Nonparametric variable importance using an augmented neural network with multi-task learning

Jean Feng [*1] **Brian D. Williamson** [*1] **Marco Carone** [1 2] **Noah Simon** [1]

Abstract

In predictive modeling applications, it is often of interest to determine the relative contribution of subsets of features in explaining the variability of an outcome. It is useful to consider this variable importance as a function of the unknown, underlying data-generating mechanism rather than the specific predictive algorithm used to fit the data. In this paper, we connect these ideas in nonparametric variable importance to machine learning, and provide a method for efficient estimation of variable importance when building a predictive model using a neural network. We show how a single augmented neural network with multi-task learning simultaneously estimates the importance of many feature subsets, improving on previous procedures for estimating importance. We demonstrate on simulated data that our method is both accurate and computationally efficient, and apply our method to both a study of heart disease and for predicting mortality in ICU patients.

1. Introduction

Machine learning-based techniques are increasingly used to make decisions about allocating scientific resources in HIV vaccine studies (Rolland & Gilbert, 2012) and to improve patient care (Avati et al., 2017), among other high-impact areas of research and practice. Understanding the importance of measured features in prediction may make such algorithms more interpretable (Guidotti et al., 2018). In other words, are some features worth focusing on in future scientific studies? What information is most important for predicting a patient's prognosis? Answering these questions requires a rigorous definition of an estimable variable importance measure, with valid statistical inference for the true importance of the features under study; this is critical if we plan to use machine learning-based methods to make healthcare or resource allocation decisions.

Neural networks have been very effective in complicated domains (Krizhevsky et al., 2012; Mikolov et al., 2013), but existing variable importance measures for neural networks (Garson, 1991; Bach et al., 2015) suffer from two major drawbacks: they (1) are typically defined in terms of the network parameters, whereas importance is naturally a property of the data-generating mechanism; and (2) do not yield assessments of variability, as their definitions make classical statistical analysis difficult.

To address these two issues, we use the model-agnostic definition of variable importance in Williamson et al. (2017): given the true joint distribution of outcome Y and features $X_1, ..., X_p$, the importance of the feature subset with indices $s \subseteq \{1, .., p\}$ is based on the squared difference between the full and reduced conditional means

$$\left\{E\left(Y|X_1, ..., X_p\right) - E\left(Y|\left\{X_j : j \in \{1, ..., p\} \setminus s\right\}\right)\right\}^2.$$

If the squared difference is large on average, then the feature subset has high importance. If the squared difference is always zero, then the feature subset has zero importance. An analysis through influence functions (Bickel et al., 1998) yields a procedure that results in both a statistically efficient estimator and an asymptotically valid confidence interval.[1]

We describe a novel neural network structure and objective function that efficiently computes this variable importance measure for many feature subsets. For each feature index subset s, estimating its importance depends on an estimate of the reduced conditional mean of Y. Our method simultaneously estimates all required conditional means with a single neural network by augmenting the input features with an indicator vector for the subset s. We train the network by minimizing the multi-task loss of predicting Y given the feature subsets of interest using stochastic gradient descent.

[*]Equal contribution [1]Department of Biostatistics, University of Washington, Seattle, Washington, USA [2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. Correspondence to: Jean Feng <jeanfeng@uw.edu>, Brian Williamson <brianw26@uw.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[1] Koh & Liang (2017) use influence functions from a parametric perspective to analyze the dependence of predictions from neural networks on training observations. Our work uses influence functions from a semiparametric perspective instead.

This paper makes two contributions. First, to our knowledge, this is the first nonparametric variable importance measure applied to neural networks. Second, our proposed augmented neural network with multi-task learning estimates variable importance for many feature subsets simultaneously. We demonstrate empirically that our method accurately estimates variable importance and yields confidence intervals with asymptotically correct coverage. We also analyze data from a study of heart disease and data from a study of mortality prediction in ICU patients. Code and Supplementary Material (SM) are available at https://github.com/jjfeng/nnet_var_import.

2. Related work

Most variable importance measures applied to neural networks are intimately tied to the network itself, and are defined in terms of the weights between the nodes. There are both overall measures of variable importance (Garson, 1991; Olden & Jackson, 2002; Lipovetsky & Conklin, 2001) as well as local measures, i.e., how much to attribute an individual prediction to a particular feature (Bach et al., 2015; Shrikumar et al., 2017; Sundararajan et al., 2017; Murdoch et al., 2018). There are no statistically rigorous proposals for evaluating uncertainty of these estimates; it is unclear that a statistically valid evaluation of uncertainty is even possible. Moreover, these methods typically quantify the importance only of individual variables. In contrast, our method estimates individual and group importance, and provides confidence intervals with correct coverage.

Measures of variable importance defined independently from the estimation procedure have been proposed, and include a nonparametric extension of R^2 (Doksum & Samarov, 1995); the risk difference, $E(Y \mid A = a, X = x) - E(Y \mid A = 0, X = x)$ (van der Laan, 2006; Chambaz et al., 2012; Sapp et al., 2014); and the mean absolute difference $E\left\{|Y - E(Y \mid X)| - |Y - E(Y \mid X_{(-j)})|\right\}$ (Lei et al., 2017). All of these methods allow formal inference, but suffer from computational issues. They are all currently estimated by refitting separate models – our proposal jointly estimates the conditional means and can be used to speed up calculations for these importance measures. In this paper, we focus on the squared difference, as it is commonly used in regression. In practice, the most appropriate variable importance measure depends on the scientific goal.

Sundararajan et al. (2017) proposed that variable importance measures for neural networks should satisfy two axioms: (1) if two inputs differ in one feature and have different outputs, then the differing feature should have nonzero importance; and (2) if two networks have the same output for all inputs, the variable importance measures should be the same. We show that our measure satisfies a similar set of axioms generalized to a nonparametric setting.

3. Variable importance

Consider the random vector $(X = (X_1, ..., X_p), Y)$ with probability distribution P over $\mathcal{X} \times \mathbb{R}$, where $\mathcal{X}$ denotes the possible realizations of X, and the outcome, Y, is a real-valued variable with a natural ordering. For example, Y may be a binary variable or a continuous variable, corresponding to a classification or prediction problem, respectively. We measure the importance of $\{X_j\}_{j \in s}$ for any $s \subseteq \{1, ..., p\}$ under the distribution P.

Denote the conditional means with respect to distribution P based on the full set and a reduced set of features as

$$\mu_P(x) \coloneqq E_P(Y \mid X = x) \text{ and}$$
$$\mu_{P,s}(x) \coloneqq E_P(Y \mid X_{(-s)} = x_{(-s)}) \,,$$

respectively, where $x_{(-s)}$ denotes the feature vector after removing features with indices in s – when $s = \emptyset$, then $\mu_{P,\emptyset} \equiv \mu_P$. The variable importance of $\{X_j\}_{j \in s}$ is

$$\Psi_s(P) \coloneqq \frac{\int \{\mu_P(x) - \mu_{P,s}(x)\}^2 dP(x)}{\mathrm{Var}_P(Y)} \,. \qquad (1)$$

For convenience, let $\Phi_s(P)$ denote the numerator of (1). We may interpret (1) as the additional proportion of variability in the outcome explained by the features in the set s, as it is the difference in population R^2 obtained from using the full set of features or the reduced set of features; (1) also generalizes the parametric R^2 to a nonparametric setting (Williamson et al., 2017).

We improve upon the original axioms of Sundararajan et al. (2017) by formulating a nonparametric version that handles groups of variables and properly takes into account the probability measure P:

A1. For a nonempty feature subset X_s with indices $s \subseteq \{1, ..., p\}$, if μ_P and $\mu_{P,s}$ differ over a subset with non-zero probability measure under P, then X_s has non-zero importance.

A2. If P and P' have the same cumulative distribution function, then the variable importance of any subset of features is the same under P and P'.

The importance measure (1) we focus on clearly satisfies these axioms.

We envision this variable importance measure as an integral part of the scientific method. The hypothesis-generation step often involves testing for effects across many features. However, after multiple hypothesis correction, one may have low statistical power to detect effects. Variable importance may be useful in such cases: we define feature subsets of interest using domain knowledge, estimate their importance, and use the importance ranking to prioritize features for future confirmatory studies. In practice, this means that

one is generally not interested in estimating the variable importance of all possible feature subsets but rather no more than a polynomial number of them.

3.1. Statistical inference

Suppose we observe n iid observations $\{(X^{(i)}, Y^{(i)})\}_{i=1}^n$ from an unknown distribution P_0 known only to lie in a fully unrestricted (nonparametric) model $\mathcal{M}$, and wish to estimate the importance $\psi_{0,s} := \Psi_s(P_0)$ of the features X_s. An obvious approach to estimating $\psi_{0,s}$ is to estimate the relevant components of P_0 and plug these estimates into the mapping Ψ_s. If $\widehat{P}_n$ is a consistent estimator of P_0, with $\hat{\mu}$ and $\hat{\mu}_s$ denoting corresponding estimators of μ_{P_0} and $\mu_{P_0,s}$, respectively, a natural plug-in estimator is given by

$$\Psi_s(\widehat{P}_n) := \frac{\frac{1}{n}\sum_{i=1}^n \{\hat{\mu}(X^{(i)}) - \hat{\mu}_s(X^{(i)})\}^2}{\frac{1}{n}\sum_{i=1}^n \left(Y^{(i)} - \frac{1}{n}\sum_{i=1}^n Y^{(i)}\right)^2} . \quad (2)$$

This simple estimator converges to $\psi_{0,s}$ but usually at a rate slower than $n^{-1/2}$ because of excess finite-sample bias. Unfortunately, optimal estimators for the conditional means μ_{P_0} and $\mu_{P_0,s}$ are not optimal for estimating $\psi_{0,s}$. In statistical terms, the plug-in estimator is not a regular and asymptotically linear estimator.

Analyzing the efficient influence function (EIF) of (1) suggests a method for constructing an asymptotically efficient estimator of $\Psi_s(P_0)$ via the elimination of this excess bias (Bickel et al., 1998). Specifically, a one-step correction yields the following estimator of $\psi_{0,s}$:

$$\hat{\psi}_{n,s} := \Psi_s(\widehat{P}_n) + \underbrace{\frac{1}{n}\sum_{i=1}^n D_s^*(\widehat{P}_n)(X^{(i)}, Y^{(i)})}_{\text{one-step correction}} , \quad (3)$$

where $D_s^*(P)$ denotes the EIF of Ψ_s at P relative to $\mathcal{M}$, and the evaluation $D_s^*(P)(x, y)$ is given by

$$\frac{2\{y - \mu_P(x)\}\{\mu_P(x) - \mu_{P,s}(x)\}}{\mathrm{Var}_P(Y)} + \frac{\{\mu_P(x) - \mu_{P,s}(x)\}^2}{\mathrm{Var}_P(Y)} - \Phi_s(P)\left\{\frac{y - E_P(Y)}{\mathrm{Var}_P(Y)}\right\}^2 .$$

Plugging in $\hat{\mu}$, $\hat{\mu}_s$ and the empirical variance, our final estimator (3) simplifies to

$$\frac{\sum_{i=1}^n A_{i,n,s}\{\hat{\mu}(X^{(i)}) - \hat{\mu}_s(X^{(i)})\}}{\sum_{i=1}^n \left(Y^{(i)} - \frac{1}{n}\sum_{i=1}^n Y^{(i)}\right)^2} \quad (4)$$

with $A_{i,n,s} := 2Y_i - \hat{\mu}(X^{(i)}) - \hat{\mu}_s(X^{(i)})$. If (i) we can estimate the conditional means at sufficiently fast rates, i.e.,

$$E_{P_0}\{\mu_{P_0}(X) - \hat{\mu}(X)\}^2 = o_p(n^{-1/2}) , \quad (5)$$

$$E_{P_0}\{\mu_{P_0,s}(X) - \hat{\mu}_s(X)\}^2 = o_p(n^{-1/2}) , \quad (6)$$

(ii) the true variable importance value is not a boundary value of 0 or 1 – the behavior at the boundary is difficult to analyze – and (iii) $\hat{\mu}$ and $\hat{\mu}_s$ eventually lie in a Donsker class (van der Vaart, 2000), then the corrected estimator (4) converges at the desired rate to a non-trivial distribution:

$$\sqrt{n}(\hat{\psi}_{n,s} - \psi_{0,s}) \to_d N\left[0, \mathrm{Var}_{P_0}\{D_s^*(P_0)(Y, X)\}\right] . \quad (7)$$

Since D_s^* is the EIF, $\hat{\psi}_{n,s}$ is an asymptotically efficient estimator. We can build confidence intervals using any consistent estimator of $\mathrm{Var}_{P_0}\{D_s^*(P_0)(Y, X)\}$, such as $\frac{1}{n}\sum_{i=1}^n \{D_s^*(\widehat{P}_n)(X^{(i)}, Y^{(i)})\}^2$.

3.2. Local variable importance

Until now, we have focused on a global measure of importance by integrating over the entire distribution P, answering questions such as "in general, what factors are most important in predicting survival?" For certain settings, we may be interested in a local version of variable importance, answering questions like "for subjects ages 65 and up, what factors are most important in predicting survival?" A simple extension of (1) allows us to define a local version of variable importance. For feature subset X_s and subpopulation $A \subseteq \mathcal{X}$, the importance of X_s in subpopulation A is $\Psi_s\left(P_{0|X\in A}\right)$, where we plugged the conditional distribution $P_{0|X\in A}$ into (1). We only estimate variable importance when A has positive probability measure since arbitrary restricted conditional distributions are typically difficult to estimate without further regularity assumptions (see, e.g., Hall et al., 2004). To estimate local importance, one can estimate this conditional mean directly by restricting to the subpopulation A *or* determine the restricted conditional mean from an estimate over the entire population.

4. Estimating conditional means

We now present a computationally efficient method for estimating the conditional means required by (2) and (4) for many, possibly overlapping, feature index sets $\mathcal{S}$, where $s \subseteq \{1, ..., p\}$ for each $s \in \mathcal{S}$. The procedure in Williamson et al. (2017) is computationally prohibitive if the cardinality $|\mathcal{S}|$ of $\mathcal{S}$ is large: it estimates the conditional mean for each feature subset of interest separately.

Estimating the conditional means is an independent step from estimating variable importance. The proposed augmented network structure may have applications outside variable importance, such as settings with missing data.

Let $\Theta^{(q)}$ be the set of all possible parameters for neural networks with q input nodes and one output node. The neural network parameterized by $\theta \in \Theta^{(q)}$ is denoted by $f(\cdot; \theta)$. The training data is the set of observations $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$.

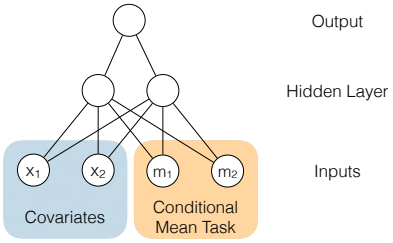

Figure 1. The augmented neural network structure for data (X_1, X_2, Y) has two input nodes, X_1 and X_2, as well as binary inputs m_1 and m_2 that indicate the conditional mean task at hand. To predict the conditional mean $E(Y|X_1, X_2)$, set both $m_1 = m_2 = 0$. To predict the reduced conditional mean $E(Y|X_1)$, set x_2 to any value (e.g., 0) and $m_1 = 0, m_2 = 1$.

4.1. Multiple network approach

An obvious approach to estimating multiple conditional means is to fit separate neural networks. That is, we estimate the full conditional mean μ_{P_0} with a neural network parameterized by

$$\hat{\theta} \in \arg\min_{\theta \in \Theta^{(p)}} \frac{1}{n} \sum_{i=1}^{n} \left\{ y^{(i)} - f(x^{(i)}; \theta) \right\}^2,$$

and estimate the reduced conditional mean $\mu_{P_0,s}$ for variable index set $s \in \mathcal{S}$ with a neural network parameterized by

$$\hat{\theta}_s \in \arg\min_{\theta \in \Theta^{(p-|s|)}} \frac{1}{n} \sum_{i=1}^{n} \left\{ y^{(i)} - f(x^{(i)}_{(-s)}; \theta) \right\}^2.$$

The above procedure requires training a total of $|\mathcal{S}| + 1$ neural networks. Though these neural networks can be trained in parallel, this quickly becomes computationally prohibitive even for a polynomial number of feature subsets.

4.2. Augmented network with multi-task learning

Rather than fitting multiple neural networks, we train a single network to jointly learn the conditional means using multi-task learning (MTL) (Collobert & Weston, 2008; Ruder, 2017). MTL tends to be advantageous when the tasks are related, which is clearly true in our case: for any $x \in \mathcal{X}$ and indices $s, t \subseteq \{1, ..., p\}$ where $s \subseteq t$, the conditional means are related by

$$\mu_{P_0,t}(x) = E_{P_0} \left\{ \mu_{P_0,s}(X) \middle| X_{(-t)} = x_{(-t)} \right\}.$$

MTL is best used when all conditional means can be compactly represented by a single neural network. As an extreme example, suppose μ_{P_0} is the sum of the univariate reduced conditional means $E_{P_0}(Y|X_j)$ for $j = 1, \ldots, p$ so that any reduced conditional mean $\mu_{P_0,s}$ is the sum of $E_{P_0}(Y|X_k)$ for $k \in s$. We can approximate the conditional means for all nonempty $s \subseteq \{1, ..., p\}$ using a single neural network with sub-networks approximating the univariate reduced conditional means.

More specifically, we propose using a single augmented neural network with $2p$ input nodes and one output node, parameterized by $\theta \in \Theta^{(2p)}$, to approximate the conditional means $\mu_{P_0,s}$ for all $s \in \mathcal{S}$; Figure 1 provides an example of the proposed node structure. We augment the input features x with the binary vector $e_s \in \{0, 1\}^p$ with ones only in positions s.[2] The conditional mean $\mu_{P_0,s}(x)$ is approximated by the output $f(x, e_s; \theta)$.

Moreover, this augmented network structure approximates all of the conditional means arbitrarily well so long as the activation functions at the hidden nodes are not polynomials. Using the classic result from Leshno et al. (1993), we show that there is a neural network that approximates the function

$$g_{P_0}(x, m) := \mu_{P_0}(x) \mathbb{1}\{m = 0\} + \sum_{s \in \mathcal{S}} \mu_{P_0,s}(x) \mathbb{1}\{m = e_s\}$$

arbitrarily well, where g_{P_0} combines the conditional means into a single function. Since g_{P_0} contains all the information from the conditional means, this neural network is also a good approximation of the conditional means.

Lemma 1. *Let the activation function σ be any locally bounded function where the set of points at which σ is discontinuous has zero Lebesgue measure. Let*

$$\mathcal{F} := span\{\sigma(\beta v + b) : \beta \in \mathbb{R}^{2p}, b \in \mathbb{R}\}.$$

Consider any $\mathcal{S} \subseteq \{s : s \subseteq \{1, ..., p\}\}$. For any random vector (X, Y) with conditional means $\mu_{P_0,s}$ for all $s \in \mathcal{S}$, there exists a sequence of neural networks $\{f_j\}_{j=1}^{\infty} \in \mathcal{F}$ such that for any compact set $K \subseteq \mathbb{R}^p$,

$$\lim_{j \to \infty} \max_{s \subseteq \mathcal{S}} \| f_j(x, e_s) - \mu_{P_0,s}(x) \|_{L^\infty(K)} = 0,$$

where $\|h\|_{L^\infty(K)} = \sup_{x \in K} |h(x)|$.

A proof is given in the SM.

4.2.1. Multi-task learning via stochastic gradient descent

We train the augmented neural network by minimizing the sum of the losses for estimating all conditional means:

$$\hat{\theta} \in \arg\min_{\theta \in \Theta^{(2p)}} \frac{1}{n} \sum_{i=1}^{n} \left[\left\{ y^{(i)} - f(x^{(i)}, 0; \theta) \right\}^2 + \sum_{s \in \mathcal{S}} \mathbb{E}_{W_s} \left(\left[y^{(i)} - f(\xi(x^{(i)}, W_s; s), e_s; \theta) \right]^2 \right) \right], \tag{8}$$

[2] There are different ways to perform MTL using neural networks. One approach is to use a separate output node for each task; however, the number of parameters in our model would grow linearly in $|\mathcal{S}|$. Instead, we augment the input with a binary vector e_s that represents all possible conditional mean "tasks," which only adds p nodes to the input layer.

where W_s is a random variable with value in $\mathbb{R}^{|s|}$; and the function $\xi(x, w; s)$ maps (x, w) to $\mathbb{R}^p$ by defining $\{\xi(x, w; s)\}_{(-s)} = x^{(i)}_{(-s)}$ and $\{\xi(x, w; s)\}_s = W_s$. The distribution of W_s is a hyperparameter of the training procedure, but as the number of samples increases to infinity, the fitted neural network will not depend on the distribution of W_s. While W_s can be set to a constant, we found that in practice it is better to choose use a stochastic W_s – introducing noise teaches the network to be invariant to the values that are marginalized out in the reduced conditional mean.

At first glance, it seems as if the multi-task loss in (8) simply shifts the computational load of minimizing $|\mathcal{S}| + 1$ neural networks to calculating a loss function with $|\mathcal{S}|+1$ terms per observation. However, we can minimize this multi-task loss efficiently using (mini-batch) stochastic gradient descent.

At each iteration, we construct the stochastic objective by first replacing each expectation $\mathbb{E}_{W_s}(\cdot)$ in (8) with its value at a single sample w_s, yielding

$$\frac{1}{n}\sum_{i=1}^{n}\left(\left\{y^{(i)} - f(x^{(i)}, m = 0; \theta)\right\}^2 + \sum_{s\in\mathcal{S}}\left[y^{(i)} - f\{\xi(x^{(i)}, w_s; s), m = e_s; \theta\}\right]^2\right). \tag{9}$$

We can alternatively think of (9) as the (scaled) mean squared error over a newly constructed dataset D':

$$\left\{(x^{(i)}, y^{(i)})\right\}_{i=1}^{n} \cup \left[\cup_{s\in\mathcal{S}}\left\{\left(\xi(x^{(i)}, w_s; s), y^{(i)}\right)\right\}_{i=1}^{n}\right].$$

If the number of summands in the multi-task loss is large (i.e., $|\mathcal{S}|$ is large), then we can also sample the number of summands to keep in the multi-task loss. That is, we uniformly sample a mini-batch D'' from D' and the stochastic objective is equal to the mean squared error over D''.

These two stochastic approximations ensure that the expectation of the stochastic objective is equal to (8), so stochastic gradient descent with properly chosen or adapted step sizes converges to a stationary point of (8) (Bottou, 1998). Here we use *Adam* to train the augmented neural network (Kingma & Ba, 2014).

Our method may be more easily understood through a comparison between the proposed procedure and dropout (Hinton et al., 2012; Srivastava et al., 2014). Dropout regularizes neural networks by randomly dropping nodes, encoding the prior belief that the true function does not overly depend on one node. The training algorithm for the augmented network also uses random inputs that encode a similar prior belief: the full conditional mean is similar to the reduced conditional means. If the prior belief holds, our augmented neural network is likely to better estimate the conditional means compared to the multiple network procedure.

4.2.2. Applications of the augmented network

The primary problem of interest is variable importance; however, our method may be used in settings with missing data, illustrating the broader applicability of our method.

Variable importance: Two issues arise when using estimates of the conditional means from the augmented network in the plug-in estimator (2) or the corrected estimator (4). First, the estimated full and reduced conditional means are encouraged to be similar by sharing parameters in the augmented neural network. Our empirical results suggest that the variable importance estimates tend to be biased downwards in small samples due to this parameter sharing. Second, confidence intervals for the true variable importance are based on the asymptotic distribution in (7), and rely on the assumption that the neural network estimates the true conditional means at sufficiently fast rates given in (5) and (6). To prove that we attain these rates, we must balance the approximation and estimation error. If the true data-generating mechanism falls in a function class that is easy to approximate with neural networks, such as those with finite total variation (Bach, 2017), then we achieve the desired rates and our confidence intervals are valid. We show empirical evidence that our procedure is valid in general in Section 5.

Missing data: To our knowledge, few papers address how to flexibly and accurately estimate the conditional means using a neural network in the presence of missing covariates (García-Laencina et al., 2010). Much of the literature focuses on imputation of the missing covariates using, e.g., recurrent neural networks (Bengio & Gingras, 1996) or multi-task learning (García-Laencina et al., 2007). However, imputation does not, in general, minimize the mean squared error for predicting Y when covariates in s are missing; one should instead directly estimate $\mu_{P_0,s}$, as done in our augmented network. Others have proposed parametric approaches, estimating the joint distribution of the covariate vector X and then estimating $\mu_{P_0,s}$ (Tresp et al., 1994). However, if the model is incorrect, this does not yield valid estimates for the reduced conditional means.

5. Experiments on simulated data

We compare performance of the multiple networks and the single augmented network approaches for estimating variable importance, using both the plug-in (2) and corrected (4) estimators. We compute the difference between the estimate and the truth – the empirical bias – and the empirical coverage of nominal 95% confidence intervals. The asymptotic distribution of the plug-in estimator is difficult to analyze; while a bootstrap approach is tempting, Williamson et al. (2017) show empirically that this does not yield satisfactory coverage.

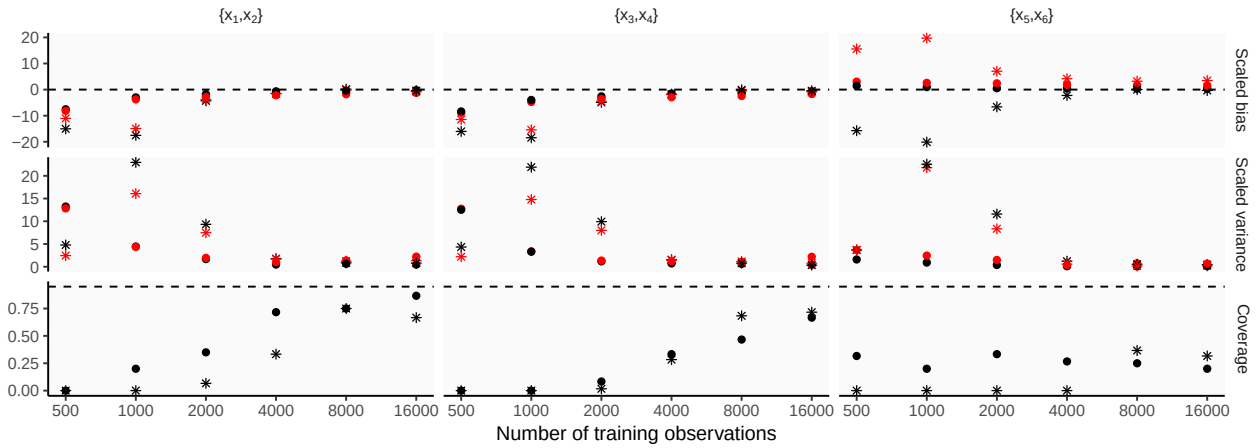

Figure 2. Simulation results where $E_{P_0}(Y|X)$, given in (10), only depends on $x_1, ..., x_4$. We estimate variable importance of $\{x_1, x_2\}$ (left col), $\{x_3, x_4\}$ (middle col), and $\{x_5, x_6\}$ (right col) using the plug-in (red) and corrected estimator (black) with multiple networks (stars) and an augmented network (circles). Top row: bias of the estimates multiplied by $\sqrt{n}$ for training set size n. Middle row: variance of the estimates multiplied by n. Bottom row: coverage of the 95% nominal conf. interval. Dashed line is the desired 95% coverage level.

In both simulation studies, data are generated from a model of the form $y = f(\boldsymbol{x}) + \sigma\epsilon$, where f is the full conditional mean, $\epsilon \sim N(0, 1)$, and σ was chosen such that the signal-to-noise ratio was roughly 2. Each covariate X_j for $j = 1, ...p$ was independently drawn from the $U(-2, 2)$ distribution.

We fit fully-connected neural networks with rectified linear hidden units by minimizing the penalized multi-task loss

$$\underbrace{\mathcal{L}(\{(x^{(i)}, y^{(i)})\}_{i=1}^n, \theta)}_{\text{multi-task loss in (8)}} + \lambda\|\theta\|_2^2,$$

where W_s in (8) is a standard normal random variable. Each experiment was run on training set sizes ranging from 500 to 16000, with sixty replicates each.

5.1. A non-additive six-variable function

We consider here that X is composed of six features, and the conditional mean only depends on the first four features:

$$f(x_1, ..., x_6) = x_1 \sin(x_1 + 2x_2) \cos(x_3 + 2x_4). \quad (10)$$

We are interested in estimating the variable importance of groups $\{x_1, x_2\}$, $\{x_3, x_4\}$, and $\{x_5, x_6\}$, given by 0.820, 0.838, and zero, respectively. Unsurprisingly, the variable importance values here sum to more than one, since (1) measures additional variability explained by the feature subset. Since $\{x_5, x_6\}$ are not important at all, the asymptotic result (7) does not apply and we do not expect the empirical coverage to reach 95%. However, the asymptotic behavior of our estimator for $\{x_1, x_2\}$ and $\{x_3, x_4\}$ should still follow (7).

The empirical bias for the estimators scaled by $\sqrt{n}$ shrinks towards zero as the training set size increases (Fig 2, top row), which is expected given the asymptotic result (7). Bias tends to be smallest when we estimate variable importance using a single augmented neural network and the corrected estimator. The difference between the plug-in and corrected estimators tends to be smaller than the difference between using multiple networks and a single augmented neural network. This is somewhat expected as the plug-in and corrected estimators differ by a small correction term that tends to zero. The variance of the estimators scaled by n converges towards a positive constant as the training set size increases (Fig 2, middle row), which is also expected according to (7). The scaled variance tends to be smallest when we use the augmented network and a corrected estimator. Finally, the coverage for groups $\{x_1, x_2\}$ and $\{x_3, x_4\}$ approaches the desired 95% level as sample size grows (Fig 2, bottom row). As expected, the coverage for the group $\{x_5, x_6\}$ is poor, never going above 50%; however, this is an improvement over the zero coverage seen in Williamson et al. (2017).

The overall differences between fitting multiple networks and a single network are most apparent in small samples: as the number of samples grows, the differences between the two model-fitting procedures are negligible. A possible explanation is that the prior belief plays a less influential role in larger sample sizes.

Fitting a single network takes approximately the same amount of time in the multiple networks or augmented network approach (60–160 seconds), but the total computational power is much higher for the multiple network approach. When estimating the conditional means separately, one must cross-validate over a larger range of network struc-

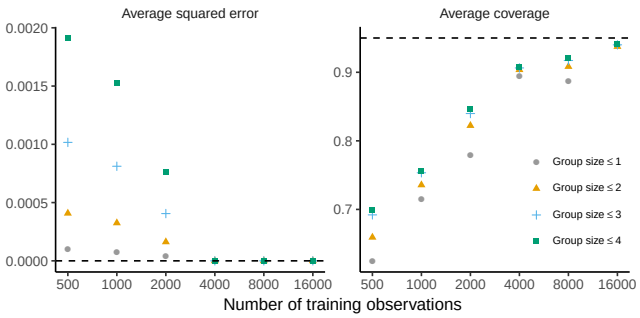

Figure 3. We estimate importance of variable groups up to size four when $E(Y|X)$ is the sum of univariate functions. Left: average squared error using the plug-in and corrected estimators. Both estimators had nearly identical errors so each point represents both. Right: average coverage of the nominal 95% confidence intervals from the corrected estimator. Dashed line is the desired 95% level.

tures since different covariates have different optimal structures (Table B.1 in the SM).

We found that the mean squared error of the fitted augmented network is smaller when W_s in (8) is a standard normal random variable versus constant at zero (Figure B.1 in the SM). We hypothesize that using a constant does not train the network to be invariant to the "missing" input nodes for the reduced conditional mean task.

5.2. A sum of univariate functions

Here, we consider the sum of eight univariate functions:

$$f(\boldsymbol{x}) = x_1 + x_2^2 + \sin(x_3) + \cos(x_4) + (x_5 + 1)^2 - 2x_6 + \max(x_7, 0) + x_8. \quad (11)$$

We estimate the importance of all groups with cardinality up to four, for a total of 162 groups. Since fitting multiple networks is computationally prohibitive, we estimate importance using the augmented network with the plug-in and corrected estimators. We assess accuracy using average squared error of the variable importance estimates and empirical coverage of the confidence intervals over variable groups up to sizes one to four.

The mean average squared error for all variable groups decreases as the number of training observations increases (Fig 3). Note that the importance of groups increases with its cardinality so it is expected that the error is higher for larger groups. The squared error from the plug-in and corrected estimators were very close for this example; the main advantage in using the corrected estimator is that we obtain asymptotically valid confidence intervals.

We find that the coverage converges to 95% as the training set grows in size. It is somewhat surprising to achieve such good coverage when we simultaneously estimate 162 conditional means with a single network. This may be due to the univariate sum structure in (11).

6. A heart disease study

We analyze the importance of features for predicting myocardial infarction (MI) using data from a retrospective cross-sectional sample of 462 white men aged 15–64 in a region of the Western Cape, South Africa (Rosseauw et al., 1983). Measurements on two sets of features are available: behavioral features, including cumulative tobacco consumption, current alcohol consumption, and type A behavior; and biological features, including systolic blood pressure (SBP), LDL cholesterol, adiposity, family history of heart disease, obesity, and age. We estimated importance of individual features and the two feature groups for predicting the presence of MI, using neural networks with a sigmoid output node.

The estimated variable importance are shown in the left two columns of Figure 4. Family history, age, and LDL have the highest estimated importance. The biological features were estimated to be more important than the behavioral features.

The plug-in and corrected estimates were much more similar in the augmented network approach compared to the multiple network approach. The lack of agreement in the multiple network is not desired; since the correction in (3) has mean zero, we expect similar results using the two estimators.

Our results using the augmented network were the most similar to those in Williamson et al. (2017), though the rankings differ slightly. However, we observed a striking difference in computation time – we obtained results in 15 minutes, whereas Williamson et al. (2017) needed 24 hours.

7. Predicting mortality of ICU patients

We analyze the importance of variables measured during the first two days of patients' ICU stays for predicting in-hospital mortality using data from the PhysioNet/CinC Challenge 2012 (Silva et al., 2012). We have 4000 records, with five general descriptors collected upon admission and 37 features measured over the course of the first 48 hours after admission to the ICU, such as Glasgow Coma Score (GCS), blood urea nitrogen (BUN), and heart rate.

We extracted 55 summary features for prediction, including the general descriptors and the min/max, mean, and last measured value of variables used in the SAPS I/II (Le et al., 1984; Le Gall et al., 1993) and Xia et al. (2012) (Table C.2 in the SM). We estimate the importance of 25 variable groups which fall into two categories: "medical test groups" contain summary features for variables measured by the same medical test and "individual variable groups" contain summary features from the same variable. Here, we discuss our results for the medical test groups; individual variable groups are discussed in the SM.

The metabolic panel group consists of the summary features of bicarbonate, BUN, sodium, potassium, and glucose. The

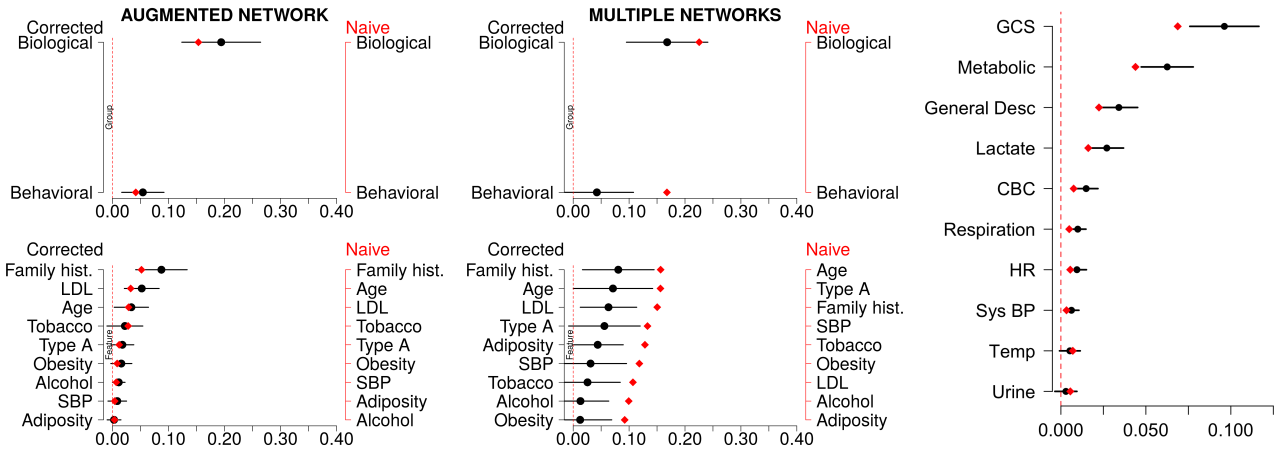

Figure 4. Naive (red diamonds) and corrected (black circles) variable importance estimates for both individual features and the biological and behavioral feature groups in the heart study, augmented network structure with MTL (left) and multiple networks (center); and importance estimates of variables from medical tests in the ICU data estimated using the augmented network structure with MTL (right). Confidence intervals for the true variable importance, based on the corrected estimator only, are displayed as black bars. For heart study figures (left and center), the y-axis labels on the left correspond to the corrected estimates, while the y-axis labels on the right correspond to the plug-in estimates.

complete blood count test (CBC) group consists of the summary features for white blood cells and hematocrit. The respiration group consists of the summary features for respiration rate, mechanical ventilation, fraction of inspired oxygen, and partial pressure of oxygen. The general descriptors group consists of age, sex, height, weight, and ICU admission type. Each of the remaining variables – GCS, systolic blood pressure, temperature, lactate, heart rate, and urine – were treated as a medical test groups.

We estimated the conditional means using the augmented network approach. For features that seem informative for the outcome, we imputed missing values by drawing uniformly from the healthy normal range at each step of stochastic gradient descent. For features that seem uninformative (i.e., missing at random), we indicated that the covariate was missing via the binary task vector and replaced missing values using a standard normal distribution.

GCS assesses consciousness based on the patient's ability to open their eyes, talk, and move – it had the highest estimated importance (Fig 4 right). This conclusion is sensible as GCS measures a patient's immediate risk of dying; GCS is also the highest scoring item on SAPS II (up to 26 points). The metabolic panel scored next highest in terms of importance, which also aligns with SAPS II – the metabolic panel items may contribute up to 24 points.

We estimate the lowest importance for urine and temperature, in line with previous studies. Temperature is one of the lowest-scoring items in SAPS II (at most 3 points). Urine can contribute up to 11 points in SAPS II, but other mortality scoring methods like APACHE IV (Zimmerman et al., 2006) do not use urine output to predict mortality.

8. Discussion

We propose using neural networks to estimate the nonparametric variable importance measure (1), allowing us to interpret the relationship between subsets of features and the outcome. We show how to perform statistical inference on this variable importance measure. To estimate the importance of many feature subsets efficiently, we propose fitting an augmented neural network with multi-task learning that simultaneously estimates the relationship between the outcome and feature subsets. This network structure may also be used for problems with missing data.

The variable importance measure considered here is restricted to outcomes that have a natural ordering; however, there are many problems where the outcome space is not ordered. We plan to derive an alternate measure to address categorical outcomes. In addition, we are working to understand the behavior of our estimator when the true importance measure is a boundary value of zero or one.

We believe that other machine-learning methods may also be modified to estimate variable importance efficiently. It seems particularly important to focus on popular methods that are difficult to interpret, such as gradient boosted trees (Freund & Schapire, 1997; Friedman, 2001). Interpretation is crucial when using these methods to make important decisions. We hope that statistical inference for variable importance is a useful tool in this direction.

Acknowledgments

The authors gratefully acknowledge the support of NIH/NIAID grant 5UM1AI068635 (MC) and NIH Grant DP5OD019820 (NS).

References

Avati, A, Jung, K, Harman, S, Downing, L, Ng, A, and Shah, NH. Improving palliative care with deep learning. *IEEE International Conference on Bioinformatics and Biomedicine*, 2017. arXiv:1711.06402.

Bach, F. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

Bach, S, Binder, A, Montavon, G, Klauschen, F, Müller, K-R, and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, 10(7):e0130140, 2015.

Bengio, Y and Gingras, F. Recurrent neural networks for missing or asynchronous data. In *Advances in Neural Information Processing Systems*, 1996.

Bickel, PJ, Klaasen, CAJ, Ritov, Y, and Wellner, JA. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998.

Bottou, L. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

Chambaz, A, Neuvial, P, and van der Laan, MJ. Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic Journal of Statistics*, 6: 1059–1099, 2012.

Collobert, R and Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

Doksum, K and Samarov, A. Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *The Annals of Statistics*, 23(5):1443–1473, 1995.

Freund, Y and Schapire, RE. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Friedman, JH. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pp. 1189–1232, 2001.

García-Laencina, PJ, Serrano, J, Figueiras-Vidal, AR, and Sancho-Gómez, J-L. Multi-task neural networks for dealing with missing inputs. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 2007.

García-Laencina, PJ, Sancho-Gómez, J-L, and Figueiras-Vidal, AR. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2):263–282, 2010.

Garson, DG. Interpreting neural network connection weights. *Artificial Intelligence Expert*, 1991.

Guidotti, R, Monreale, A, Turini, F, Pedreschi, D, and Giannotti, F. A survey of methods for explaining black box models. *arXiv:1802.01933*, 2018.

Hall, P, Racine, J, and Li, Q. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.

Hinton, GE, Srivastava, N, Krizhevsky, A, Sutskever, I, and Salakhutdinov, R R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580*, 2012.

Kingma, D and Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Koh, PW and Liang, P. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

Le, JRG, Loirat, P, Alperovitch, A, Glaser, P, Granthil, C, Mathieu, D, Mercier, P, Thomas, R, and Villers, D. A simplified acute physiology score for icu patients. *Critical Care Medicine*, 12(11):975–977, 1984.

Le Gall, J-R, Lemeshow, S, and Saulnier, F. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *JAMA*, 270(24): 2957–2963, 1993.

Lei, J, G'Sell, M, Rinaldo, A, Tibshirani, RJ, and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, (just-accepted), 2017.

Leshno, M, Lin, VY, Pinkus, A, and Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

Lipovetsky, Stan and Conklin, Michael. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

Mikolov, T, Sutskever, I, Chen, K, Corrado, GS, and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

Murdoch, WJ, Liu, PJ, and Yu, B. Beyond word importance: Contextual decomposition to extract interactions from lstms. *arXiv:1801.05453*, 2018.

Olden, JD and Jackson, DA. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling*, 154(1):135–150, 2002.

Rolland, M and Gilbert, P. Evaluating immune correlates in HIV type 1 vaccine efficacy trials: What RV144 may provide. *AIDS Research and Human Retroviruses*, 28(4): 400–404, 2012.

Rosseauw, J, Du Plessis, J, Benade, A, Jordann, P, Kotze, J, Jooste, P, and Ferreira, J. Coronary risk factor screening in three rural communities. *South African Medical Journal*, 64(12):430–436, 1983.

Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv:1706.05098*, 2017.

Sapp, S, van der Laan, MJ, and Page, K. Targeted estimation of binary variable importance measures with interval-censored outcomes. *The International Journal of Biostatistics*, 10(1):77–97, 2014.

Shrikumar, A, Greenside, P, and Kundaje, A. Learning important features through propagating activation differences. *arXiv:1704.02685*, 2017.

Silva, I, Moody, G, Scott, DJ, Celi, LA, and Mark, RG. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *Computing in Cardiology (CinC), 2012*. IEEE, 2012.

Srivastava, N, Hinton, GE, Krizhevsky, A, Sutskever, I, and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

Sundararajan, M, Taly, A, and Yan, Q. Axiomatic attribution for deep networks. *arXiv:1703.01365*, 2017.

Tresp, V, Ahmad, S, and Neuneier, R. Training neural networks with deficient data. In *Advances in Neural Information Processing Systems*, 1994.

van der Laan, MJ. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

van der Vaart, AW. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Williamson, BD, Gilbert, PB, Simon, N, and Carone, M. Nonparametric variable importance assessment using machine learning techniques. *UW Biostatistics Working Paper Series*, Working Paper 422, 2017.

Xia, H, Daley, BJ, Petrie, A, and Zhao, X. A neural network model for mortality prediction in icu. In *Computing in Cardiology (CinC), 2012*. IEEE, 2012.

Zimmerman, JE, Kramer, AA, McNair, DS, and Malila, FM. Acute physiology and chronic health evaluation (apache) iv: hospital mortality assessment for todays critically ill patients. *Critical Care Medicine*, 34(5):1297–1310, 2006.