Journal of Machine Learning Research 25 (2024) 1-27 Submitted 7/22; Revised 11/23; Published 1/24

Decorrelated Variable Importance

Isabella Verdinelli ISABELLA@STAT.CMU.EDU
Department of Statistics
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213, USA

Larry Wasserman LARRY@STAT.CMU.EDU
Department of Statistics
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213, USA

Editor: Eric Laber

Abstract

Because of the widespread use of black box prediction methods such as random forests and neural nets, there is renewed interest in developing methods for quantifying variable importance as part of the broader goal of interpretable prediction. A popular approach is to define a variable importance parameter — known as LOCO (Leave Out COvariates) — based on dropping covariates from a regression model. This is essentially a nonparametric version of R^2. This parameter is very general and can be estimated nonparametrically, but it can be hard to interpret because it is affected by correlation between covariates. We propose a method for mitigating the effect of correlation by defining a modified version of LOCO. This new parameter is difficult to estimate nonparametrically, but we show how to estimate it using semiparametric models.

Keywords: Correlation, Nonparametric Estimators, Prediction, Variable Importance

1 Introduction

Due to the increasing popularity of black box prediction methods like random forests and neural nets, there has been renewed interest in the problem of quantifying variable importance in regression. Consider predicting $Y \in \mathbb{R}$ from covariates (X, Z) where $X \in \mathbb{R}^g$ and $Z \in \mathbb{R}^h$. We have separated the covariates into X and Z where X represents the covariates whose importance we wish to assess. In what follows, we let $U = (X, Z, Y)$ denote all the variables. Define $\mu(x, z) = \mathbb{E}[Y|X = x, Z = z]$ so that

$$Y = \mu(X, Z) + \epsilon$$

where $\mathbb{E}[\epsilon|X, Z] = 0$.

A popular measure of the importance of X is

$$\psi_L = \mathbb{E}[(\mu(Z) - \mu(X, Z))^2] = \mathbb{E}[(Y - \mu(Z))^2] - \mathbb{E}[(Y - \mu(X, Z))^2]. \quad (1)$$

where $\mu(Z) = \mathbb{E}[Y|Z = z]$. Up to scaling, ψ_L is a nonparametric version of the usual R^2 from standard regression. This was called LOCO (**L**eave **O**ut **CO**variates) in Lei et al. (2018)

and Rinaldo et al. (2019) and has been further studied recently in Williamson et al. (2021), Williamson et al. (2020) and Zhang and Janson (2020). The parameter ψ_L is appealing because it is very general and easy to interpret. But it suffers from some problems. In particular, the value of ψ_L depends on the correlation between X and Z. When X and Z are highly correlated, ψ will be near 0 since removing X has little effect. In some applications, this might be undesirable as it obscures interpretability. We refer to this problem as *correlation distortion*. Another, more technical problem with LOCO, is its quadratic nature which causes some issues when constructing confidence intervals.

In this paper, we define a modified version of ψ_L denoted by ψ_0 that is invariant to the correlation between X and Z. There is a tradeoff: the modified parameter ψ_0 is free from correlation distortion but it is more difficult to estimate than ψ_L. In a sense, we remove the correlation from the estimand at the expense of larger confidence intervals. This is similar to estimating a coefficient in a linear regression where the value of the regression coefficient does not depend on the correlation between X and Z while the width of the confidence interval does. To reduce the difficulties in estimating ψ_0, we approximate $\mu(x, z)$ with the semiparametric model $\mu(x, z) = \beta(z)^T x + f(z)$.

Related Work. Assessing variable importance is an active area of research. Recent papers on LOCO include Lei et al. (2018); Rinaldo et al. (2019); Williamson et al. (2021, 2020); Zhang and Janson (2020). Another approach is to use derivatives of the regression function as suggested in Samarov (1993), and has received renewed attention in the machine learning literature (Ribeiro et al., 2016). There has been a surge of interest in an approach based on Shapley values, see for example, Messalas et al. (2019); Aas et al. (2019); Lundberg and Lee (2016); Covert et al. (2020); Fryer et al. (2020); Covert and Lee (2020); Israeli (2007); Bénard et al. (2021). We discuss derivatives and Shapley values in Section 5. Another paper that uses semiparametric models for intepretability is Sani et al. (2020) but that paper does not focus on variable importance. Loh and Zhou (2021) contains a review of several feature importance methods and, in particular, discusses the importance of missing data.

Paper Outline. In Section 2 we describe some issues related to LOCO and this leads us to define a few modified versions of the parameter. In Section 3 we discuss inference for the parameters. Section 4 contains some simulation studies. Section 5 discusses other issues and other measures of variable importance. A concluding discussion is in Section 6. Technical details and proofs are in an appendix.

2 Issues With LOCO

The parameter ψ_L is general and it is easy to obtain point estimates for it; see Section 3.1. But it does have two shortcomings which we now discuss.

2.1 Issue 1: Inference For Quadratic Functionals

The first, and less serious issue, is that ψ_L is a quadratic parameter and it is difficult to get confidence intervals for quadratic parameters because their limiting distribution and rate of

convergence change as ψ_L approaches 0. This is actually a common problem but it receives little attention. Many other parameters have this problem, including distance correlation (Székely et al., 2007), RKHS correlations (Sejdinovic et al., 2013) and kernel two-sample statistics (Gretton et al., 2012) among others.

To illustrate, consider the following toy example. Let $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ and consider estimating $\psi = \mu^2$ with $\widehat{\psi} = \overline{Y}_n^2$. When $\mu \neq 0$, we have $\sqrt{n}(\widehat{\psi} - \psi) \rightsquigarrow N(0, \tau^2)$ for some τ^2. When $\mu = 0$, $\widehat{\psi} \sim \sigma^2 \chi_1^2 / n$. When μ is close to 0, its distribution is neither Normal nor chi-squared, and the rate of convergence can be anything between $1/n$ and $1/\sqrt{n}$.

More generally, when dealing with a quadratic functional ψ, it is often the case that an estimator $\widehat{\psi}$ converges to a Normal at a $n^{-1/2}$ rate when $\psi \neq 0$ but at the null, where $\psi = 0$, the influence function for the parameter vanishes, the rate becomes n^{-1} and the limiting distribution is typically a combination of χ^2 random variables. Near the null, we get behavior in between these two cases. A valid confidence interval C_n should satisfy $P(\psi_n \in C_n) \to 1 - \alpha$ even if ψ_n is allowed to change with n. In particular, we want to allow $\psi_n \to 0$. Finding a confidence interval with this uniformly correct coverage, with length $n^{-1/2}$ away from the null and length n^{-1} at the null is, to the best of our knowledge, an unsolved problem.

Our proposal is to construct a conservative confidence interval that does not have length $O(1/n)$ at the null. We replace the standard error se of $\widehat{\psi}$ with $\sqrt{se^2 + c^2/n}$ where c is a constant. We take $c = (\text{Var}[Y])^2$ to put the quantity on the right scale, but other constants could be used. This leads to valid confidence intervals but they are conservative near the null as they shrink at rate $n^{-1/2}$ instead of n^{-1}.

We are only aware of two other attempts to address this issue. Both involve expanding the width of the confidence interval to be $O(n^{-1/2})$. Dai et al. (2021) added noise of the form $cZ/\sqrt{n}$ to the estimator, where $Z \sim N(0,1)$. They choose c by permuting the data many times and finding a c that gives good coverage under the simulated permutations. However, this is computationally expensive and adding noise seems unnecessary. Williamson et al. (2020) deal with this problem by writing ψ as a sum of two parameters $\psi = \psi_1 + \psi_2$ such that neither ψ_1 nor ψ_2 vanish when $\psi = 0$. Then, they estimate ψ_1 and ψ_2 on separate splits of the data. This again amounts to adding noise of size $O(1/\sqrt{n})$.

All three approaches are basically the same; they have the effect of expanding the confidence interval by $O(n^{-1/2})$ which maintains validity at the expense of efficiency at the null. Our approach has the virtue of being simple and fast. It does not require adding noise, extra calculations or doing an extra split of the data.

To see that expanding the standard error does lead to an interval with correct coverage, let $\widehat{\psi}$ denote an estimator of a parameter ψ_n which we allow to change with n. We are concerned with the case were the bias b_n satisfies $b_n = o(n^{-1/2})$ and the variance v_n satisfies $v_n = o(1/n)$. (The variance would be of order $1/n$ in the non-degenerating case.) Then, by Markov's inequality, the non-coverage of the interval $\widehat{\psi}_n \pm z_{\alpha/2}\sqrt{se^2 + c^2/n}$ is

$$\begin{aligned} P\Big(|\widehat{\psi}_n - \psi_n| > z_{\alpha/2}\sqrt{se^2 + c^2/n}\Big) &\leq P\Big(|\widehat{\psi}_n - \psi_n| > z_{\alpha/2}\sqrt{c^2/n}\Big) \\ &\leq \frac{n}{c z_{\alpha/2}^2}\mathbb{E}[|\widehat{\psi}_n - \psi_n|^2] = \frac{n}{c z_{\alpha/2}^2}(b_n^2 + v_n) = o(1). \end{aligned}$$

2.2 Issue 2: Correlation Distortion

The second and more pernicious problem is that ψ_L depends on the correlation between X and Z. In particular, if X and Z are highly correlated, then ψ_L will typically be close to 0. We call this, *correlation distortion*. There may be applications where this is acceptable. But in some cases we may want to alleviate this distortion and that is the focus of this paper.

To appreciate the effect of correlation distortion, consider the linear model $Y = \beta X + \theta Z + \epsilon$. In this case, a natural measure of variable importance is β which is unaffected by correlation between X and Z. The standard error of the estimate $\widehat{\beta}$ is affected by the correlation but the estimand itself is not. For this model, $\psi_L = \beta^2\gamma^2$ where $\gamma^2 = \mathbb{E}[(X - \nu(Z))^2]$ and $\nu(z) = \mathbb{E}[X|Z = z]$. This makes it clear that $\psi_L \to 0$ as X and Z become more correlated. The same fate befalls the partial correlation ρ between Y and X which in this model is $\rho = (1 + \frac{\beta^2\sigma^2}{\gamma^2})^{-1/2}$ where $\sigma^2 = \text{Var}[\epsilon]$. Again, $\rho \to 0$ as $\gamma \to 0$.

To deal with this problem, we define a modified LOCO parameter ψ_0 which is unaffected by the dependence between X and Z. Let $p_0(x, y, z) = p(y|x, z)p(x)p(z)$. Then p_0 is the distribution that is closest to p in Kullback-Leibler distance subject to making X and Z independent. We define

$$\psi_0 = \mathbb{E}_0[(\mu_0(X, Z) - \mu_0(Z))^2]. \tag{2}$$

A simple calculation shows that $\mu_0(z) = \mathbb{E}_0[Y|Z = z] = \int \mu(x, z)p(x)dx$ and so

$$\psi_0 = \int (\mu_0(x, z) - \mu_0(z))^2 p(x)p(z)dxdz. \tag{3}$$

We can think of ψ_0 as a counterfactual quantity answering the question: what would the change in $\mu(X, Z)$ be if we dropped X and had X and Z been independent.

This parameter completely eliminates the correlation distortion but, as we show in our simulations, it can be hard to get an accurate estimate of ψ_0. In particular, nonparametric confidence intervals are wide. A simple, but somewhat ad-hoc solution, is to first remove $Z_j's$ that are highly correlated with X. That is, define $\psi_1 = \mathbb{E}[(\mu(V) - \mu(X, V))^2]$ where $V = (Z_j : \ |\rho(X, Z_j)| \leq t)$ for some t where ρ is a measure of dependence.

The main solution we propose is to use the semiparametric model $\mu(x, z) = x^T\beta(z) + f(z)$. Under this model, one can show that ψ_0 takes the form $\text{tr}\left(\Sigma_X\mathbb{E}[\beta(Z)\beta(Z)^T]\right)$ where $\Sigma_X = \text{Var}[X]$. (See appendix 8.4 for details). However, this parameter is still difficult to estimate so we propose the following two simpler models. First, let $\mu(x, z) = \beta^T x + f(z)$. Then ψ_0 becomes

$$\psi_2 = \beta^T\Sigma_X\beta. \tag{4}$$

The second model is

$$\mu(x, z) = \beta^T x + \sum_j \sum_j \gamma_{jk}x_j z_k + f(z). \tag{5}$$

In Section 3.5 we show that ψ_0 then becomes

$$\psi_3 = \theta^T\Omega\theta \tag{6}$$

where

$$\theta = \left\{\mathbb{E}[\widetilde{Z}\widetilde{Z}^T \otimes (X - \nu(Z))(X - \nu(Z))^T]\right\}^{-1} \mathbb{E}\left[\Big(Y - \mu(Z)\Big)\ \Big(\widetilde{Z} \otimes (X - \nu(Z))\Big)\right].$$

$$\psi_0 = \int\int (\mu(x,z) - \mu_0(z))^2 p(x)p(z)dxdz \qquad \psi_1 = \mathbb{E}[(\mu(X,V) - \mu(V))^2]$$
$$\psi_2 = \beta^T \Sigma_X \beta \qquad \psi_3 = \theta^T \Omega \theta$$

$$\mu_0(z) = \int \mu(x,z)p(x)dx \qquad V = (Z_j : \ |\rho(X,Z_j)| \leq t)$$
$$\widetilde{Z}^T = (1, Z^T) \qquad \Omega = \Sigma_X \otimes \begin{bmatrix} 1 & m_Z^T \\ m_Z & \Sigma_Z + m_Z m_Z^T \end{bmatrix}$$
$$\beta = \mathbb{E}[(Y - \mu(Z))(X - \nu(Z))]/\mathbb{E}[(X - \nu(Z))^2]$$
$$\theta = \left\{ \mathbb{E}[\widetilde{Z}\widetilde{Z}^T \otimes (X - \nu(Z))(X - \nu(Z))^T] \right\}^{-1} \mathbb{E}\left[\Big(Y - \mu(Z)\Big) \ \Big(\widetilde{Z} \otimes (X - \nu(Z))\Big) \right]$$

Table 1: **Summary of Decorrelated Parameters**

.

$\nu(z) = \mathbb{E}[X|Z = z]$, $\widetilde{Z} = (1, Z)$ and

$$\Omega = \Sigma_X \otimes \mathbb{E}[\widetilde{Z}\widetilde{Z}^T] = \Sigma_X \otimes \begin{pmatrix} 1 & m_Z^T \\ m_Z & \Sigma_Z + m_Z m_Z^T \end{pmatrix},$$

$m_Z = \mathbb{E}[Z]$ and $\Sigma_Z = \text{Var}[Z]$. Table 1 summarizes the expressions for the parameters.

Remark: *Using the semiparametric model simplifies statistical inference for ψ_0. Of course, using a model always carries risks. In particular, if the model is not a reasonable approximation to $\mu(x,z)$ then we could be introducing bias. Therefore, as in all cases where a model is used, one should be aware that if the model is wrong then we are actually estimating the projection of $\mu(x,z)$ onto the model and then ψ_0 captures the importance of X in the projected model.*

Remark: *In all the above definitions, we can replace X with $b(X) = (b_1(X), \ldots, b_k(X))$ for a given set of basis functions $b_1, \ldots, b_k$ to make the model more flexible. For example, we can take $b(X) = (X, X^2, X^3)$ or an orthogonalized version of the polynomials, which is what we use in several of our examples.*

In these semiparametric models, we can estimate the nuisance functions $\nu(z) = \mathbb{E}[X|Z = z]$ and $\mu(z)$ either nonparametrically or parametrically.

3 Inference

In this section we discuss estimation of $\psi \in \{\psi_L, \psi_0, \psi_1, \psi_2, \psi_3\}$. For ψ_0, ψ_2 and ψ_3 we use one-step estimation which we now briefly review. See Hines et al. (2021) for a recent tutorial on one-step estimators. Let $\psi(\gamma)$ be a parameter with efficient influence function $\phi(u, \gamma, \psi)$ where γ denotes nuisance functions. We split the data into two groups $\mathcal{D}_0$ and $\mathcal{D}_1$

and we estimate γ from $\mathcal{D}_0$. Splitting the data is a common technique in semiparametric inference as it leads to central limit theorems under weaker conditions than would otherwise be necessary. The one-step estimator is

$$\widehat{\psi} = \widehat{\psi}_{\text{pi}} + \frac{1}{n}\sum_i \phi(U_i, \widehat{\gamma}, \widehat{\psi}_{\text{pi}})$$

where $\widehat{\psi}_{\text{pi}} = \psi(\widehat{\gamma})$ is the plug-in estimator and the average is over $\mathcal{D}_1$. This estimator comes from the von Mises expansion of $\psi(\gamma)$ around a point $\overline{\gamma}$ given by $\psi(\gamma) = \psi(\overline{\gamma}) + \int \phi(u, \overline{\gamma})dP(u) + R$ where R is the remainder. Alternatively, we can define $\widehat{\psi}$ as the solution to the estimating equation $n^{-1}\sum_i \phi(U_i, \widehat{\gamma}, \psi) = 0$.

Both estimators have second order bias $||\widehat{\gamma} - \gamma||^2$. Under appropriate conditions, both estimators satisfy $\sqrt{n}(\widehat{\psi} - \psi) \rightsquigarrow N(0, \tau^2)$ where $\tau^2 = \mathbb{E}[\phi^2(U, \gamma, \psi)]$. The key condition for this central limit theorem to hold is that $||\widehat{\gamma} - \gamma||^2 = o_P(n^{-1/2})$ which holds under standard smoothness assumptions. For example, if γ is in a Holder class of smoothness s, then an optimal estimator $\widehat{\gamma}$ satisfies $||\widehat{\gamma} - \gamma||^2 = O_P(n^{-2s/(2s+d)}) = o_P(n^{-1/2})$ when $s > d/2$. The plugin estimator has first order bias $||\widehat{\gamma} - \gamma||$ which will never be $o_P(n^{-1/2})$.

The usual confidence interval is $\widehat{\psi} \pm z_{\alpha/2}\text{se}$ where $\text{se}^2 = \widehat{\tau}^2/n$ and $\widehat{\tau}^2 = n^{-1}\sum_i \phi^2(U_i, \widehat{\gamma})$. But we find that this often underestimates the standard error. Instead, we use a different approach described in Section 3.6. We consider three different estimators for the nuisance functions $\mu(z)$ and $\nu(z)$: (i) linear, (ii) additive and (iii) random forests.

3.1 Estimating ψ_L

Williamson et al. (2021) found the efficient influence function for ψ_L. However, in Williamson et al. (2020) the authors note that one can avoid having to use the influence function by rewriting ψ_L as

$$\psi_L = \mathbb{E}[(Y - \mu(Z))^2] - \mathbb{E}[(Y - \mu(X, Z))^2].$$

It is easy to check that the corresponding plugin estimator

$$\widehat{\psi}_L = \frac{1}{n}\sum_i (Y_i - \widehat{\mu}(Z_i))^2 - \frac{1}{n}\sum_i (Y_i - \widehat{\mu}(X_i, Z_i))^2$$

already has second order bias $O(||\widehat{\mu} - \mu||^2)$ so that using the influence function is unnecessary.

3.2 Estimating ψ_0

We first derive the efficient, nonparametric estimator of ψ_0 and then we discuss some issues. Recall that $U = (X, Y, Z)$.

Theorem 1 *Let $\psi_0 = \psi_0(\mu, p) = \int\int (\mu(x, z) - \mu_0(z))^2 p(x)p(z)dxdz$. The efficient influence function is*

$$\begin{aligned}\phi(U, \mu, p) &= \int (\mu(x, Z) - \mu_0(Z))^2 p(x)dx + \int (\mu(X, z) - \mu_0(z))^2 p(z)dz \\ &\quad + 2\tfrac{p(X)p(Z)}{p(X,Z)}(\mu(X, Z) - \mu_0(Z))(Y - \mu(X, Z)) - 2\psi(p).\end{aligned}$$

In particular, we have the following von Mises expansion. Let $(\overline{\mu}, \overline{p})$ *be arbitrary and let* (μ, p) *denote the true functions. Then*

$$\psi_0(\mu, p) = \psi_0(\overline{\mu}, \overline{p}) + \int\int \phi(u, \overline{\mu}, \overline{p}) dP(u) + R$$

where the remainder R *satisfies*

$$R = O(||\overline{p}_X - p_X|| \times ||\overline{\delta} - \delta||) + O(||\overline{p}_Z - p_Z|| \times ||\overline{\delta} - \delta||) + O(||\overline{p}_X - p_X|| \times ||\overline{p}_Z - p_Z||) + O(||\overline{\delta} - \delta||^2)$$

and $\delta = \mu(x, z) - \mu_0(z)$. *Hence, if* $||\overline{p}_X - p_X|| = o_P(n^{-1/4})$, $||\overline{p}_Z - p_Z|| = o_P(n^{-1/4})$, $||\overline{\delta} - \delta|| = o_P(n^{-1/4})$ *then* $\sqrt{n}R = o_P(1)$.

The one-step estimator is

$$\widehat{\psi}_0 = \psi_0(\widehat{\mu}, \widehat{p}) + \frac{1}{n}\sum_i \phi(U_i, \widehat{\mu}, \widehat{p}).$$

The estimator from solving the estimating equation is $\widehat{\psi} = (2n)^{-1}\sum_i L(U_i, \widehat{\mu}, \widehat{p})$ where

$$\begin{aligned} L(U, \mu, p) &= \textstyle\int (\mu(x, Z) - \mu_0(Z))^2 p(x) dx + \int (\mu(X, z) - \mu_0(z))^2 p(z) dz \\ &\quad + 2\frac{p(X)p(Z)}{p(X,Z)}(\mu(X, Z) - \mu_0(Z))(Y - \mu(X, Z)). \end{aligned} \tag{7}$$

Corollary 2 *Suppose that* $||\widehat{p} - p|| = o_P(n^{-1/4})$ *and* $||\widehat{\mu} - \mu|| = o_P(n^{-1/4})$. *When* $\psi_0 \neq 0$, *for either of the two estimators above,*

$$\sqrt{n}(\widehat{\psi}_0 - \psi_0) \rightsquigarrow N(0, \sigma^2)$$

where $\sigma^2 = \mathbb{E}[\phi^2(U, \mu, p)]$.

In our implementation, we estimate $p(x, z), p(x), p(z)$ with kernel density estimators. We estimate integrals with respect to the densities by sampling from the kernel estimators. Specifically,

$$\widehat{\mu}_*(z) = \frac{1}{N}\sum_{j=1}^N \widehat{\mu}(X_j^*, z) \quad \text{where } X_1^*, \ldots, X_N^* \sim \widehat{p}(x).$$

Similarly, $\int (\mu(X, z) - \widehat{\mu}_0(z))^2 p(z)$ is estimated by

$$\frac{1}{N}\sum_j (\mu(X, Z_j^*) - \widehat{\mu}_0(Z_j^*))^2 \quad \text{where } Z_1^*, \ldots, Z_N^* \sim \widehat{p}(z)$$

and $\int (\mu(x, Z) - \widehat{\mu}_0(Z))^2 p(x)$ is estimated by $N^{-1}\sum_j (\mu(X_j^*, z) - \widehat{\mu}_0(z))^2$. Thus

$$\begin{aligned} \widehat{\psi}_0 &= \tfrac{1}{2n}\textstyle\sum_i L(U_i, \widehat{\mu}, \widehat{p}) \\ &= \tfrac{1}{nN}\textstyle\sum_i \sum_j \left(\widehat{\mu}(X_j^*, Z_i) - \frac{1}{N}\sum_{s=1}^N \widehat{\mu}(X_s^*, z)\right)^2 + \frac{1}{nN}\sum_i \sum_j \left(\widehat{\mu}(X_i, Z_j^*) - \frac{1}{N}\sum_{s=1}^N \widehat{\mu}(X_i, Z_s^*)\right)^2 \\ &\quad + \tfrac{2}{n}\textstyle\sum_i \frac{\widehat{p}(X_i)\widehat{p}(Z_i)}{\widehat{p}(X_i, Z_i)}\left(\widehat{\mu}(X_i, Z_i) - \frac{1}{N}\sum_j \widehat{\mu}(X_j^*, Z_i)\right)(Y_i - \widehat{\mu}(X_i, Z_i)). \end{aligned}$$

Finite Sample Problems. In principle, $\widehat{\psi}_0$ is fully efficient. In practice, $\widehat{\psi}_0$ can behave poorly as we now explain. One of the terms in the von Mises remainder is $||\widehat{\mu}_0(z) - \mu_0(z)||^2$. Now $\mu_0(z) = \int \mu(x,z)p(x)dx$. When X and Z are highly correlated, there will be a large set A_z of x values, where there are no observed data and so $\widehat{\mu}_0(z)$ will be quite far from $\mu_0(z)$ because $\widehat{\mu}(x,z)$ must suffer large bias or variance (or both) over that region. This is known as extrapolation error. For this reason we now consider alternative versions of ψ_0.[1]

3.3 Estimating ψ_1

Recall that $\psi_1 = \mathbb{E}[(\mu(X,V) - \mu(V))^2]$ where $V = (Z_j : \ |\rho(X,Z_j)| \leq t)$ for some t. We take $\rho(X,Z_j) = \sum_{i=1}^g |\rho(X_i,Z_j)|$ where $\rho(X_i,Z_j)$ is the Pearson correlation. We use $t = .5$ in our examples. For simplicity we assume that the values $\rho(X,Z_j)$ are distinct. In this case $P(\widehat{V} = V) \to 1$ as $n \to \infty$ where $\widehat{V} = (Z_j : \ |\widehat{\rho}(X,Z_j)| \leq t)$ and the randomness of $\widehat{V}$ can be ignored asymptotically and ψ_1 can be estimated in the same way as ψ_L with $\widehat{V}$ replacing Z.

Lemma 3 *If* $||\widehat{\mu}(x,v) - \mu(x,v)|| = o_P(n^{-1/4})$ *and* $\psi_1 \neq 0$ *then* $\sqrt{n}(\widehat{\psi}_1 - \psi_1) \rightsquigarrow N(0,\tau^2)$.

An alternative to removing correlated variables is to group together highly correlated variables and only report the variable importance of the group.

3.4 Estimating ψ_2

Consider the partially linear model $Y = \beta^T X + f(Z) + \epsilon$. Then $\mu_0(z) = \int \mu(x,z)p(x)dx = \beta^T m_X + f(z)$ where $m_X = \mathbb{E}[X]$ and so

$$\psi_2 \equiv \int\int (\mu(x,z) - \mu_0(z))^2 p(x)p(z)dxdz = \beta^T \Sigma_X \beta$$

and $\beta = \mathbb{E}[(Y - \mu(Z))(X - \nu(Z))]/\mathbb{E}[(X - \nu(Z))^2]$.

The efficient influence function for ψ_2 is

$$\phi = 2\beta^T \Sigma_X \phi_\beta + \beta^T((X - m_X)(X - m_X)^T)\beta - \psi_2$$

where

$$\phi_\beta = \Sigma_X^{-1}(X - \nu(Z))\Big\{(Y - \mu(Z)) - (X - \nu(Z))^T\beta)\Big\}$$

and we have the von Mises expansion $\psi_2(\mu,\nu,\beta,\Sigma_X) = \psi_2(\overline{\mu},\overline{\nu},\overline{\beta},\overline{\Sigma}_X) + \int \phi(u,\overline{\mu},\overline{\nu},\overline{\beta},\overline{\Sigma}_X)dP + R$ where the remainder R satisfies

$$\begin{aligned} R \quad &= O(||\mu(\overline{P}) - \mu(P)|| \times ||\nu(\overline{P}) - \nu(P)||) + O(||\text{vec}(\Sigma_X(\overline{P})) - \text{vec}(\Sigma_X(P))||^2) \\ &\quad + O(||\beta(\overline{P}) - \beta(P)||^2) + O(||\beta(\overline{P}) - \beta(P)|| \times ||\text{vec}(\Sigma_X(\overline{P})) - \text{vec}(\Sigma_X(P))||). \end{aligned}$$

1. Readers familiar with causal inference will recognize that, formally, $\mu_0(z)$ is the average treatment effect if we think of Z as a treatment and X as a confounder. But the role of treatment and confounder is switched with the treatment being the multivariate vector Z. The difficulty in estimating $\mu_0(z)$ when X and Z are highly correlated is known as the overlap problem in causal inference (D'Amour et al., 2021).

We omit the calculation of the influence function and remainder as they are standard. Hence, if $||\mu(\overline{P}) - \mu(P)|| \times ||\nu(\overline{P}) - \nu(P)||) = o(n^{-1/2})$, $||\beta(\overline{P}) - \beta(P)|| = o(n^{-1/4})$, and $||\text{vec}(\Sigma_X(\overline{P})) - \text{vec}(\Sigma_X(P))|| = o(n^{-1/4})$, then $\sqrt{n}R = o(1)$. It is easy to verify that $||\beta(\overline{P}) - \beta(P)|| = O(||\mu(\overline{P}) - \mu(P)|| \times ||\nu(\overline{P}) - \nu(P)||)$ and so ψ_2 satisfies the double robustness property, namely, that the bias involves the product of two quantities. It suffices to estimate either μ or ν accurately to get a consistent estimator.

The one-step estimator is given by

$$\widehat{\psi}_2 = \frac{1}{n}\sum_i \widehat{\beta}^T(X_i - \widehat{\mu}(Z_i))(X_i - \widehat{\mu}(Z_i))^T\widehat{\beta} + \frac{2}{n}\sum_i \widehat{\beta}^T\widehat{\Sigma}_X\phi_\beta(X_i, Z_i)$$

where

$$\widehat{\beta} = \left\{\frac{1}{n}\sum_i (X_i - \widehat{\nu}(Z_i))(X_i - \widehat{\nu}(Z_i))^T\right\}^{-1}\frac{1}{n}\sum_i (X_i - \widehat{\nu}(Z_i))(Y_i - \widehat{\mu}(Z_i))$$

and the sums are over $\mathcal{D}_1$.

3.5 Estimating ψ_3

Consider the partially linear model with interactions:

$$Y = \beta^T X + \sum_{j=1}^{g}\sum_{k=1}^{h}\gamma_{jk}X_jZ_k + f(Z) + \epsilon.$$

Define

$$\Theta = \begin{bmatrix} \beta_1 & \gamma_{11} & \cdots & \gamma_{1h} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_g & \gamma_{g1} & \cdots & \gamma_{gh} \end{bmatrix}, \quad \mathbb{W} = \begin{bmatrix} X_1 & X_1Z_1 & \cdots & X_1Z_h \\ \vdots & \vdots & \vdots & \vdots \\ X_g & X_gZ_1 & \cdots & X_gZ_h \end{bmatrix} = X\ \widetilde{Z}^T$$

where $\widetilde{Z}^T = (1, Z^T)$. Then we can write

$$Y = \theta^T W + f(Z) + \epsilon$$

where $\theta = \text{vec}(\Theta)$ and $W = \text{vec}(\mathbb{W}) = \text{vec}(X\widetilde{Z}^T) = \widetilde{Z} \otimes X$.

Lemma 4 *We have*

$$\theta = \left\{\mathbb{E}[\widetilde{Z}\widetilde{Z}^T \otimes (X - \nu(Z))(X - \nu(Z))^T]\right\}^{-1}\mathbb{E}\left[\Big(Y - \mu(Z)\Big)\ \Big(\widetilde{Z} \otimes (X - \nu(Z))\Big)\right]$$

and under this model, ψ_0 is equal to $\psi_3 = \theta^T\Omega\theta$ where

$$\Omega = \Sigma_X \otimes \mathbb{E}[\widetilde{Z}\widetilde{Z}^T] = \Sigma_X \otimes \begin{pmatrix} 1 & m_Z^T \\ m_Z & \Sigma_Z + m_Z m_Z^T \end{pmatrix},$$

$m_Z = \mathbb{E}[Z]$ *and* $\Sigma_Z = \mathrm{Var}[Z]$. *The efficient influence function for* ψ_3 *is*

$$\phi = 2\theta^T \Omega \phi_\theta + \theta^T \mathring{\Omega} \theta - \psi_3 \tag{8}$$

where

$$\phi_\theta = \Big\{ \mathbb{E}[R_{XZ} R_{XZ}^T] \Big\}^{-1} R_{XZ}(R_Y - R_{XZ}^T \theta),$$

$R_Y = Y - \mu(Z)$, $R_{XZ} = \mathrm{vec}[(X - \nu(Z))\widetilde{Z}^T]$,

$$\mathring{\Omega} = \left\{ [(X - m_X)(X - m_X)^T - \Sigma_X] \otimes \begin{bmatrix} 1 & m_Z^T \\ m_Z & \Gamma \end{bmatrix} \right\} + \left\{ \Sigma_X \otimes \begin{bmatrix} 0 & (Z - m_Z)^T \\ Z - m_Z & \mathring{\Gamma} \end{bmatrix} \right\},$$

(the influence function of Ω*)* $\Gamma = \Sigma_Z + m_Z m_Z^T$, *and*

$$\mathring{\Gamma} = (Z - m_Z)(Z - m_Z)^T - \Sigma_Z + m_Z(Z - m_Z)^T + (Z - m_Z)m_Z^T$$

(the influence function of Γ*).*

Then $\psi_3(u, \theta, \Omega) = \psi_3(u, \overline{\theta}, \overline{\Omega}) + \int \phi(u, \overline{\theta}, \overline{\Omega}) dP(u) + R$ where the remainder R satisfies

$$\begin{aligned} R \;\; &= O(||\theta(\overline{P}) - \theta(P)||^2) + O(||\mathrm{vec}(\Omega(\overline{P})) - \mathrm{vec}(\Omega(P))||^2) \\ &\quad + O(||\theta(\overline{P}) - \theta(P)|| \times ||\mathrm{vec}(\Omega(\overline{P})) - \mathrm{vec}(\Omega(P))||). \end{aligned}$$

Thus if $||\theta(\overline{P}) - \theta(P)|| = o(n^{-1/4})$ and $||\mathrm{vec}(\Omega(\overline{P})) - \mathrm{vec}(\Omega(P))|| = o(n^{-1/4})$ then $\sqrt{n}R = o(1)$. Again, we have the double robustness property.

The sample estimate of θ is $\widehat{\theta} = (\mathbb{R}_{XZ}^T \mathbb{R}_{XZ})^{-1} \mathbb{R}_{XZ}^T \mathbb{R}_Y$ where the i^{th} row of $\mathbb{R}_{XZ}$ is $\mathrm{vec}[(X_i - \widehat{\nu}(Z_i))\widetilde{Z}_i^T]$ and $\mathbb{R}_Y(i) = Y_i - \widehat{\mu}(Z_i)$. Let $\widehat{\Omega}$ be the sample version of Ω. The one-step estimator is

$$\widehat{\psi}_3 = \frac{1}{n} \sum_i \widehat{\theta}^T \widehat{\phi}_\Omega(U_i) \widehat{\theta} + \frac{2}{n} \sum_i \widehat{\theta}^T \widehat{\Omega} \phi_\theta(U_i)$$

where the sums are over $\mathcal{D}_1$.

3.6 Confidence Intervals

Now we describe the construction of the confidence intervals using a method we refer to as t-Cross. Let ψ denote a generic parameter. We combine two ideas: cross-fitting (Newey and Robins, 2018) and t-inference (Ibragimov and Müller, 2010). Here are the steps:

1. Divide the data into B disjoint sets $\mathcal{D}_1, \ldots, \mathcal{D}_B$; we take $B = 5$ in the examples.
2. Estimate the nuisance functions using all the data except $\mathcal{D}_j$ and compute $\widehat{\psi}_j$ on $\mathcal{D}_j$. Here, $\widehat{\psi}_j$ is the estimate of ψ using the data in $\mathcal{D}_j$.
3. Let $\overline{\psi} = B^{-1} \sum_{j=1}^B \widehat{\psi}_j$. When $\psi \neq 0$, each $\widehat{\psi}_j$ is asymptotically Normal so that $\overline{\psi}$ is asymptotically t_{B-1}.

4. The confidence interval is

$$\overline{\psi} \pm t_{B-1,\alpha/2} \text{ se}$$

where $\text{se}^2 = (s^2/B + c^2/n)$ where $s^2 = (B-1)^{-1}\sum_{j=1}^{B}(\widehat{\psi}_j - \overline{\psi})^2$.

The t-method loses some efficiency because it divides the data into groups. The rate of convergence does not change but the interval could be slightly larger. But the advantage is that s^2 is an unbiased estimate of the variance of $\widehat{\psi}$ which does not depend on the accuracy of the estimated influence function. So we are trading efficiency for robustness.

Remark: *Nonparametric and semiparametric confidence intervals require fairly strict assumptions. For example, we need to assume fast rates for the nuisance functions. An alternative is to use variability intervals which are centered at the mean of the estimator rather than at the true value. This might be less informative but requires much weaker assumptions.*

4 Simulations

In this section, we compare the behavior of the different parameters in some synthetic examples. For each example, we estimate all the parameters $\psi_L, \psi_0, \psi_1, \psi_2, \psi_3$. To estimate the parameters we need to estimate the nuisance functions $\mu(z)$ and $\nu(z)$. As mentioned above, we consider three approaches to estimating these functions: linear models, additive models and random forests. For the additive models we use the R package `mgcv`. For random forests we use the R package `grf`. We always use the default settings making no attempt to tune the methods to achieve good coverage.

Example 1. We start with a very simple scenario where $Y = 2X + \epsilon$, $\epsilon \sim N(0,1)$, $Z_1 = \delta X + \xi$, $\xi \sim N(0,1)$, and $(Z_2, \ldots, Z_5) \sim N(0, I)$. Figure 2 shows the coverage as a function of the correlation between X and Z_1. As expected, ψ_L has poor coverage as the correlation increases. The parameter ψ_0 partially corrects the correlation distortion while the other parameters do a much better job. The coverage for ψ_1 decreases as correlation increases. However, when the correlation is large enough, it becomes easier to identify correlated variables and then the coverage increases. The true values of the parameters are plotted in Figure 1.

Examples 2-5. Now we consider four multivariate examples. In each case, $n = 10,000$, $h = 5$ and $\epsilon \sim N(0,1)$. The distributions are defined as follows:

Example 2: X is standard Normal, $Z_1 = X + N(0, .4^2)$, $(Z_2, \ldots, Z_h)$ is standard multivariate Normal. The regression function is $Y = 2X^3 + \epsilon$. Hence $\text{Cor}(X_1, Z) = .93$.

Example 3: Here, $Z \sim N(0, I)$, $X_1 = 2Z_1 + \epsilon_1$, $X_2 = 2Z_2 + \epsilon_2$, $Y = 2X_1X_2 + \epsilon$ where $\epsilon, \epsilon_1, \epsilon_2 \sim N(0,1)$. Hence $\text{Cor}(X_1, Z_1) = \text{Cor}(X_2, Z_2) = .89$.

Example 4: Let $X \sim \text{Unif}(-1,1)$, $Z \sim \text{Unif}(-1,1)$, and $Y = X^2(X + (7/5)) + (25/9)Z^2 + \epsilon$. This example is from Williamson et al. (2021). Our coverage for ψ_L is similar but slightly less than that in Williamson et al. (2021) but we are using a different nonparametric estimator. In this case, X and Z are uncorrelated.

Example 5: $X \sim N(0,1)$, $Z_1 = X + N(0, .4^2)$ $(Z_2, \ldots, Z_d) \sim N(0, I)$ and $Y = 2X^2 + XZ_1 + \epsilon$. In this case $\text{Cor}(X, Z_1) = .93$

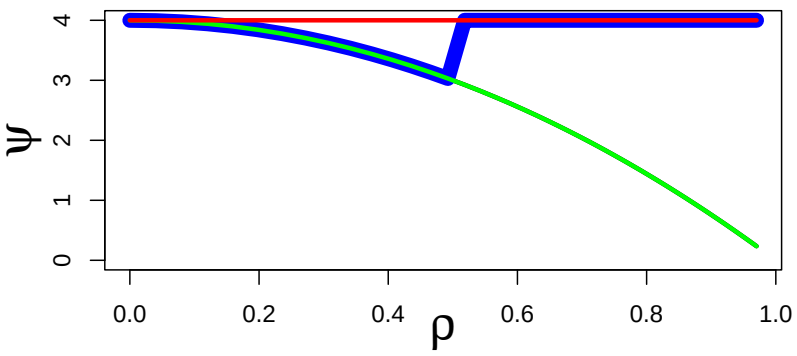

Figure 1: This plot shows the true values of the parameters in Example 1 as a function of the correlation ρ between X and Z. The top red line is $\psi_0 = \psi_2 = \psi_3$. The green line is ψ_L. The blue line is ψ_1 which equals ψ_L for $\rho < .5$ and equals ψ_0 for $\rho > .5$.

In examples 2,4 and 5, we replaced X with orthogonal polynomials $b_1(X), b_2(X), b_3(X)$.

The results from 100 simulations are summarized in Figures 2 and 3 and in Table 2. The standard error of the coverage is 0.03. Figure 2 shows how often the confidence interval contains the target parameter ψ_0 as a function of the correlation which varies from 0 to 1. In other words, we treat $\psi_0 = \beta^2 = 4$ as the truth and we evaluate how well an interval based on estimating ψ_j covers ψ_0. They all cover well except ψ_L and ψ_1. This is to be expected as $\psi_0 = \psi_2 = \psi_3$ in this example. However, ψ_L decreases as a function of the correlation. In fact, we evaluated how often the interval for ψ_L contains the true value of ψ_L. It turns out that the coverage of the interval based on ψ_L does cover ψ_L at the nominal level (although, as with many examples, the forest based method tends to sometimes undercover). The coverage for ψ_1 goes down and then up because Z_1, which is correlated with X, gets removed when the correlation is large enough. Essentially, when the correlation is less than .5, $\psi_1 = \psi_L$ but after that, Z_1 is removed and $\psi_1 = \psi_0$. This shows the inherent instability of trying to remove correlated variables.

Figure 3 shows the average of the left and right endpoints of the confidence intervals. The vertical line marks our target which is ψ_0. The first thing to notice is that no method does uniformly well. Inferences for ψ_3 are mostly pretty good, but the others are not and this is to be expected. The coverage of ψ_L is poor because it is not targeting the right parameters. Similarly for ψ_1. The poor coverage of ψ_0 in some cases is due to the difficulty of estimating the parameter nonparametrically. ψ_2 does not include interactions and does poorly when there are interactions. The random forest method has a tendency to undercover. However, what is not shown here, is that each method does cover its own target at the nominal level.

Estimating variable importance well is surprisingly difficult. Generally, we find that ψ_3 works best. However, it does poorly in two cases: in Example 5, with linear regressions, and in Example 2 using random forests. ψ_0 rarely does well. Apparently, the functional is

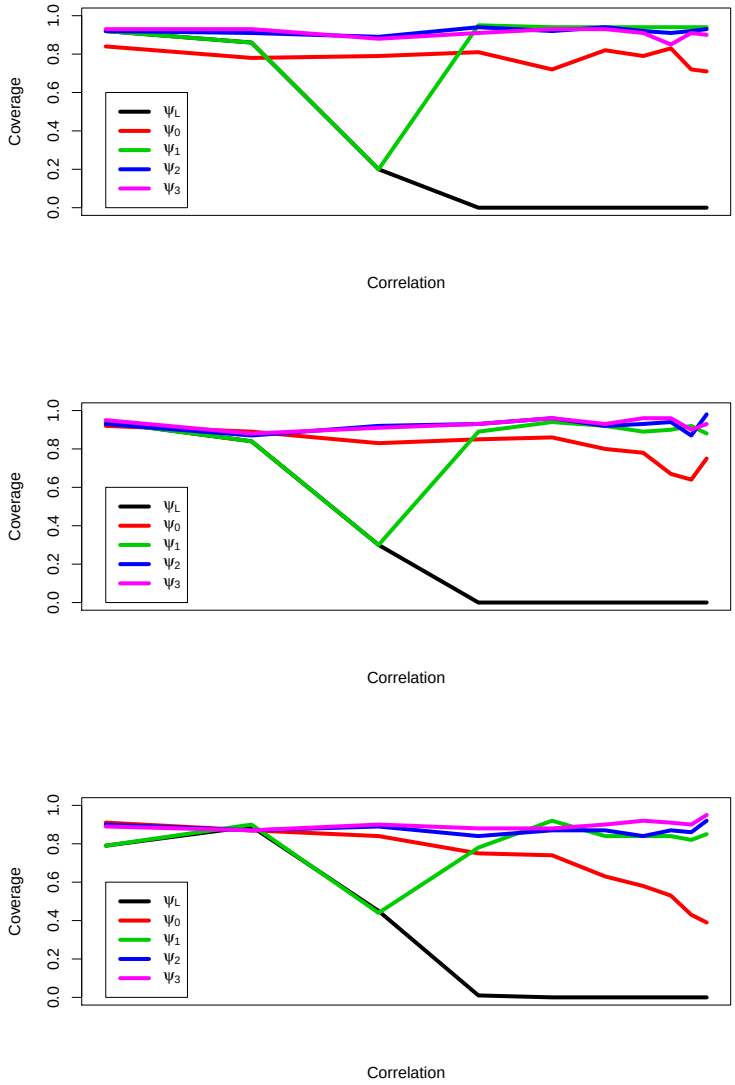

Figure 2: *Example 1: Coverage as a function of correlation. Top: linear. Middle: additive. Bottom: forests.*

too difficult to estimate nonparametrically. ψ_1 works well in a few cases, but is not reliable enough in general. Similar behavior occurs for ψ_2. Except for a few cases, ψ_L never does well. This is not unexpected due to the correlation distortion.

However, it should be noted that these methods are all doing well in the sense of covering the value of ψ in the projected model at the nominal level. For example, when using linear models for μ and ν, we are really estimating the value of ψ for the projection of the distribution onto the space of linear models. The parameter estimate may capture useful information even if it is not estimating ψ_0.

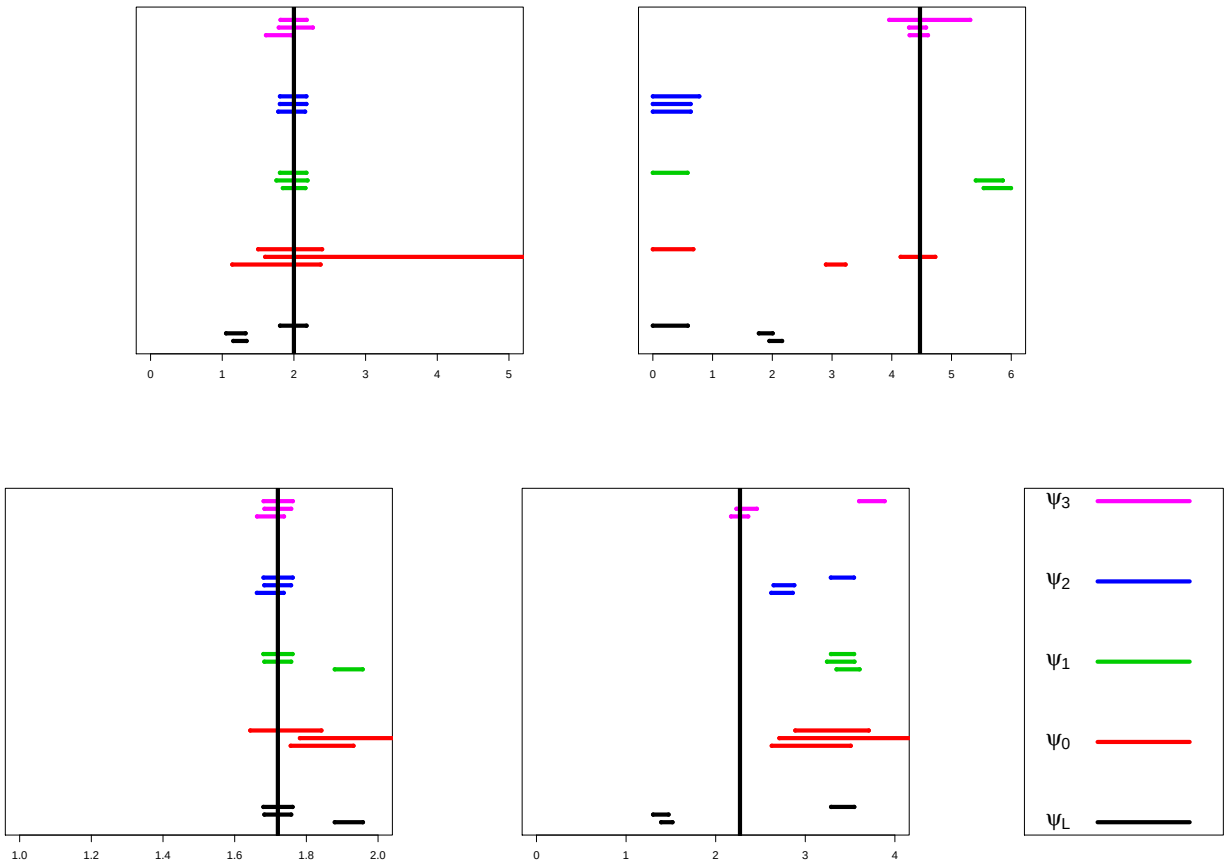

Figure 3: *The average of the left and right endpoints of the confidence intervals over 100 simulations for Examples 2,3,4,5. The vertical line is ψ_0. The plot shows how the confidence intervals of each parameter compare to the true value of ψ_0. Top left is Example 2. Top right is Example 3. Bottom left is Example 4. Bottom middle is Example 5. The bottom right shows the legend for all the plots. In each panel, the groups of three line segments correspond to the three different models: the top is based on linear models, the middle is based on additive models and the bottom is based on random forests.*

5 Other Issues

In this section we discuss two further topics: other variable importance parameters, and Shapley values.

	Linear					Additive					Forest				
	ψ_L	ψ_0	ψ_1	ψ_2	ψ_3	ψ_L	ψ_0	ψ_1	ψ_2	ψ_3	ψ_L	ψ_0	ψ_1	ψ_2	ψ_3
Example 2	1	0.84	1	1	1.00	0.00	0.79	1.00	1.00	0.97	0.00	0.75	1.00	0.99	0.30
Example 3	0	0.00	0	0	0.99	0.00	0.88	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.91
Example 4	1	0.87	1	1	1.00	0.98	0.20	0.98	0.98	0.98	0.00	0.21	0.00	0.86	0.85
Example 5	0	0.01	0	0	0.00	0.00	0.83	0.00	0.00	0.85	0.00	0.05	0.00	0.00	1.00

Table 2: *Coverage results for Examples 2,3,4 and 5. The standard error on the estimates coverage is 0.03. Overall, ψ_3 performs best in these examples. But when linear regressions are used, ψ_3 fails. For random forests, ψ_3 does poorly in Example 2. The most robust behavior is given by the additive model.*

5.1 Other Parameters

We have focused on LOCO in this paper but there are many other variable importance parameters all of which can be estimated in a manner similar to the methods in this paper. Samarov (1993) suggested $\psi = \int (\partial\mu(x,z)/\partial x)^T(\partial\mu(x,z)/\partial x)dP$. This parameter is not subject to correlation distortion. Estimating derivatives can be difficult but in the semi-parametric case, ψ takes a simple form. In the partially linear model we have $\psi = ||\beta||^2$ and in the partially linear model with interactions (5) we have

$$\psi = ||\beta||^2 + 2\beta G^T m_Z + G^T \Sigma_Z G$$

where $G_{jk} = \gamma_{jk}$.

Another parameter is inspired by causal inference. If we viewed X as a treatment and Z as confounding variables, then (under some conditions) the causal effect, that is the mean of Y had X been set to x, is given by Robins' g-formula $g(x) = \int \mu(x,z)dP(z)$. We could then define ψ as the variance $\mathrm{Var}[g(X)]$ or the average squared derivative of $\int (\partial g(x)/\partial x)^T(\partial g(x)/\partial x)dP$. These parameters do not suffer from correlation distortion. Now $\mathrm{Var}[g(X)]$ equals $\beta^T\Sigma_X\beta$ under the partially linear model and is $(\beta+\Gamma m_Z)^T\Sigma_X(\beta+\Gamma m_Z)$ under the partially linear model with interactions. Using the derivative, in the partially linear model we get $\psi = ||\beta||^2$ and in partially linear model with interactions we get

$$\psi = ||\beta||^2 + 2\beta\Gamma^T m_Z + \Gamma^T m_Z m_Z^T \Gamma.$$

The nonparametric partial correlation is defined by

$$\rho = \frac{\mathbb{E}[(Y-\mu(Z))(X-\nu(Z))]}{\sqrt{\mathbb{E}(Y-\mu(Z))^2\mathbb{E}(X-\nu(Z))^2}}.$$

Under p_0 we get a decorrelated version

$$\rho_0 = \frac{\mathbb{E}_0[(Y-\mu_0(Z))(X-\nu_0(Z))]}{\sqrt{\mathbb{E}_0(Y-\mu_0(Z))^2\mathbb{E}_0(X-\nu_0(Z))^2}} = \frac{\int\int(\mu(x,z)-\mu_0(z))(x-m_X)p(x)p(z)dxdz}{\sigma_X\sqrt{\int\int\int(y-\mu_0(z))^2p(y|x,z)p(x)p(z)}}.$$

More detail about ρ_0 are in the appendix.

5.2 Shapley Values

A method for defining variable importance that has attracted much attention lately is based on Shapley values (Messalas et al., 2019; Aas et al., 2019; Lundberg and Lee, 2016; Covert et al., 2020; Fryer et al., 2020; Covert and Lee, 2020; Israeli, 2007; Mase et al., 2019; Bénard et al., 2021). This is an idea from game theory where the goal is to define the importance of each player in a cooperative game. While Shapley values can be useful in some settings, for example, computer experiments (Owen and Prieur, 2017) we argue here that Shapley values do not solve the decorrelation issue and LOCO or decorrelated LOCO may be preferable for routine regression problems. However, this is an active area of research and the issue is far from settled. Shapley values may indeed have some other advantages.

The Shapley value is defined as follows. Suppose we have covariates $(Z_1, \ldots, Z_d)$ and that we want to measure the importance of Z_j. For any subset $S \subset \{1, \ldots, d\}$ let $Z_S = (Z_j : \ j \in S)$ and let $\mu(S) = \mathbb{E}[Y|Z_S]$. The Shapley value for Z_j is

$$s_j = \frac{1}{d!} \sum_{\pi} [V(S_j^+(\pi)) - V(S_j(\pi))]$$

where the sum is over of permutations of $(Z_1, \ldots, Z_d)$, $S_j(\pi)$ denotes all variables before Z_j in permutation π, $S_j^+(\pi) = \{S_j(\pi) \bigcup \{j\}\}$ and $V(S)$ is some measure of fit the regression model with variables S. If $V(S) = -\mathbb{E}[(Y - \mu(S))^2]$, then

$$s_j = \frac{1}{d!} \sum_{\pi} \mathbb{E}[(\mu(S_j) - \mu(S_j^+))^2].$$

This is just the LOCO parameter averaged over all possible submodels. The Shapley value for a group of variables can be defined similarly.

It is clear that this parameter is difficult to compute and inference, while possible (Williamson and Feng, 2020) is very challenging. The appeal of the Shapley value is that it has the following nice properties:

(A1): $\sum_j s_j = \mathbb{E}[(Y - \mu(Z))^2]$.
(A2) If $\mathbb{E}[(Y - \mu(S \bigcup \{i\}))^2] = \mathbb{E}[(Y - \mu(S \bigcup \{j\}))^2]$ for every S not containing i or j, then $s_i = s_j$.
(A3) If we treat $\{Z_j, Z_k\}$ as one variable, then its Shapley value s_{jk} satisfies $s_{jk} = s_j + s_k$.
(A4) If $\mathbb{E}[(Y - \mu(S \bigcup \{j\}))^2] = \mathbb{E}[(Y - \mu(S))^2]$ for all S then $s_j = 0$.

However, we see two problems with Shapley values applied to regression. First, it defines variable importance with respect to all submodels. But most of those submodels are not of interest. Indeed, most of them would be a bad fit to the data and are not relevant. So it is not clear why we should involve them in any definition of variable importance or in the axioms. (An intriguing idea might be to weight the submodels according to their predictive value). Second, they succumb to correlation distortion. To see this, suppose that $Y = \beta Z_1 + \epsilon$, that the Z_j's have variance 1 and that they are perfectly correlated, that is, $P(Z_j = Z_k) = 1$ for every j and k. The Shapley value for Z_1 turns out to be $s_1 = \beta^2/d$ which is close to 0 when d is large. In contrast, $\psi_0 = \beta^2$, which seems more a appropriate.

The confidence interval for ψ_0 would have infinite length since the design is singular which also seems appropriate, since estimating the importance of a single variable among a set of perfectly correlated variables should be an impossible inferential task. For these reasons, we feel that decorrelated LOCO may have some advantages over Shapley values.

6 Conclusion

We showed that correlation distortion can be removed from LOCO by modifying the definition appropriately. This leads to the parameter ψ_0. As we have seen, getting valid inferences for ψ_0 nonparametrically is difficult even in fairly simple examples. This is mainly because the parameter involves the function $\mu_0(z) = \int \mu(x,z)p(x)dx$ which requires estimating $\mu(x,z)$ in regions where there is little data due to the dependence between x and z. The easiest remedy is to remove correlated variables as we did for ψ_1 but this led to disappointing behavior. The other remedy was to use a semiparametric model for $\mu(x,z)$ which led to ψ_2 and ψ_3. This appears to be the best approach. We emphasize that even when the coverage for ψ_2 and ψ_3 is low, (when the semiparametric model is misspecified), these parameters are still useful if we interpret them as projections. For example, ψ_2 measures the variable importance of X in the regression function of the form $\beta x + f(z)$ that best approximates $\mu(x,z)$. In the sense ψ_2 still captures part of the variable importance. Graham and de Xavier Pinto (2021) discuss in detail the interpretation of misspecified semiparametric models.

We only dealt with low dimensional models. The methods extend to high dimensional models by using the usual sparsity based estimators for the nuisance functions $\mu(z)$ and $\nu(z)$. We plan to explore this in future work.

Finally, we briefly discussed the role of Shapley values which have become popular in the literature on variable importance. The motivation for using Shapley values appears to be that they might alleviate correlation distortion. Indeed, if the variables were independent, Shapley values would probably not be considered. But we argued that they do not adequately address the problem. Instead, we believe that some form of decorrelation might be preferred.

Acknowledgments

We would like to acknowledge two referees, whose comments helped to improve the paper.

7 Appendix

In this appendix we have proofs and details for a few other parameters.

7.1 Proofs

Theorem 1. *Let* $\psi_0(\mu,p) = \int\int(\mu(x,z)-\mu_0(z))^2p(x)p(z)dxdz$. *The efficient influence function is*

$$\begin{aligned}\phi(X,Y,Z,\mu,p) &= \int(\mu(x,Z)-\mu_0(Z))^2p(x)dx + \int(\mu(X,z)-\mu_0(z))^2p(z)dz \\ &+ 2\frac{p(X)p(Z)}{p(X,Z)}(\mu(X,Z)-\mu_0(Z))(Y-\mu(X,Z)) - 2\psi(p).\end{aligned}$$

In particular, we have the following von Mises expansion

$$\psi_0(\mu,p) = \psi_0(\overline{\mu},\overline{p}) + \int\int\phi(x,y,z,\overline{\mu},\overline{p})dP(x,y,z) + R$$

where the remainder R *satisfies*

$$\begin{aligned}||R|| &= O(||\overline{p}(x,z)-p(x,z)||^2) + O(||\overline{\mu}(x,z)-\mu(x,z)||^2) \\ &+ O(||\overline{p}(x,z)-p(x,z)||\times||\overline{\mu}(x,z)-\mu(x,z)||).\end{aligned}$$

Proof. To show that $\phi(X,Y,Z,\mu,p)$ is the efficient influence function we verify that $\phi(X,Y,Z,\mu,p)$ is the Gateuax derivative of ψ and that it has the claimed second order remainder. We will use the symbol $'$ to denote the Gateuax derivative defined by

$$\lim_{\epsilon\to 0}\frac{\psi_0((1-\epsilon)P+\epsilon\delta_{XYZ})-\psi_0(P)}{\epsilon}$$

where δ_{XYZ} is a point mass at (X,Y,Z). Also, let δ_X denote a point mass at X, δ_{XY} a point mass at (X,Y) etc. Let $w(x,z)=p(x)p(z)$. Then $\psi_0=\int\int(\mu(x,z)-\mu_0(z))^2w(x,z)dxdz$. Now

$$\psi' = \int\int(\mu(x,z)-\mu_0(z))^2w'(x,z)dxdz + 2\int\int w(x,z)(\mu(x,z)-\mu_0(z))(\mu'(x,z)-\mu_0'(z))dxdz$$

First, note that $w'(x,z)=p(x)(\delta_Z(z)-p(z))+p(z)(\delta_X(x)-p(x))$. Next

$$\mu(x,z) = \int yp(y|x,z)dy = \int y\frac{p(x,y,z)}{p(x,z)}dy$$

and

$$\mu_\epsilon(x,z) = \int y\frac{p(x,y,z)+\epsilon(\delta_{XYZ}-p(x,y,z))}{p(x,z)+\epsilon(\delta_{XZ}-p(x,z))}dy$$

So

$$\begin{aligned}\mu'(x,z) &= \int y\left\{\frac{p(x,z)(\delta_{XYZ}-p(x,y,z))-p(x,y,z)(\delta_{XZ}-p(x,z))}{p^2(x,z)}\right\}dy \\ &= \frac{Y}{p(X,Z)}I(x=X,z=Z)-\mu(x,z)-\frac{\mu(x,z)I(x=X,z=Z)}{p(x,z)}+\mu(x,z) \\ &= \frac{(Y-\mu(x,z))}{p(x,z)}I(x=X,z=Z)\end{aligned}$$

Now $\mu_0(z) = \int \mu(x,z)p(x)\,dx$ so

$$\begin{aligned}\mu_0'(z) &= \int \mu(x,z)(\delta_X(x) - p(x))dx + \int p(x)\mu'(x,z)dx \\ &= \mu(X,z) - \mu_0(z) + \frac{(Y-\mu(X,z))p(X)}{p(X,z)} I(z=Z)\end{aligned}$$

so

$$\begin{aligned}\phi(X,Y,Z,\mu,p) &= \int (\mu(x,Z)-\mu_0(Z))^2 p(x)dx - \psi + \int (\mu(X,z)-\mu_0(z))^2 p(z)dz - \psi \\ &+ 2\, w(X,Z)(\mu(X,Z)-\mu_0(Z))\frac{(Y-\mu(X,Z))}{p(X,Z)} \\ &- 2\int\int w(x,z)(\mu(x,z)-\mu_0(z))(\mu(X,z)-\mu_0(z))dxdz \\ &- 2\,\frac{(Y-\mu(X,Z))p(X)}{p(X,Z)}\int w(x,Z)(\mu(x,Z)-\mu_0(Z))dx \\ &= \int (\mu(x,Z)-\mu_0(Z))^2 p(x)dx + \int (\mu(X,z)-\mu_0(z))^2 p(z)dz - 2\psi \\ &+ 2\, w(X,Z)(\mu(X,Z)-\mu_0(Z))\frac{(Y-\mu(X,Z))}{p(X,Z)} \\ &- 2\int\int w(x,z)(\mu(x,z)-\mu_0(z))(\mu(X,z)-\mu_0(z))dxdz \\ &- 2\,\frac{(Y-\mu(X,Z))p(X)p(Z)}{p(X,Z)}\int p(x)(\mu(x,Z)-\mu_0(Z))dx \\ &= \int (\mu(x,Z)-\mu_0(Z))^2 p(x)dx + \int (\mu(X,z)-\mu_0(z))^2 p(z)dz \\ &+ 2\,\frac{p(X)p(Z)}{p(X,Z)}(\mu(X,Z)-\mu_0(Z))(Y-\mu(X,Z)) - 2\psi(p)\end{aligned}$$

which has the claimed form.

Now we consider the von Mises remainder. The remainder at (p,μ) in the direction of $(\overline{p},\overline{\mu})$ is

$$R = \psi(p,\mu) - \psi(\overline{p},\overline{\mu}) - \int \phi(u,\overline{\mu},\overline{p})dP(u).$$

Now

$$\begin{aligned}-R &= \psi(\overline{p},\overline{\mu}) - \psi(p,\mu) \\ &+ \int\int \overline{p}(x)p(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))^2\,dx\,dz + \int\int p(x)\overline{p}(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))^2\,dx\,dz \\ &+ 2\int\int\int p(x,y,z)\frac{\overline{p}(x)\overline{p}(z)}{\overline{p}(x,z)}(\overline{\mu}(x,z)-\overline{\mu}_0(z))(y-\overline{\mu}(x,z))\,dx\,dy\,dz - 2\psi(p) \\ &= \int\int \overline{p}(x)p(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))^2\,dx\,dz + \int\int p(x)\overline{p}(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))^2\,dx\,dz \\ &+ 2\int\int\int p(x,y,z)\frac{\overline{p}(x)\overline{p}(z)}{\overline{p}(x,z)}(\overline{\mu}(x,z)-\overline{\mu}_0(z))(y-\overline{\mu}(x,z))\,dx\,dy\,dz - \psi(\overline{p},\overline{\mu}) - \psi(p,\mu)\end{aligned}$$

$$
\begin{aligned}
&= \int\int \overline{p}(x)p(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))^2\ dx\,dz + \int\int p(x)\overline{p}(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))^2\ dx\,dz \\
&+ 2\int\int p(x,z)\frac{\overline{p}(x)\overline{p}(z)}{\overline{p}(x,z)}(\overline{\mu}(x,z)-\overline{\mu}_0(z))(\mu(x,z)-\overline{\mu}(x,z))\ dx\,dz - \psi(\overline{p},\overline{\mu}) - \psi(p,\mu) \\
&= \int\int \overline{p}(x)\ p(z)\ (\overline{\mu}(x,z)-\overline{\mu}_0(z))^2\ dx\,dz + \int\int p(x)\overline{p}(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))^2\ dx\,dz \\
&- \int\int \overline{p}(x)\overline{p}(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))^2\ dx\,dz - \int\int p(x)\,p(z)\,(\mu-\mu_0)^2\ dx\,dz + 2S
\end{aligned}
$$

where

$$S = 2\int\int (p(x,z)-\overline{p}(x,z))(\mu(x,z)-\overline{\mu}(x,z))\overline{p}(x)\overline{p}(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))dx\,dz.$$

Now consider the term $m = \int\int \overline{p}(x)\overline{p}(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))(\mu(x,z)-\overline{\mu}(x,z))\ dx\,dz$. We have

$$
\begin{aligned}
m &= \int\int \overline{p}(x)\overline{p}(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))(\mu(x,z)-\overline{\mu}(x,z))\ dx\,dz \\
&= \int\int \overline{p}(x)\overline{p}(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))(\mu(x,z)-\mu_0(z)+\mu_0(z)-\overline{\mu}_0(z)+\overline{\mu}_0(z)-\overline{\mu}(x,z))\ dx\,dz \\
&= \int\int \overline{p}(x)\overline{p}(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))(\mu(x,z)-\mu_0(z))\ dx\,dz \\
&+ \int\int \overline{p}(x)\overline{p}(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))(\mu_0(z)-\overline{\mu}_0(z))\ dx\,dz \\
&+ \int\int \overline{p}(x)\overline{p}(z)(\overline{\mu}(x,z)-\overline{\mu}_0(z))(\overline{\mu}_0(z)-\overline{\mu}(x,z))\ dx\,dz \\
&= \int\int \overline{p}(x)\overline{p}(z)\sqrt{\delta}\sqrt{\overline{\delta}}\ dx\,dz + 0 - \int\int \overline{p}(x)\overline{p}(z)\overline{\delta}\ dx\,dz,
\end{aligned}
$$

where $\delta = \mu(x,z)-\mu_0(z)$ and $\overline{\delta} = \overline{\mu}(x,z)-\overline{\mu}_0(z)$. Hence,

$$
\begin{aligned}
-R &= \int\int \overline{p}(x)p(z)\overline{\delta}\ dx\,dz + \int\int p(x)\overline{p}(z)\overline{\delta}\ dx\,dz + 2\int\int \overline{p}(x)\overline{p}(z)\sqrt{\delta}\sqrt{\overline{\delta}}\ dx\,dz \\
&- 2\int\int \overline{p}(x)\overline{p}(z)\overline{\delta}\ dx\,dz - \int\int \overline{p}(x)\overline{p}(z)\overline{\delta}\ dx\,dz \\
&- \int\int p(x)p(z)\delta\ dx\,dz - \int\int \overline{p}(x)\overline{p}(z)\delta\ dx\,dz + \int\int \overline{p}(x)\overline{p}(z)\delta\ dx\,dz \\
&= \int\int \overline{p}(x)p(z)\overline{\delta}\ dx\,dz + \int\int p(x)\overline{p}(z)\overline{\delta}\ dx\,dz - \int\int \overline{p}(x)\overline{p}(z)(\sqrt{\overline{\delta}}-\sqrt{\delta})^2\ dx\,dz \\
&- 2\int\int \overline{p}(x)\overline{p}(z)\overline{\delta}\ dx\,dz - \int\int p(x)p(z)\delta\ dx\,dz + \int\int \overline{p}(x)\overline{p}(z)\delta\ dx\,dz \\
&= \int\int (p(x)-\overline{p}(x))\overline{p}(z)(\overline{\delta}-\delta)\ dx\,dz + \int\int \overline{p}(x)(p(z)-\overline{p}(z))(\overline{\delta}-\delta)\ dx\,dz \\
&+ \int\int (\overline{p}(x)-p(x))(\overline{p}(z)-p(z))\delta\ dx\,dz - \int\int \overline{p}(x)\overline{p}(z)(\sqrt{\overline{\delta}}-\sqrt{\delta})^2\ dx\,dz.
\end{aligned}
$$

And hence

$$||R|| = O(||p(x)-\overline{p}(x)||\ ||\overline{\delta}-\delta||) + O(||p(z)-\overline{p}(z)||\ ||\overline{\delta}-\delta||)$$

$$
\begin{aligned}
&+ \quad O(||p(x)-\overline{p}(x)||\ ||p(z)-\overline{p}(z)||) + O(||\overline{\delta}-\delta||^2) \\
&= \quad O(||\overline{p}(x,z)-p(x,z)||^2) + O(||\overline{\mu}(x,z)-\mu(x,z)||^2) \\
&+ \quad O(||\overline{p}(x,z)-p(x,z)|| \times ||\overline{\mu}(x,z)-\mu(x,z)||). \ \square
\end{aligned}
$$

Lemma 3. Suppose that $||\widehat{\mu}(x,v)-\mu(x,v)|| = o_P(n^{-1/4})$. Then, when $\psi_1 \neq 0$, we have that $\sqrt{n}(\widehat{\psi}_1 - \psi_1) \rightsquigarrow N(0,\tau^2)$ for some τ^2.

Proof We have

$$
\begin{aligned}
Y_i - \widehat{\mu}(\widehat{V}_i) &= (Y_i - \mu(V_i)) + (\mu(V_i) - \mu(\widehat{V}_i)) + (\mu(\widehat{V}_i) - \widehat{\mu}(\widehat{V}_i)) \\
&= (Y_i - \mu(V_i)) - (\widehat{V}_i - V_i)^T \nabla\mu(\widetilde{V}_i) + (\mu(\widehat{V}_i) - \widehat{\mu}(\widehat{V}_i))
\end{aligned}
$$

for some $\widetilde{V}_i$ between V_i and $\widehat{V}_i$. Squaring, summing and letting $\epsilon_i = Y_i - \mu(V_i)$,

$$
\begin{aligned}
\frac{1}{n}\sum_i (Y_i - \widehat{\mu}(\widehat{V}_i))^2 &= \frac{1}{n}\sum_i \epsilon_i^2 + \frac{1}{n}\sum_i ((\widehat{V}_i - V_i)^T \nabla\mu(\widetilde{V}_i))^2 + \frac{1}{n}\sum_i (\mu(\widehat{V}_i) - \widehat{\mu}(\widehat{V}_i)) \\
&+ \frac{2}{n}\sum_i \epsilon_i (\widehat{V}_i - V_i)^T \nabla\mu(\widetilde{V}_i)^2 + \frac{2}{n}\sum_i \epsilon_i (\mu(\widehat{V}_i) - \widehat{\mu}(\widehat{V}_i)) \\
&+ \frac{2}{n}(\widehat{V}_i - V_i)^T \nabla\mu(\widetilde{V}_i)(\mu(\widehat{V}_i) - \widehat{\mu}(\widehat{V}_i)) \\
&= \frac{1}{n}\sum_i \epsilon_i^2 + \frac{2}{n}\sum_i \epsilon_i (\widehat{V}_i - V_i)^T \nabla\mu(\widetilde{V}_i) + \frac{2}{n}\sum_i \epsilon_i (\mu(\widehat{V}_i) - \widehat{\mu}(\widehat{V}_i)) + R_n
\end{aligned}
$$

where $R_n = O(||\widehat{\delta}-\delta||^2) + O(||\widehat{\mu}-\mu||^2) + O(||\widehat{\delta}-\delta||\ ||\widehat{\mu}-\mu||^2) = o_P(n^{-1/2})$. The mean of the first three terms is $\mathbb{E}[(Y-\mu(V))^2]$. By a similar argument,

$$
\begin{aligned}
\frac{1}{n}\sum_i (Y_i - \widehat{\mu}(X_i,\widehat{V}_i))^2 &= \frac{1}{n}\sum_i \widetilde{\epsilon}_i^2 + \frac{2}{n}\sum_i \widetilde{\epsilon}_i (\widehat{V}_i - V_i)^T \nabla\mu(X_i,\widetilde{V}_i) \\
&+ \frac{2}{n}\sum_i \widetilde{\epsilon}_i (\mu(X_i,\widehat{V}_i) - \widehat{\mu}(X_i,\widehat{V}_i)) + \widetilde{R}_n
\end{aligned}
$$

where $\widetilde{\epsilon}_i = Y_i - \mu(X_i,V_i)$, $\widetilde{R}_n = O(||\widehat{\delta}-\delta||^2) + O(||\widehat{\mu}-\mu||^2) + O(||\widehat{\delta}-\delta||\ ||\widehat{\mu}-\mu||^2) = o_P(n^{-1/2})$ and the mean of the first three terms is $\mathbb{E}[(Y-\mu(X,V))^2]$. The result follows from the CLT and the fact that $\sqrt{n}(R_n + \widetilde{R}_n) = o_P(1)$. ■

Lemma 4. We have that ψ_0 under the partially linear model with interactions, is equal to $\psi_3 = \theta^T \Omega \theta$ where

$$
\Omega = \Sigma_X \otimes \begin{pmatrix} 1 & m_Z^T \\ m_Z & \Sigma_Z \end{pmatrix}.
$$

Proof. Let us write

$$
\mu(x,z) = \theta^T W \equiv \theta_0^T X + \sum_{j=1}^h \theta_j^T X Z_j
$$

where we have written $\theta = (\theta_0, \theta_1, \ldots, \theta_h)$ and so $\mu_0(z) = \theta_0^T m_X + \sum_{j=1}^h \theta_j^T m_X Z_j$. Thus

$$\begin{aligned}(\mu(x,z) - \mu_0(z))^2 &= \theta_0^T (X - m_X)(X - m_X)^T \theta_0 + \sum_{j=1}^h \theta_j^T (X - m_X)(X - m_X)^T Z_j^2 \, \theta_j \\ &+ 2\sum_{j=1}^h \theta_0^T (X - m_X)(X - m_X)^T Z_j \, \theta_j + 2\sum_{j \neq k} \theta_j^T (X - m_X)(X - m_X)^T Z_j Z_k \theta_k\end{aligned}$$

and so

$$\begin{aligned}E_0[(\mu(x,z) - \mu_0(z))^2] &= \theta^T \Sigma_X \; \theta_0 + \sum_{j=1}^h \theta_j^T \Sigma_X (\Sigma_Z(j,j) + m_Z^2(j)) \, \theta_j \\ &+ 2\sum_{j=1}^h \theta_0^T \Sigma_X \; m_Z(j) \, \theta_j + 2\sum_{j \neq k} \theta_j^T \theta_k (\Sigma_Z(j,k) + m_Z(j) \; m_Z(k)) \\ &= \theta^T \Omega \; \theta. \; \square\end{aligned}$$

7.2 ψ_L Under the Semiparametric Model

Here we give the form that ψ_L takes under the semiparametric model. Under the model $\mu(x,z) = f(z) + x^T \beta(z)$, we have $\psi_L = \mathbb{E}[\beta^T(Z)(X - \nu(Z))(X - \nu(Z))^T \beta(Z)]$ which has efficient influence function

$$\begin{aligned}\phi &= 2\beta(Z)^T (X - \nu(Z))(X - \nu(Z))^T V^{-1}(Z) XY \\ &- 2\beta(Z)^T (X - \nu(Z))(X - \nu(Z))^T V^{-1}(Z)(X - \nu(Z))(X - \nu(Z))^T \beta \\ &- \beta^T (X - \nu(Z))(X - \nu(Z))^T \beta - \psi_L.\end{aligned}$$

When $\mu(x,z) = \beta^T x + \sum_{jk} \gamma_{jk} x_j z_k + f(z)$ then

$$\psi_L = \theta^T (\Omega_{11} + \Omega_{12} + \Omega_{21} + \Omega_{22})$$

where

$$\Omega_{11} = \begin{pmatrix} 1 & m_Z^T \\ m_Z & \Sigma_Z + m_Z m_Z^T \end{pmatrix} \otimes \Sigma_X,$$

$$\Omega_{12} = \begin{pmatrix} 1 & m_Z^T \\ m_Z & \Sigma_Z + m_Z m_Z^T \end{pmatrix} \otimes \mathbb{E}[(X - m_X)(m_X - \nu(Z))^T],$$

$$\Omega_{21} = \begin{pmatrix} 1 & m_Z^T \\ m_Z & \Sigma_Z + m_Z m_Z^T \end{pmatrix} \otimes \mathbb{E}[(X - m_X)(m_X - \nu(Z))^T],$$

$$\Omega_{12} = \begin{pmatrix} 1 & m_Z^T \\ m_Z & \Sigma_Z + m_Z m_Z^T \end{pmatrix} \otimes \mathbb{E}[(m_X - \nu(Z))(X - m_X)^T],$$

$$\Omega_{22} = \begin{pmatrix} 1 & m_Z^T \\ m_Z & \Sigma_Z + m_Z m_Z^T \end{pmatrix} \otimes \mathbb{E}[(m_X - \nu(Z))(m_X - \nu(Z))^T].$$

We omit the expression for influence function.

7.3 Partial Correlation

In this section, we give the decorrelated version of the partial correlation. Recall that

$$\rho_0 = \frac{\mathbb{E}_0[(Y-\mu_0(Z))(X-\nu_0(Z))]}{\sqrt{\mathbb{E}_0(Y-\mu_0(Z))^2\mathbb{E}_0(X-\nu_0(Z))^2}} = \frac{\int\int(\mu(x,z)-\mu_0(z))(x-m_X)p(x)p(z)dxdz}{\sigma_X\sqrt{\int\int\int(y-\mu*(z))^2p(y|x,z)p(x)p(z)}}.$$

Theorem 5 *The efficient influence function for ρ_0 is*

$$\phi = \frac{1}{\sqrt{\phi_2\phi_3}}\left\{\phi_1 - \frac{\psi_1}{2\psi_2}\phi_2 - \frac{\psi_2}{2\psi_3}\phi_3\right\}$$

where, in this section, we define

$$\begin{aligned}
\psi_1 &= \int\int(\mu(x,z)-\mu_0(z))(x-m_X)p(x)p(z)dxdz \\
\psi_2 &= \sigma_X^2 \\
\psi_3 &= \int\int\int(y-\mu_0(z))^2p(y|x,z)p(x)p(z)dxdzdy
\end{aligned}$$

and

$$\begin{aligned}
\phi_1 &= \mu_0(X)(X-m) + (X-m)\frac{Y-\mu(X,Z)}{p(X,Z)}p(X)p(Z) + (X-m)p(X)\mu(X,Z) - \mu_0(z) - 2\psi_1 \\
\phi_2 &= (X-m)^2 - \sigma_X^2 \\
\phi_3 &= (Y-v(Z))^2 - \psi_3 - 2\frac{p(X)p(Z)}{p(X,Z)}(Y-\mu(X,Z)).
\end{aligned}$$

Proof Let us write $\rho_0 = f(\psi_1,\psi_2,\psi_3)$ where $f(a,b,c) = a/\sqrt{bc}$ and

$$\begin{aligned}
\psi_1 &= \mathbb{E}_0[(Y-\mu_0(Z))(X-\nu_0(Z))] \\
\psi_2 &= \sigma_X^2 \\
\psi_3 &= \int\int\int(y-\mu*(z))^2p(y|x,z)p(x)p(z).
\end{aligned}$$

So the influence function is

$$f_1(\psi_1,\psi_2,\psi_3)\phi_1 + f_2(\psi_1,\psi_2,\psi_3)\phi_2 + f_3(\psi_1,\psi_2,\psi_3)\phi_3$$

where $f_j = \partial f/\partial\psi_j$ and ϕ_j is the influence function for ψ_j. Hence,

$$\phi = \frac{1}{\sqrt{\phi_2\phi_3}}\left\{\phi_1 - \frac{\psi_1}{2\psi_2}\phi_2 - \frac{\psi_2}{2\psi_3}\phi_3\right\}.$$

Now

$$\psi_1 = \int(\mu_0(x)-\psi_0)(x-m_X)p(x)dx = \int\mu_0(x)(x-m_X)p(x)dx$$

where $\mu_0(x) = \int \mu(x,z)p(z)$. So

$$\begin{aligned}
\phi_1 &= \int \mu_0(x)'(x-m_X)p(x)\ dx - \int \mu_0(x)m_X'p(x)\ dx + \int \mu_0(x)(x-m_X)p(x)'\ dx \\
&= 7\int \mu_0(x)'(x-m_X)p(x)\ dx - \int \mu_0(x)(X-m_X)p(x)\ dx + \mu_0(X)(X-m_X) - \psi_1 \\
&= \mu_0(X)(X-m_X) + \int \mu_0(x)'(x-m_X)p(x)\ dx - 2\psi_1.
\end{aligned}$$

Now

$$\begin{aligned}
\mu_0(x)' &= \int \mu'(x,z)p(z)\ dz + \mu(x,Z) - \mu_0(z) \\
&= \int \frac{Y-\mu(x,z)}{p(x,z)} I(X=x, Z=z)p(z)\ dz + \mu(x,Z) - \mu_0(z) \\
&= I(x=X)\frac{Y-\mu(x,Z)}{p(x,Z)}p(Z) + \mu(x,Z) - \mu_0(z).
\end{aligned}$$

Thus,

$$\begin{aligned}
\int \mu'(x,z)p(z)\ dz &= \int (x-m)p(x)\left\{ I(x=X)\frac{Y-\mu(x,Z)}{p(x,Z)}p(Z) + \mu(x,Z) - \mu_0(z)\right\} \\
&= (X-m)\frac{Y-\mu(X,Z)}{p(X,Z)}p(X)p(Z) + (X-m)p(X)\mu(X,Z) - \mu_0(z)
\end{aligned}$$

So

$$\phi_1 = \mu_0(X)(X-m) + (X-m)\tfrac{Y-\mu(X,Z)}{p(X,Z)}p(X)p(Z) + (X-m)p(X)\mu(X,Z) - \mu_0(z) - 2\psi_1.$$

Also

$$\phi_2 = (X-m)^2 - \sigma^2.$$

Now we turn to $\psi_3 = \int\int\int (y-\mu*(z))^2 p(x,y,z)$. Then

$$\begin{aligned}
\phi_3 &= (Y-v(Z))^2 - \psi_3 - 2\int p(x,y,z)(y-v(z))v'(z)dz \\
&= (Y-v(Z))^2 - \psi_3 - 2\int p(x,z)(\mu-v(z))v'(z)dz
\end{aligned}$$

and

$$v'(z) = \mu(X,z) - v(z) + I(z=Z)\frac{p(X)(Y-\mu(X,z))}{p(X,z)}$$

so that

$$\begin{aligned}
\phi_3 &= (Y-v(Z))^2 - \psi_3 - 2\textstyle\int p(x,z)(\mu-v(z))v'(z)dz \\
&= (Y-v(Z))^2 - \psi_3 - 2\tfrac{p(X)p(Z)}{p(X,Z)}(Y-\mu(X,Z)).
\end{aligned}$$

The remainder can be shown to be second order in a similar way to ψ_0. We omit the details. ■

7.4 Varying Coefficient Model

Let $\mu(x,z) = x^T\beta(z) + f(z)$. In this case ψ_0 becomes $\psi_4 = \text{tr}(\Sigma_X H)$. Define

$$\begin{aligned} V(z) &= \text{Var}[X|Z=z]\ C(z) = \text{Cov}[X,Y|Z=z] \\ f(z) &= \mu(z) - \nu(z)^T\beta(z)\ \beta(z) = V^{-1}(Z)C(z) \\ M &= \mathbb{E}[\beta(Z)]\ S = \text{Var}[\beta(Z)]. \end{aligned}$$

Lemma 6 *The efficient influence function for ψ_4 is*

$$\phi = \text{tr}(\Sigma_X\phi_H) + (X - m_X)^T H(X-m) - \psi_4$$

where $H = \mathbb{E}[\beta(Z)\beta(Z)^T]$,

$$\begin{aligned} \phi_H &= \beta(Z)\beta(Z)^T - H + \beta(Z)[YX^T - \beta(Z)^T(X-\nu(Z))(X-\nu(Z))^T]V^{-1}(Z) \\ &+ V^{-1}(Z)[XY - (X-\nu(Z))(X-\nu(Z))^T\beta(Z)]\beta(Z)^T. \end{aligned}$$

Hence, the estimator is

$$\widehat{\psi}_4 = \frac{1}{n}\sum_i \text{tr}(\widehat{\Sigma}_X\widehat{\phi}_H(U_i)) + \frac{1}{n}\sum_i (X_i - \overline{X})^T H(U_i)(X_i - \overline{X}).$$

References

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *arXiv preprint arXiv:1903.10464*, 2019.

Clément Bénard, Gérard Biau, Sébastien Da Veiga, and Erwan Scornet. Shaff: Fast and consistent shapley effect estimates via random forests. *arXiv preprint arXiv:2105.11724*, 2021.

Ian Covert and Su-In Lee. Improving kernelshap: Practical shapley value estimation via linear regression. *arXiv preprint arXiv:2012.01536*, 2020.

Ian Covert, Scott Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *arXiv preprint arXiv:2004.00668*, 2020.

Ben Dai, Xiaotong Shen, and Wei Pan. Significance tests of feature relevance for a blackbox learner. *arXiv preprint arXiv:2103.04985*, 2021.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2): 644–654, 2021.

Daniel Fryer, Inga Strumke, and Hien Nguyen. Shapley value confidence intervals for variable selection in regression models. 2020.

Bryan S Graham and Cristine Campos de Xavier Pinto. Semiparametrically efficient estimation of the average linear regression function. *Journal of Econometrics*, 2021.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying statistical learning based on efficient influence functions. *arXiv preprint arXiv:2107.00681*, 2021.

Rustam Ibragimov and Ulrich K Müller. t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, 28(4):453–468, 2010.

Osnat Israeli. A shapley-based decomposition of the r-square of a linear regression. *The Journal of Economic Inequality*, 5(2):199–212, 2007.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Wei-Yin Loh and Peigen Zhou. Variable importance scores. *arXiv preprint arXiv:2102.07765*, 2021.

Scott Lundberg and Su-In Lee. An unexpected unity among methods for interpreting model predictions. *arXiv preprint arXiv:1611.07478*, 2016.

Masayoshi Mase, Art B Owen, and Benjamin Seiler. Explaining black box decisions by shapley cohort refinement. *arXiv preprint arXiv:1911.00467*, 2019.

Andreas Messalas, Yiannis Kanellopoulos, and Christos Makris. Model-agnostic interpretability with shapley values. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–7. IEEE, 2019.

Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Art B Owen and Clémentine Prieur. On shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

Alessandro Rinaldo, Larry Wasserman, and Max G'Sell. Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438–3469, 2019.

Alexander M Samarov. Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association*, 88(423):836–847, 1993.

Numair Sani, Jaron Lee, Razieh Nabi, and Ilya Shpitser. A semiparametric approach to interpretable machine learning. *arXiv preprint arXiv:2006.04732*, 2020.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

Brian Williamson and Jean Feng. Efficient nonparametric statistical inference on population feature importance using shapley values. In *International Conference on Machine Learning*, pages 10282–10291. PMLR, 2020.

Brian D Williamson, Peter B Gilbert, Noah R Simon, and Marco Carone. A unified approach for inference on algorithm-agnostic variable importance. *arXiv preprint arXiv:2004.03683*, 2020.

Brian D Williamson, Peter B Gilbert, Marco Carone, and Noah Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1): 9–22, 2021.

Lu Zhang and Lucas Janson. Floodgate: inference for model-free variable importance. *arXiv preprint arXiv:2007.01283*, 2020.