi_{\omega}) = 1 - \frac{\xi_{n,\omega}(\Psi_{\omega})}{\xi_{n,\omega}(\bar{Y}_{\omega,n,k})}.$$

Step 3d. *Maximize cross-validated R^2 .* Compute $\omega_{n,k} = \arg\max_{\omega} \rho_{n,k,\omega}^2(\Psi_{\omega})$ using standard optimization software. In our simulation and data analysis, we used an interior algorithm with augmented Lagrange multipliers [22, 23].

Step 4. *Combine learners based on outcome weights.* Given weights $\omega_{n,k}$, compute the prediction function $\psi_{n,k} = \Psi_{\omega_{n,k}}(T_{n,k})$ for predicting $Y_{\omega_{n,k}}$.

Step 5. *Compute cross-validated nonparametric R^2 .* Using each validation sample, compute the mean squared-error for the combined learner,

$$\xi_{n,k}(\psi_{n,k}) = \frac{1}{|V_{n,k}|} \sum_{i \in V_{n,k}} \{Y_{\omega_{n,k},i} - \psi_{n,k}(X_i)\}^2,$$

which provides an evaluation of the performance of the composite learner $\psi_{n,k}$ for predicting the composite outcome $Y_{\omega_{n,k}}$ based on the training sample-specific weights. We define $\xi_{n,\Omega}(\Psi_{\Omega}) = \frac{1}{K} \sum_{k=1}^K \xi_{n,k}(\psi_{n,k})$ as the average measure across the K splits. Similarly, estimate the mean squared-error of the composite empirical mean

$$\bar{Y}_{\omega_{n,k}} = \frac{1}{|T_{n,k}|} \sum_{i \in T_{n,k}} Y_{\omega_{n,k},i},$$

via

$$\xi_{n,k}(\bar{Y}_{\omega_{n,k}}) = \frac{1}{|V_{n,k}|} \sum_{i \in V_{n,k}} \{Y_{\omega_{n,k},i} - \bar{Y}_{\omega_{n,k}}\}^2,$$

and define $\xi_{n,\Omega}(\bar{Y}_{\Omega}) = \frac{1}{K} \sum_{k=1}^K \xi_{n,k}(\bar{Y}_{\omega_{n,k}})$. The estimate of $\rho_{0n,\Omega}^2(\Psi)$ is

$$\rho_{n,\Omega}^2(\Psi) = 1 - \frac{\xi_{n,\Omega}(\Psi_{\Omega})}{\xi_{n,\Omega}(\bar{Y}_{\Omega})}.$$

2.2 Inference

An asymptotically justified confidence interval can be constructed for $\rho_{0n,\Omega}^2(\Psi)$ using influence function-based methodology [17]. We propose to construct a Wald-style $100(1-\alpha)\%$ confidence interval based on $\log\{\xi_{n,\Omega}(\Psi_{\Omega})/\xi_{n,\Omega}(\bar{Y}_{\Omega})\}$ and back-transform this interval to obtain an interval on the original scale. Thus, our interval is of the form

$$1 - \exp \left[\log \left\{ \frac{\xi_{n,\Omega}(\Psi_{\Omega})}{\xi_{n,\Omega}(\bar{Y}_{\Omega})} \right\} \pm z_{1-\alpha/2} \frac{\sigma_n}{n^{1/2}} \right],$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution and σ_n^2 is a consistent estimate of the asymptotic variance of $\log\{\xi_{n,\Omega}(\Psi_{\Omega})/\xi_{n,\Omega}(\bar{Y}_{\Omega})\}$. Similarly, a level α one-sided test of the null hypothesis of no association between X and Y can be performed by rejecting the null hypothesis whenever

$$\frac{\log\{\xi_{n,\Omega}(\Psi_{\Omega})/\xi_{n,\Omega}(\bar{Y}_{\Omega})\}}{\sigma_n/n^{1/2}} < z_{\alpha}.$$

A closed-form variance estimator is constructed as follows. For $k = 1, \dots, J$, we define

$$D_{n,k,\Psi,\Omega}(O) = \{Y_{\omega_{n,k}} - \psi_{n,k}(X)\}^2 - \xi_{n,k}(\psi_{n,k}), \text{ and}$$

$$D_{n,k,\bar{Y},\Omega}(O) = \{Y_{\omega_{n,k}} - \bar{Y}_{\omega_{n,k}}\}^2 - \xi_{n,k}(\bar{Y}_{\omega_{n,k}}).$$

These equations represent cross-validated estimates of the influence function of $\xi_{n,k}(\psi_{n,k})$ and $\xi_{n,k}(\bar{Y}_{\omega_{n,k}})$, respectively. We define $I_{n,k} = (D_{n,k,\Psi,\Omega}, D_{n,k,\bar{Y},\Omega})^T$ as the bivariate estimated influence function and use the delta method to obtain a cross-validated estimate of the influence function of $\log\{\xi_{n,\Omega}(\Psi_{\Omega})/\xi_{n,\Omega}(\bar{Y}_{\Omega})\}$. The estimated gradient of this transformation at the parameter estimates is $g_n = (\xi_{n,\Omega}^{-1}(\Psi_{\Omega}), -\xi_{n,\Omega}^{-1}(\bar{Y}_{\Omega}))^T$. A consistent variance estimator is

$$\sigma_n^2 = g_n^T \left\{ \frac{1}{K} \sum_{k=1}^K \frac{1}{|V_{n,k}|} \sum_{i \in V_{n,k}} I_{n,k}(O_i) I_{n,k}^T(O_i) \right\} g_n.$$

3 Variable importance

Often we are not only be interested in an overall summary of the association between X and Y , but also in the relative importance of each component of X . Machine learning is often criticized as a “black box” approach, in which it is difficult to understand the relative contributions of different variables [24]. To provide better understanding of the “black box,” we propose to study differences in the estimated association between X and Y when considering different subsets of variables. Our proposal is similar to variable importance measures proposed for specific machine learning algorithms, such as random forests [25], because they measure the change in predictive performance with and without each predictor variable considered. Although existing

approaches may have poorly behaved statistical inference [26], the present approach yields straightforward, asymptotically justified inference.

To assess the importance of a subset $S \subset \{1, \dots, D\}$ of variables $\tilde{X} = \{X_s : s \in S\}$, we propose to repeat our procedure, but restrict only to variables not in this subset. We obtain an estimate $\rho_{n, \hat{\Omega}}^2(\tilde{\Psi}_{\hat{\Omega}})$ of the cross-validated performance measure for the composite prediction algorithm based on this subset of variables. The joint importance of the variables in $\tilde{X}$ is defined by a contrast between $\rho_{0n, \Omega}^2(\Psi_{\Omega})$ and $\rho_{0n, \hat{\Omega}}^2(\Psi_{\hat{\Omega}})$ such as

$$\Delta_{0n}(\Psi_{\Omega}, \tilde{\Psi}_{\hat{\Omega}}) = \rho_{0n, \Omega}^2(\Psi_{\Omega}) - \rho_{0n, \hat{\Omega}}^2(\Psi_{\hat{\Omega}}). \quad (3)$$

Point estimates for (3) are constructed by plugging in the estimates, and asymptotically justified confidence intervals and hypothesis tests about these estimates are constructed using influence functions as in the previous section.

4 Simulations

We evaluated the finite-sample performance of the proposed estimators in two simulations. In the first, we studied the operating characteristics of our estimator. This simulation confirms that the asymptotic theory developed in previous sections leads to reasonable finite-sample performance for the estimators. In the second simulation, we compared our proposed method to existing methods for assessing associations between X and Y . This simulation shows that our method correctly identifies associations in situations where existing methods do not.

4.1 Simulation 1: operating characteristics

We generated 1000 simulated data sets by independently sampling n copies of $X_1, X_2, X_3, X_7, X_8, X_9$ from a Uniform(0,4) distribution and n copies of X_4, X_5, X_6 from a Bernoulli distribution with success probabilities of 0.75, 0.25, and 0.5. We sampled $\epsilon_1, \epsilon_2, \epsilon_3$ independently from a Normal (0,5) distribution and let

$$Y_1 = X_1 + 2X_2 + 4X_3 + X_4 + 2X_5 + 4X_6 + 2X_7 + \epsilon_1,$$

$$Y_2 = X_1 + 2X_2 + 4X_3 + X_4 + 2X_5 + 4X_6 + 2X_8 + \epsilon_2, \text{ and}$$

$$Y_3 = X_1 + 2X_2 + 4X_3 + X_4 + 2X_5 + 4X_6 + 2X_9 + \epsilon_3.$$

The optimal R -squared for each outcome is about 0.60. The true optimal weighting scheme for the outcomes is $\omega_0 = (1/3, 1/3, 1/3)$ and the true optimal R -squared for the optimally composite outcome is approximately 0.81. We considered sample sizes of $n = 100, 500, 1000, 5000$. For computational simplicity, we considered a small learner library consisting of a main terms linear regression (function `SL.glm` in `SuperLearner` package), an intercept-only regression (`SL.mean`), and a forward-stepwise selection linear regression model (`SL.step.forward`). We used 10 folds for each layer of cross-validation, $K=K^* = J = 10$.

Figure 3: Results of the first simulation study. Left: Bias of $\rho_{n, \Omega}^2(\Psi_{\Omega})$. Right: Estimated coverage of nominal 95% intervals (with Monte Carlo confidence intervals) for the proposed confidence interval.

The estimates of predictive performance were biased downwards in smaller sample sizes (Figure 3). However, even with $n = 100$ the average estimated performance was only about 5% too small. The confidence interval coverage was less than nominal for small sample sizes, but had nominal coverage in larger samples.

We excluded X_2 and then excluded X_7 to estimate the additive difference in R -squared values (equation (3)) for these two variables. Examining the data-generating mechanism, we note that X_2 was important for predicting each individual outcome and should be important for predicting the composite outcome. The optimal R -squared for predicting each individual outcome without X_2 was 0.52, and the additive importance of X_2 was $0.60 - 0.52 = 0.08$. For the composite outcome, the optimal weightings were unchanged (1/3, 1/3, 1/3), but the optimal R -squared without X_2 decreased to 0.68 and the additive importance of X_2 for the composite outcome was $0.81 - 0.68 = 0.13$. In contrast, X_7 was important only for predicting Y_1 and had no effect on predicting Y_2 or Y_3 . Therefore, the optimal composite outcome was expected to upweight Y_2 and Y_3 when excluding X_7 , and the importance of X_7 for predicting the optimal composite outcome may be minimal. The optimal weights without X_7 were (0.24, 0.38, 0.38), whereas the optimal R -squared without X_7 was 0.79, leading to an additive importance of $0.81 - 0.79 = 0.02$.

The importance measures for X_2 exhibited substantial bias (25% truth) when $n = 100$, but both measures were unbiased with sample sizes >1000 (Figure 4). The nominal 95% confidence interval coverage was less than nominal in small samples, but the coverage was $>90\%$ for sample sizes >500 .

4.2 Simulation 2: power of associative hypothesis tests

In this simulation, X was simulated as above. The outcome Y was simulated as a 10-dimensional multivariate normal variate with mean $\mu = (\mu_1(x), \dots, \mu_{10}(x))^T$ and covariance matrix Σ . Letting $\Sigma_{i,j}$ denote the (i, j) entry in Σ , we set $\Sigma_{i,i} = 5$ for $i = 1, \dots, 10$. We also set $\Sigma_{1,2} = \Sigma_{1,3} = \Sigma_{4,7} = -2$ and $\Sigma_{3,9} = \Sigma_{4,9} = \Sigma_{5,9} = 2$, while also appropriately setting the corresponding lower triangle elements to these values. All other off-diagonal elements were set to zero. We studied three settings. The first setting had no association between X and Y , and we set $\mu_j(x) = 0$ for $j = 1, \dots, 10$. The second setting had a linear association between X and one component of Y . Here, we set $\mu_6(x) = -2 + 0.75x_1$ and $\mu_j(x) = 0$ for $j \neq 6$. The third setting had a nonlinear association between X and one component of Y . Here, we set $\mu_6(x) = -2 + 0.75(x - 2)^2$ and $\mu_j(x) = 0$ for $j \neq 6$.

We studied the power of level 0.05 tests of the hypothesis of no association between X and Y . Power was estimated as 100 times the proportion of the 1,000 total simulations where the null hypothesis was rejected. We tested this hypothesis with the proposed one-sided hypothesis test with all layers of cross validation set to five-fold. We used a super learner library that included an intercept-only regression, a main terms regression, and a generalized additive model [27]. We also tested this hypothesis using four common tests of canonical correlation: Wilks' Λ , the Hotelling-Lawley trace, the Pillai-Bartlett trace, and Roy's largest root. Rather than comparing these statistics to their respective approximate asymptotic distributions, we used permutation tests to compute p -values [28]. We also studied the power of a test based on a principal components approach. For this test, we constructed a composite outcome based on the first principal component of Y and subsequently fit

Figure 4: Results of second simulation study. Left: Bias of the estimated variable importance measure comparing a super learning that uses all of the variable to one that omits either X_2 (triangle) or X_7 (diamond). Right: Estimated coverage for nominal 95% intervals (with Monte Carlo confidence intervals) for the proposed confidence intervals.

Figure 5: Power of various level 0.05 tests for rejecting the null hypothesis of no association between X and Y under the null hypothesis (left), linear association (middle), and nonlinear association (right).

a main-terms linear regression on X . The null hypothesis of no association was rejected whenever the p -value associated with the global F -test for this linear regression was less than 0.05.

Under the null hypothesis, each canonical correlation-based and the PCA-based test had approximately nominal type I error (Figure 5, left panel). However, our proposed test was found to be conservative, falsely rejecting the null hypothesis less than 1% of the time. In the second scenario, the canonical correlation-based tests had greater power to reject the null hypothesis than our proposed test at the two smallest sample sizes (middle panel). The PCA-based test had poor power, as the first principal component fails to adequately reflect the component of Y for which a relationship with X exists. In the third scenario, only our proposed test had power to correctly reject the null hypothesis (right panel).

4.3 Additional simulations

In the web supplementary material, we include several additional simulations that explore various settings. In particular, we repeat simulation one using a more aggressive super learner library and using a more complex data generating process. The results in terms of bias and confidence interval coverage are similar as with the results presented here. We also repeated simulation 2 using covariates that each have a Normal distribution, a setting classically associated with canonical correlation analysis. Our results are again largely similar to the results of simulation 2 presented here.

5 Data analysis

Neurocognitive impairment may affect 250 million children under five years globally [30]. The Healthy birth, growth, and development knowledge integration initiative was established, in part, to inform global public health programs toward optimizing neurocognitive development [31]. Neurocognitive development is often studied through studies that enroll pregnant women and follow their children through early childhood and adolescence. Covariate information about the child's parents, environment, and somatic growth is recorded at regular intervals and in early adolescence, children complete tests that measure diverse domains of neurocognitive development such as motor, mathematics, and language skills. Researchers are often interested in assessing the correlation between covariate information and neurocognitive development. Such an assessment may be useful for developing effective prediction algorithms that identify children at high risk for neurocognitive deficits [32].

The ongoing Cebu Longitudinal Health and Nutrition Study (CLHNS) enrolled Filipino women who gave birth in 1983–1984, and the children born to these women have been followed prospectively [33, 34]. The long-term follow-up with these children has enabled researchers to quantify the long-term effects of prenatal and early childhood nutrition and health on outcomes (adolescent and adult health, economics, and development) [35, 36]. We are interested in assessing the strength of correlation between prenatal and early childhood data and later schooling achievement. We are also interested in understanding the extent to which somatic growth (height and body weight) early in life associates with later neurocognitive outcomes [37, 38]. In 1994, achievement tests were administered to 2166 children in three subjects: mathematics, and English and Cebuano languages.

We applied the present methods to assess association of the three test scores with variables collected from birth to age two years (Table 1). For variables that had missing values, we created an indicator of missingness, and missing values of the original variable were set = 0. The library of candidate super learner algorithms included a random forest, gradient boosted machines, and elastic net regression algorithms. The tuning parameters for each algorithm were chosen via nested five-fold cross-validation. We used 10-fold cross-validation for each layer of cross-validation.

Table 1: Variables used to predict achievement test scores in Cebu Longitudinal Health and Nutrition Study analysis.

Group	Variables
Health care	Health care access, use of preventive health care
Household	Child: adult ratio, child dependency ratio, crowding index, urban score
Socioeconomic status	Total income, socioeconomic status
Water and sanitation	Sanitation, access to clean water
Parental	Mother age, father age, mother height, mother education (y), Father education (y), marital status, mother age first child, parity
Growth	Weight-for-age Z-score, height-for-age z-score [29] (0, 6, 12, 18, 24 mo)
Other	Mother smoked during pregnancy, child's sex, gestational age at birth

Figure 6: The change in R -squared for predicting each outcome and the composite outcome when eliminating groups of variables. Abbreviations: CI, confidence interval; SES, socioeconomic status.

The estimated cross-validated R -squared for predicting test scores was 0.24 (95% CI: 0.21, 0.27) for mathematics, 0.31 (95% CI: 0.28, 0.34) for English, and 0.23 (95% CI: 0.20, 0.26) for Cebuano. The estimated association with the composite outcome was 0.32 (95% CI: 0.28, 0.35). The estimated association with the optimally weighted outcome was only slightly higher than predicting an equally weighted outcome (0.30, 95% CI: 0.27, 0.34), as well as an outcome weight based on a linear combination based on the first principal component of the test scores (0.28, 95% CI: 0.25, 0.31).

We computed variable importance measures by repeating the procedure, eliminating groups of variables and estimating the additive change in performance. We found that the child's sex and parental information were responsible for the largest proportion of the association with achievement test score performance (Figure 6). However, the somatic growth variables, as a group, modestly increased the association, with an estimated change in association of 0.01 (95% CI: 0.00, 0.02; p -value = 0.02).

6 Discussion

Our proposed method provides a new means of summarizing associations with multivariate outcomes and can provide powerful tests for association in large samples when complex and nonlinear relationships are present. We found that existing tests provide greater power in small samples against alternatives with linear relationships between covariates and outcomes. This is unsurprising as our approach relies on several layers of cross-validation, which may stretch small samples too thin. Thus, we suggest that in small samples, it may be preferable to adopt traditional approaches to assessing associations in multivariate settings.

We also found that our test was conservative under the null hypothesis. This can be explained by the fact that the true value of our data-adaptive target parameter was often less than zero, while our test was based on testing a value of zero for this parameter. We suggest that a permutation test could be used to construct a hypothesis test with better operating characteristics; however, this may often prove to be computationally intractable in practice. In such cases, we may instead opt for the more conservative, but less computationally intensive test.

A possible criticism of our approach is the data-adaptive nature of the parameter $\rho_{0n,\Omega}^2(\Psi_\Omega)$. One could argue that this parameter is only a useful measure of association insofar as the relevant prediction functions approximate ψ_{0,ω_0} , the unknown true conditional mean of the optimally combined outcome. This strengthens the argument for using super learning to a prediction function that is as accurate as possible. It would be interesting to extend the approach to estimate target parameters where rather than conditioning on the training-sample-specific learners in the parameter definition, we instead estimate some correlation that depends on ψ_{0,ω_0} , as in [39]. However, this would require a more involved estimation procedure, such as one-step estimation or targeted minimum loss estimation. Similarly, our test of the null hypothesis of no association will only be powered against alternatives where linear combinations of outcomes are associated with X . In the presence of more complex relationships, our test may not have power.

An interesting extension suggested by a review is to use cross-validation to select between our proposed measure of association and traditional canonical correlation analysis (CCA). This would require developing an objective criteria that could be used in a given training sample to determine whether there was sufficient evidence to suggest that CCA provides an adequate summary of the association between X and Y . For example, if a linear model for each component of Y has the highest (nested) cross-validated risk in a given training sample, then we may believe that CCA will appropriately describe the association between X and Y . Then, we could use the corresponding validation sample to compute the canonical correlation. Such a modification of our procedure may yield increased power in situations where linear relationships truly are present, while providing an adaptive way to switch to a nonlinear approach. We leave to future work the theoretical and practical study of such an approach. Future extensions may

consider nonlinear combinations of the outcome, perhaps via alternating conditional expectation [40], which would bring our proposal closer in scope to that of kernel CCA methods.

Though our proposal is computationally intensive, the vast majority of the computational burden lies in fitting the candidate learners, which can be implemented in a parallelized framework. The computational burden of the procedure may be further reduced by clever choices of number of cross-validation folds. For example, in the context of using super learning to separately develop a prediction function of each of the J outcomes, if $K = K'$ and we also build a super learner based on K -fold cross-validation, then the procedure requires fitting $K^3 \sum_{j=1}^J M_j$ total candidate learners. However, if $K' = K - 1$ and we use $(K - 2)$ -fold cross-validation to construct the super learner, it is possible to re-use candidate learner fits at different points in the procedure, thereby drastically decreasing the computational burden of the procedure. In fact, we can show that this approach requires only $(K^3 + 5K)/6 \times \sum_{j=1}^J M_j$ candidate learner fits. Thus, for $K = 10$, the latter approach is expected to execute in 17.5% the time it would take if all layers of cross-validation are set equal to 10. An R package implementing our methods is available (<https://github.com/benkeser/cvma>).

It may also be of interest to consider extensions for estimating covariate-adjusted effects of variables on data-adaptively-combined outcome; estimation of these effects and associated inference may be facilitated via cross-validated targeted maximum likelihood [17, 41].

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: This research was funded by National Heart, Lung, and Blood Institute (grant no. 1R01HL137808-01A1) and Bill and Melinda Gates Foundation (grant no. OPP1147962).

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

Appendix: Super learner

An appealing approach for developing a prediction function is regression stacking [19, 20], also known as super learning [21]. Here we describe how to develop a super learner separately for each of the J outcomes, though in theory, this approach could be extended to simultaneous prediction of the vector Y . We first construct a library of M_j candidate learners for each of the J outcomes. The library of estimators can include parametric model-based estimators as well as machine learning algorithms. Parametric model-based methods could include linear regression of Y_j on X , while machine learning-based methods could include random forests [25], gradient boosted machines [42], or deep neural networks [43]. If X is high-dimensional, learners can incorporate dimension-reduction strategies. For example, we may include a main-terms linear regression of Y_j on all variables in X , while also including a linear regression using only a subset of X chosen based on their univariate association with Y_j . The library of learners can also include different choices of tuning parameters for machine learning algorithms, an important consideration for methods that require proper selection of many tuning parameters (e.g., deep learning).

Because there is no way to know a-priori which of these myriad estimators will be best, a cross-validated empirical criterion is used to adaptively combine, or ensemble, the various learners. This approach is appealing in the present application in that the best learner for each outcome might be different. By allowing the data to determine the best ensemble of learners for each outcome, we expect to obtain a better fit over all outcomes and thereby obtain a more accurate summary of the association between X and Y . Indeed, the super learner is known to be optimal in the sense that, under mild assumptions, its goodness-of-fit is essentially equivalent to the (unknown) best-fitting candidate estimator [44, 45]. A full treatment of super learning [21] and the R package [46] are available. Here, we provide a brief overview of the procedure.

Super learning is often based on K' -fold cross validation, and we illustrate the method for $K' = 2$ in Figure 7. The super learner for a given outcome Y_j may be implemented in the following steps:

Figure 7: A flowchart illustrating a two-fold super learner procedure, which maps a set of observed data indices, F_n^* into $\Psi_{j,\beta_{n,j}}(F_n^*)$ an ensemble prediction function for outcome Y_j . Notation: $T_{n,k}$, the k -th training sample; $V_{n,k}$, the k -th validation sample; $\Psi_{j,m}(T_{n,k})$, the m -th learner for the j -th outcome fit in the k -th training sample; $\xi_{n,k}(\Psi_{j,m})$, the k -th cross-validated mean squared-error; $\xi_n(\Psi_{j,m})$, cross-validated mean squared-error; $\beta_{n,j}$, a J -length vector of super learner weights; $\Psi_{j,m}(F_n^*)$, the m -th learner for the j -th outcome fit using $\{O_i : i \in F_n^*\}$.

Step 1: Partition the data. Starting with a data set consisting of observations $\{O_i : i \in F_n^*\}$ for $F_n^* \subseteq \{1, \dots, n\}$, we randomly partition the data into K^* splits of approximately equal size. We define the k^* -th training sample as the observed data with observations in the k^* -th split removed. We use $T_{n,k}^* \subset F_n^*$ to denote the indices of observations in the k^* -th training sample and $V_{n,k}^*$ to denote the indices of observations in the k^* -th validation sample for $k^* = 1, \dots, K^*$.

Step 2: Fit learners in training data. We now use the observations in the training sample to fit each of the M_j candidate learners. We use $\Psi_{j,m}$ to denote the m -th candidate learner for the j -th outcome. The learner $\Psi_{j,m}$ takes as input a set of indices and returns a prediction function for Y_j . We use $\Psi_{j,m}(T_{n,k}^*)$ to denote the m -th candidate learner for the j -th outcome fit in the k^* -th training sample. For example, $\Psi_{j,m}$ might correspond with fitting an ordinary least squares linear regression of Y_j on X , in which case $\Psi_{j,m}(T_{n,k}^*)$ corresponds to the fitted linear predictor based on the regression fit in the k^* -th training sample and $\Psi_{j,m}(T_{n,k}^*)(x)$ corresponds to the linear predictor of the fitted regression evaluated at the observation x .

Step 3: Evaluate learners in validation data. Next, we use the data in the k^* -th validation sample to evaluate each of the learners $\Psi_{j,m}(T_{n,k}^*)$, $m = 1, \dots, M_j$ via

$$\xi_{n,k^*}(\Psi_{j,m}) := \frac{1}{|V_{n,k^*}^*|} \sum_{i \in V_{n,k^*}^*} \{Y_{j,i} - \Psi_{j,m}(T_{n,k^*}^*)(X_i)\}^2,$$

where we use $|V_{n,k^*}^*|$ to denote the number of observations in the k^* -th validation sample. We obtain an overall estimate of the performance of the m -th learner via

$$\xi_n(\Psi_{j,m}) := \frac{1}{K^*} \sum_{k^*=1}^{K^*} \xi_{n,k^*}(\Psi_{j,m}).$$

Step 4: Find ensemble weights. The cross-validation selector $\tilde{m}$ is defined as the single learner with the smallest value of ξ_n and one may use $\Psi_{j,\tilde{m}}(F_n^*)$ as the learner for the j -th outcome, i.e., we refit the cross-validation-selected learner using all the data and use this as prediction function. However, it is often possible to improve on the fit of the cross-validation selector by considering ensemble learners of the form $\Psi_{j,\beta} = \sum_m \beta_{j,m} \Psi_{j,m}$, where the weights $\beta_{j,m}$ are non-negative for all m and sum to 1. The prediction function $\Psi_{j,\beta}$ takes as input X , computes the prediction of Y_j using each of the M_j learners, and returns a convex combination of these predictions. We can compute $\beta_{n,j} := \arg \min_{\beta} \xi_n(\Psi_{j,\beta})$, the choice of weights that minimizes cross-validated mean squared-error, using standard optimization software.

Step 5: Refit learners using full data. We next refit the learners that received non-zero weight in $\beta_{n,j}$ using all observations in F_n^* . We denote these fits $\Psi_{j,m}(F_n^*)$ for $m = 1, \dots, M_j$.

Step 6: Combine learners into super learner. The super learner is the convex combination of the M_j learners using weights $\beta_{n,j}$ and is denoted by $\Psi_j(F_n^*) = \Psi_{j,\beta_{n,j}}(F_n^*)$.

References

- Hotelling H. The most predictable criterion. *J Educ Psychol* 1935;26:139.
- Hotelling H. Relations between two sets of variates. *Biometrika* 1936;28:321–77.
- Wilks SS. Certain generalizations in the analysis of variance. *Biometrika* 1932;24:471–94.
- Bartlett M. The statistical significance of canonical correlations. *Biometrika* 1941;32:29–37.
- Pillai KS. On the distribution of the largest or the smallest root of a matrix in multivariate analysis. *Biometrika* 1956;43: 122–7.
- Roy SN. The individual sampling distribution of the maximum, the minimum and any intermediate of the p-statistics on the null-hypothesis. *Sankhyā: Indian J Stat* 1945;7:133–58.
- Nandy RR, Cordes D. Novel nonparametric approach to canonical correlation analysis with applications to low cnr functional mri data. *Magn Reson Med* 2003;50:354–65.
- Andrew G, Arora R, Bilmes J, Livescu K. Deep canonical correlation analysis. In: *International conference on machine learning*; 2013:1247–55 pp.
- Michaeli T, Wang W, Livescu K. Nonparametric canonical correlation analysis. In: *International conference on machine learning*; 2016:1967–76 pp.
- Glymour C, Scheines R, Spirtes P, Kelly K. *Discovering causal structure: artificial intelligence, philosophy of science, and statistical modeling*. New York: Academic Press; 1987.
- Sobel M. Causal inference in latent variable models. In: von Eye A, Clogg C, editors. *Latent variables analysis: applications for developmental research*. Thousand Oaks, CA, USA: Sage Publications, Inc.; 1994:3–35 pp.
- Edwards JR, Bagozzi RP. On the nature and direction of relationships between constructs and measures. *Psychol Methods* 2000;5:155.
- Häggglund G. Milestones in the history of factor analysis. In: *Structural equation modeling: present and future*. Lincolnwood, IL, USA: Scientific Software Inc.; 2001.
- Bollen K. Latent variables in psychology and the social sciences. *Annu Rev Psychol* 2002;53:605–34.
- Skinner BF. *About behaviorism*: Vintage; 1976.
- Harman HH. *Modern factor analysis*. University of Chicago Press; 1960.
- Hubbard AE, Kherad-Pajouh S, van der Laan MJ. Statistical inference for data adaptive target parameters. *Int J Biostat* 2016;12: 3–19.
- Izenman AJ. *Modern multivariate statistical techniques: regression, classification and manifold learning*. New York, USA: Springer; 2008.
- Wolpert DH. Stacked generalization. *Neural Network* 1992;5:241–59.
- Breiman L. Stacked regressions. *Mach Learn* 1996;24:49–64.
- van der Laan MJ, Polley EC. Super learner. *Stat Appl Genet Mol Biol* 2007;6:1–23.
- Ye Y. Interior algorithms for linear, quadratic, and linearly constrained non-linear programming, Ph.D. thesis: Department of ESS, Stanford University; 1987.
- Ghalanos A, Theussl S. Rsolnp: general non-linear optimization using augmented lagrange multiplier method, r package version 1.16; 2015.
- Freitas AA. Comprehensible classification models: a position paper. *ACM SIGKDD Explor Newslett* 2013;15:1–10.
- Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
- Strobl C, Zeileis A. Danger: high power! – exploring the statistical properties of a test for random forest variable importance: Department of Statistics: University of Munich Technical Reports Accessed: November 17, 2016.
- Hastie T, Tibshirani R. *Generalized additive models*: Wiley Online Library; 1990.
- Menzel U. CCP: Significance tests for canonical correlation analysis (CCA), r package version 1.1; 2012. Available from: <https://CRAN.R-project.org/package=CCP>.
- WHO Multicentre Growth Reference Study Group. WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development. Geneva: World Health Organization; 2006.
- Black M, Walker S, Fernald L, Andersen C, DiGirolamo A, Lu C, et al. Lancet early childhood development series steering committee, early childhood development coming of age: science through the life course. *Lancet* 2017;389:77–90.
- Jumbe N, Murray JC, Kern S. Data sharing and inductive learning – toward healthy birth, growth, and development. *N Engl J Med* 2016;374:2415–7.

32. Walker SP, Wachs TD, Gardner JM, Lozoff B, Wasserman GA, Pollitt E, et al. Child development: risk factors for adverse outcomes in developing countries. *Lancet* 2007;369:145–57.
33. Cebu Longitudinal Health and Nutrition Survey. Available from: <http://www.cpc.unc.edu/projects/cebu> [Accessed: November 17, 2016].
34. Adair LS, Popkin BM, Akin JS, Guilkey DK, Gultiano S, Borja J, et al. Cohort profile: the cebu longitudinal health and nutrition survey. *Int J Epidemiol* 2011;619–25. <https://doi.org/10.1093/ije/dyq085>.
35. Daniels MC, Adair LS. Growth in young Filipino children predicts schooling trajectories through high school. *J Nutr* 2004;134:1439–46.
36. Daniels MC, Adair LS. Breast-feeding influences cognitive development in Filipino children. *J Nutr* 2005;135:2589–95.
37. Carba DB, Tan VL, Adair LS. Early childhood length-for-age is associated with the work status of Filipino young adults. *Econ Hum Biol* 2009;7:7–17.
38. Smithers LG, Lynch JW, Yang S, Dahhou M, Kramer MS. Impact of neonatal growth on IQ and behavior at early school age. *Pediatrics* 2013;132:e53–60.
39. Feng J, Williamson B, Simon N, Carone M. Nonparametric variable importance using an augmented neural network with multi-task learning. In: *International conference on machine learning*; 2018:1496–505 pp.
40. Breiman L, Friedman JH. Estimating optimal transformations for multiple regression and correlation. *J Am Stat Assoc* 1985;80:580–98.
41. Zheng W, van der Laan MJ. Asymptotic theory for cross-validated targeted maximum likelihood estimation, UC Berkeley Division of Biostatistics Working Paper Series; 2010 Paper 273.
42. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001;1189–232. <https://doi.org/10.1214/aos/1013203451>.
43. Ripley BD. *Pattern recognition and neural networks*. Cambridge, UK: Cambridge University Press; 2007.
44. van der Laan MJ, Dudoit S, Keles S. Asymptotic optimality of likelihood-based cross-validation. *Stat Appl Genet Mol Biol* 2004;3:1–23.
45. van der Laan MJ, Dudoit S, van der Vaart A. The cross-validated adaptive epsilon-net estimator. *Stat Decis* 2006;24:373–95.
46. Polley E, LeDell E, Kennedy C, Lendle S, van der Laan M. Superlearner: super learner prediction, r package version 2.0-21; 2013. Available from: <http://CRAN.R-project.org/package=SuperLearner>.

Supplementary Material: The online version of this article offers supplementary material (<https://doi.org/10.1515/ijb-2019-0061>).