Statistical Inference Using Machine Learning Methods

Contents

1 Introduction

The goal of semiparametric inference is to learn about a low-dimensional parameter of interest in the presence of a possibly infinite-dimensional 'nuisance' parameter. To obtain valid inference we need to be able to estimate this nuisance parameter sufficiently well, and it is only in recent work that the power of machine learning methods has been used to achieve this goal.

In this essay, we review current state of the art procedures that allow for inference using machine learning methods. Including this one, our work is separated into five sections. In Section 2, we provide a broad overview of semiparametric statistics, extending the notions of information and efficiency from the familiar parametric setting, with the goal of developing the necessary context to understand the success of approaches taken in recent work. Section 3 then introduces a paper that sets out a general framework for using machine learning to yield valid inference in a variety of contexts. Using what we learned in Section 2, we motivate the proofs and interpret the results in a semiparametric context. In Section 4 we specialise and look at a particular method known as regression splines for estimating the nuisance parameter under a framework that is similar but subtly different to what we see in Section 3, and we suggest reasons for why better results are obtained in a particular case. In Section 5, we summarise the key points made throughout and suggest areas for future work.

2 Semiparametric Models

In applications, semiparametric models can naturally be viewed as an intermediate strategy between purely parametric and nonparametric models, thereby balancing precision in inference and robustness of the model.

Let $X_1, \ldots, X_n$ be an identically, independently distributed sample of realisations from the distribution P on $(\mathcal{X}, \mathcal{B})$, where $\mathcal{X}$ is a Euclidean sample space and $\mathcal{B}$ is its Borel sigma-field. We say that the *model* is the set $\mathcal{P}$ of all possible values of P, and the problem of statistical inference is estimation of the value of a functional $\nu : \mathcal{P} \to \mathbb{R}^k$. We typically view $\nu(P) = (\nu_1(P), \ldots, \nu_k(P))$ as a vector of features of P that are perhaps of specific interest.

In the classical parametric setting, we assume that P is determined by ν and possibly a 'nuisance parameter' η. For instance, a normal linear model can be parameterised by the mapping $\theta \to P_\theta$ where $\theta = (\mu, \sigma)$. Here $\nu(P) = \mu$ is the parameter of interest and σ is a nuisance parameter in the sense that it is necessary to describe P and affects variability in estimates of μ. On the other extreme, a nonparametric model is one with no restrictions: $\mathcal{P}$ is an infinite-dimensional model consisting of the collection of all probability measures on $(\mathcal{X}, \mathcal{B})$. Semiparametric models are somewhere inbetween, and can often be described in terms of a parametrisation $(\theta, \eta) \to P_{\theta,\eta}$, where θ is a Euclidean parameter but η ranges over an infinite-dimensional set (such as a set of functions, or a nonparametric class of distributions). A simple example is the partially linear model

$$\mathbb{E}(Y|X, Z) = \theta^T X + \eta(Z),$$

where (X, Y, Z) is an $\mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^q$-valued observable random variable, $\theta \in \mathbb{R}^p$, and η is an unknown real function.

How might we constuct an estimator $\hat{\theta}$ of θ in this setting? Things are complicated by the presence of η. Ordinary least squares (OLS) regression of Y on X will not provide a consistent estimator unless X and $\eta(Z)$ are orthogonal, that is $\mathbb{E}[X\eta(Z)] = 0$, and while such models exist, it is not generally a reasonable assumption to make. Furthermore, it is appealing for our estimators to be not just consistent but root-N consistent, by which we mean $\sqrt{N}(\hat{\theta} - \theta) = O_P(1)$, as uniformly in possible distributions P as is feasible. Efficiency is another consideration: the Cramér-Rao lower bound tells us that in a parametric model there is a notion of optimality for the limiting variance of a regular estimator, and indeed such a notion can be extended to semiparametric models.

One approach is to use a so-called two-step estimator: form an estimate of η, and then use OLS as before to construct $\hat{\theta}$. Such a procedure could be implemented iteratively until convergence. However, much depends on the quality of the estimate of the nuisance parameter. There is no general recipe for estimating η, and in modern high-dimensional settings η can be a highly complex function and difficult to get right; for these reasons we are not just interested in efficiency, but also in robustness with respect to first-step estimation.

The central questions in this section are, therefore:

(1) How well can we estimate Eucliean parameters in a semiparametric model?

(2) How well can we estimate infinite dimensional nuisance parameters, and how can we ensure robustness?

We begin by developing the notion of information in a semiparametric context.

2.1 Tangent spaces

To get a handle on a notion of information in a semiparametric model $\mathcal{P}$, we explore one-dimensional parametric submodels $\mathcal{P}_0 = \{P_t : 0 \leq t < \epsilon\} \subset \mathcal{P}$ indexed by a real parameter t, with $P_0 := P$ the true distribution. We can view each such submodel as a 'smooth' path of distributions in the model that explore slight deviations from P, the idea being that estimating a parameter $\nu(P)$ in this smaller space is easier than under the whole space. Indeed, for each smooth parametric submodel we can calculate the Fisher information for estimating $\nu(P)$, and by considering all submodels we can get an upper bound on the information (equivalently, a lower bound on the variance) associated with estimating $\nu(P)$ given the model $\mathcal{P}$.

Suppose μ is an arbitrary measure relative to which P has density p and P_t has density p_t. We explore paths $\{p_t(x) : 0 \leq t < \epsilon\}$ along the direction of a bounded $S(x)$,

$$p_t(x) = p(x)(1 + tS(x)).$$

Note that $\mathbb{E}_P[S(X)] = 0$ because $p_t(x)$ must integrate to 1, and furthermore,

$$S(x) = \frac{\partial}{\partial t} \log\, p_t(x)|_{t=0},$$

is the tangent of the path at the 'point' P. We call this the *score function* for the parametric submodel, as we do in the classical context. For such a submodel to be well-defined the path must be 'smooth' in the sense that the above derivative condition holds in quadratic mean; in particular $\mathbb{E}_P[S(X)^2] < \infty$. By considering a collection of one-dimensional submodels, we accumulate a collection of score functions $\dot{\mathcal{P}}_P$ which we call a *tangent set*. We can identify this with a subset of $L_2^0(P)$, the space of mean zero functions in $L_2(P)$: in the fully nonparametric case this subset is all of $L_2^0(P)$, whereas in the parametric case it is a finite-dimensional space spanned by the finite number of score functions. In this sense, semiparametric models restrict the size of the tangent space.

This discussion could be formalised to allow for models that cannot be dominated. If the tangent set is a linear space we call it a *tangent space*. It will be convenient from now on to assume we are working with a tangent space, and indeed in all known examples this is the case so we are not losing much generality (Newey 1990).

Let us now look at the parameter of interest $\nu(P)$ along these submodels. Each of these submodels is a pertubation in $\mathcal{P}$ away from the true distribution in a certain direction given by the score. It is then natural to look at the local effect on ν of this pertubation, for which we will need a notion of differentiability of parameters in a model. Informally, we could then look at the pertubation that is in some sense least favourable for the purpose of estimating ν, which would give an information upper bound. Suppose there exists a continuous linear map $\dot{\mu}_P : L_2(P) \to \mathbb{R}^k$ such that for every $S \in \dot{\mathcal{P}}_P$ and a submodel $\{P_t : 0 \leq t < \epsilon\}$ with score function S

$$\frac{d\nu(P_t)}{dt}|_{t=0} = \dot{\mu}_P(S).$$

We then say that ν is a *differentiable parameter* at P relative to a given tangent space $\dot{\mathcal{P}}_P$. This definition requires differentiability of the map $\chi(t) := \nu(P_t)$ at $t = 0$ in the usual way, but also that it can be expressed in a particular form.

The Riesz representation theorem for a Hilbert space H states for any continuous linear map L on H there exists a unique element h_0 of H such that for all $h \in H$ we can write $L(h) = \langle h, h_0 \rangle$. Since $L_2(P)$ is a Hilbert space and $\dot{\mu}_P$ is continuous by assumption we can apply this here, giving the relationship

$$\dot{\mu}_P(S) = \langle \psi_P \, , S \rangle_P = \mathbb{E}_P[\psi_P(X)S(X)], \text{ for all } S \in \dot{\mathcal{P}}_P,$$

where $\psi_P : \mathcal{X} \to \mathbb{R}^k$ is a measurable function that we will call an *influence function*. Combining these statements, for every $S \in \dot{\mathcal{P}}_P$ and a submodel $\{P_t : 0 \leq t < \epsilon\}$ with score function S

$$\frac{dv(P_t)}{dt}|_{t=0} = \mathbb{E}_P[\psi_P(X)S(X)].$$

Note that this function is not necessarily unique: we are only specifying the inner products of those S in the tangent space, which is not all of $L_2(P)$. In particular, we can add to ψ_P any function orthogonal to the tangent space. Thus we can uniquely specify the function by orthogonal projection of a candidate ψ onto the tangent space. We will denote this projection $\tilde{\psi}$ and we refer to it as the *efficient influence function*.

From here onwards we will drop subscripts indicating dependence on the true distribution P, so that $\mathbb{E} = \mathbb{E}_P$.

Example: Parametric model
There is a relation between the classical score function and the notion of tangent spaces. Consider the parametric model indexed by the parameter θ_0 belonging to an open subset Θ of $\mathbb{R}^k$. Let $\dot{\ell}_{\theta_0} = \frac{\partial}{\partial \theta} \log \, \mathrm{d}P_\theta|_{\theta=\theta_0}$ be the classical score, and for each $h \in \mathbb{R}^k$ consider the one-dimensional submodels $P_t = P_{\theta_0 + th}$ for t small enough such that $\{P_t\} \subset \mathcal{P}$. Then by the chain rule,

$$\frac{\partial}{\partial t} \log \, \mathrm{d}P_{\theta_0+th}|_{t=0} = h^T \frac{\partial}{\partial \theta} \log \, \mathrm{d}P_\theta|_{\theta=\theta_0} = h^T \dot{\ell}_{\theta_0}.$$

Therefore, for each $h \in \mathbb{R}^k$ the submodel $t \to P_{\theta_0+th}$ has the score $S(x) = h^T \dot{\ell}_{\theta_0}$, so the tangent space at P_{θ_0} is given by $\dot{\mathcal{P}}_{P_{\theta_0}} = \{h^T \dot{\ell}_{\theta_0} : h \in \mathbb{R}^k\}$, the linear span of the score functions.

Furthermore, any map $\chi(\theta_0) := \nu(P_{\theta_0})$ that is differentiable in the normal way as a map from Θ to $\mathbb{R}^k$ is also a differentiable parameter, since

$$\begin{aligned}
\frac{\partial}{\partial t}\chi(\theta_0 + th)|_{t=0} &= h^T \frac{\partial \chi(\theta_0)}{\partial \theta} = \frac{\partial \chi(\theta_0)}{\partial \theta^T} h \\
&= \mathbb{E}[\frac{\partial \chi(\theta_0)}{\partial \theta^T} (\mathbb{E}[\dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}^T])^{-1} \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}^T h] \\
&= \mathbb{E}\left[\left(\frac{\partial \chi(\theta_0)}{\partial \theta^T} I(\theta_0)^{-1} \dot{\ell}_{\theta_0} \right) h^T \dot{\ell}_{\theta_0} \right] \\
&= \mathbb{E}[\tilde{\psi}(X)S(X)],
\end{aligned}$$

where the Fisher information matrix is given by $I(\theta_0) = \mathbb{E}[\dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}^T]$. This calculation shows that $\tilde{\psi} = \frac{\partial \chi(\theta_0)}{\partial \theta^T} I(\theta_0)^{-1} \dot{\ell}_{\theta_0}$ is the efficient influence function, since it is an

element of the tangent space satisfying the required property. Notice that (by the Delta method) $\text{Var}[\tilde{\psi}] = \frac{\partial \chi(\theta_0)}{\partial \theta^T} I^{-1}(\theta_0) \frac{\partial \chi(\theta_0)}{\partial \theta}$ is the classical Cramér-Rao lower bound for estimating $\chi(\theta_0)$: though we are yet to define a notion of efficiency in the semiparametric setting, the naming of the 'efficient influence function' somewhat gives away the trick that this property holds in a more general sense.

2.2 Information bounds

Recall that the Cramér-Rao lower bound for a parametric model provides an asymptotically sharp bound for the performance of an unbiased estimator. Suppose for simplicity that θ is scalar. If $\hat{\chi}$ is a regular unbiased estimator of a differentiable function $\chi(\theta)$ then

$$\text{Var}(\hat{\chi}) \geq \frac{\chi'(\theta)^2}{I(\theta)}.$$

It would be useful to have a precise formulation of what 'asymptotically sharp' means in a semiparametric context. We can motivate this by looking at the information associated with each parametric submodel.

Assume the parameter $\nu(P)$ is differentiable and scalar. The Fisher information about t in a smooth submodel $\{P_t\}$ with score S at $t = 0$ is $\mathbb{E}[S(X)^2]$. Therefore, the Cramér-Rao lower bound for estimating $\nu(P)$ in this submodel is

$$\frac{(d\nu(P_t)/dt|_{t=0})^2}{\mathbb{E}[S(X)^2]} = \frac{\mathbb{E}[\tilde{\psi}(X)S(X)]^2}{\mathbb{E}[S(X)^2]}.$$

The supremum of this expression over all submodels (equivalently, over the tangent space since the bound only depends on the score) is a lower bound for estimating $\psi(P)$ given model $\mathcal{P}$, when the true distribution is P. As explained, this is intuitively clear: in the full model, we certainly cannot do any better than in the worst submodel. We can use the Cauchy-Schwarz inequality to simply bound this expression,

$$\sup_{S \in \dot{\mathcal{P}}_P} \frac{\mathbb{E}[\tilde{\psi}(X)S(X)]^2}{\mathbb{E}[S(X)^2]} \leq \frac{\mathbb{E}[\tilde{\psi}(X)^2]\mathbb{E}[S(X)]^2}{\mathbb{E}[S(X)^2]} = \mathbb{E}[\tilde{\psi}(X)^2].$$

So we see that $\text{Var}[\tilde{\psi}(X)]$ is the optimal asymptotic variance in the semiparametric sense, achieved when $S = \tilde{\psi}$. In the multivariate case, we instead have the optimal covariance matrix $\mathbb{E}[\tilde{\psi}(X)\tilde{\psi}^T(X)]$. Nothing in this construction indicates that this is sharp, though we will see that for some estimators it can be achieved under certain conditions.

For a geometric perspective and to roughly formalise our intuition we refer to Mukhin (2018). Recall that the length of a curve is the sum of the lengths of the tangent vectors along the curve, so that in the model $\mathcal{P}$ we can speak of a distance between the true measure P and a pertubation P_ϵ along the path $t \to P_t$. Since the tangent of the path is exactly the score and we have the covariance inner-product,

$$\|P - P_\epsilon\|^2 = \int_0^\epsilon \mathbb{E}[S(X)^2]dt = \epsilon \mathbb{E}[S(X)^2].$$

With this interpretation, the Fisher information about t in a smooth submodel $\{P_t\}$ quantifies the deviation from the true distribution.

Therefore, one way of viewing our result is that the path in the model that travels away from the true distribution in the direction of the efficient influence function is a 'least favourable path' in the following sense: it makes the rate of change in value of the parameter with respect to the deviation from P maximum.

2.3 Semiparametric efficiency

It is important to clarify what we mean by 'optimal'. In fact, we are only considering those estimators that are regular in the following sense.

Regularity. An estimator sequence $\hat{\nu}_n$ is *regular* at P for estimating $\nu(P)$ if there exists a probability measure L such that

$$\sqrt{n}(\hat{\nu}_n - \nu(P_{1/\sqrt{n},S})) \overset{P_{1/\sqrt{n},S}}{\to} L, \text{ every } S \in \dot{\mathcal{P}}_P.$$

Here we have also indexed the parametric submodel with its score function for clarity. This is a type of uniform convergence condition: it requires that the limiting distribution does not depend on $\hat{\theta}_n$ for each n. In particular, the mean of the limiting distribution is zero. Our previous discussion on an optimal asymptotic variance holds for the class of regular estimators because 'superefficient' estimators with smaller asymptotic variance, the classic example being Hodges' estimator, are excluded.

With this class of estimators, we are ready to formulate semiparametric efficiency. Let $V = \mathbb{E}[\tilde{\psi}(X)\tilde{\psi}^T(X)]$ be the semiparametric efficiency bound we previously derived.

Theorem. If $\hat{\nu}_n$ is regular then the limiting distribution of $\sqrt{n}(\hat{\nu}_n - \nu(P))$ is equal to $L + U$, where $L \sim N(0, V)$ and U is independent of L.

Since the asymptotic covariance matrix of $\hat{\nu}_n$ is then $V + \mathbb{E}[UU^T]$, where $\mathbb{E}[UU^T]$ is a positive semi-definite matrix, the semiparametric notion of efficiency follows naturally.

Asymptotic Efficiency. We say that $\hat{\nu}_n$ is *asymptotically efficient* at P if it is regular at P with limit distribution $L = N(0, V)$.

It is possible to frame this result in a more constructive way.

Asymptotic Linearity. We say that an estimator $\hat{\nu}_n$ is asymptotically linear if there is a function ψ of the data such that

$$\sqrt{n}(\hat{\nu}_n - \nu(P)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i) + o_P(1),$$
$$\mathbb{E}[\psi(X)] = 0, \ \text{Var}[\psi(X)] \text{ finite and non-singular.}$$

Assuming $\nu(P)$ is a differentiable parameter, then as in Newey (1990) it can be shown that if $\hat{\nu}_n$ is asymptotically linear in ψ then it is regular if and only if ψ is the influence function we defined previously. It follows as a consequence that $\hat{\nu}_n$ is asymptotically efficient if and only it is asymptotically linear in the efficient influence function, since the central limit theorem implies $\hat{\nu}_n \to N(0, V)$. As an aside, this representation justifies the name 'influence function' as ψ can be interpreted as the first-order effect of an observation on $\hat{\nu}_n$.

Importantly, given an influence function we can compute an explicit form of the

remainder term for an estimator $\hat{\theta} = \hat{\theta}_n$ of θ_0,

$$\sqrt{n}(\hat{\theta} - \theta_0) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i) = R_n,$$

allowing us to then formulate conditions for R_n to be small. If we can establish conditons under which $R_n = o_p(1)$, then we know the limiting distribution of $\hat{\theta}$ by the central limit theorem. From there, we can construct confidence intervals or perform hypothesis tests in the usual way. The assumption we are making here is that there is a general recipe for constructing an influence function, and this is a topic for the end of this section.

2.4 Strictly semiparametric models

In the applications we will consider, the semiparametric model can be conveniently divided into a parameter of interest and a nuisance parameter. Often the parameter of interest is the parametric component $\nu(P_{\theta,\eta}) = \theta$ in the model $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in T\}$, where Θ is an open subset of $\mathbb{R}^k$ and T is arbitrary, for example a set of functions or a nonparametric class of distributions.

We consider the submodels $t \to P_{\theta+th,\eta_t}$ for paths $\{\eta_t\}$ in T and $h \in \mathbb{R}^k$. We expect the score function of such a submodel to be of the form

$$\frac{\partial}{\partial t} \log \mathrm{d}P_{\theta+th,\eta_t}|_{t=0} = h^T \dot{\ell}_{\theta,\eta} + g.$$

Here $\dot{\ell}_{\theta,\eta}$ is the familiar score for θ in the model where η is fixed; similarly, g is a score function for η if θ is fixed. Specifically, g will belong to the tangent set $\dot{\mathcal{P}}_\eta$ for η (note this depends on θ, but for ease of notation we will drop this dependence here). In order to calculate an information bound, $\dot{\mathcal{P}}_\eta$ should be infinite-dimensional so as to span the set of all scores for parametric submodels.

Recall that $\nu(P_{\theta+th,\eta_t}) = \theta + th$ is differentiable as a parameter if and only if there exists $\tilde{\psi}$ such that

$$\begin{aligned} h &= \frac{\partial}{\partial t} \nu(P_{\theta+th,\eta_t})|_{t=0} \\ &= \mathbb{E}[\tilde{\psi}(h^T \dot{\ell}_{\theta,\eta} + g)], \ h \in \mathbb{R}^k, g \in \dot{\mathcal{P}}_\eta. \end{aligned}$$

In particular, observe that $h = 0$ implies $\tilde{\psi}$ is orthogonal to the set $\dot{\mathcal{P}}_\eta$. When $\dot{\mathcal{P}}_\eta$ is linear, let $\Pi(\cdot|\dot{\mathcal{P}}_\eta)$ be the orthogonal projection onto the closed linear span of $\dot{\mathcal{P}}_\eta$ in $L_2(P_{\theta,\eta})$. Then the *efficient score function* for θ is $\tilde{\ell}_{\theta,\eta} = \dot{\ell}_{\theta,\eta} - \Pi(\dot{\ell}_{\theta,\eta}|\dot{\mathcal{P}}_\eta)$, which is the orthogonal projection onto the orthogonal complement of the nuisance tangent space. In addition, define the *efficient information matrix* for θ to be $\tilde{I}_{\theta,\eta} = \mathbb{E}[\tilde{\ell}_{\theta,\eta}\tilde{\ell}_{\theta,\eta}^T]$. Now, provided $\tilde{I}_{\theta,\eta}$ is invertible, the efficient influence function for estimation of θ is $\tilde{\psi} = \tilde{I}_{\theta,\eta}^{-1}\tilde{\ell}_{\theta,\eta}$, since

$$\begin{aligned} \mathbb{E}[\tilde{I}_{\theta,\eta}^{-1}\tilde{\ell}_{\theta,\eta}(h^T \dot{\ell}_{\theta,\eta} + g)] &= \tilde{I}_{\theta,\eta}^{-1}\mathbb{E}[\tilde{\ell}_{\theta,\eta}\dot{\ell}_{\theta,\eta}^T]h \\ &= \tilde{I}_{\theta,\eta}^{-1}\mathbb{E}[\tilde{\ell}_{\theta,\eta}(\tilde{\ell}_{\theta,\eta}^T + \Pi(\dot{\ell}_{\theta,\eta}^T|\dot{\mathcal{P}}_\eta))]h = h. \end{aligned}$$

where we used the definition of an orthogonal projection in the second line. In an attempt to provide some intuition, recall that tangents (scores) to the model at P are

deviations in a neighbourhood of P which arise from θ and η deviating from their true values. Projecting the original score for θ (for fixed η) $\dot{\ell}_{\theta,\eta}$ in the direction orthogonal to the directions that η deviates gives us the component of the deviation that is due to θ only, which is smaller than $\dot{\ell}_{\theta,\eta}$ by an amount $\Pi(\dot{\ell}_{\theta,\eta}|\dot{\mathcal{P}}_\eta)$, representing a part of the information for θ that is lost due to the presence of η.

2.5 Estimating equations

Adopting the same notation as previous sections, except now denoting data by $Z \sim P$, let $m(z, \theta, \eta)$ be a vector of functions with the same dimension as θ. The idea of a semiparametric m-estimator is that we impose the moment condition

$$\mathbb{E}[m(Z, \theta_0, \eta_0)] = 0$$

for the true values θ_0, η_0 (and importantly not for others values of θ). We will refer to m as a *moment function*. Next, we 'plug-in' an estimator $\hat{\eta}$ of η and seek an estimator $\hat{\theta}$ that solves

$$\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \theta, \hat{\eta}) = o_P(n^{-1/2}),$$

since for root-n consistency it is enough to solve this sample moment condition up to $o_P(n^{-1/2})$. We can view such an approach as a natural generalisation of the method of maximum likelihood for a parametric model, with m replacing the classical score function.

For now we assume η is known, and so set $\eta = \eta_0$. We will use a simple mean-value expansion argument to show $\hat{\theta}$ is asymptotically linear.

Expanding $m(Z, \hat{\theta}, \eta_0)$ in $\hat{\theta}$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(Z_i, \hat{\theta}, \eta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(Z_i, \theta_0, \eta_0) + \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} m(Z_i, \theta, \eta_0)|_{\theta=\tilde{\theta}}\right] \sqrt{n}(\hat{\theta} - \theta_0).$$

Since the left-hand side is $o_P(1)$ by assumption, then provided the sample average of the Jacobian converges uniformly in a neighbourhood of θ_0 to $J_0 := \frac{\partial}{\partial\theta}\mathbb{E}[m(z, \theta, \eta_0)]|_{\theta=\theta_0}$, and J_0 is invertible,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} -J_0^{-1} m(Z_i, \theta_0, \eta_0) + o_P(1).$$

Under our conditions, $\hat{\theta}$ is asymptotically linear with influence function $\psi(Z) = -J_0^{-1} m(Z, \theta_0, \eta_0)$. As a result, $\sqrt{n}(\hat{\theta} - \theta_0) \to N(0, V)$ where $V = J_0^{-1}\mathbb{E}[m(Z, \theta_0, \eta_0) m(Z, \theta_0, \eta_0)^T](J_0^{-1})^T$. If m is the efficient score then the estimator is semiparametric efficient, which motivates constructing an m to approximate the efficient score; however, computing the orthogonal projection onto the orthocomplement of the nuisance tangent set is tricky in practice.

The situation where η is not known is much more complicated. Conditions must be imposed on the bias of nuisance function estimators to ensure root-N consistency. In early semiparametric efficiency literature this is achieved by assuming the nuisance estimators belong to a so-called Donsker class and using maximal inequalities to bound

the bias tems. Such an assumption may not be appropriate in a modern setting where such estimators can be highly complex and may have unknown statistical properties, making the checking of entropic conditions difficult. We will postpone discussion along these lines until Section 3. For now, we handle the preliminary task of working out how estimation of an unknown η affects the asymptotic variance of estimators based on moment equatons.

2.6 Influence function adjustments

Let $Z_1, \ldots, Z_n$ denote i.i.d. realisations of $P \in \mathcal{P}$. Let the $k \times 1$ vector $\theta_0 := \nu(P)$ be the true value of a differentiable parameter, and let $\hat{\theta} = \hat{\theta}_n$ be an estimator which we asssume is regular and asymptotically linear with influence function ψ, that is

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Z_i) + o_P(1),$$

$$\mathbb{E}[\psi(Z)] = 0, \ \mathrm{Var}[\psi(Z)] \text{ finite.}$$

We are interested in the form of ψ when we have to account for a first-step estimator $\hat{\eta}$ of a nuisance parameter η. As before, introduce the moment condition $\mathbb{E}[m(Z, \theta_0, \eta_0)] = 0$ for the true values θ_0, η_0, so our two-step estimator $\hat{\theta}$ solves

$$\frac{1}{n} \sum_{i=1}^{n} m(Z_i, \theta, \hat{\eta}) = 0. \tag{2.1}$$

Proposition. The influence function of $\hat{\theta}$ is

$$\psi(Z) = -J_0^{-1}\{m(Z, \theta_0, \eta_0) + \alpha(Z)\},$$

where α is the influence function of $\int m(z, \theta_0, \hat{\eta}) dP(z) dz$.

Proof. Let $\{P_t\}$ be a smooth path in $\mathcal{P}$ with score S. Also, let η_t, θ_t be the limit of the estimators $\hat{\eta}, \hat{\theta}$ respectively when Z has distribution P_t. Then by the law of large numbers and (2.1), if $\mathbb{E}_t$ is the expectation under P_t then

$$\mathbb{E}_t[m(Z, \theta_t, \eta_t)] = 0. \tag{2.2}$$

Assume in what follows that all derivatives are to be evaluated at zero. Firstly, observe that

$$\frac{\partial}{\partial t} \mathbb{E}_t[m(Z, \theta_0, \eta_t)] = \frac{\partial}{\partial t} \mathbb{E}_t[m(Z, \theta_0, \eta_0)] + \frac{\partial}{\partial t} \mathbb{E}[m(Z, \theta_0, \eta_t)].$$

Working on the first term and supposing we can differentiate under the integral sign,

$$\begin{aligned}
\frac{\partial}{\partial t} \mathbb{E}_t[m(Z, \theta_0, \eta_0)] &= \frac{\partial}{\partial t} \int m(z, \theta_0, \eta_0) dP_t(z) dz \\
&= \int m(z, \theta_0, \eta_0) \frac{\partial dP_t(z)}{\partial t} \frac{1}{dP_0(z)} dP_0(z) dz \\
&= \int m(z, \theta_0, \eta_0) \frac{\partial}{\partial t} \log[dP_t(z)] dP_0(z) dz \\
&= \int m(z, \theta_0, \eta_0) S(z) dP_0(z) dz = \mathbb{E}[m(Z, \theta_0, \eta_0) S(Z)].
\end{aligned}$$

Next, provided $J_0 := \frac{\partial}{\partial \theta} \mathbb{E}[m(Z, \theta, \eta_0)]|_{\theta=\theta_0}$ is invertible it follows from (2.2) and the implicit function theorem that

$$\frac{d\nu(P_t)}{dt} = -J_0^{-1}\frac{\partial}{\partial t}\mathbb{E}_t[m(Z,\theta_0,\eta_t)],$$

Combining results, we have

$$\frac{d\nu(P_t)}{dt} = -J_0^{-1}\mathbb{E}[m(Z,\theta_0,\eta_0)] + \frac{\partial}{\partial t}\mathbb{E}[m(Z,\theta_0,\eta_t)]. \tag{2.3}$$

Recall that since $\hat{\theta}$ is regular and asymptotically linear in ψ,

$$\frac{d\nu(P_t)}{dt} = \mathbb{E}[\psi(Z)S(Z)],$$

so we want to put the right-hand side of (2.3) into this form. In other words, we want to put them into an asymptotically linear form with respect to some influence function. Therefore, if we suppose there exists a function $\alpha \in L_2^0(P)$ such that

$$\frac{\partial}{\partial t}\mathbb{E}[m(Z,\theta_0,\eta_t)] = \mathbb{E}[\alpha(Z)S(Z)], \tag{2.4}$$

we can plug this into (2.3) and move $-J_0^{-1}$ inside the expectation to obtain

$$\frac{d\nu(P_t)}{dt} = \mathbb{E}[-J_0^{-1}\{m(Z,\theta_0,\eta_0) + \alpha(Z)\}S(Z)],$$

from which it follows that

$$\psi(Z) = -J_0^{-1}\{m(Z,\theta_0,\eta_0) + \alpha(Z)\}.$$

□

The first term is the usual formula for the influence function of an m-estimator when we assume the nuisance function is equal to its true value, and so the second term can be interpreted as an adjustment to account for estimation of η_0. Additionally, an interesting conclusion of this result is that the asymptotic variance of semiparametric estimators depends only on the limit of the estimator $\hat{\eta}$, that is the nonparametric quantity $\hat{\eta}$ estimates, and not on the type of estimator.

A particular case where the adjustment term has a more explicit form is when η is a mean-square projection, for instance a conditional expectation, and this is what we turn to next. A conditional expectation is a common nuisance parameter in applications, and many methods exist for its estimation, one of which we will see in Section 3.

Proposition. Let X be an $r \times 1$ vector and let $H = \{\eta : \mathbb{E}[\eta(X)^2] < \infty\}$. In addition, suppose $Y \in L_2(P)$. If $\eta_0(x) = \mathbb{E}[Y|X]$ then the adjustment term is

$$\alpha(Z) = \delta(X)[Y - \eta_0(X)],$$

where δ belongs to H and is the solution of a certain functional equation (so long as a solution exists).

Proof. Again we consider a smooth path $\{P_t\}$ with score S. Starting from equation (2.4), assume there is a function $D(Z,\theta,\eta)$ linear in η such that

$$\frac{\partial}{\partial t}\mathbb{E}[m(Z,\theta_0,\eta_t)] = \frac{\partial}{\partial t}\mathbb{E}[D(Z,\theta_0,\eta_t)].$$

Then our aim is to show the right-hand side is equal to $\mathbb{E}[\delta(X)\{Y - \eta_0(x)\}S(Z)]$.

Proceeding as before, we get this term into the appropriate form by assuming there is $\delta \in H$ such that

$$\mathbb{E}[D(Z, \theta_0, \eta)] = \mathbb{E}[\delta(X)\eta(X)], \text{ for all } \eta \in H. \tag{2.5}$$

Recall that the Riesz representation theorem tells us such a δ exists if and only if $\mathbb{E}[D(Z, \theta_0, \eta)]$ is mean-square continuous as a function of η. It can be shown that mean-square continuity is necessary for an estimator to be root-N consistent, and so this assumption is justified.

Let $\eta_t(X) = \mathbb{E}_t[Y|X]$. Then we can compute (again all derivatives are evaluted at zero),

$$\begin{aligned}
\frac{\partial}{\partial t}\mathbb{E}[D(Z, \theta_0, \eta_t)] &= \frac{\partial}{\partial t}\mathbb{E}[\delta(X)\eta_t(X)] \\
&= \int \delta(x)\frac{\partial \eta_t(x)}{\partial t} dP(z)dz \\
&= \frac{\partial}{\partial t}\int \delta(x)\eta_t(x)dP_t(z)dz - \frac{\partial}{\partial t}\int \delta(x)\eta_0(x)dP_t(z)dz \\
&= \frac{\partial}{\partial t}\mathbb{E}_t[\delta(X)\{Y - \eta_0(X)\}], \text{using } \mathbb{E}_t[\delta(X)\eta_t(X)] = \mathbb{E}_t[\delta(X)Y] \\
&= \mathbb{E}[\delta(X)\{Y - \eta_0(X)\}S(Z)].
\end{aligned}$$

Therefore, by (2.5) the result follows since
$\frac{\partial}{\partial t}\mathbb{E}[m(Z, \theta_0, \eta_t)] = \mathbb{E}[\delta(X)\{Y - \eta_0(x)\}S(Z)]$, so $\alpha(Z) = \delta(X)\{Y - \eta_0(X)\}$.

□.

As a consequence, the limiting distribution of the estimator with an influence function adjustment will be the same as if we replaced the nuisance parameter with its true value, which then makes asymptotic variance estimation straightforward. This property is sometimes called 'orthogonality', which is related to robustness.

2.7 Robustness

Chernozhukov et al. (2016) show how in a more general setting we can calculate the adjustment term as the limit of a certain derivative, rather than attempting to solve a functional equation as we did previously. Moreover, importantly it is shown that the influence function adjustment eliminates terms in the expansion $\sqrt{n}(\hat{\theta} - \theta_0)$ that are first-order in the bias of nuisance function estimates, making the remainder term second-order which allows for weaker conditions on the quality of a first-step ML estimator.

Orthogonality
Suppose η belongs to a convex subset $\mathcal{T}$ of a Banach space. The functional $m = (m_1, \ldots, m_k)$ satisfying $\mathbb{E}[m(Z, \theta_0, \eta_0)] = 0$ is said to be *orthogonal* with respect to $\mathcal{T}$ if

$$\frac{\partial}{\partial t}\mathbb{E}[m(Z, \theta_0, \eta_0 + t(\eta - \eta_0)] = 0, \text{ for all } \eta \in \mathcal{T},$$

assuming this derivative (the *Gateaux derivative*) exists and is continuous.

Heuristically, we can interpret this condition as saying that small errors in first-step estimation of the nuisance functions will not invalidate the moment conditions for estimation of θ_0. This is related to our influence function adjustment calculations: if orthogonality holds then the adjustment term α is zero, so in a sense what we were doing was constructing orthogonal moment functions. As discussed, a key advantage of orthogonal moment functions is bias reduction. In some cases, bias reduction so large that the estimator based on the original moment function is not root-N consistent while the estimator based on the orthogonal moment function is.

Observe that the efficient score automatically satisfies orthogonality because it is orthogonal to the nuisance tangent set, and so its expectation should be insensitive to changes in η. The converse need not hold: any element of the orthocomplement is orthogonal, but it is the projection of the classical score onto the space that gives the efficient score.

3 A Generic Approach for Inference using Machine Learning

We now turn to the work of Chernozhukov et al. (2017) who propose a set of methods they call double or debiased machine learning (DML) that allow for root-N consistent inference on a Euclidean parameter θ_0 after first-step machine learning estimation of a high dimensional nuisance parameter η_0. Their approach is generic, and depending on the assumptions placed on η_0 (such as sparsity), allow for the use of a wide range of ML methods provided they estimate η_0 at the $o_P(n^{-1/4})$ rate.

Before we analyse the DML estimator in a general setting, we first follow the example given by the authors to illustrate the impact of the two core components in its construction: sample splitting and orthogonal moment functions.

3.1 Partially linear regression

Consider the partially linear regression model with data $Z = (D, Y, X)$ where D is a treatment variable (for instance, it may be binary to represent whether a treatment or policy has been applied), Y is the outcome of interest, $X \in \mathbb{R}^p$ are covariates and U, V are errors.

$$Y = D\theta_0 + g_0(X) + U,\ \mathbb{E}(U|X, D) = 0,$$

$$D = m_0(X) + V,\ \mathbb{E}(V|X) = 0.$$

This model and its extensions has seen a wide range of applicability in econometrics, for instance to analyse the expected impact of a certain policy or treatment on an outcome. Under this particular model, θ_0 is the parameter of interest, and $\eta_0 = (m_0, g_0)$.

The naive estimator
A naive estimation procedure would be to fit the outcome model directly using ML:

$$Y = D\hat{\theta} + \hat{g}(X) + \hat{U}.$$

This corresponds to the moment condition $\mathbb{E}[m(Z, \theta_0, \eta_0)] = 0$ where $m(Z, \theta, \eta) = D(Y - D\theta - g(X))$. By OLS we have the estimator

$$\begin{aligned}\hat{\theta} &= \left(\frac{1}{n}\sum_{i=1}^{n} D_i^2\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} D_i(Y_i - \hat{g}(X_i)) \\ &= (\mathbb{E}[D^2])^{-1}\frac{1}{n}\sum_{i=1}^{n} D_i(Y_i - \hat{g}(X_i)) + o_P(1)\end{aligned}$$

We will show that this estimator suffers from two sources of bias.

(1) Regularisation bias: in high-dimensional settings, machine learning estimators are used to prevent overfitting. However, this regularisation trades variance of our estimates $\hat{\eta}$ for bias. As we have suggested in Section 2, orthogonal moment conditions can help correct this.

(2) Overfitting: trying to prevent overfitting using regularised estimators is not guaranteed to work. Using the same data in our estimates of the nuisance parameter to estimate θ_0 can introduce significant bias terms. By basing different estimates on different subsamples of the data we can avert this. In addition, the use of sample splitting removes the need for those difficult to verify assumptions regarding the complexity of ML estimators we discussed in Section 2.

Suppose we split our sample into a main sample of observations indexed by $i \in I$ that is used to form $\hat{\theta}$, and an auxiliary sample indexed by $i \in I^c$ that is used to form $\hat{g}(X)$. For now, we assume that this deals with the problem of overfitting, and we will see how regularisation bias prevents $\hat{\theta}$ from being root-N consistent.

We are interested in whether it is possible that the naive estimator satisfies $\sqrt{n}(\hat{\theta} - \theta_0) = O_P(1)$. Substitute the definition of Y_i into our expression for $\hat{\theta}$ to obtain $\sqrt{n}(\hat{\theta} - \theta_0) = a + b + o_P(1)$, where

$$a := (\mathbb{E}[D^2])^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I} D_i U_i, \qquad b := (\mathbb{E}[D^2])^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I} D_i(g_0(X_i) - \hat{g}(X_i)).$$

Notice that a is the normalised sum of n terms of mean zero, thus converging to a centred normal by the central limit theorem. On the other hand, $g_0(X) - \hat{g}_0(X)$ is uncentered since $\hat{g}_0$ is biased. Therefore, in general b is not centered so our expression diverges in probability, unless our estimate converges at a $o_P(n^{-1/2})$ rate. In general, this rate is not feasible for an infinite-dimensional parameters, so $\sqrt{n}(\hat{\theta} - \theta_0) \neq O_P(1)$ and we do not have root-N consistency.

The DML estimator

The solution proposed by the authors is so-called 'double machine learning': we use ML to fit two models instead of one. Specifically, we estimate

$$D = \hat{m}(X) + \hat{V},$$

$$Y = D\hat{\theta} + \hat{g}(X) + \hat{U}.$$

We then regress $Y - \hat{g}(X)$ on $\hat{V}$ to obtain the estimator

$$\check{\theta}_0 = \left(\frac{1}{n} \sum_{i \in I} \hat{V}_i D_i \right)^{-1} \frac{1}{n} \sum_{i \in I} \hat{V}_i \left(Y_i - \hat{g}(X_i) \right).$$

This corresponds to the moment conditon $\mathbb{E}[m(Z, \theta_0, \eta_0)] = 0$ where $m(Z, \theta, \eta) = (D - m(X)(Y - D\theta - g(X))$. In this case one can see that the moment condition is simply imposing that the errors U, V are orthogonal, or equivalently, uncorrelated.

We can decompose the estimation error as $\sqrt{n}(\hat{\theta} - \theta_0) = a^* + b^* + c^* + o_P(1)$, where

$$\begin{aligned}
a^* &= (\mathbb{E}V^2)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} V_i U_i, \\
b^* &= (\mathbb{E}V^2)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} \left(m_0(X_i) - \hat{m}_0(X_i) \right) \left(g_0(X_i) - \hat{g}_0(X_i) \right), \\
c^* &= (\mathbb{E}V^2)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} \Big[U_i \left(m_0(X_i) - \hat{m}_0(X_i) \right) + V_i \left(g_0(X_i) - \hat{g}_0(X_i) \right) \Big].
\end{aligned}$$

Firstly, just as before a^* is the sum of mean zero terms and is well-behaved.

On the other hand, notice the difference between the bias term of our naive estimator and b^*, which now depends on the product of the errors from estimating g_0 and m_0: orthogonalising the moment function has created a second order remainder term. Therefore, it is enough that each be estimated at the $o_P(n^{-1/4})$ rate.

It remains to show that the terms in $c^* = o_P(1)$. This is where we see the impact of sample splitting: recall that we formed our estimates of g_0 and m_0 using the auxiliary sample $(X_j)_{j \in I^c}$, and so if we condition on this sample $\hat{g}, \hat{m}$ are non-stochastic. Then, since $\mathbb{E}(V_i | X_i) = \mathbb{E}(U_i | X_i) = 0$, each term in c^* has mean zero. Chebyshev's inequality then confirms the variance of c^* vanishes in probability.

A problem with sample-splitting is that our estimates lose efficiency because only a portion of the data is used. However, efficiency can be recovered using cross-fitting: in two-fold cross-fitting, we swap the role of the auxiliary sample and the main sample to obtain a new estimate, and then our final estimate is the average of the two. K-fold cross-fitting is the natural extension where the final estimate is the average of the estimates from each K folds.

With the intuition in place, we now formalise the discussion.

3.2 Properties of the DML estimator

Suppose our nuisance parameters belong to $\eta \in T$, where T is a convex subset of a Banach space. Additionally, we assume our estimators $\hat{\eta}$ belong to a subset $\mathcal{T}_{\mathcal{N}} \subset T$ with high-probability, where $\eta_0 \in \mathcal{T}_{\mathcal{N}}$. This will be a sequence of 'nusiance realisation sets' that shrinks around a neighbourhood of the true value η_0. As usual we suppose that the true value θ_0 of the parameter of interest satisfies the moment condition

$$\mathbb{E}[m(Z, \theta_0, \eta_0)] = 0,$$

where m is a vector of known (orthogonal) moment functions $(m_i)_{i=1}^k$ with $m_i : \mathcal{Z} \times \Theta \times T \to \mathbb{R}$.

As discussed, the estimator will use sample splitting: we randomly partition $\{1, \ldots, N\}$ into K indexing sets $(I_k)_{k=1}^K$ of equal size $n = N/K$. Let $\mathbb{E}_{n,k}$ be the empirical expectation over the k-th indexing set of the data, so that $\mathbb{E}_{n,k}[\phi(Z)] = n^{-1} \sum_{i \in I_k} \phi(Z_i)$. For our purposes, we will also suppose that the moment function is linear in θ such that $m(z, \theta, \eta) = m_a(z, \eta)\theta + m_b(z, \eta)$.

Procedure
(1) For each $k \in \{1, \ldots, K\} =: [K]$ construct an ML estimator of the nuisance parameter using the auxiliary sample I_k^c

$$\hat{\eta}_k := \hat{\eta}((Z_i)_{i \in I_k^c}).$$

(2) For each $k \in [K]$, construct the estimator $\tilde{\theta}_k$ of the parameter of interest as the solution, or an $\epsilon_N = o(\delta_N N^{-1/2})$ approximate solution where $\delta_N \to 0$, to the sample moment condition

$$\mathbb{E}_{n,k}[m(Z, \tilde{\theta}_k, \hat{\eta}_k)] = 0,$$

where m is the orthogonal moment function.
(3) We then aggregate the K estimators to obtain the cross-fit estimator

$$\hat{\theta} = \frac{1}{K} \sum_{k=1}^{K} \tilde{\theta}_k.$$

For robustness (accounting for model misspecification) it is desirable for our results to be valid over a wide class of possible underlying true distributions of the data Z. We let $\{\mathcal{P}_N\}_{N \geq 1}$ be a sequence of families of distributions P of the data. We will show our results hold uniformly in $P \in \mathcal{P}_N$, meaning for any sequence $(P_N)_{N \geq 1}$ such that $P_N \in \mathcal{P}_N$ for each N.

Theorem. Under regularity assumptions, the DML estimator achieves root-N consistency and is asymptotically normal:

$$\sqrt{N}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(z_i) + o_P(1) \to N(0, \sigma^2).$$

where $\psi(z) := J_0^{-1} m(z, \theta_0, \eta_0)$ is the influence function, and the asymptotic variance is

$$\sigma^2 = \mathbb{E}[\psi(Z, \theta_0, \eta_0)\psi(Z, \theta_0, \eta_0)^T]$$

Therefore, we can construct confidence intervals that are uniformly valid for a wide class of distributions. If we are interested in the jth component of θ_0 we have the asymptotic confidence interval of size $1 - \alpha$:

$$\mathcal{C} := \left[\hat{\theta}_j \pm \Phi^{-1}(1 - \alpha/2) \sqrt{\sigma_{jj}^2 / N} \right]$$

which satisfies

$$\sup_{P \in \mathcal{P}_N} |\mathbb{P}_P(\theta_{0,j} \in \mathcal{C}) - (1 - \alpha)| \to 0$$

We need an estimator for the variance in these confidence intervals, and we can use

$$\hat{\sigma}^2 = \frac{1}{K} \sum_{k=1}^{K} \hat{J}^{-1} \mathbb{E}[\psi(Z, \theta_0, \hat{\eta}_k)\psi(Z, \theta_0, \hat{\eta}_k)^T](\hat{J}^{-1})^T,$$

where

$$\hat{J} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{n,k}[m_a(Z, \hat{\eta}_k)].$$

We note in particular if m is the efficient score for estimation of θ_0 as discussed in Section 2, then $\hat{\theta}$ is asymptotically efficient at P and so achieves the semiparametric efficiency bound relative to the model $\mathcal{P}$.

For root-N consistent estimation of θ_0 to hold we must place regularity conditions on the moment function as well as ensure that we estimate the nuisance parameter sufficiently well in the first step. Formally, we can state these conditions as follows:

Regularity conditions for the DML estimator
Let $\{\delta_N\}_{N\geq 1}$ and $\{\Delta_N\}_{N\geq 1}$ be sequences of positive constants tending to zero, with $\delta_N \geq N^{-1/2}$, and let $q > 2$.
(i) Given an indexing set I, we require that the nuisance parameter estimator $\hat{\eta}$ constructed using I^c belongs to the set $\mathcal{T}_N$ with probability at least $1 - \Delta_N$, where $\eta_0 \in \mathcal{T}_N$.
(ii) The variance of m is non-degenerate: we insist all eigenvalues of σ^2 are greater than $c_0 > 0$.
(iii) Denote the Jacobian $J_0 = \dfrac{\partial}{\partial\theta}\mathbb{E}[m(Z, \theta, \eta_0)]|_{\theta=\theta_0} = \mathbb{E}[m_a(Z, \eta_0)]$. For identifiability, we stipulate that there exists $c_0 > 0, c_1 > 0$ such that each eigenvalue of J_0 is between c_0 and c_1 (in particular, they are bounded away from zero and so J_0 is invertible).
(iv) Assume the moment conditions hold:

$$m_N := \sup_{\eta\in\mathcal{T}_N} \left(\mathbb{E}\left[\left\|m(Z, \theta_0, \eta)\right\|^q\right]\right)^{1/q} \leq c_1,$$
$$m'_N := \sup_{\eta\in\mathcal{T}_N} \left(\mathbb{E}\left[\left\|m_a(Z, \eta)\right\|^q\right]\right)^{1/q} \leq c_1.$$

(v) We need to be able to estimate the nuisance parameter 'well enough',

$$r_N := \sup_{\eta\in\mathcal{T}_N} \left\|\mathbb{E}[m_a(Z, \eta)] - \mathbb{E}_P[m_a(Z, \eta_0)]\right\| \leq \delta_N,$$
$$r'_N := \sup_{\eta\in\mathcal{T}_N} \left(\mathbb{E}\left[\left\|m(Z, \theta_0, \eta) - m(Z, \theta_0, \eta_0)\right\|^2\right]\right)^{1/2} \leq \delta_N,$$

(vi) The moment function is (nearly) orthogonal: it is twice Gateaux continuously differentiable, and the first and second derivatives are uniformly bounded by $\lambda_N, \lambda'_N = o(\delta_N N^{-1/2})$ respectively.

We will see that these conditions arise naturally from the proofs.

In applications where the moment function is sufficiently smooth, suppose we have the upper bound ϵ_N on the rate of convergence of $\hat{\eta}$ to η_0:

$$\|\hat{\eta} - \eta_0\|_{L_2(P)} = O(\epsilon_N)$$

then we have the bounds

$$r_N = O(\epsilon_N), \qquad r'_N = O(\epsilon_N), \qquad \lambda'_N = O(\epsilon_N^2),$$

and since it follows from our assumptions that $\lambda'_N = o(N^{-1/2})$ we have the rate requirement

$$\epsilon_N = o(N^{-1/4}).$$

This is a common condition found in the semiparametric literature when classical nonparametric estimators are employed, and is achievable for many ML methods under certain assumptions on η_0.

Proof. As we discussed in Section 2, working with estimators based on influence functions is useful because given a candidate influence function ψ, we can explicitly calculate the remainder term

$$\sqrt{N}(\hat{\theta} - \theta_0) - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(z_i),$$

allowing us to then straightforwardly formulate conditions for it to be small.

In the case of a linear score we can obtain an explicit expression for our estimator $\hat{\theta}$ using the sample moment conditions $\mathbb{E}_{n,k}[m(Z, \tilde{\theta}_k, \hat{\eta}_k)] = 0$. For each k, these imply $\hat{J}_k \tilde{\theta_k} = \mathbb{E}_{n,k}[m_b(Z, \hat{\eta_k})]$. Therefore to identify $\tilde{\theta}_k$ we want a unique solution to this equation, which is equivalent to saying $\hat{J}_k$ is invertible (at least with probability approaching 1).

If we are in such a situation, we have

$$\tilde{\theta}_k = -\hat{J}_k \mathbb{E}_{n,k}[m_b(z, \hat{\eta_k})]$$

and so

$$\begin{aligned}
\sqrt{n}(\tilde{\theta}_k - \theta_0) &= -\sqrt{n}\hat{J}_k^{-1}(\underbrace{\mathbb{E}_{n,k}[m_b(Z, \hat{\eta}_k)] + \hat{J}_k\theta_0}_{=m(Z,\theta_0,\hat{\eta}_k)}) \\
&= -\hat{J}_k^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I_k} m(Z_i, \theta_0, \hat{\eta}_k).
\end{aligned}$$

We want to show that if we replace the estimates of the Jacobian and the nuisance parameter with their true values, the bias remainder term converges to zero. To that end, define the following quantities:

$$\begin{aligned}
&\hat{J}_k := \mathbb{E}_{n,k}[m_a(Z, \hat{\eta}_k)], \\
&R_{1,k} := \hat{J}_k - J_0, \\
&R_{2,k} := \mathbb{E}_{n,k}[m(Z, \theta_0, \hat{\eta}_k)] - \mathbb{E}_{n,k}[m(Z, \theta_0, \eta_0)].
\end{aligned}$$

To get a grip on the remainder term of the expression, we must first look at the remainders of $R_{1,k}$ and $R_{2,k}$.

Claim. The following bounds hold:

$$\begin{aligned}
&(1)\ \max_{k \in [K]} \|R_{1,k}\| = O_{P_N}(N^{-1/2} + r_N), \\
&(2)\ \max_{k \in [K]} \|R_{2,k}\| = O_{P_N}(N^{-1/2} r'_N + \lambda_N + \lambda'_N).
\end{aligned}$$

Proof.

(1) Firstly, fix k and decompose $\left\|R_{1,k}\right\|$ in the following manner:

$$\begin{aligned}\left\|R_{1,k}\right\| &= \left\|\mathbb{E}_{n,k}[m_a(Z,\hat{\eta}_k)] - \mathbb{E}_{P_N}[m_a(Z,\eta_0)]\right\| \\ &\leq \mathcal{I}_{1,k} + \mathcal{I}_{2,k}\end{aligned}$$

by the triangle inequality, where

$$\begin{aligned}\mathcal{I}_{1,k} &:= \left\|\mathbb{E}_{n,k}[m_a(Z,\hat{\eta}_k)] - \mathbb{E}_{P_N}[m_a(Z,\eta_k)|(Z_i)_{i\in I_k^c}]\right\|, \\ \mathcal{I}_{2,k} &:= \left\|\mathbb{E}_{P_N}[m_a(Z,\eta_k)|(Z_i)_{i\in I_k^c}] - \mathbb{E}_{P_N}[m_a(Z,\eta_0)]\right\|.\end{aligned}$$

Now notice that

$$\mathcal{I}_{1,k} = \frac{1}{n}\left\|\sum_{i\in I_k}\left\{m_a(Z_i,\hat{\eta}_k) - \mathbb{E}_{P_N}[m_a(Z_i,\hat{\eta}_k)|(Z_i)_{i\in I_k^c}]\right\}\right\|,$$

and so, since the estimator $\hat{\eta}_k$ is formed from the auxiliary sample I_k^c and hence is non-random conditional on it:

$$\begin{aligned}\mathbb{E}_{P_N}[\mathcal{I}_{1,k}^2|(Z_i)_{i\in I_k^c}] &\leq \frac{1}{n}\mathbb{E}_{P_N}\left[\left\|m_a(Z,\hat{\eta}_k) - \mathbb{E}_{P_N}[m_a(Z,\hat{\eta}_k)|(Z_i)_{i\in I_k^c}]\right\|^2\right] \\ &\leq \frac{4}{n}\mathbb{E}_{P_N}\left[\left\|m_a(Z,\hat{\eta}_k)\right\|^2|(Z_i)_{i\in I_k^c}\right] \\ &\leq \sup_{\eta\in\mathcal{T}_N}\frac{4}{n}\mathbb{E}_{P_N}\left[\left\|m_a(Z,\eta)\right\|^2|(Z_i)_{i\in I_k^c}\right] \\ &= \sup_{\eta\in\mathcal{T}_N}\frac{4}{n}\mathbb{E}_{P_N}\left[\left\|m_a(Z,\eta)\right\|^2\right] \\ &\leq 4c_1^2/n.\end{aligned}$$

Therefore, since conditional convergence implies unconditional convergence and $n = O(N)$, we have that $\mathcal{I}_{1,k} = O_{P_N}(N^{-1/2})$. Turning to $\mathcal{I}_{2,k}$, we work on the event $\mathcal{E}_N = \{\hat{\eta}_k \in \mathcal{T}_N \text{ for all } k \in [K]\}$, which can be shown to hold with P_N probability tending to 1. Therefore, we can simply bound $\mathcal{I}_{2,k} = O_{P_N}(r_N)$, since,

$$\mathcal{I}_{2,k} \leq \sup_{\eta\in\mathcal{T}_N}\left\|\mathbb{E}_{P_N}[m_a(Z,\eta)|(Z_i)_{i\in I_k^c}] - \mathbb{E}_{P_N}[m_a(Z,\eta_0)]\right\| = r_N$$

Since these bounds hold uniformly over k, we obtain (1).

(2) As before, fix $k \in [K]$, and introduce the notation $\mathbb{G}_{n,k}$ where

$$\begin{aligned}\mathbb{G}_{n,k}[\phi(Z)] &= \frac{1}{\sqrt{n}}\sum_{i\in I_k}\left(\phi(Z_i) - \int\phi(z)dP_N\right) \\ &= \sqrt{n}\left(\mathbb{E}_{n,k}[\phi(Z)] - \mathbb{E}_{P_N}[\phi(Z)]\right)\end{aligned}$$

We use this to decompose

$$\begin{aligned}&\mathbb{E}_{n,k}[m(Z,\theta_0,\hat{\eta}_k)] - \mathbb{E}_{n,k}[m(Z,\theta_0,\eta_0)] \\ &= \frac{1}{\sqrt{n}}\mathbb{G}_{n,k}[m(Z,\theta_0,\hat{\eta}_k) - m(Z,\theta_0,\eta_0)] + \mathbb{E}_{P_N}[m(Z,\theta_0,\hat{\eta}_k) - m(Z,\theta_0,\eta_0)|(Z_i)_{i\in I_k^c}]\end{aligned}$$

and proceeding in the same way as before we use the triangle inequality to obtain

$$\left\|R_{2,k}\right\| \leq \frac{1}{\sqrt{n}}\mathcal{I}_{3,k} + \mathcal{I}_{4,k}$$

where

$$\begin{aligned}
\mathcal{I}_{3,k} :=& \left\|\mathbb{G}_{n,k}[m(Z,\theta_0,\hat{\eta}_k) - m(Z,\theta_0,\eta_0)]\right\|, \\
\mathcal{I}_{4,k} :=& \left\|\mathbb{E}_{P_N}[m(Z,\theta_0,\hat{\eta}_k)|(Z_i)_{i\in I_k^c}] - \mathbb{E}_{P_N}[m(Z,\theta_0,\eta_0)]\right\|, \\
=& \left\|\mathbb{E}_{P_N}[m(Z,\theta_0,\hat{\eta}_k)|(Z_i)_{i\in I_k^c}]\right\|,
\end{aligned}$$

since $\mathbb{E}_{P_N}[m(Z,\theta_0,\eta_0)] = 0$ for $P_N \in \mathcal{P}_N$ by assumption. Note that naively using the convexity of the norm and the first moment condition yields

$$\mathcal{I}_{4,k} \leq c_1,$$

and thus the bound $\mathcal{I}_{4,k} = O_{P_N}(1)$. This is not sufficient for the remainder to be $o_P(1)$, and we will use orthogonality in order to attain a tighter bound.

Consider the function

$$f_k(r) := \mathbb{E}_{P_N}[m(Z,\theta_0,\eta_0 + r(\hat{\eta}_k - \eta_0))|(Z_i)_{i\in I_k^c}], \qquad r \in [0,1].$$

Since we have assumed the moment function is twice Gateaux continuously differentiable, we can Taylor expand to obtain

$$f_k(1) = \mathbb{E}_{P_N}[m(Z,\theta_0,\hat{\eta}_k)|(Z_i)_{i\in I_k^c}] = f_k'(0) + \frac{1}{2}f_k''(\tilde{r}), \qquad \tilde{r} \in (0,1),$$

where again we have used the fact that $f_k(0) = \mathbb{E}_{P_N}[m(Z,\theta_0,\eta_0)] = 0$.

Note that our assumptions allow us to control the norm of these two terms on the right-hand side. We see that $f_k'(0)$ is the Gateaux derivative of the score evaluated at zero, and so (near) orthogonality is precisely the condition that $\left\|f_k'(0)\right\|$ is small. In our case, we have imposed $\left\|f_k'(0)\right\| \leq \lambda_N$. On the other hand, we have assumed that the second Gateaux derivative of the moment function is bounded uniformly over $r \in (0,1)$, in particular $\left\|f_k''(\tilde{r})\right\| \leq \lambda_N'$.

Using the fact that $n = O(N)$, we conclude that

$$\begin{aligned}
\left\|\mathcal{I}_{4,k}\right\| = \left\|f_k'(0) + f_k''(\tilde{r})/2\right\| &\leq \left\|f_k'(0)\right\| + \left\|f_k''(\tilde{r})/2\right\| \\
&\leq \lambda_N + \frac{1}{2}\lambda_N'
\end{aligned}$$

and so

$$\mathcal{I}_{4,k} = O_{P_N}(\lambda_N + \lambda_N').$$

In precisely the same manner as we did for $\mathcal{I}_{2,k}$, we work on the event $\hat{\eta}_k \in \mathcal{T}_N$ to obtain $\mathcal{I}_{3,k} = O_{P_N}(r_N')$.

Combining these results yields

$$\max_{k\in[K]}\left\|R_{2,k}\right\| = O_{P_N}(N^{-1/2}r_N' + \lambda_N + \lambda_N'),$$

which is what we wanted to show.

□

Armed with these bounds, we can now proceed with the proof.

We return to the expression

$$\sqrt{n}(\tilde{\theta}_k - \theta_0) = -(R_{1,k} + J_0)^{-1} \left[\frac{1}{\sqrt{n}} \sum_{i \in I_k} m(Z_i, \theta_0, \eta_0) + \sqrt{n} R_{2,k} \right]. \qquad (*)$$

Starting with the first term, we look at the error when we replace it with the true value J_0^{-1}:

$$\begin{aligned}(J_0 + R_{1,k})^{-1} - J_0^{-1} &= (J_0 + R_{1,k})^{-1}(J_0 - (J_0 + R_{1,k}))J_0^{-1} \\ &= -(J_0 + R_{1,k})^{-1} R_{1,k} J_0^{-1}.\end{aligned}$$

Therefore, we can bound the error by

$$\begin{aligned}\left\| (J_0 + R_{1,k})^{-1} - J_0^{-1} \right\| &\leq \left\| (J_0 + R_{1,k})^{-1} \right\| \times \left\| R_{1,k} \right\| \times \left\| J_0^{-1} \right\| \\ &= O_{P_N}(1) O_{P_N}(N^{-1/2} + r_N) O_{P_N}(1) \\ &= O_{P_N}(N^{-1/2} + r_N).\end{aligned}$$

Now looking at the second term, we have

$$\left\| \frac{1}{\sqrt{n}} \sum_{i \in I_k} m(Z_i, \theta_0, \eta_0) + \sqrt{n} R_{2,k} \right\| \leq \left\| \frac{1}{\sqrt{n}} \sum_{i \in I_k} m(Z_i, \theta_0, \eta_0) \right\| + \left\| \sqrt{n} R_{2,k} \right\|$$

Note that $\left\| \sqrt{n} R_{2,k} \right\| = o_P(1)$ by construction of the constants r'_N, λ_N, and λ'_N. To get a bound on the sum, note that by Markov's inequality, for $a > 0$

$$\begin{aligned}\mathbb{P}_{P_N} \left(\left\| \frac{1}{\sqrt{n}} \sum_{i \in_k} m(Z_i, \theta_0, \eta_0) \right\|^2 > a \right) &\leq \frac{n^{-1}}{a} \mathbb{E}_{P_N} \left[\left\| \sum_{i \in I_k} m(Z_i, \theta_0, \eta_0) \right\|^2 \right] \\ &\leq \frac{1}{a} \mathbb{E}_{P_N} \left[\left\| m(Z, \theta_0, \eta_0) \right\|^2 \right] \leq \frac{c_1^2}{a},\end{aligned}$$

and so by definition we attain,

$$\left\| \frac{1}{\sqrt{n}} \sum_{i \in_k} m(Z_i, \theta_0, \eta_0) \right\| = O_{P_N}(1).$$

Therefore, the second term is $O_{P_N}(1)$.

We have shown that if we replace $\hat{J}_k$ with its true value J_0 in $(*)$, the resulting error is $O_{P_N}(N^{-1/2} + r_N)$. This is a good thing, since in particular it is $o_{P_N}(1)$.

Note that for the construction of confidence intervals we need σ^{-1} to exist almost surely. We have assumed that the eigenvalues of J_0 are less than c_1, while the eigenvalues of $\mathbb{E}_{P_N}[m(Z,\theta_0,\eta_0)m(Z,\theta_0,\eta_0)^T]$ are greater than c_0. Therefore the eigenvalues of

$$\sigma^2 = J_0^{-1}\mathbb{E}_{P_N}[m(Z,\theta_0,\eta_0)m(Z,\theta_0,\eta_0)^T](J_0^{-1})^T$$

are greater than $1/c_1 \times c_0 \times 1/c_1 = c_0/c_1^2$. Since σ^{-1} is a symmetric matrix, its norm is given by its largest eigenvalue, hence

$$\left\|\sigma^{-1}\right\| \leq c_1/\sqrt{c_0}$$

and so $\left\|\sigma^{-1}\right\| = O_{P_N}(1)$.

Combining these last two results with our rate result on the other remainder term $R_{2,k}$ establishes for all $k \in [K]$

$$\sqrt{n}\sigma^{-1}(\tilde{\theta}_k - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i\in I_k}\overline{\psi}(Z_i) + o_{P_N}(1),$$

where $\overline{\psi} := \sigma^{-1}\psi$. Hence,

$$\begin{aligned}
\sqrt{N}\sigma^{-1}(\hat{\theta} - \theta_0) &= \sqrt{N}\sigma^{-1}\left(\frac{1}{K}\sum_{k=1}^{K}\tilde{\theta}_k - \theta_0\right) \\
&= \frac{\sqrt{n}}{\sqrt{N}}\sum_{k=1}^{K}\sqrt{n}\sigma^{-1}(\tilde{\theta}_k - \theta_0) \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\overline{\psi}(Z_i) + o_{P_N}(1),
\end{aligned}$$

so the result follows by applying a central limit theorem.

□

4 Regression Splines: A Case Study

Moving away from this generic framework, we now turn to a particular machine learning method known as regression splines and see how they can be used to construct an estimator of the expected conditonal covariance that is semiparametrically efficient under minimal conditions. This work is highly related to that of Chernozhukov et al. (2017) that we have just considered: orthogonal moment conditions are constructed and cross-fitting is used in the same way to achieve bias reduction. However, the regression spline estimators have not simply been plugged into the DML framework: one extra step of sample-splitting has been used so that estimates of different nuisance functions may be based on distinct subsamples. When an estimator is nonlinear in its nonparametric estimators, this deals with a 'nonlinearity bias' that is of the same order as own-observation bias in the case of regression splines, and results in an estimator with fast remainder rates; indeed, when the estimand is an expected conditional covariance they are the fastest known possible. Before exploring this quantitatively we briefly introduce regression splines.

4.1 Regression spline theory

Regression splines are a type of series estimator that partition the support of a quantity of interest into segments, and in each segment the function is approximated by a continuous polynomial in such a way that globally the approximating function is quite smooth.

To be concrete, following Powell (1981) let $[a, b]$, $a, b \in \mathbb{R}$, be the support of a real function f, and say we have data $a \leq x_1 < x_2 < \cdots < x_N \leq b$, as well as 'knots' $a = \xi_1 < \xi_1 < \cdots < \xi_M = b$ that connect the segments of the support. We say s is a spline function of order κ if s is a polynomial of degree at most κ on each of the intervals $\{[\xi_{i-1}, \xi_i] : i = 1, \ldots, M\}$, and in addition s is $\kappa - 1$ times continuously differentiable. The spline function is determined by these conditions: it must equal f at the data-points and satisfy the continuity conditions. The positioning of the knots could be data-dependent, for instance it is advantageous to concentrate knots where f varies most rapidly.

A function spline estimator using K terms will have the form

$$s_K(x) = \sum_{k=1}^{K} \lambda_k p_k(x), \ a \leq x \leq b,$$

where $(p_k)_k$ are basis functions. For instance,

$$p_k(x) = \begin{cases} x^k, & 1 \leq k \leq \kappa, \\ (x - \xi_{k-\kappa})_+^{\kappa}, & \kappa + 1 \leq k \leq K. \end{cases} \tag{4.1}$$

is the truncated power basis. The basis functions are chosen so that their closed linear span can approximate well an element in the space of functions to which f is assumed to belong. In addition, for computational reasons it is often convenient to choose basis functions that are identically zero over a large part of the support; such functions exist and are called B-splines.

We can extend this to data with dimension $r > 1$. For example, following Newey (1994) we can construct a tensor product spline basis by defining, for $j = 1, \ldots, r$,

$$p_k^j(x_j) = \begin{cases} (x_j)^k, & 1 \leq k \leq \kappa, \\ (x_j - \xi_{j,k-\kappa})_+^{\kappa}, & \kappa + 1 \leq k \leq K. \end{cases} \tag{4.2}$$

Let λ be an $r \times 1$ vector of non-negative integers (a 'multi-index') and define $x^\lambda = x_1^{\lambda_1} \cdots x_r{}^{\lambda_r}$. Let $(\lambda(1), \ldots, \lambda(K))$ be a sequence of such multi-indices. We can then construct the multivariate spline basis by letting the univariate spline bases above interact,

$$p_k(x) = \prod_{j=1}^{r} p_{\lambda_j(k)}^j(x_j), \ k = 1, \ldots, K.$$

If we are interested in the functional $g_0(x) = \mathbb{E}[Y|X = x]$ given data $X_1, \ldots, X_N$, we can let $p^K(x) = (p_1(X), \ldots, p_K(X))^T$ be a $K \times 1$ vector of spline basis functions and

estimate g_0 by regressing observed values of $\mathbf{Y} := (Y_i)_{i=1}^N$ on $(p(X_i))_{i=1}^N$. Denote $P = (P^K(X_1), \ldots, P^K(X_N))$ the 'design matrix' for this regression, then

$$\hat{\gamma}(x) = p^K(x)^T (P^T P)^{-1} P^T \mathbf{Y}.$$

The quality of this approximation depends on three things: the 'smoothness' of the function g_0, the dimension K of the vector of splines, and the dimension r of each observation X. Firstly, following Newey et al. (2018) we classify smoothness in the following way.

Hölder class. We say that a function $x \to f(x)$ with domain a compact subset $D \subset \mathbb{R}^r$ is *Hölder* of order s if there exists a constant C such that f is $\overline{s} := \lfloor s \rfloor$ times continuously differentiable and all partial derivatives $\nabla^{\overline{s}}$ of order $\overline{s}$ are C-Lipschitz, meaning

$$\sup_{x,\tilde{x} \in D} |\nabla^{\overline{s}} f(\tilde{x}) - \nabla^{\overline{s}} f(x)| \leq C \|\tilde{x} - x\|^{s - \overline{s}}.$$

.

We now quote standard results for spline approximations under this notion of smoothness, see for example Powell (1982).

Convergence rates for regression splines
Let g_0 be Holder of order s_g and let $g_K(x) = p(x)^T \{\mathbb{E}[p(x)p(x)^T]\}^{-1} \mathbb{E}[p(x)g_0(x)]$ be the population approximation to g_0 using splines of order κ with K terms. Then we have the following rates of convergence,

$$\sup_{x \in [0,1]^r} |g_0(x) - g_K(x)| = O(K^{-\zeta_g}),\ \mathbb{E}[\{g_0(X) - g_K(X)\}^2] = O(K^{-2\zeta_g}).$$

where $\zeta_g = \min\{1 + \kappa, s_g\}/r$. In the examples we consider we can always choose the order of the splines $\kappa > s_g - 1$, so we assume this is the case for simplicity.

4.2 The DCDR estimator

We have the familiar motivation of seeking an asymptotically linear estimator $\hat{\theta}$ satisfying

$$\sqrt{N}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(Z_i) + o_P(1).$$

In particular the remainder term is $O_P(\Delta_N)$ where $\Delta_N \to 0$.

Just as before, we are considering semiparametric m-estimators with population moment condition $\mathbb{E}[m(Z, \theta_0, \eta_0)] = 0$, and hence a plug-in estimator $\hat{\theta}$ will satisfy the sample moment condition

$$\frac{1}{N} \sum_{i=1}^N m(Z_i, \theta, \hat{\eta})|_{\theta = \hat{\theta}} = 0.$$

In particular, the paper focuses on nuisance functions $\eta_0(x) = \mathbb{E}[Y|X = x]$ and functionals $m(Z, \theta, \eta) = m(Z, \eta) - \theta$ where $m(Z, \eta)$ is affine in η for every Z, meaning $m(Z, \eta) - m(Z, 0)$ is linear in η. In this situation, $\theta_0 = \mathbb{E}[m(Z, \eta_0)]$. As indicated, we

will specialise to the case $m(Z, \eta) = \text{Cov}(W, Y|X)$ so that our parameter of interest θ_0 is the expected conditional covariance.

Introducing the functions $\eta_0(X) = \mathbb{E}[Y|X]$ and $\gamma_0(X) = \mathbb{E}[W|X]$, we notice that

$$\begin{aligned}\mathbb{E}[\text{Cov}(W, Y|X)] &= \mathbb{E}[\mathbb{E}(WY|X) - \mathbb{E}(W|X)\mathbb{E}(Y|X)] \\ &= \mathbb{E}[W\{Y - \eta_0(X)\}],\end{aligned}$$

so that the the plug-in estimator is

$$\tilde{\theta} = \frac{1}{N}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} W_i[Y_i - \hat{\eta}_\ell(X_i)],$$

where as usual $I_\ell, \ell = 1, \ldots, L$ is a partition of the observation index set $[N]$ into L distinct subsets of roughly equal size and $\hat{\eta}_\ell$ is formed using the observations $(Z_i)_{i\in I_\ell^c}$.

We let $p(X) = (p_1(X), \ldots, p_K(X))^T$ be a $K \times 1$ vector of regression splines, and let P be the $N \times K$ matrix with ith row $P_i = p(X_i)$. Let $\hat{I}_\ell, \tilde{I}_\ell$ denote the indexing sets of observations used to form $\hat{\eta}_\ell, \tilde{\gamma}_\ell$ respectively, where $\hat{I}_\ell \cap \tilde{I}_\ell = \emptyset$, and let $\overline{I}_\ell$ be the indexing set for $\hat{\theta}$. Since the quantities we are estimating are conditional expectations, as before we can form,

$$\begin{aligned}\hat{\eta}_\ell(x) &= p(x)^T(P_{\hat{I}_\ell}^T P_{\hat{I}_\ell})^{-1} P_{\hat{I}_\ell}^T Y_{\hat{I}_\ell}, \\ \tilde{\gamma}_\ell(x) &= p(x)^T(P_{\tilde{I}_\ell}^T P_{\tilde{I}_\ell})^{-1} P_{\tilde{I}_\ell}^T W_{\tilde{I}_\ell},\end{aligned}$$

which are simply the fitted values from the regression of Y and W on $p(X)$ respectively over the relevant set of observations.

The proposed DCDR estimator is

$$\hat{\theta} := \frac{1}{N}\sum_{\ell=1}^{L}\sum_{i\in\overline{I}_\ell}[W_i - \tilde{\gamma}_\ell(X_i)][Y_i - \hat{\eta}_\ell(X_i)].$$

Looking at the form of the estimator more closely, we see that the summand can be split into two terms of independent interest,

$$[W_i - \tilde{\gamma}_\ell(X_i)][Y_i - \hat{\eta}_\ell(X_i)] = W_i[Y_i - \hat{\eta}_\ell(X_i)] - \tilde{\gamma}_\ell(X_i)][Y_i - \hat{\eta}_\ell(X_i)].$$

The first term is the plug-in estimator of $\mathbb{E}[\text{Cov}(W, Y|X)]$ which does not account for estimation of η_0, resulting in bias. The second term is the appropriate adjustment for this first-order bias which we can easily show using the work we have done in Section 2.

Recall that when $\eta \in H$ is a conditional expectation we have the correction term $\alpha(Z) = \delta(X)(Y - \eta_0(X))$. In addition, we can let $D(z, \eta) = m(z, \eta) - m(z, 0)$ by assumption of linearity of this quantity. The function $\delta(X)$ is characterised by the property,

$$\mathbb{E}[D(Z, \eta)\eta(X)] = \mathbb{E}[\delta(X)\eta(X)], \text{ for all } \eta \in H.$$

Since in our case $D(z, \eta) = -W\eta(X)$, by the tower property of conditional expectations this is the same as saying $\mathbb{E}[-\mathbb{E}[W|X]\eta(X)] = \mathbb{E}[\delta(X)\eta(X)]$ for all $\eta \in H$, from which it follows that $\delta(X) = -\mathbb{E}[W|X]$. Therefore, the bias adjustment term is

$\alpha(Z) = -\gamma_0(X)[Y - \eta_0(X)]$, which is precisely what the additional term above estimates. This justifies the form of the DCDR estimator; as we have discussed and will explicitly show, the first-order bias term is eliminated, giving it a faster remainder rate compared to the plug-in estimator.

Chernozhukov et al. (2016) show for orthogonal moment functions that are affine in each first-step (nonparametric) component the estimator will be 'doubly robust' in the sense that the moment conditions will still approximately hold even when we get one of our first-step estimates wrong. This is a stronger form of orthogonality. By assumption, we are precisely in this setting here, and so $\hat{\theta}$ is doubly robust.

Indeed, letting $\phi(Z, \theta, \gamma, \eta)$ be our orthogonalised moment function, ϕ satisfies $\mathbb{E}[\phi(Z, \theta_0, \gamma, \eta)] = -\mathbb{E}[\{\eta(X) - \eta_0(X)\}\{\gamma(X) - \gamma_0(X)\}]$ and so is zero for either $\eta = \eta_0$ or $\gamma = \gamma_0$. This means, for example, that we can afford to estimate η_0 somewhat poorly so long as we can ensure a good estimate of γ_0. The estimator is also doubly cross-fit: a different subsample is used to estimate each first-step component, and a final subsample is used to form $\hat{\theta}$. If this were not the case, we would simply be plugging-in regression splines into the DML framework; however, we will show this procedure achieves further bias reduction. Following Newey et al. (2018), we will refer to $\hat{\theta}$ as the doubly cross-fit, doubly robust (DCDR) estimator.

To prove some technical lemmas, we provide the assumptions of the DCDR estimator here, which are standard for regression splines.

Assumptions
(1)$\text{Var}[Y|X] \leq C, K \to \infty$, and $K\text{log}K/N \to 0$.

(2) i)$p(X) = Wq(U)$ where i) the support of U is $[0,1]^r$, W is continuously distributed with bounded density that is bounded away from zero;

ii) $q(W)$ are tensor product B-splines of order κ with knot spacing approximately proportional to the number of knots;

iii) $q(U)$ is normalised so that the eigenvalues of the matrix $\mathbb{E}[q(U)q(U)^T]$ are strictly greater than some positive constant $C > 0$, and $\sup_{u\in[0,1]^r} \|q(u)\| \leq C\sqrt{K}$;

iv) W is bounded and $\mathbb{E}[W^2|U]$ is bounded away from zero.

(3) $(\mathbb{E}[\{m(Z, \eta_0) - m(Z, \eta_K)\}^2])^{1/2} = O_P(K^{-s_\eta/r})$.

Linearisation of the DCDR estimator
We begin with a result that is not specific to regression splines and does not require that $\tilde{\gamma}$ and $\hat{\eta}$ are estimated using different subsamples, which will enable comparison with the DML estimator. Let $\psi(Z) = m(Z, \theta_0, \eta_0) + \eta_0(X)[Y - \gamma_0(X)]$. Then under our assumptions the DCDR estimator $\hat{\theta}$ satisfies

$$\sqrt{N}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Z_i) - \frac{1}{\sqrt{N}} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} [\tilde{\gamma}(X_i) - \gamma_0(X_i)][\hat{\eta}_\ell(X_i) - \eta_0(X_i)] + O_P(R_N),$$

where $R_N = \Delta_N^m + \Delta_N^\gamma + \Delta_N^\eta$ are remainder terms involving mean-square errors of the functions m, $\tilde{\gamma}$ and $\hat{\eta}$. Importantly the remainder is second order: it depends on the product of the estimation remainders for the nonparametric estimators $\tilde{\gamma}$ and $\hat{\eta}$.

Newey et al. (2018) show that $\max(\Delta_N^m, \Delta_N^\gamma, \Delta_N^\eta) = O_P(\Delta_N^*)$ in the case of the expected conditional covariance under certain assumptions, and so if $\Delta_N^* = o_P(1)$ the estimator is asymptotically normal with mean $\theta_0 - \mathbb{E}[\{\gamma_0(X) - \tilde{\gamma}(X)\}\{\eta_0(X) - \hat{\eta}(X)\}]$.

4.3 The DML estimator versus the DCDR estimator

Let $\mathbb{E}_{\overline{n},\ell}$ denote the empirical expectation over the $\overline{n}$ terms in $\overline{I}_\ell$.

Lemma. Suppose our assumptions are satisfied.

(1) If $\tilde{\gamma}$ and $\hat{\eta}$ are computed using the same subsample then,

$$\mathbb{E}_{\overline{n},\ell}[\{\gamma_0(X) - \tilde{\gamma}(X)\}\{\eta_0(X) - \hat{\eta}(X)\}] = O_P\left(K^{-\frac{s_\gamma + s_\eta}{r}} + \frac{K}{N}\right)$$

(2) If $\tilde{\gamma}$ and $\hat{\eta}$ are computed using distinct subsamples then,

$$\mathbb{E}_{\overline{n},\ell}[\{\gamma_0(X) - \tilde{\gamma}(X)\}\{\eta_0(X) - \hat{\eta}(X)\}] = O_P\left(K^{-\frac{s_\gamma + s_\eta}{r}} + \frac{\sqrt{K}}{N}\right)$$

This result unambiguously illustrates the benefit of three-way sample splitting in achieving further bias reductions. Looking at our linearisation expression, we see that a term of order $\sqrt{K/N}$ is added if we use three-way sample splitting, otherwise it is of order $K/\sqrt{N}$. For a regression spline first-step, this is the same order as own-observation bias.

As an illustration, suppose we choose $K = K_N$ to minimise each upper bound in the lemma. For (1) we choose $K = N^{\frac{r}{s_\gamma + s_\eta + r}}$ and hence the best bias bound is $O_P(N^{-\frac{s_\gamma + s_\eta}{s_\gamma + s_\eta + r}})$. This implies that $s_\gamma + s_\eta > r/2$ is needed for the bias to be $o_P(N^{-1/2})$. Robins et al. (2008) show that this is in fact a minimal requirement for efficiency: if $s_\gamma + s_\eta < r/2$ then no semiparametric efficient estimators exist.

On the other hand, for (2) the minimal bound is achieved by taking $K = N^{\frac{2r}{2(s_\gamma + s_\eta) + r}}$, giving a bias of order $N^{-\frac{2(s_\gamma + s_\eta)}{2(s_\gamma + s_\eta) + r}}$. Therefore, it is only necessary for $s_\gamma + s_\eta > r/4$ to obtain a $o_P(N^{-1/2})$ bias term.

We follow Lin Liu et al. (2019) to give another angle to help explain the better performance of the DCDR estimator. Suppose for simplicity that $s = s_\eta = s_\gamma$. We prove below that the mean-square error of our nonparametric estimators is $O_P(K/N + K^{-2s/r})$, consisting of variance and squared bias terms respectively.

Looking first at the DML estimator, observe that if we let $K \gg N^{\frac{r}{2s+r}}$ then the squared bias $K^{-2s/r}$ of $\tilde{\gamma}, \hat{\eta}$ shrinks faster than their variance K/N. This is known as an undersmoothed estimator. Conversely, these terms are of the same order when $K = N^{\frac{r}{2s+r}}$. However, in the case of the DML it is not optimal to use such an estimator since the result is an increase in bias of $\hat{\theta}$. On the contrary, the DCDR estimator uses undersmoothed $\tilde{\gamma}, \hat{\eta}$ to achieve the optimal rate given above, since in this case $K = N^{\frac{2r}{4s+r}} > N^{\frac{r}{2s+r}}$. To summarise, the DCDR estimator inherits its improved properties from double cross-fitting and undersmoothing the nonparametric estimators.

Proof. It is sufficient to consider one set of subsamples only. In addition, we will work on the event that $\hat{\Sigma} := \mathbb{E}_{\hat{n}}[p(X)p(X)^T], \tilde{\Sigma} := \mathbb{E}_{\tilde{n}}[p(X)p(X)^T]$ have smallest eigenvalue larger than $1/2$, so in particular they are invertible and $\hat{\Sigma}, \tilde{\Sigma} \geq \frac{1}{2}I$ in the positive semi-definite sense. It can be shown that the probability of this event tends to 1. Further, define the residuals $u_i = Y_i - \eta_0(X_i)$ and the approximation errors $v_i = \eta_0(X_i) - \eta_K(X_i)$, as well as quantities $\hat{\beta}_1 = \hat{\Sigma}^{-1}\mathbb{E}_{\overline{n}}[p(X)u]$, $\hat{\beta}_2 = \hat{\Sigma}^{-1}\mathbb{E}_{\overline{n}}[p(X)v]$. Denote $\delta = \Sigma^{-1}\mathbb{E}[p(X)\eta_0(X)]$.

Since $\mathbb{E}_{\overline{n},\ell}[\{\eta_0(X) - \hat{\eta}(X)\}^2] = \mathbb{E}[\{\eta_0(X) - \hat{\eta}(X)\}^2] + O_P\left(\frac{1}{\sqrt{N}}\right)$ (provided all subsamples are roughly the same size), if we can show the bounds

$$\mathbb{E}[\{\eta_0(X) - \hat{\eta}(X)\}^2]^{1/2} = O_P\left(\sqrt{\frac{K}{N}} + K^{-\frac{s_\eta}{r}}\right)$$
$$\mathbb{E}[\{\gamma_0(X) - \tilde{\gamma}(X)\}^2]^{1/2} = O_P\left(\sqrt{\frac{K}{N}} + K^{-\frac{s_\gamma}{r}}\right)$$

then we are done, since by the Cauchy-Schwarz inequality

$$\begin{aligned}\mathbb{E}_{\overline{n},\ell}[\{\gamma_0(X) - \tilde{\gamma}(X)\}\{\eta_0(X) - \hat{\eta}(X)\}] &\leq \mathbb{E}_{\overline{n},\ell}[\{\gamma_0(X) - \tilde{\gamma}(X)\}^2]^{1/2}\mathbb{E}_{\overline{n},\ell}[\{\eta_0(X) - \hat{\eta}(X)\}^2]^{1/2} \\ &= O_P\left(\sqrt{\frac{K}{N}} + K^{-\frac{s_\gamma}{r}}\right) O_P\left(\sqrt{\frac{K}{N}} + K^{-\frac{s_\eta}{r}}\right) \\ &= O_P\left(\frac{K}{N} + K^{-\frac{s_\gamma + s_\eta}{r}}\right).\end{aligned}$$

Let us try to bound $\mathbb{E}[\{\eta_0(X) - \hat{\eta}(X)\}^2]$. The proof for $\tilde{\gamma}$ will be identical.

Using a useful property of series estimators we can decompose this quantity in the following manner,

$$\begin{aligned}\mathbb{E}[\{\hat{\eta}(X) - \eta_0(X)\}^2] &= \mathbb{E}[\{\hat{\eta}(X) - \eta_K(X) + \eta_K(X) - \eta_0(X)\}^2] \\ &= \mathbb{E}[\Big(\{\eta_0(X) - \eta_K(X)\} + p(X)^T\{\hat{\delta} - \delta\}\Big)^2] \\ &= \mathbb{E}[\{\eta_0(X) - \eta_K(X)\}^2] + \mathbb{E}[\Big(p(X)^T\{\hat{\delta} - \delta\}\Big)^2]\end{aligned}$$

The cross-terms cancel since

$$\mathbb{E}[p(X)\{\eta_0(X) - \eta_K(X)\}] = \mathbb{E}\left[p(X)\{p(X)^T\Sigma^{-1}\mathbb{E}[p(X)\eta_0(X)] - \eta_0(X)\}\right] = 0.$$

Referring to our convergence results for splines at the start of this section, the first term in this decomposition is $O_P\left(K^{-\frac{2s_\eta}{r}}\right)$. We place a preliminary bound on the second term by observing

$$\begin{aligned}\mathbb{E}[\Big(p(X)^T\{\hat{\delta} - \delta\}\Big)^2] &= (\hat{\delta} - \delta)^T\Sigma(\hat{\delta} - \delta) \\ &= O\left(\left\|\hat{\delta} - \delta\right\|^2\right).\end{aligned}$$

Next, by adding and subtracting terms we have the useful result $\delta - \hat{\delta} = \hat{\beta}_1 + \hat{\beta}_2 + O_P\left(\frac{1}{\sqrt{N}}\right)$, so we proceed with looking at bounding these two quantities on the right-hand side.

Firstly,

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\beta}_1\right\|^2\right] &= \mathbb{E}\left[\mathbb{E}_{\overline{n}}[p(X)u]^T\hat{\Sigma}^{-2}\mathbb{E}_{\overline{n}}[p(X)u]\right] \\
&\leq 4\mathbb{E}\left[\mathbb{E}_{\overline{n}}[p(X)u]^T\mathbb{E}_{\overline{n}}[p(X)u]\right], \text{ using } \hat{\Sigma}^{-2} \leq 4I \\
&= \frac{4}{\hat{n}^2}\sum_{i,j\in\overline{I}}\mathbb{E}[p(X_i)^T p(X_j)u_i u_j] \\
&= \frac{4}{\hat{n}}\mathbb{E}\left[\left\|p(X)\right\|^2 u^2\right] = O\left(\frac{K}{N}\right), \text{ since } \mathbb{E}[u^2|X] = \mathrm{Var}[Y|X] \leq C \text{ by assumption.}
\end{aligned}$$

The corresponding bound on $\hat{\beta}_2$ goes through in the same way, except we need to bound $\mathbb{E}[v^2|X]$ instead. Again referring to our convergence results for regression splines,

$$\begin{aligned}
\mathbb{E}[v^2|X] &= \mathbb{E}[\{\eta_0(X) - \eta_K(X)\}^2] \\
&\leq \sup_{x\in[0,1]^r}|\eta_0(x) - \eta_K(x)|^2 = O\left(K^{-\frac{2s_\eta}{r}}\right)
\end{aligned}$$

Therefore $\mathbb{E}\left[\left\|\hat{\beta}_2\right\|^2\right] = O\left(K^{-\frac{2s_\eta}{r}}\frac{K}{N}\right)$, and so applying the triangle inequality yields $\left\|\hat{\delta} - \delta\right\|^2 \leq \left(\left\|\hat{\beta}_1\right\| + \left\|\hat{\beta}_2\right\|\right)^2 = O\left(\frac{K}{N}\right)$.

In total, $\mathbb{E}[\{\hat{\eta}(X) - \eta_0(X)\}^2] = O_P\left(K^{-\frac{2s_\eta}{r}}\right) + O\left(\frac{K}{N}\right)$, which is the stated result.

□

We can prove the second result using similar calculations, and it follows from Newey et al. (2018) which we omit here for brevity. The idea behind the improved factor is that instead of simply using a Cauchy-Schwarz bound we can split the product into factors that have all been estimated using different subsamples. By then conditioning on certain subsamples we can extract tighter bounds.

We conclude with a statement of the main result of the paper.

4.4 Fast remainder rate of the DCDR

Theroem. Let $\psi(Z) = m(Z, \theta_0, \eta_0) + \eta_0(X)[Y - \gamma_0(X)]$, then under our assumptions the DCDR estimator achieves the fastest known remainder rate,

$$\sqrt{N}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(Z_i) + O_P(\Delta_N^*),$$

where $\Delta_N^* = \sqrt{N}K^{-2(s_\eta+s_\gamma)/r} + K^{-s_\eta/r} + K^{-s_\gamma/r} + \sqrt{\frac{K}{N}}$. In particular, if $s_\eta + s_\gamma > r/2$ then $\Delta_N^* = o_P(1)$, and so the DCDR estimator is root-N consistent and asymptotically normal under minimal conditions. In fact, ψ is the efficient influence function for estimating θ_0, so $\hat{\theta}$ will be semiparametric efficient. Previously, the only known estimator to achieve semiparametric efficiency under minimal conditions used

higher order influence functions in its construction, which are far less simple than what is offered here.

In summary, in the sense of bias reduction the DCDR estimator can provide substantial improvements compared to the DML estimator, at least for the estimation of the expected conditional covariance. In other cases, the DCDR estimator is shown to have a faster remainder rate relative to the simple cross-fit plug-in estimator when the estimand is an average linear functional of a conditional expectation, but it is not generally known whether the fast remainder rate Δ_N^* can be achieved.

5 Conclusion

In this essay, we have seen how concepts from semiparametric statistics have been used to improve the feasbility of first-step ML estimators of nonparametric functions for inference on a low-dimensional parameter. For instance, the use of an influence function adjustment such as those calculated in Newey (1994) can create a remainder that is second-order in the bias of the nuisance function estimates, enabling valid inference dependent upon a convergence rate of $o(n^{-1/4})$, which is achievable by a range of ML methods under structured assumptions. It is highly related and in some cases equivalent to the concept of robustness of moment conditions that are characterised by a degree of resilience to errors in the first-step.

In high-dimensional settings, the regularisation property of many ML methods is highly desirable to prevent overfitting and allow for feasible learning, making them indispensible for modern statistics. By creating a generic framework for the use of ML through sample splitting and orthogonalisation, the DML estimator provides a means of taking a modern approach to problems in causal inference for the statistical practitioner.

We also showed how, using regression splines, we can achieve semiparametric efficiency under minimal conditions for the expected conditional covariance uisng a simple DCDR estimator. We compared the DCDR approach to simply plugging-in regression splines into the DML framework, and showed that in terms of bias reduction the DCDR estimator was superior for this estimand for reasons related to the three-way sample splitting to deal with non-linearity bias. It would be interesting to see whether we could achieve the same bias reductions by adapting the DML estimator to incorporate three-way sample splitting when we have products of nonparametric estimators. It would also be interesting to compare regression splines with other specific machine learning methods to be able to properly assess the benefits and drawbacks of regression splines as a nonparametric estimator.

Moreover, the DCDR estimator for the expected conditional covariance represents a significant improvement in terms of simplicity. Previously, the only known estimator of the expected conditional covariance to achieve semiparametric efficiency under minimal conditions was based on higher order influence functions (HOIFs), a generalisation of the tangent spaces and influence functions we have been considering in order to obtain consistent estimates of higher order bias terms and thus achieve significant bias reduction. It would be useful to know whether similar simplications could be obtained for other estimates based on HOIFs.

References

(1) van der Vaart, A. *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, 1998
(2) Bickel, P. et al., *Efficient and Adaptive Estimation for Semiparametric Models*, Springer, 1993
(3) Newey, W. *The Asymptotic Variance of Semiparametric Estimators*, Econometrica, Vol. 62, No. 6 (Nov., 1994), pp.1349-1382
(4) Newey, W. *Semiparametric Efficiency Bounds*, Journal of Applied Econometrics, Vol. 5, No. 2 (Apr. - Jun., 1990), pp. 99-135
(5) Chernozhukov, V. et al. (2018), *Locally Robust Semiparametric Estimation*, arXiv:1608.00033
(6) Chernozhukov, V. et al. (2017), *Double Machine Learning for Treatment and Causal Parameters* arXiv:1608.00060
(7) Newey, W. and Robins, J. (2018), *Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation* arXiv:1801.09138
(8) Mukhin, Y. (2018), *Sensitivity of Regular Estimators*, arXiv:1805.08883
(9) Liu, L. (2019), *On assumption-free tests and confidence intervals for causal effects estimated by machine learning*, arXiv:1904.04276
(10) Powell, M., *Approximation Theory and Methods*, Cambridge University Press, 1981