# Learning L2 Continuous Regression Functionals Via Regularized Riesz Representers

Victor Chernozhukov, Whitney K Newey, Rahul Singh

North American Meeting of the Econometric Society
January 2019

# INTRODUCTION

Many interesting objects are mean square continuous functionals of a regression.

–Average consumer surplus bound.

–Parameters in imperfect information games.

–Average treatment effects.

–Average conditional covariance.

The regression may be high dimensional.

There may be many prices that should be included as covariates to the own price in estimating demand for some good.

There may be many covariates for the treatment effect.

May want to condition on many covariates in analysis of covariance.

This paper develops machine learning (ML) estimators of these objects and gives asymptotic theory.

Estimation based on debiased moment conditions, where first step has zero first order effect on moments.

Debiased moment conditions depend on two unknown functions, the regression and a Riesz representer.

We develop Lasso and Dantzig regularized estimators of the Riesz representer (RR) and derive convergence rates.

We plug-in regularized Riesz representer into debiased moment functions to do double/debiased machine learning (DML).

We give conditions for root-n consistency and asymptotic normality of estimator of the regression functional.

Debiasing via DML is based on zero derivative of the estimating equation with respect to each nonparametric component, as in Belloni, Chernozhukov, and Hansen (2014), Farrell (2015), and Robins et al. (2013).

Different than bias correcting the regression learner, as in Zhang and Zhang (2014), Belloni Chernozhukov, and Wang (2014), Belloni, Chernozhukov, and Kato (2015), Javanmard and Montanari (2014a,b; 2015), van de Geer et al. (2014), Neykov et al. (2015), Ren et al. (2015), Jankova and van de Geer (2015, 2016a,b), Bradic and Kolar (2017); Zhu and Bradic (2018).

Two debiasing approaches similar when the functional of interest is a coefficient of a partially linear model (as discussed in Chernozhukov et al., 2018a), but quite different for other functionals.

Differences analogous to those between root-n consistent semiparametric estimation and nonparametric estimation. Inference for a nonparametric regression requires bias correcting or undersmoothing the regression estimator while root-n consistent functional estimation can be based on learners that are not debiased or undersmoothed (see Newey 1994 for series regression).

DML does not require debiased first step learners, because remainder is second order when use debiased moment functions.

The functionals here are different than those analyzed in Cai and Guo (2017). We consider nonlinear functionals and linear functionals where the linear combination coefficients are estimated; not in Cai and Guo (2017). Also the $L_2$ continuity of linear functionals provides additional structure that we exploit, involving the RR, which is not exploited in Cai and Guo (2017).

Targeted maximum likelihood (van der Laan and Rubin, 2006) based on ML in van der Laan and Rose (2011); large sample theory in Luedtke and van der Laan (2016), Toth and van der Laan (2016), and Zheng et al. (2016). Here we provide DML learners via regularized RR, which are relatively simple to implement and analyze and directly target functionals of interest. No need to set it up as targeted maximum likelihood.

$L_2$ continuity is setting where root-n consistent, efficient semiparametric estimation of the object of interest is possible under sufficient regularity conditions; see Jankova and Van De Geer (2016a). Our results apply to different objects than considered by Ning and Liu (2017). We do not work with an explicit semiparametric form for the distribution of the data. Instead we learn functionals of a regression. Our estimators are DML of a functional of interest rather than estimating the efficient score for a parameter of interest in an explicit form of a semiparametric model.

We build on previous work on debiased estimating equations constructed by adding an influence function. Hasminskii and Ibragimov (1979) and Bickel and Ritov (1988) suggested such estimators for functionals of a density. Doubly robust estimating equations as in Robins, Rotnitzky, and Zhao (1995) and Robins and Rotnitzky (1995) have this structure. Newey, Hsieh, and Robins (1998, 2004) and Robins et al. (2008) further developed theory. For an affine functional the doubly robust learner we consider is given in Chernozhukov et al. (2016). We use simple and general regularity conditions in Chernozhukov et al. (2018b) that only require $L_2$ convergence of nonparametric learners.

The RR learners we consider are linear in a dictionary of functions. Such RR learners were previously used in Newey (1994) for asymptotic variance estimation and in Robins et al. (2007) for estimation of the inverse of the propensity score with missing data. Recently Newey and Robins (2017) considered such RR learning in efficient semiparametric estimation of linear regression functionals with low dimensional regressors. Hirshberg and Wager (2018) gave different RR estimators when the regression is restricted to a Donsker class. None of these works are about machine learning.

The Athey, Imbens, and Wager (2018) learner of the average treatment effect is based on a specific regression learner and on approximate balancing weights when the regression is linear and sparse. Our estimator allows for a wide variety of regression learners and does not restrict the regression to be sparse or linear. We do this via regularized RR learning that can also be interpreted as learning of balancing weights or inverse propensity scores.

Zhu and Bradic (2017) obtained root-n consistency for partially linear model when the regression function is dense. Our results apply to a wide class of affine and nonlinear functionals and allow slow rates for the regression learner.

Chernozhukov, Newey, and Robins (2018) have previously given the Dantzig learner of the RR. We innovate here by giving Lasso learner of the RR, allowing the functional to depend on nonregressors, by deriving convergence rates for both Lasso and Dantzig as learners of the true RR rather than a sparse approximation to it, allowing a general regression learner rather than just Dantzig, and by providing learners for nonlinear functionals. Also innovative relative to other previous work in the ways described in previous paragraphs.

# LEARNING AFFINE FUNCTIONALS

Start with object of interest

$$\theta_0 = E[m(W, \gamma_0)],\ m(W, \gamma) - m(W, 0) \text{ is linear in } \gamma.$$

$$\gamma_0(x) = E[Y|X = x],\ \text{or more generally } \gamma_0 \text{ is a mean square projection.}$$

$$E[m(W, \gamma) - m(W, 0)] \text{ is mean square continuous in } \gamma.$$

The last condition is necessary for finiteness of the semiparametric variance bound for $\theta_0$.

By Riesz representation theorem there is (unknown) $\alpha_0(x)$ such that for all $\gamma$ with finite second moment

$$E[m(W, \gamma) - m(W, 0)] = E[\alpha_0(X)\gamma(X)].$$

The Riesz representer $\alpha_0(X)$ is important in what follows.

Some examples:

–Bound on average consumer surplus:

$Y$ is share of income spent on a good, $X = (P_1, Z)$, $P_1$ is the good's price, $Z$ includes income $Z_1$, prices of other goods, and covariates.

Let $\check{p}_1 < \bar{p}_1$ be lower and upper prices, $\kappa$ a bound on the income effect, and $\omega(z)$ some weight function.

Object of interest is

$$\theta_0 = E[\omega(Z) \int_{\check{p}_1}^{\bar{p}_1} (\frac{Z_1}{u}) \gamma_0(u, Z) \exp(-\kappa[u - \check{p}_1]) du],$$

When consumer preferences statistically independent of $X$ and $\kappa$ is a lower (upper) bound on the income effect then $\theta_0$ is an upper (lower) bound on the weighted average over consumers of equivalent variation for a change in the price of the first good from $\check{p}_1$ to $\bar{p}_1$; see Hausman and Newey (2016).

For average surplus $\theta_0 = E[m(W, \gamma_0)]$,

$$m(w, \gamma) = \omega(z) \int_{\check{p}_1}^{\bar{p}_1} (z_1/u)\gamma(u, z) \exp(-\kappa[u - \check{p}_1])du$$

The RR is

$$\alpha_0(x) = f_0(p_1|z)^{-1}\omega(z)1(\check{p}_1 < p_1 < \bar{p}_1)(z_1/p_1) \exp(-\kappa[p_1 - \check{p}_1]),$$

where $f_0(p_1|z)$ is the conditional pdf of $P_1$ given $Z$.

–Average Treatment Effect: $X = (D, Z)$ and $\gamma_0(x) = \gamma_0(d, z)$, where $D \in \{0, 1\}$ is treatment indicator and $Z$ are covariates.

$$\theta_0 = E[\gamma_0(1, Z) - \gamma_0(0, Z)].$$

When treatment effect is mean independent of the treatment $D$ conditional on covariates $Z$ then $\theta_0$ is the average treatment effect, Rosenbaum and Rubin (1983).

$$m(w, \gamma) = \gamma(1, z) - \gamma(0, z),$$

$$\alpha_0(x) = d/\pi_0(z) - (1 - d)/[1 - \pi_0(z)], \pi_0(z) = \mathsf{Pr}(D = 1|Z = z).$$

$E[m(W, \gamma)]$ is mean square continuous when $E[1/\pi_0(Z)^2] < \infty$ and $E[1/\{1 - \pi_0(Z)\}^2] < \infty$.

–Expected Conditional Covariance:

Average conditional covariance between $Y$ and some other variable, say $W_1$.

Object of interest is

$$\theta_0 = E[Cov(Y, W_1|X)] = E[W_1\{Y - \gamma_0(X)\}].$$

Useful for analysis of covariance while controlling for regressors $X$.

An important component in the coefficient $\beta_0$ of $W_1$ for a partially linear regression of $Y$ on $W_1$ and unknown functions of $X$.

Differs from previous examples in $m(w, \gamma)$ depending on $w$ not just $x$.

$$m(w, \gamma) = w_1\{y - \gamma(x)\},\ \alpha_0(x) = -E[W_1|X = x].$$

Learn $\theta_0$ using the doubly robust moment function

$$\psi(w, \theta, \gamma, \alpha) = m(w, \gamma) - \theta + \alpha(x)[y - \gamma(x)],$$

given in Chernozhukov et al. (2016, "Locally Robust Semiparametric Estimation").

This function is doubly robust; by Riesz representation,

$$E[\psi(W, \theta_0, \gamma, \alpha)] = -E[\{\alpha(X) - \alpha_0(X)\}\{\gamma(X) - \gamma_0(X)\}],$$

for all $\gamma$ and $\alpha$; equal to zero if either $\gamma = \gamma_0$ or $\alpha = \alpha_0$.

$\psi(w, \theta, \gamma, \alpha)$ is debiased; functional derivative of $E[\psi(W, \theta_0, \gamma, \alpha)]$ with respect to $\alpha$ or $\gamma$ is zero at $\gamma_0$, $\alpha_0$.

A debiased/double machine learner $\hat{\theta}$ can be constructed from machine learners $\hat{\gamma}$ and $\hat{\alpha}$ by setting sample moment of $\psi(w, \theta, \hat{\gamma}, \hat{\alpha})$ to zero.

To avoid own observation bias and Donsker conditions for $\hat{\gamma}$ and $\hat{\alpha}$ we cross-fit, i.e. sample split.

Machine learners not known to satisfy Donsker conditions.

Suppose $W_i$ $(i = 1, ..., n)$ is i.i.d.. Let $I_\ell$, $(\ell = 1, ..., L)$ be a partition of the observation index set $\{1, ..., n\}$ into $L$ distinct subsets of about equal size.

Let $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$ be constructed from observations not in $I_\ell$.

Average over $i \in I_\ell$ while using other observations for $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$.

The estimator $\hat{\theta}$ is

$$\hat{\theta} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \{m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\}.$$

$$\hat{\theta} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \{ m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)] \}.$$

A variety of regression learners $\hat{\gamma}_\ell$ could be used here; Lasso, Dantzig, Neural Nets, Boosting.

Need $\hat{\gamma}_\ell$ with $n^{-d_\gamma}$ convergence rate for some $d_\gamma > 0$.

We give here Lasso learner $\hat{\alpha}_\ell$ and conditions for Dantzig learner $\hat{\alpha}_D$. These learners use $p \times 1$ dictionary of functions $b(x)$ where $p$ can be much bigger than $n$.

RR learners take form

$$\hat{\alpha}_L(x) = b(x)' \hat{\rho}_L, \ \hat{\alpha}_D(x) = b(x)' \hat{\rho}_D,$$

for $\hat{\rho}$ given below, dropping $\ell$ subscript for now.

$\hat{\rho}$ only uses $m(W, \gamma)$; don't need to know form of $\alpha_0(x)$.

Learn Riesz representer (RR) using $m(w,b) = (m(w,b_1),...,m(w,b_p))'$ and

$$M = E[m(W,b) - m(W,0)] = E[\alpha_0(X)b(X)].$$

Cross moments $M$ between the true, unknown RR $\alpha_0(x)$ and the dictionary $b(x)$ are equal to the expectation of a known vector of functions $m(w,b) - m(w,0)$.

Unbiased estimator of $M = E[\alpha_0(X)b(X)]$ is

$$\hat{M} = \frac{1}{n}\sum_{i=1}^{n}\{m(W_i,b) - m(W_i,0)\}.$$

Unbiased estimator of $G = E[b(X)b(X)']$ is

$$\hat{G} = \frac{1}{n}\sum_{i=1}^{n} b(X_i)b(X_i)'.$$

$$\hat{M} = \frac{1}{n}\sum_{i=1}^{n}\{m(W_i, b) - m(W_i, 0)\}, \hat{G} = \frac{1}{n}\sum_{i=1}^{n} b(X_i)b(X_i)'.$$

$\hat{M}$ analogous to $\sum_{i=1}^{n} Y_i b(X_i)/n$ in Lasso and Dantzig regression.

$$E[\sum_{i=1}^{n} Y_i b(X_i)/n] = E[\gamma_0(X)b(X)]$$

Analogous to Lasso and Dantzig regression,

$$E[\hat{M}] = E[m(W, b) - m(W, 0)] = E[\alpha_0(X)b(X)].$$

Lasso for RR replaces $\sum_{i=1}^{n} Y_i b(X_i)/n$ in regression objective function by $\hat{M}$ and drops $\sum_{i=1}^{n} Y_i^2/n$. Let $|\rho|_1 = \sum_{j=1}^{p} |\rho_j|$, $|\rho|_\infty = \mathsf{max}_{j \leq p} |\rho_j|$.

$$\hat{\rho}_L = \mathsf{arg}\min_{\rho}\{-2\hat{M}'\rho + \rho'\hat{G}\rho + 2r_L|\rho|_1\}.$$

Dantzig similar,

$$\hat{\rho}_D = \mathsf{arg}\min_{\rho} |\rho|_1 \, s.t. |\hat{M} - \hat{G}\rho|_\infty \leq \lambda_D,$$

$$\hat{\alpha}_L(x) = b(x)'\hat{\rho}_L, \quad \hat{\rho}_L = \arg\min_{\rho}\{-2\hat{M}'\rho + \rho'\hat{G}\rho + 2r_L|\rho|_1\}.$$

Does not use form of $\alpha_0(x)$; only depends on $m(w, \gamma)$ and $b(x)$.

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\}.\}.$$

Base asymptotic variance estimator on

$$\hat{\psi}_i = m(W_i, \hat{\gamma}_\ell) - \hat{\theta} + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)].$$

Asymptotic variance estimator is

$$\hat{V} = \frac{1}{(n-1)}\sum_{i=1}^{n}(\hat{\psi}_i - \bar{\psi})^2, \ \bar{\psi} = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i.$$

Give conditions for

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V), \quad \hat{V} \xrightarrow{p} V.$$

Based on conditions of "Locally Robust Semiparametric Estimation," Chernozhukov et al. (2018b). Let $\|A\| = \sqrt{E[A(W)^2]}$ be the mean-square norm.

Assumption A: $Var(Y|X)$ and $\alpha_0(X)$ bounded,

$$\int [m(W, \hat{\gamma}) - m(W, \gamma)]^2 F_0(dW) \xrightarrow{p} 0, \ \|\hat{\alpha}_{L\ell} - \alpha_0\| \xrightarrow{p} 0, \ \|\hat{\gamma}_\ell - \gamma_0\| \xrightarrow{p} 0.$$

Assumption B: For $\ell = 1, ..., L$,

$$\frac{1}{\sqrt{n}} \sum_{i \in I_\ell} [\hat{\alpha}_{L\ell}(X_i) - \alpha_0(X_i)][\hat{\gamma}_\ell(X_i) - \gamma_0(X_i)] \xrightarrow{p} 0.$$

Very simple and general. It is sufficient for Assumption B that

$$\sqrt{n}\|\hat{\alpha}_{L\ell} - \alpha_0\|\|\hat{\gamma}_\ell - \gamma_0\| \xrightarrow{p} 0,$$

i.e. product of $L2$ rates for $\hat{\alpha}$ and $\hat{\gamma}$ is faster than $1/\sqrt{n}$.

# Convergence Rates

Here derive $L2$ convergence rates for Lasso $\hat{\alpha}$; results for Dantzig similar. Give general conditions and primitive examples.

Assumption 1: *There is $B_n^b$ such that with probability one,*

$$max_{1 \leq j \leq p} |b_j(X)| \leq B_n^b.$$

This standard condition implies that

$$|\hat{G} - G|_\infty = O_p(\varepsilon_n^G),\ \varepsilon_n^G = (B_n^b)^2 \sqrt{\frac{\ln(p)}{n}}.$$

Assumption 2: *There is $\varepsilon_n^M$ such that*

$$|\hat{M} - M|_\infty \leq O_p(\varepsilon_n^M),$$

Allows for $m(w, \gamma)$ nonlinear in $\gamma$. For $|m(W, b)|_\infty \leq B_n^m A(W)$,

$$\varepsilon_n^M = B_n^m \sqrt{\frac{\ln(p)}{n}}$$

Explicitly treat bias. First type of bias condition is

Assumption 3: *There is* $\rho_n$ *such that* $\|\alpha_0 - b'\rho_n\|^2 = O(\max\{\varepsilon_n^G, \varepsilon_n^M\})$.

Sparsity plays no role in this condition. Holds when $\rho_n$ such that $\|\alpha_0 - b'\rho_n\|^2$ shrinks faster than some power of $1/p$ and $p$ grows than large enough power of $n$.

Let $B_n = |\rho_n|_1$ for $\rho_n$ from Assumption 3.

Theorem 1: *If Assumptions 1 - 3 are satisfied then for any* $r_L$ *such that* $\varepsilon_n^M + \varepsilon_n^G(1 + B_n) = o(r_L)$,

$$\|\alpha_0 - \hat{\alpha}_L\|^2 = O_p((1 + B_n)r_L), \;\; |\hat{\rho}_L|_1 = O_p(1 + B_n).$$

If $B_n \leq C$ and $\varepsilon_n^M = \sqrt{\ln(p)/n}$ then if $r_L$ is chosen so that $\sqrt{\ln(p)/n} = o(r_L)$,

$$\|\alpha_0 - \hat{\alpha}_L\|^2 = O_p(r_L).$$

The rate is about $(\ln(p)/n)^{1/4}$; no condition on $G = E[b(X)b(X)']$.

Faster rates under sparsity conditions. Let $\varepsilon_n = \max\{\varepsilon_n^G, \varepsilon_n^M\}$.

Assumption 4: There exists $C > 0$ and $\bar{\rho}$ with $\bar{s}$ nonzero elements such that

$$\left\|\alpha_0 - b'\bar{\rho}\right\|^2 \leq C\bar{s}\varepsilon_n^2$$

$\|\alpha_0 - b'\bar{\rho}\|^2$ is squared bias term; $\bar{s}\varepsilon_n^2$ is a variance like term; Assumption 4 specifies $\bar{s}$ large enough that squared bias is no larger than variance.

$b'\bar{\rho}$ is a sparse approximation because $\bar{\rho}$ has only $\bar{s}$ nonzero components

Make $\bar{s}$ as small as possible to get fastest rate in our results; sets variance equal to squared bias.

If $\alpha_0(x)$ is sparse then Assumption 4 holds with $\bar{s}$ being number of nonzero coefficients.

Approximately sparse $\alpha_0(x)$.

$(\tilde{b}_1(x), \tilde{b}_2(x), ...)$ and $C > 0$ with $|\tilde{b}_j(X)| \leq C$ and

$$\alpha_0(x) = \sum_{j=1}^{\infty} \tilde{b}_j(x)\tilde{\rho}_j, \quad \left|\tilde{\rho}_j\right| \leq Cj^{-d}.$$

For each $p$ and $\bar{s}$ small enough $(\tilde{b}_1(x), ..., \tilde{b}_{\bar{s}}(x))$ is a subvector of $b(x)$.

Choose $\bar{\rho}_k = \tilde{\rho}_j$ if $b_k(x) = \tilde{b}_j(x)$ for some $j \leq \bar{s}$ and otherwise let $\bar{\rho}_k = 0$. Then for some $\bar{C}$

$$\left|\alpha_0(X) - b(X)'\bar{\rho}\right| = \left|\sum_{j=s+1}^{\infty} \tilde{b}_j(X)\tilde{\rho}_j\right| \leq \bar{C} \sum_{j=\bar{s}+1}^{\infty} j^{-d} \leq \bar{C}(\bar{s})^{-d},$$

Here $\varepsilon_n = \sqrt{\ln(p)/n}$ the best $\bar{s}$ gives

$$\bar{s}\varepsilon_n^2 = \tilde{C}\left(\frac{\ln p}{n}\right)^{2d/(1+2d)}.$$

Also use sparse eigenvalue conditions.

Let $J = \{1, ..., p\}$, $J_\rho$ be the subset of $J$ with $\rho_j \neq 0$, and $J_\rho^c$ be the complement of $J_\rho$ in $J$.

Assumption 5: *$G$ is nonsingular and has largest eigenvalue uniformly bounded in $n$. Also there is $k > 3$ such such that*

$$\inf_{\{\delta:\delta\neq 0, \sum_{j\in\mathcal{J}_{\rho_L}^c} |\delta_j| \leq k \sum_{j\in\mathcal{J}_{\rho_L}} |\delta_j|\}} \frac{\delta' G \delta}{\sum_{j\in\mathcal{J}_{\rho_L}} \delta_j^2} > 0.$$

Population version of Bickel, Ritov, and Tsybakov (2009).

Let $\bar{B}_n = |\bar{\rho}|_1$ for $\bar{\rho}$ in Assumption 4.

Theorem 3: *If Assumptions 1, 2, 4, and 5 are satisfied and $\varepsilon_n^M + \varepsilon_n^G(1 + \bar{B}_n) = o(r_L)$ then $\|\alpha_0 - \hat{\alpha}_L\|^2 = O_p(\bar{s} r_L^2)$.*

Approximately sparse rate for $\|\hat{\alpha}_L - \alpha_0\|$ is $(\mathsf{ln}(p)/n)^{d/(1+2d)}$.

# Asymptotic Normality of $\hat{\theta}$

For approximately sparse case:

Theorem: If i) $Var(Y|X)$ and $\alpha_0(X)$ are bounded; ii) $\int[m(W,\gamma) - m(W,\gamma_0)]^2 F_0(dW)$ is continuous at $\gamma_0$ in $\|\hat{\gamma} - \gamma_0\|$; iii) Assumption 2 is satisfied; iv) there is $C$ and subgaussian $A(W)$ with $\left|b_j(X)\right| \leq C$ and $\left|m(W,b_j)\right| \leq A(W)$ for all $j$; iv) $\alpha_0(x)$ is approximately sparse with $|\rho_{0j}| \leq Cj^{-d}$, $r > 1$; v) $\|\hat{\gamma} - \gamma_0\| = O_p(n^{-d_\gamma})$,

$$vi) \ \frac{d}{2d+1} + d_\gamma > \frac{1}{2}.$$

Then for $r_L$ close enough to $\sqrt{\ln(p)/n}$, $\sqrt{\ln(p)/n} = o(r_L)$, $\psi_i = m(W_i,\gamma_0) - \theta_0 + \alpha_0(X_i)[Y_i - \gamma_0(X_i)]$,

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i + o_p(1), \\ \hat{V} \xrightarrow{p} V &= E[\psi_i^2]. \end{aligned}$$

Average Treatment Effect:

$b(x) = [dq(z)', (1-d)q(z)']'$ where $q(z)$ is covariate dictionary.

Note that $m(w, b) = [q(z)', -q(z)']'$, so

$$\hat{M} = \begin{pmatrix} \bar{q} \\ -\bar{q} \end{pmatrix}, \bar{q} = \frac{1}{n}\sum_i q(Z_i).$$

Let $\hat{\rho}^d$ be coefficients of $dq(z)$ and $\hat{\rho}^{1-d}$ coefficients of $(1-d)q(z)$.

RR learner

$$\hat{\alpha}(X_i) = D_i\hat{\omega}_i^d + (1-D_i)\hat{\omega}_i^{1-d},\ \hat{\omega}_i^d = q(Z_i)'\hat{\rho}^d,\ \hat{\omega}_i^{1-d} = q(Z_i)'\hat{\rho}^{1-d},$$

$\hat{\omega}_i^d$ and $\hat{\omega}_i^{1-d}$ sum to one but need not be nonnegative.

Lasso enforces approximate dictionary balance:

$$\left|\frac{1}{n}\sum_i q_j(Z_i)[1 - D_i\hat{\omega}_i^d]\right| \leq r_L, \left|\frac{1}{n}\sum_i q_j(Z_i)[1 + (1-D_i)\omega_i^{1-d}]\right| \leq r_L.$$

Lasso RR implies approximate balance

$$\left|\frac{1}{n}\sum_i q_j(Z_i)[1 - D_i\hat{\omega}_i^d]\right| \leq r_L, \left|\frac{1}{n}\sum_i q_j(Z_i)[1 + (1 - D_i)\omega_i^{1-d}]\right| \leq r_L.$$

Estimator is

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{\hat{\gamma}_\ell(1, Z_i) - \hat{\gamma}_\ell(0, Z_i) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\}.$$

Get asymptotic normality and consistent variance estimator for ANY regression learner $\hat{\gamma}$ (e.g. boosting, neural net, ...) when $\|\hat{\gamma} - \gamma_0\| = O_p(n^{-d_\gamma})$, $\alpha_0(x)$ approximately sparse with $\left|\tilde{\rho}_j\right| \leq Cj^{-d}$, $r_L = \tilde{C}\,\mathsf{ln}(n)\sqrt{\mathsf{ln}(p)/n}$,

$$\frac{d}{2d+1} + d_\gamma > \frac{1}{2}.$$

# Nonlinear Functionals

DML of

$$\theta_0 = E[m(W, \gamma_0)],$$

for nonlinear $m(w, \gamma)$.

Use linearization and different $\hat{M}$ that depends on $\hat{\gamma}$.

Give $\hat{M}$, rates for $\hat{\alpha}$, and conditions for root-n consistency and asymptotic normality of nonlinear functionals.

Again use a RR.

Here the RR is for linearization of the functional.

Suppose $m(w, \gamma)$ has Gateaux derivative $D(w, \zeta, \gamma)$ where $\zeta$ represents a deviation from $\gamma$ and $D(w, \zeta, \gamma)$ is linear in $\zeta$.

That is

$$\frac{d}{d\tau} m(w, \gamma + \tau\zeta)\Big|_{\tau=0} = D(w, \zeta, \gamma),$$

where $\tau$ is a scalar.

Assume that $E[D(W, \gamma, \gamma_0)]$ is a linear mean square continuous functional of $\gamma$ so a RR $\alpha_0(x)$ with

$$E[D(W, \gamma, \gamma_0)] = E[\alpha_0(X)\gamma(X)],$$

for all $\gamma(x)$ with finite second moment.

$$E[D(W,\gamma,\gamma_0)] = E[\alpha_0(X)\gamma(X)],$$

Analogous to previous with $m(w,\gamma) - m(w,0)$ replaced by first order approximation $D(w,\gamma,\gamma_0)$.

The Riesz representation implies that for

$D(w,b,\gamma_0) = (D(w,b_1,\gamma_0),...,D(w,b_p,\gamma_0))'$, we have

$$M = E[D(W,b,\gamma_0)] = E[\alpha_0(X)b(X)].$$

A learner $\hat{\theta}$ can be constructed from learner $\hat{\alpha}_\ell(x)$ of the RR $\alpha_0(x)$ and a learner $\hat{\gamma}_\ell(x)$ of $E[Y|X=x]$ exactly before. Not doubly robust but is "locally robust" with zero first-order effect of ML.

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{m(W_i,\hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\}.$$

Zero first order bias because $\alpha_0(x)[y-\gamma_0(x)]$ is influence function for $E[m(W,\gamma)]$; Newey (1994), Chernozhukov et al. (2016).

Use different estimator $\hat{\alpha}_\ell(x)$ based on a different $\hat{M}_\ell$.

Let $\hat{\gamma}_{\ell,\ell'}$ learn $E[Y|X]$ observations not in either $I_\ell$ or $I_{\ell'}$

$$\begin{aligned}
\hat{M}_\ell &= (\hat{M}_{\ell 1}, ..., \hat{M}_{\ell p})', \\
\hat{M}_{\ell j} &= \frac{d}{d\tau}\left(\frac{1}{n - n_\ell}\right) \sum_{\ell' \neq \ell} \sum_{i \in I_{\ell'}} m(W_i, \hat{\gamma}_{\ell,\ell'} + \tau b_j) \\
&= \left(\frac{1}{n - n_\ell}\right) \sum_{\ell' \neq \ell} \sum_{i \in I_{\ell'}} D(W_i, b_j, \hat{\gamma}_{\ell,\ell'}).
\end{aligned}$$

Further sample splitting that allows for $p$ large.

To get convergence rate for $\hat{M}$.

Assumption: *There is* $\varepsilon > 0$, $B_n^D$, $B_n^\Delta$ *and sub-Gaussian* $A(W)$ *such that for all* $\gamma$ *with* $\|\gamma - \gamma_0\| \leq \varepsilon$, i)

$$\max_j |D(W, b_j, \gamma)| \leq B_n^D A(W),$$

*ii)* $\max_j |E[D(W, b_j, \gamma) - D(W, b_j, \gamma_0)]| \leq B_n^\Delta \|\gamma - \gamma_0\|$.

Lemma: *If Assumption and* $\|\hat{\gamma} - \gamma_0\| = O_p(\varepsilon_n^\gamma)$ then

$$|\hat{M} - M|_\infty = O_p(\varepsilon_n^D),\ \varepsilon_n^D = (B_n^D \sqrt{\frac{\ln(p)}{n}} + B_n^\Delta \varepsilon_n^\gamma).$$

Also

Assumption: *There are* $\varepsilon$, $C > 0$ *such that for all* $\gamma$ *with* $\|\gamma - \gamma_0\| \leq \varepsilon$,

$$|E[m(W, \gamma) - m(W, \gamma_0) - D(W, \gamma - \gamma_0, \gamma_0)]| \leq C\|\gamma - \gamma_0\|^2.$$

Stronger than $E[m(W, \gamma)]$ being an $L_2$ differentiable functional.

Under conditions in paper including $d_\gamma > 1/4$ (regression estimator converges faster than $n^{-1/4}$)

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{m(W_i, \gamma_0) - \theta_0 + \alpha_0(X_i)[Y_i - \gamma_0(X_i)]\} + o_p(1).$$

# APPLICATION TO BOUND ON AVERAGE SURPLUS FROM TAXING SODA

Empirical application to demand analysis in Chernozhukov, Hausman, Newey (2018, "Demand Analysis with Many Prices," in preparation). Use Nielsen scanner data. Metropolitan Houston.

There are 1483 households and up to 36 months in the sample. Data is from shopping trips to a number of retailers, based on what the individuals purchased. We calculated total expenditures to be the total amount in purchases during the month that the individuals reported. The data includes race, marital status, household composition, as well as male and female employment status. We use family size; found other family composition covariates not important.

Included prices for 15 goods. Cross price effects are small; suggest approximate sparsity might be reasonable way to think about this.

Regression of soda expenditure share on ln(prices), ln(total expenditure), powers of logs of own price and total expenditure up to order 4, quadratic in other prices, about 140 regressors.

Object of interest is

$$\theta_0 = E[\omega(Z) \int_{\check{p}_1}^{\bar{p}_1} (\frac{Z_1}{u}) \gamma_0(u, Z) \exp(-\kappa[u - \check{p}_1]) du],$$

If $\kappa$ is a lower (upper) bound on the income effect then $\theta_0$ is an upper (lower) bound on the weighted average over consumers of equivalent variation for a change in the price of the first good from $\check{p}_1$ to $\bar{p}_1$;

Use $\omega(Z)$ for high and low total expenditure groups.

Allow endogeneity of total expenditure using a control function.

Instrument using earnings and use residual from earnings regression as control variable.

# AEV Bounds Estimates

Income effect lower bound taken to be zero and upper bound to be 20 times the maximum of the income effect over .1, .25, .5, .75, .9 quantile effect regressions.

Consider 10 percent tax at mean price in data.

## Equivalent Variation Bounds Over Lowest Expenditure Quartile

| | | |
|---|---|---|
| Lowest Quartile of Expenditure | 4.66 | .30 |
| | | – |

## Equivalent Variation Bounds Over Highest Expenditure Quartile

| | | |
|---|---|---|
| Highest Quartile of Expenditure | 23.29 | 1.75 |
| | | – |