cemmap

centre for microdata methods and practice

Double/de-biased machine learning using regularized Riesz representers

Victor Chernozhukov
Whitney Newey
James Robins

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP15/18

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

Double/De-Biased Machine Learning Using Regularized Riesz Representers

Victor Chernozhukov VCHERN@MIT.EDU
Massachusetts Institute of Technology

Whitney Newey WNEWEY@MIT.EDU
Massachusetts Institute of Technology

James Robins ROBINS@HSPH.HARVARD.EDU
Harvard Universiy

Abstract

We provide adaptive inference methods for linear functionals of ℓ_1-regularized linear approximations to the conditional expectation function. Examples of such functionals include average derivatives, policy effects, average treatment effects, and many others. The construction relies on building Neyman-orthogonal equations that are approximately invariant to perturbations of the nuisance parameters, including the Riesz representer for the linear functionals. We use ℓ_1-regularized methods to learn the approximations to the regression function and the Riesz representer, and construct the estimator for the linear functionals as the solution to the orthogonal estimating equations. We establish that under weak assumptions the estimator concentrates in a $1/\sqrt{n}$ neighborhood of the target with deviations controlled by the normal laws, and the estimator attains the semi-parametric efficiency bound in many cases. In particular, either the approximation to the regression function or the approximation to the Rietz representer can be "dense" as long as one of them is sufficiently "sparse". Our main results are non-asymptotic and imply asymptotic uniform validity over large classes of models.

Keywords: Approximate Sparsity vs. Density, Double/De-biased Machine Learning, Regularized Riesz Representers, Linear Functionals

1. Introduction

Consider a random vector $(Y, X')'$ with distribution P and finite second moments, where the outcome Y takes values in $\mathbb{R}$ and the covariate X taking values $x \in \mathcal{X}$, a Borel subset of $\mathbb{R}^d$. Denote the conditional expectation function map $x \mapsto \mathrm{E}[Y \mid X = x]$ by γ_0^*. We consider a function γ_0, given by $x \mapsto b(x)'\beta_0$, as a sparse linear approximation to γ_0^*, where b is a p-dimensional vector, a dictionary of basis functions, mapping $\mathcal{X}$ to $\mathbb{R}^p$. The dimension p here can be large, potentially much larger than the sample size.

Our goal is to construct high-quality inference methods for a real-valued linear functional of γ_0 given by:

$$\theta_0 = \mathrm{E}m(X, \gamma_0) = \int m(x, \gamma_0) dF(x),$$

where F is the distribution of X under P, for example the average derivative and other functionals listed below. (See Section 2 below regarding formal requirements on m). When the approximation error $\gamma_0 - \gamma_0^*$ is small, our inference will automatically re-focus on a more ideal target – the linear functional of the conditional expectation function,

$$\theta_0^* = \mathrm{E}m(X, \gamma_0^*).$$

Example 1 (Average Derivative) *Consider an average derivative in the direction a:*

$$\theta_0 = \mathrm{E}a'\nabla_x \gamma_0(X), \quad a \in \mathbb{R}^d.$$

This functional corresponds to an approximation to the effect of policy that shifts the distribution of covariates via the map $x \mapsto x + a$, so that

$$\int (\gamma_0(x+a) - \gamma_0(x))dF(x) \approx \int a'\nabla_x \gamma_0(x)dF(x).$$

Example 2 (Policy Effect from a Distribution Shift) *Consider the effect from a counterfactual change of covariate distribution from F_0 to F_1:*

$$\theta_0 = \int \gamma_0(x)d(F_1(x) - F_0(x)) = \int \gamma_0(x)[d(F_1(x) - F_0(x))/dF(x)]dF(x).$$

Example 3 (Average Treatment Effect) *Consider the average treatment effect under unconfoundedness. Here $X = (Z, D)$ and $\gamma_0(X) = \gamma_0(D, Z)$, where $D \in \{0, 1\}$ is the indicator of the receipt of the treatment, and*

$$\theta_0 = \int (\gamma_0(1, z) - \gamma_0(0, z))dF(z) = \int \gamma_0(d, z)(1(d = 1) - 1(d = 0))dF(x).$$

We consider $\gamma \mapsto \mathrm{E}m(X, \gamma)$ as a continuous linear functional on Γ_b, the linear subspace of $L^2(F)$ spanned by the given dictionary $x \mapsto b(x)$. Such functionals can be represented via the Riesz representation:

$$\mathrm{E}m(X, \gamma) = \mathrm{E}\gamma_0(X)\alpha_0(X),$$

where the Riesz representer $\alpha_0(X) = b(X)'\rho_0$ is identified by the system of equations:

$$\mathrm{E}m(X, b) = \mathrm{E}b(X)b(X)'\rho_0,$$

where $m(X, b) := \{m(X, b_j)\}_{j=1}^p$ is a componentwise application of linear functional $m(X, \cdot)$ to $b = \{b_j\}_{j=1}^p$. Having the Riesz representer allows us to write the following "doubly robust" representation for θ_0:

$$\theta_0 = \mathrm{E}[m(X, \gamma_0) + \alpha_0(X)(Y - \gamma_0(X))].$$

This representation is approximately invariant to small perturbations of parameters γ_0 and α_0 around their true values (see Lemma 2 below for details), a property sometime referred to as the Neyman-type orthogonality (see, e.g., Chernozhukov et al. (2016a)), making this representation a good one to use as a basis for estimation and inference in modern high-dimensional settings. (See also Proposition 5 in Chernozhukov et al. (2016b) for a formal characterization of scores of this sort as having the double robustness property in the sense of Robins and Rotnitzky (1995)).

Our estimation and inference will explore an empirical analog of this equation, given a random sample $(Y_i, X_i')_{i=1}^n$ generated as i.i.d. copies of (Y, X'). Instead of the unknown γ_0 and β_0 we will plug-in estimators obtained using ℓ_1- regularization. We shall use sample-splitting in the form of cross-fitting to obtain weak assumptions on the problem, requiring only approximate sparsity of either β_0 or ρ_0, with some weak restrictions on the sparsity indexes. For example, if both parameter values are sparse, then the product of effective dimensions has to be much smaller than n, the sample size. Moreover, one of the parameter values, but not both, can actually be "dense" and estimated at the so called "slow" rate, as long as the other parameter is sparse, having effective dimension smaller than $\sqrt{n}$.

We establish that that the resulting "double" (or de-biased) machine learning (DML) estimator $\hat{\theta}$ concentrates in a $1/\sqrt{n}$ neigborhood of the target with deviations controlled by the normal laws,

$$\sup_{t \in \mathbb{R}} \left| \mathrm{P}(\sqrt{n}\sigma^{-1}(\hat{\theta} - \theta_0) \leq t) - \Phi(t) \right| \leq \varepsilon_n,$$

where the non-asymptotic bounds on ε_n are also given.

As the dimension of b grows, suppose that the subspace Γ_b becomes larger and approximates the infinite-dimensional linear subspace $\Gamma^* \subseteq L^2(P)$, that contains the true γ_0^*. We assume the functional $\gamma \mapsto \mathrm{E}m(X, \gamma)$ is continuous on Γ^* in this case. If the approximation bias is small,

$$\sqrt{n}(\theta_0 - \theta_0^*) \to 0,$$

our inference will automatically focus on the "ideal" target θ_0^*. Therefore our inference can be interpreted as targeting the functionals of the conditional expectation function in the regimes where we can successfully approximate them. However our approach does not hinge on this property and retains interpretability and good properties under misspecification.

It is interesting to note that in the latter case α_0 will approximate the true Riesz representer α_0^* for the linear functionals on Γ^*, identified by the system of equations:

$$\mathrm{E}m(X, \gamma) = \mathrm{E}\gamma(X)\alpha_0^*(X), \quad \forall \gamma \in \Gamma^*.$$

Note that θ_0^* has a "doubly robust" representation:

$$\theta_0^* = \mathrm{E}[m(X, \gamma_0^*) - \alpha_0^*(X)(Y - \gamma_0^*(X))], \tag{1}$$

which is invariant to perturbations of γ_0^* and α_0^*. Hence, our approach can be viewed as approximately solving the empirical analog of these equations, in the regimes where γ_0 does approximate γ_0^*. In such cases our estimator attains the semi-parametric efficiency bound, because its influence function is in fact the efficient score for θ_0; see van der Vaart (1991); Newey (1994).

When $\Gamma^* = L^2(F)$ and the functional $\gamma \mapsto \mathrm{E}m(X, \gamma)$ is continuous on Γ^*, the Riesz representer $\alpha_0^*(X)$ belongs to $L^2(F)$ and can be stated explicitly in many examples:

in Example 1: $\alpha_0^*(x) = -\partial_x \log f(x)$, where $f(x) = dF(x)/dx$,

in Example 2: $\alpha_0^*(x) = d(F_1(x) - F_0(x))/dF(x)$,

in Example 3: $\alpha_0^*(x) = (1(d = 1) - 1(d = 0))/P(d \mid z)$.

However, such closed-form solutions are not available in many other examples, or when Γ^* is smaller than $L^2(F)$, which is probably the most realistic situation occurring in practice.

Using closed-form solutions for Riesz representers α_0^* in several leading examples and their machine learning estimators, Chernozhukov et al. (2016a) defined DML estimators of θ_0^* in high-dimensional settings and established their good properties. Compared to this approach, the new approach proposed in this paper has the following advantages and some limitations:

1. It automatically estimates the Riesz representer α_0 from the empirical analog of equations that implicitly characterize it.

2. It does not rely on closed-form solutions for α_0^*, which generally are not available.

3. When closed-form solutions for α_0^* are available, it avoids directly estimating α_0^*. For example, it avoids estimating derivatives of densities in Example 1 or inverting estimated propensity scores $P(d \mid z)$ in Example 3. Rather it estimates the projections α_0 of α_0^* on the subspace Γ_b, which is a much simpler problem when the dimension of X is high.

4. Our approach remains interpretable under misspecification – when approximation errors are not small, we simply target inference on θ_0 instead of θ_0^*.

5. While the current paper focuses only on sparse regression methods, the approach readily extends to cover other machine learning estimators $\hat{\gamma}$ as long as we can find (numerically) dictionaries b that (approximately span) the realizations of $\hat{\gamma} - \gamma_0$, where γ_0 is the probability limit of $\hat{\gamma}$.

6. The current approach is limited to linear functionals, but in an ongoing work we are able to extend the approach by first performing a linear expansion and then applying our new methods to the linear part of the expansion.

The paper also builds upon ideas in classical semi-parametric learning theory with low-dimensional X, which focused inference on ideal θ_0^* using traditional smoothing methods for estimating nuisance parameters γ_0 and α_0 [van der Vaart (1991); Newey (1994); Bickel et al. (1998); Robins and Rotnitzky (1995); van der Vaart (1998)], that do not apply to the current high-dimensional setting. Our paper also builds upon and contributes to the literature on the modern orthogonal/debiased estimation and inference [Zhang and Zhang (2014); Belloni et al. (2014a,b, 2015); Javanmard and Montanari (2014a,b, 2015); van de Geer et al. (2014); Ning and Liu (2014); Chernozhukov et al. (2015); Neykov et al. (2015); Ren et al. (2015); Jankova and Van De Geer (2015, 2016a,b); Bradic and Kolar (2017); Zhu and Bradic (2017b,a)], which focused on inference on the coefficients in high-dimensional linear and generalized linear regression models, without considering the general linear functionals analyzed here.

Notation. Let $W = (Y, X')'$ be a random vector with law P on the sample space $\mathcal{W}$, and $W_1^n = (Y_i, X_i)_{i=1}^n$ denote the i.i.d. copies of W. All models and probability measure P can be indexed by n, a sample size, so that the models and their dimensions can change with n, allowing any of the dimensions to increase with n.

We use the notation from the empirical process theory, see Van Der Vaart and Wellner (1996). Let $\mathbb{E}_I f$ denote the empirical average of $f(W_i)$ over $i \in I \subset \{1, ..., n\}$:

$$\mathbb{E}_I f := \mathbb{E}_I f(W) = |I|^{-1} \sum_{i \in I} f(W_i).$$

Let $\mathbb{G}_I$ denote the empirical process over $f \in \mathcal{F} : \mathcal{W} \to \mathbb{R}^p$ and $i \in I$, namely

$$\mathbb{G}_I f := \mathbb{G}_I f(W) := |I|^{-1/2} \sum_{i \in I} (f(W_i) - Pf),$$

where $Pf := Pf(W) := \int f(w)dP(w)$. Denote by the $L^q(P)$ norm of a measurable function f mapping the support of W to the real line and also the $L^q(P)$ norm of random variable $f(W)$ by $\|f\|_{P,q} = \|f(W)\|_{P,q}$. We use $\|\cdot\|_q$ to denote ℓ_q norm on $\mathbb{R}^d$.

For a differentiable map $x \mapsto f(x)$, mapping $\mathbb{R}^d$ to $\mathbb{R}^k$, we use $\partial_{x'} f$ to abbreviate the partial derivatives $(\partial/\partial x')f$, and we correspondingly use the expression $\partial_{x'} f(x_0)$ to mean $\partial_{x'} f(x) \mid_{x=x_0}$, etc. We use x' to denote the transpose of a column vector x.

2. The DML with Regularized Riesz Representers

2.1. Sparse Approximations for the Regression Function and the Riesz Representer

We work with the set up above. Consider a conditional expectation function $x \mapsto \gamma_0^*(x) = \mathrm{E}[Y \mid X = x]$ such that $\gamma_0 \in L^2(F)$ and a p-vector of dictionary terms $x \mapsto b(x) = (b_j(x))_{j=1}^p$ such that $b \in L^2(F)$. The dimension p of the dictionary can be large, potentially much larger than n.

We approximate γ_0^* as

$$\gamma_0^* = \gamma_0 + r_\gamma := b'\beta_0 + r_\gamma,$$

where r_γ is the approximation error, and $\gamma_0 := b'\beta_0$ is the "best sparse linear approximation" defined via the following Dantzig Selector type problem (Candes and Tao (2007)).

Definition 1 (Best Sparse Linear Predictor) *Let β_0 be a minimal ℓ_1-norm solution to the approximate best linear predictor equations*

$$\beta_0 \in \arg\min \|\beta\|_1 : \|\mathrm{E}[b(X)(Y - b(X)'\beta)]\|_\infty \leq \lambda_0^\beta.$$

When $\lambda_0^\beta = 0$, β_0 becomes the best linear predictor parameter (BLP).

We refer to the resulting approximation as "sparse", since solutions β_0 often are indeed sparse. Note that since $\mathrm{E}[Y \mid X] = \gamma_0^*(X)$, the approximation error r_γ is approximately orthogonal to b:

$$\|\mathrm{E}[b(X)(Y - b(X)'\beta_0)\|_\infty = \|\mathrm{E}[b(X)(\gamma_0^*(X) - b(X)'\beta_0)]\|_\infty = \|\mathrm{E}[b(X)r_\gamma(X)]\|_\infty \leq \lambda_0^\beta.$$

Consider a linear subspace $\Gamma^* \subset L^2(F)$ that contains Γ_b, the linear subspace generated by b. In some of the asymptotic results that follow, we can have $\Gamma_b \uparrow \Gamma^*$ as $p \to \infty$.

C. *For each $x \in \mathcal{X}$, consider a linear map $\gamma \mapsto m(x, \gamma)$ from Γ^* to $\mathbb{R}$, such that for each $\gamma \in \Gamma^*$, the map $x \mapsto m(x, \gamma)$ from $\mathcal{X}$ to $\mathbb{R}$ is measurable, and the functional $\gamma \mapsto \mathrm{E}m(X, \gamma)$ is continuous on Γ^* with respect to the $L^2(P)$ norm.*

Under the continuity condition in C, this functional admits a Riesz representer $\alpha_0^* \in L^2(F)$. We approximate the Riesz representer via:

$$\alpha_0^*(X) = b(X)'\rho_0 + r_\alpha(X),$$

where $r_\alpha(X)$ is the approximation error, and $b(X)'\rho_0$ is the best sparse linear approximation defined as follows.

Definition 2 (Best Sparse Linear Riesz Representer) *Let ρ_0 be a minimal ℓ_1-norm solution to the approximate Riesz representation equations:*

$$\rho_0 \in \arg\min \|\rho\|_1 : \|\mathrm{E}m(X,b) - \mathrm{E}b(X)b(X)'\rho\|_\infty \leq \lambda_0^\rho,$$

where λ_0^ρ is a regularization parameter. When $\lambda_0^\rho = 0$, we obtain $\alpha_0(X) = b(X)'\rho_0$, a Riesz representer for functionals $\mathrm{E}m(X,\gamma)$ when $\gamma \in \Gamma_b$.

As before, we refer to the resulting approximation as "sparse", since the solutions to the problem would often be sparse. Since $\mathrm{E}\alpha^*(X)b(X) = \mathrm{E}m(X,b)$, we conclude that the approximation error $r_\alpha(X)$ is approximately orthogonal to $b(X)$:

$$\|\mathrm{E}(\alpha_0^*(X) - b(X)'\rho_0)b(X)\|_\infty = \|\mathrm{E}[r_\alpha(X)b(X)]\|_\infty \leq \lambda_0^\rho.$$

The estimation will be carried out using the sample analogs of the problems above, and is a special case of the following Dantzig Selector-type problem.

Definition 3 (Regularized Minimum Distance (RMD)) *Consider a parameter $t \in T \subset \mathbb{R}^p$, such that T is a convex set with $\|T\|_1 := \sup_{t\in T}\|t\|_1 \leq B$. Consider the moment functions $t \mapsto g(t)$ and the estimated function $t \mapsto \hat{g}(t)$, mapping T to $\mathbb{R}^p$:*

$$g(t) = Gt + M; \quad \hat{g}(t) = \hat{G}t + \hat{M},$$

where G and $\hat{G}$ are p by p non-negative-definite matrices and M and $\hat{M}$ are p-vectors. Assume that t_0 is the target parameter that is well-defined by:

$$t_0 \in \arg\min \|t\|_1 : \|g(t)\|_\infty \leq \lambda_0, \quad t \in T. \tag{2}$$

Define the RMD estimator $\hat{t}$ by solving

$$\hat{t} \in \arg\min \|t\|_1 : \|\hat{g}(t)\|_\infty \leq \lambda_0 + \lambda_1, \quad t \in T,$$

where λ_1 is chosen such that $\|\hat{g}(t_0) - g(t_0)\|_\infty \leq \lambda_1$, with probability at least $1 - \epsilon_n$.

We define the estimators of β_0 and ρ_0 over subset of data, indexed by a non-empty subset A of $\{1, ..., n\}$.

Definition 4 (RMD Estimator for Sparse BLP) *Define $\hat{\beta}_A$ as the RMD estimator with parameters $t_0 = \beta_0$, T a convex set with $\|T\|_1 \leq B$, and*

$$\hat{G} = \mathbb{E}_A bb', \quad G = \mathrm{E}bb', \quad \hat{M} = \mathbb{E}_A Yb, \quad M = \mathrm{E}Yb(X).$$

Definition 5 (RMD Estimator for Sparse Riesz Representer) *Define $\hat{\rho}_A$ as the RMD estimator with parameters $t_0 = \beta_0$, T a convex set with $\|T\|_1 \leq B$, and*

$$\hat{G} = \mathbb{E}_A bb', \quad G = \mathrm{E}bb', \quad \hat{M} = -\mathbb{E}_A m(X,b), \quad M = -\mathrm{E}m(X,b).$$

2.2. Properties of RMD Estimators

Consider sequences of constants $\ell_{1n} \geq 1$, $\ell_{2n} \geq 1$, $B_n \geq 0$, and $\epsilon_n \searrow 0$, indexed by n.

MD *We have that $t_0 \in T$ with $\|T\|_1 := \sup_{t \in T} \|t\|_1 \leq B_n$, and the empirical moments obey the following bounds with probability at least $1 - \epsilon_n$:*

$$\sqrt{n}\|\hat{G} - G\|_\infty \leq \ell_{1n} \text{ and } \sqrt{n}\|\hat{M} - M\|_\infty \leq \ell_{2n}.$$

Note that in many applications the factors ℓ_{1n} and ℓ_{2n} can be chosen to grow slowly, like $\sqrt{\log(p \vee n)}$, using self-normalized moderate deviation bounds (Jing et al. (2003); Belloni et al. (2014b)), under mild moment conditions, without requiring sub-Gaussianity.

Define the identifiablity factors for $t_0 \in T$ as :

$$s^{-1}(t_0) := \inf_{\delta \in R(t_0)} |\delta' G \delta| / \|\delta\|_1^2,$$

where $R(t_0)$ is the restricted set:

$$R(t_0) := \{\delta : \|t_0 + \delta\|_1 \leq \|t_0\|_1, \quad t_0 + \delta \in T\},$$

where $s^{-1}(t_0) := \infty$ if $t_0 = 0$. The restricted set contains the estimation error $\hat{t} - t_0$ for RMD estimators with probability at least $1 - \epsilon_n$. We call the inverse of the identifiability factor, $s(t_0)$, the "effective dimension", as it captures the effective dimensionality of t_0; see remark below.

Remark 1 (Identifiability and Effective Dimension) *The identifiability factors were introduced in Chernozhukov et al. (2013) as a generalization of the restricted eigenvalue of Bickel et al. (2009). Indeed, given a vector $\delta \in \mathbb{R}^p$, let δ_A denote a vector with the j-th component set to δ_j if $j \in A$ and 0 if $j \notin A$. Then $s^{-1}(t_0) \geq s^{-1}k/2$ or*

$$s(t_0) \leq 2s/k,$$

where k is the restricted eigenvalue: $k := \inf |\delta' G \delta| / \|\delta_M\|_2^2 : \delta \neq 0, \|\delta_{M^c}\|_1 \leq \|\delta_M\|_1$, $M =$ support(t_0), $M^c = \{1, ..., p\} \setminus M$, and $s = \|t_0\|_0$. The inequality follows since $\|t_0 + \delta\|_1 \leq \|t_0\|_1$ implies $\|\delta_{M^c}\|_1 \leq \|\delta_M\|_1$, so that $\|\delta\|_1^2 \leq 2\|\delta_M\|_1^2 \leq 2s\|\delta_M\|_2^2$. Here s is the number of non-zero components of t_0. Hence we can view $s(t_0)$ as a measure of effective dimension of t_0.

Lemma 1 (Regularized Minimum Distance Estimation) *Suppose that MD holds. Let*

$$\bar{\lambda} := \tilde{\ell}_n / \sqrt{n}, \quad \tilde{\ell}_n := (\ell_{1n} B_n + \ell_{2n}).$$

Define the RMD estimand t_0 via (2) with $\lambda_0 \vee \lambda_1 \leq \bar{\lambda}$. Then with probability $1 - 2\epsilon_n$ the estimator $\hat{t}$ exists and obeys:

$$(\hat{t} - t_0)' G (\hat{t} - t_0) \leq (s(t_0)(4\bar{\lambda})^2) \wedge (8 B_n \bar{\lambda}).$$

There are two bounds on the rate, one is the "slow rate" bound $8 B_n \bar{\lambda}$ and the other one is "fast rate". The "fast rate" bound is tighter in regimes where the "effective dimension" $s(t_0)$ is not too large.

Corollary 1 (RMD for BLP and Riesz Representer) *Suppose β_0 and ρ_0 belong to a convex set T with $\|T\|_1 \leq B_n$. Consider a random subset A of $\{1, ..., n\}$ of size $m \geq n - n/K$ where K is a fixed integer. Suppose that the dictionary $b(X)$ and Y obey with probability at least $1 - \epsilon_n$*

$$\max_{j,k \leq p} |\mathbb{G}_A b_k(X) b_j(X)| \leq \ell_{1n}, \quad \max_{j \leq p} |\mathbb{G}_A (Y b_j(X))| \leq \ell_{2n}, \quad \max_{j \leq p} |\mathbb{G}_A m(X, b_j)| \leq \ell_{2n}.$$

If we set $\lambda_0^\beta \vee \lambda_1^\beta \vee \lambda_0^\rho \vee \lambda_1^\rho \leq \bar{\lambda} = \tilde{\ell}_n/\sqrt{n}$ for $\tilde{\ell}_n = B\ell_{1n} + \ell_{2n}$, then with probability at least $1 - 4\epsilon_n$, estimation errors $u = \hat{\beta}_A - \beta_0$ and $v = \hat{\rho}_A - \rho_0$ obey

$$u \in R(\beta_0), \quad [u'Gu]^{1/2} \leq r_1 := 4[(\tilde{\ell}_n (s(\beta_0)/n)^{1/2}) \wedge (\tilde{\ell}_n^{1/2} B_n n^{-1/4})],$$

$$v \in R(\rho_0), \quad [v'Gv]^{1/2} \leq r_2 := 4[(\tilde{\ell}_n (s(\rho_0)/n)^{1/2}) \wedge (\tilde{\ell}_n^{1/2} B_n n^{-1/4})].$$

The corollary follows because the stated conditions imply condition MD by Holder inequality.

2.3. Approximate Neyman-orthogonal Score Functions and the DML Estimator

Our DML estimator of θ_0 will be based on using the following score function:

$$\psi(W_i, \theta; \beta, \rho) = \theta - m(X, b(X))'\beta - \rho' b(X)(Y - b(X)'\beta).$$

Lemma 2 (Basic Properties of the Score) *The score function has the following properties:*

$$\partial_\beta \psi(X, \theta; \beta, \rho) = -m(X, b(X)) + \rho' b(X) b(X), \quad \partial_\rho \psi(X, \theta; \beta, \rho) = -b(X)(Y - b(X)\beta),$$

$$\partial^2_{\beta\beta'} \psi(X, \theta; \beta, \rho) = \partial^2_{\rho\rho'} \psi(X, \theta; \beta, \rho) = 0, \quad \partial^2_{\beta\rho'} \psi(X, \theta; \beta, \rho) = b(X) b(X)'.$$

This score function is approximately Neyman orthogonal at the sparse approximation (β_0, ρ_0), namely

$$\|\mathrm{E}[\partial_\beta \psi(X, \theta_0; \beta, \rho_0)]\|_\infty = \|M - \rho_0{}'G\|_\infty \leq \lambda_0^\rho,$$

$$\|\mathrm{E}[\partial_\rho \psi(X, \theta_0; \beta_0, \rho)]\|_\infty = \|\mathrm{E}[b(X)(Y - b(X)'\beta_0)]\|_\infty \leq \lambda_0^\beta.$$

The second claim of the lemma is immediate from the definition of (β_0, ρ_0) and the first follows from elementary calculations.

The approximate orthogonality property above says that the score function is approximately invariant to small perturbations of the nuisance parameters ρ and β around their "true values" ρ_0 and β_0. Note that the score function is exactly invariant, and becomes the doubly robust score, whence $\lambda_0^\rho = 0$ and $\lambda_0^\beta = 0$. This approximate invariance property plays a crucial role in removing the impact of biased estimation of nuisance parameters ρ_0 and β_0 on the estimation of the main parameters θ_0. We now define the double/de-biased machine learning estimator (DML), which makes use of the cross-fitting, an efficient form of data splitting.

Definition 6 (DML Estimator) *Consider the partition of $\{1, ..., n\}$ into $K \geq 2$ blocks $(I_k)_{k=1}^K$, with n/K observations in each block I_k (assume for simplicity that K divides n). For each $k = 1, ..., K$, and and $I_k^c = \{1, ..., N\} \setminus I$, let estimator $\hat{\theta}_{I_k}$ be defined as the root:*

$$\mathbb{E}_I \psi(W_i, \hat{\theta}_{I_k}; \hat{\beta}_{I_k^c}, \hat{\rho}_{I_k^c}) = 0,$$

where $\hat{\beta}_{I_k^c}$ and $\hat{\rho}_{I_k^c}$ are RMD estimators of β_0 and ρ_0 that have been obtained using the observations with indices $A = I_k^c$. Define the DML estimator $\hat{\theta}$ as the average of the estimators $\hat{\theta}_{I_k}$ obtained in each block $k \in \{1, ..., K\}$:

$$\hat{\theta} = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_{I_k}.$$

2.4. Properties of DML with Regularized Riesz Representers

We will note state some sufficient regularity conditions for DML. Let n denote the sample size. To give some asymptotic statements, let $\ell_{1n} \geq 1$, $\ell_{2n} \geq 1$, $\ell_{3n} \geq 1$, $B_n \geq 0$, and $\delta_n \searrow 0$ and $\epsilon_n \searrow 0$ denote sequences of positive constants. Let c, C, and q be positive constants such that $q > 3$, and let $K \geq 2$ be a fixed integer. We consider sequence of integers n such that K divides n (to simplify notation). Fix all of these sequences and the constants. Consider P that satisfies the following conditions:

R.1 *Assume condition C holds and that in Definitions 1 and 2, it is possible to set λ_0^ρ and λ_0^β such that (a) $\sqrt{n}(\lambda_0^\rho + \lambda_0^\beta)B_n \leq \delta_n$ and (b) the resulting parameter values β_0 and ρ_0 are well behaved with $\rho_0 \in T$ and $\beta_0 \in T$, where T is a convex set with $\|T\|_1 \leq B_n$.*

R.2 *Given a random subset A of $\{1, ..., n\}$ of size $n - n/K$, the terms in dictionary $b(X)$ and outcome Y obey with probability at least $1 - \epsilon_n$*

$$\max_{j,k\leq p}|\mathbb{G}_A b_k(X)b_j(X)| \leq \ell_{1n}, \quad \max_{j\leq p}|\mathbb{G}_A Y b_j(X)| \leq \ell_{2n}, \quad \max_{j\leq p}|\mathbb{G}_A m(X, b_j)| \leq \ell_{2n}.$$

R.3 *Assume that $c \leq \|\psi(W, \theta_0; \beta_0, \rho_0)\|_{P,q} \leq C$ for $q = 2$ and 3 and that the following continuity relations hold for all $u \in R(\beta_0)$ and $v \in R(\rho_0)$,*

$$\sqrt{\mathrm{Var}}((m(X, b) + \rho_0{}'b(X)b(X))'u) \leq \ell_{3n}\|b(X)'u\|_{P,2}$$

$$\sqrt{\mathrm{Var}}((Y - b(X)'\beta_0)b(X)'v) \leq \ell_{3n}\|b(X)'v\|_{P,2}$$

$$\sqrt{\mathrm{Var}}(u'b(X)b(X)'v) \leq \ell_{3n}(\|b(X)'u\|_{P,2} + \|b(X)'v\|_{P,2}).$$

R.4 *For $\tilde{\ell}_n := \ell_{1n}B_n + \ell_{2n}$,*

$$\left.\begin{aligned} r_1 &:= 4[(\tilde{\ell}_n(s(\beta_0)/n)^{1/2}) \wedge (\tilde{\ell}_n^{1/2}B_n n^{-1/4})] \\ r_2 &:= 4[(\tilde{\ell}_n(s(\rho_0)/n)^{1/2}) \wedge (\tilde{\ell}_n^{1/2}B_n n^{-1/4})] \end{aligned}\right| \quad \textit{we have: } n^{1/2}r_1 r_2 + \ell_{3n}(r_1 + r_2) \leq \delta_n.$$

R.1 requires that sparse approximations of γ_0^* and α_0^* with respect to the dictionary b admit well-behaved parameters β_0 and ρ_0. R.2 is a weak assumption that can be satisfied by taking $\ell_{1n} = \ell_{2n} = \sqrt{\log(p \vee n)}$ under weak assumptions on moments, as follows from self-normalized moderate deviation bounds (Jing et al. (2003); Belloni et al. (2014b)), without requiring subgaussian tails bounds. R.3 imposes a modulus of continuity for bounding variances: if elements of the dictionary are bounded with probability one, $\|b(X)\|_\infty \leq C$, then we can select $\ell_{3n} = CB_n$ for many functionals of interest, so the assumption is plausible. R.4 imposes condition on the rates r_1

and r_2 of consistency of β_0 and ρ_0, requiring in particular that $r_1 r_2 n^{1/2}$, an upper bound on the bias of the DML estimator, is small; see also the discussion below.

Consider the oracle estimator based upon the true score functions:

$$\bar{\theta} = \theta_0 + n^{-1} \sum_{i=1}^{n} \psi_0(W_i), \quad \psi_0(W_i) := \psi(W_i, \theta_0; \beta_0, \rho_0).$$

Theorem 1 (Adaptive Estimation and Approximate Gaussian Inference) *Under R.1-R.4, we have the adaptivity property, namely the difference between the DML and the oracle estimator is small: for any $\Delta_n \in (0,1)$,*

$$|\sqrt{n}(\hat{\theta} - \bar{\theta})| \leq R_n := \bar{C}\delta_n / \Delta_n$$

with probability at least $1 - \Pi_n$ for $\Pi_n := \sqrt{2}K(4\epsilon_n + \Delta_n^2)$, where $\bar{C}$ is an absolute constant. As a consequence, $\hat{\theta}_0$ concentrates in a $1/\sqrt{n}$ neighborhood of θ_0, with deviations approximately distributed according to the Gaussian law, namely:

$$\sup_{z \in \mathbb{R}} \left| \mathrm{P}_P(\sigma^{-1}\sqrt{n}(\hat{\theta}_0 - \theta_0) \leq z) - \Phi(z) \right| \leq \bar{C}'(n^{-1/2} + R_n) + \Pi_n, \tag{3}$$

where $\sigma^2 = \mathrm{Var}(\psi_0(W_i))$, $\Phi(z) = \mathrm{P}(N(0,1) \leq z)$, and $\bar{C}'$ only depends on (c, C).

The constants $\Delta_n > 0$ can be chosen such that right-hand side of (3) converges to zero as $n \to \infty$, yielding the following asymptotic result.

Corollary 2 (Uniform Asymptotic Adaptivity and Gaussianity) *Fix constants and sequences of constants specified at the beginning of Section 2.4. Let $\mathcal{P}$ be the set of probability laws P that obey conditions R.1-R.4 uniformly for all n. Then DML estimator $\hat{\theta}$ is uniformly asymptotically equivalent to the oracle estimator $\bar{\theta}$, that is $|\sqrt{n}(\hat{\theta} - \bar{\theta})| = O_P(\delta_n)$ uniformly in $P \in \mathcal{P}$ as $n \to \infty$. Moreover, $\sqrt{n}\sigma^{-1}(\hat{\theta} - \theta_0)$ is asymptotically Gaussian uniformly in $P \in \mathcal{P}$:*

$$\lim_{n \to \infty} \sup_{P \in \mathcal{P}} \sup_{z \in \mathbb{R}} \left| \mathrm{P}_P(\sigma^{-1}\sqrt{n}(\hat{\theta}_0 - \theta_0) \leq z) - \Phi(z) \right| = 0.$$

Hence the DML estimator enjoys good properties under the stated regularity conditions. Less primitive regularity conditions can be deduced from the proofs directly.

Remark 7 (Sharpness of Conditions) *The key regularity condition imposes that the bounds on estimation errors r_1 and r_2 are small and that the product $n^{1/2} r_1 r_2$ is small. Ignoring the impact of "slow" factors ℓ_n's and assuming B_n is bounded by a constant B, this requirement is satisfied if*

one of the effective dimensions is smaller than $\sqrt{n}$, either $s(\beta_0) \ll \sqrt{n}$ or $s(\rho_0) \ll \sqrt{n}$.

The latter possibility allows one of the parameter values to be "dense", having unbounded effective dimension, in which case this parameter can be estimated at the "slow" rate $n^{-1/4}$. These types of conditions are rather sharp, matching similar conditions used in Javanmard and Montanari (2015) in the case of inference on a single coefficient in Gaussian sparse linear regression models, and those in Zhu and Bradic (2017a) in the case of point testing general linear hypotheses on regression coefficients in linear Gaussian regression models.

Proceeding further, we notice that the difference between our target and the ideal target is:

$$|\theta_0 - \theta_0^*| = |\mathrm{E} r_\alpha(X) r_\gamma(X)| \leq \|r_\alpha\|_{P,2} \|r_\gamma\|_{P,2}.$$

Hence we obtain the following corollary.

Corollary 3 (Inference targets θ_0^* when approximation errors are small) *Suppose that, in addition to R.1-R.4, the product of approximation errors $r_\gamma = \gamma_0^* - \gamma_0$ and $r_\alpha = \alpha_0^* - \alpha_0$ is small,*

$$\sqrt{n} |\mathrm{E} r_\alpha(X) r_\gamma(X)| \leq \delta_n.$$

Then conclusions of Theorem 1 and Corollary 1 hold with θ_0 replaced by θ_0^* and the constants $\bar{C}$ and $\bar{C}'$ replaced by $2\bar{C}$ and $2\bar{C}'$.

The plausibility of $\sqrt{n}|\mathrm{E} r_\alpha(X) r_\gamma(X)| \leq \sqrt{n}\|r_\alpha\|_{P,2}\|r_\gamma\|_{P,2}$ being small follows from the fact that many rich functional classes admit sparse linear approximations with respect to conventional dictionaries b. For instance, Tsybakov (2012) and Belloni et al. (2014b) give examples of Sobolev and rearranged Sobolev balls, respectively, as the function classes and elements of the Fourier basis as the dictionary b, in which sparse approximations have small errors.

Remark 8 (Sharpness of Conditions in the Context of ATE) *In the context of vanishing approximation errors, and in the context of Example 3 on Average Treatment Effects, our estimator implicitly estimates the inverse of the propensity score directly rather than inverting a propensity score estimator as in most of the literature. The approximate residual balancing estimator of Athey et al. (2016) can also be thought of as implicitly estimating the inverse propensity score. An advantage of the estimator here is its DML form allows us to tradeoff rates at which the mean and the inverse propensity score are estimated while maintaining root-n consistency. Also, we do not require that the conditional mean be linear and literally sparse with the sparsity index $s(\beta_0) \ll \sqrt{n}$; in fact we can have a completely dense conditional mean function when the approximation to the Rietz representer has the effective dimension $s(\rho_0) \ll \sqrt{n}$. More generally, when the approximation errors don't vanish, our analysis also explicitly allows for misspecification of the regression function.*

3. Proofs

Proof of Lemma 1. Consider the event $\mathcal{E}_n$ such that

$$\|\hat{g}(t_0)\| \leq \lambda_0 + \lambda_1 \text{ and } \sup_{t \in T} \|\hat{g}(t) - g(t)\|_\infty \leq \bar{\lambda} \tag{4}$$

holds. This event holds with probability at least $1 - 2\epsilon_n$. Indeed, by the choice of λ_1 and $\|g(t_0)\| \leq \lambda_0$, we have with probability at least $1 - \epsilon_n$:

$$\|\hat{g}(t_0)\|_\infty \leq \|\hat{g}(t_0) - g(t_0)\|_\infty + \|g(t_0)\| \leq \lambda_1 + \lambda_0.$$

Hence on the event $\mathcal{E}_n$ we have

$$\|\hat{t}\|_1 \leq \|t_0\|_1 \quad \|\hat{g}(\hat{t})\|_\infty \leq \lambda_0 + \lambda_1.$$

This implies, by $\|g(t_0)\| \leq \lambda_0$, by $\lambda_0 \wedge \lambda_1 \leq \bar{\lambda}$, and by (4), that

$$\|G(\hat{t} - t_0)\|_\infty = \|g(\hat{t}) - g(t_0)\|_\infty \leq 2\lambda_0 + \lambda_1 + \sup_{t \in T} \|\hat{g}(t) - g(t)\|_\infty \leq 4\bar{\lambda}.$$

Then $\delta = \hat{t} - t_0$ obeys, by definition of $s(t_0)$,

$$\|\delta\|_1^2 \leq s(t_0)\delta' G\delta \leq s(t_0)\|G\delta\|_\infty \|\delta\|_1 \leq s(t_0) 4\bar{\lambda}\|\delta\|_1,$$

which implies that

$$\|\delta\|_1 \leq s(t_0)4\bar{\lambda}, \quad \delta' G\delta \leq s(t_0)(4\bar{\lambda})^2,$$

which establishes the first part of the bound.

The second bound follows from $\|\delta\|_1 \leq 2B_n$ and $\delta' G \delta \leq \|G\delta\|_\infty \|\delta\|_1 \leq 4\bar{\lambda}2B$. ■

Proof of Theorem 1. Step 1. We have a random partition (I_k, I_k^c) of $\{1, ..., n\}$ into sets of size n/K and $m := n - n/K$. Omit the indexing by k in this step. Here we bound $|\sqrt{n}(\hat{\theta}_I - \bar{\theta}_I)|$, where

$$\bar{\theta}_I = \theta_0 + \mathbb{E}_I \psi_0(W_i).$$

Define

$$\partial_\beta \psi_0(W_i) := \partial_\beta \psi(X, \theta; \beta_0, \rho_0) = m(X, b(X)) + {\rho_0}' b(X) b(X)$$
$$\partial_\rho \psi_0(W_i) := \partial_\rho \psi(X, \theta; \beta_0, \rho_0) = -b(X)(Y - b(X)'\beta_0)$$
$$\partial^2_{\beta\rho'} \psi_0(W_i) := \partial^2_{\beta\rho'} \psi(X, \theta; \beta_0, \rho_0) = b(X)b(X)'.$$

Define the estimation errors

$$u = \hat{\beta}_{I^c} - \beta_0 \text{ and } v = \hat{\rho}_{I^c} - \rho_0.$$

Since $\partial^2_{\beta\beta'} \psi(X, \theta; \beta, \rho) = 0$ and $\partial^2_{\rho\rho'} \psi(X, \theta; \beta, \rho) = 0$, as noted in Lemma 2, we have by the exact Taylor expansion:

$$\hat{\theta}_I = \bar{\theta}_I + (\mathbb{E}_I \partial_\beta \psi_0) u + (\mathbb{E}_I \partial_\rho \psi_0) v + u' (\mathbb{E}_I \partial^2_{\beta\rho'} \psi_0) v.$$

With probability at least $1 - 4\epsilon_n$, by Corollary 1, the following event occurs:

$$\mathcal{E}_n = \{u \in R(\beta_0), \ \ v \in R(\rho_0), \ \ \sqrt{u'Gu} \leq r_1, \ \ \sqrt{v'Gv} \leq r_2\}.$$

Using triangle and Holder inequalities, we obtain that on this event:

$$\begin{aligned} |\sqrt{m}(\hat{\theta}_I - \bar{\theta}_I)| \leq \text{rem}_I \quad := & \quad |\mathbb{G}_I \partial_\beta \psi_0 u| + \sqrt{m}\|P\partial_\beta \psi_0\|_\infty \|u\|_1 \\ + & \quad |\mathbb{G}_I \partial_\rho \psi_0 v| + \sqrt{m}\|P\partial_\rho \psi_0\|_\infty \|v\|_1 \\ + & \quad |u' \mathbb{G}_I \partial^2_{\beta\rho'} \psi_0 v| + \sqrt{m}|u'[P\partial^2_{\beta\rho'} \psi_0] v|. \end{aligned}$$

Moreover, on this event, by $\|R(\beta_0)\|_1 \leq 2B_n$ and $\|R(\rho_0)\|_1 \leq 2B_n$, by Lemma 2, and by R.1:

$$\sqrt{m}\|P\partial_\beta \psi_0\|_\infty \|u\|_1 \leq \sqrt{m}\lambda_0^\beta 2B_n \leq \delta_n, \sqrt{m}\|P\partial_\rho \psi_0\|_\infty \|v\|_1 \leq \sqrt{m}\lambda_0^\rho 2B_n \leq \delta_n.$$

Note that v and u are fixed once we condition on the observations $(W_i)_{i\in I^c}$. We have that on the event $\mathcal{E}_n$, by i.i.d. sampling, R.3, and R.4,

$$\begin{aligned}
\sqrt{\mathrm{Var}}[\mathbb{G}_I \partial_\beta \psi_0 u \mid (W_i)_{i\in I^c}] &= \sqrt{\mathrm{Var}}(\partial_\beta \psi_0 u \mid (W_i)_{i\in I^c}) \leq \ell_{3n}\sqrt{u'Gu} \leq \delta_n, \\
\sqrt{\mathrm{Var}}[\mathbb{G}_I \partial_\rho \psi_0 v \mid (W_i)_{i\in I^c}] &= \sqrt{\mathrm{Var}}(\partial_\rho \psi_0' v \mid (W_i)_{i\in I^c}) \leq \ell_{3n}\sqrt{v'Gv} \leq \delta_n, \\
\sqrt{\mathrm{Var}}[u'\mathbb{G}_I \partial^2_{\beta\rho'} \psi_0 v \mid (W_i)_{i\in I^c}] &= \sqrt{\mathrm{Var}}[u'bb'v \mid (W_i)_{i\in I^c}] \leq \ell_{3n}(u'Gu + v'Gv)^{1/2} \leq \delta_n, \\
\sqrt{m}|u'[P\partial^2_{\beta\rho'}\psi_0]v| &\leq \sqrt{m}|u'G\tilde{v}| \leq \sqrt{m}(u'Guv'Gv)^{1/2} \leq \delta_n.
\end{aligned}$$

Hence we have that for some numerical constant $\bar{C}$ and any $\Delta_n \in (0,1)$:

$$\mathrm{P}(\mathrm{rem}_I > \bar{C}\delta_n/\Delta_n) \leq \mathrm{P}(\mathrm{rem}_I > \bar{C}\delta_n/\Delta_n \cap \mathcal{E}_n) + \mathrm{P}(\mathcal{E}_n^c)$$

$$\leq \mathrm{E}\mathrm{P}(\mathrm{rem}_I > \bar{C}\delta_n/\Delta_n \cap \mathcal{E}_n \mid (W_i)_{i\in I^c}) + \mathrm{P}(\mathcal{E}_n^c) \leq \Delta_n^2 + 4\epsilon_n.$$

Step 2. Here we bound the difference between $\hat{\theta} = K^{-1}\sum_{k=1}^K \hat{\theta}_{I_k}$ and $\bar{\theta} = K^{-1}\sum_{k=1}^K \bar{\theta}_{I_k}$:

$$\sqrt{n}|\hat{\theta} - \bar{\theta}| \leq \frac{\sqrt{n}}{\sqrt{m}}\frac{1}{K}\sum_{k=1}^K \sqrt{m}|\hat{\theta}_{I_k} - \bar{\theta}_{I_k}| \leq \frac{\sqrt{n}}{\sqrt{m}}\frac{1}{K}\sum_{k=1}^K \mathrm{rem}_{I_k}.$$

By the union bound we have that

$$\mathrm{P}\left(\frac{1}{K}\sum_{k=1}^K \mathrm{rem}_{I_k} > \bar{C}\delta_n/\Delta_n\right) \leq K(\Delta_n^2 + 4\epsilon_n),$$

and we have that $\sqrt{n/m} = \sqrt{K/(K-1)} \leq \sqrt{2}$, since $K \geq 2$. So it follows that

$$|\sqrt{n}(\hat{\theta} - \bar{\theta})| \leq R_n := \bar{C}\delta_n/\Delta_n$$

with probability at least $1 - \Pi_n$ for $\Pi_n := \sqrt{2}K(4\epsilon_n + \Delta_n^2)$, where $\bar{C}$ is an absolute constant.

Step 3. To show the second claim, let $Z_n := \sqrt{n}\sigma^{-1}(\bar{\theta} - \theta_0)$. By the Berry-Esseen bound, for some absolute constant A,

$$\sup_{z\in\mathbb{R}} |\mathrm{P}(Z_n \leq z) - \Phi(z)| \leq A\|\psi_0/\sigma\|^3_{P,3} n^{-1/2} \leq A(C/c)^3 n^{-1/2},$$

where $\|\psi_0/\sigma\|^3_{P,3} \leq (C/c)^3$ by R.3. Hence, using Step 2, for any $z \in \mathbb{R}$, we have

$$\begin{aligned}
&\mathrm{P}(\sqrt{n}\sigma^{-1}(\hat{\theta} - \theta_0) \leq z) - \Phi(z) \leq \mathrm{P}(Z_n \leq z + \sigma^{-1}R_n) + \Pi_n - \Phi(z) \\
&\leq A(C/c)^3 n^{-1/2} + \bar{\phi}\sigma^{-1}R_n + \Pi_n \leq \bar{C}'(n^{-1/2} + R_n) + \Pi_n,
\end{aligned}$$

where $\bar{\phi} = \sup_z \phi(z)$, where ϕ is the density of $\Phi(z) = \mathrm{P}(N(0,1) \leq z)$, and $\bar{C}'$ depends only on (C,c). Similarly, conclude that $\mathrm{P}(\sqrt{n}\sigma^{-1}(\hat{\theta} - \theta_0) \leq z) - \Phi(z) \geq \bar{C}'(n^{-1/2} + R_n) - \Pi_n$. ■

References

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*, 2016.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies*, 81:608–650, 2014a. ArXiv, 2011.

A. Belloni, V. Chernozhukov, and K. Kato. Uniform post selection inference for lad regression models and other z-estimators. *Biometrika*, 102:77–94, 2015. ArXiv, 2013.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root lasso in nonparametric regression. *Annals of Statistics (arXiv 2011)*, 42(2):757–788, 2014b.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998.

P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009. ArXiv, 2008.

Jelena Bradic and Mladen Kolar. Uniform inference for high-dimensional quantile regression: linear functionals and regression rank scores. *arXiv preprint arXiv:1702.06209*, 2017.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41 (6):2786–2819, 2013.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7: 649–688, 2015.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*, 2016a.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James Robins. Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*, 2016b.

Jana Jankova and Sara Van De Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229, 2015.

Jana Jankova and Sara Van De Geer. Semi-parametric efficiency bounds and efficient estimation for high-dimensional models. *arXiv preprint arXiv:1601.00815*, 2016a.

Jana Jankova and Sara Van De Geer. Confidence regions for high-dimensional generalized linear models under sparsity. *arXiv preprint arXiv:1610.01353*, 2016b.

A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: asymptotic theory. *IEEE Transactions on Information Theory*, 60: 6522–6554, 2014a. ArXiv, 2013.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014b.

Adel Javanmard and Andrea Montanari. De-biasing the lasso: Optimal sample size for gaussian designs. *arXiv preprint arXiv:1508.02757*, 2015.

Bing-Yi Jing, Qi-Man Shao, Qiying Wang, et al. Self-normalized cramér-type large deviations for independent random variables. *The Annals of probability*, 31(4):2167–2215, 2003.

W. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382, 1994.

Matey Neykov, Yang Ning, Jun S Liu, and Han Liu. A unified theory of confidence regions and testing for high dimensional estimating equations. *arXiv preprint arXiv:1510.08986*, 2015.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *arXiv preprint arXiv:1412.8765*, 2014.

Zhao Ren, Tingni Sun, Cun-Hui Zhang, Harrison H Zhou, et al. Asymptotic normality and optimalities in estimation of large gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

J. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc*, 90(429):122–129, 1995.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2012.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202, 2014. ArXiv, 2013.

A. W. van der Vaart. On differentiable functionals. *Annals of Statistics*, 19(1):178–204, 1991.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

C. Zhang and S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. *J. R. Statist. Soc. B*, 76:217–242, 2014. ArXiv, 2012.

Yinchu Zhu and Jelena Bradic. Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, (just-accepted), 2017a.

Yinchu Zhu and Jelena Bradic. Breaking the curse of dimensionality in regression. *arXiv preprint arXiv:1708.00430*, 2017b.