Applied Causal Inference Powered by ML and AI

Victor Chernozhukov* Christian Hansen† Nathan Kallus‡
Martin Spindler§ Vasilis Syrgkanis¶

March 5, 2025

Publisher: Online
Version 0.1.1

* MIT
† Chicago Booth
‡ Cornell University
§ Hamburg University
¶ Stanford University

Statistical Inference on Predictive and Causal Effects in High-Dimensional Linear Regression Models

4

"The partial trend regression method can never, indeed, achieve anything which the individual trend method cannot, because the two methods lead by definition to identically the same results."
(An in-words restatement of the FWL theorem.)

– Ragnar Frisch and Frederick V. Waugh [1].

Here we discuss inference on predictive effects using Double Lasso methods, where we use Lasso (at least) twice to residualize outcomes and a target covariate of interest whose predictive effect we'd like to infer. Double Lasso methods rely on the approximate sparsity of the best linear predictors for the outcome and for the target covariate. The resulting estimator concentrates in a $1/\sqrt{n}$ neighborhood of the true value and is approximately Gaussian, enabling the construction of confidence bands. We explain the low bias property of the Double Lasso method using Neyman orthogonality, and isolate the latter as a critical property for further generalizations.

4.1 Introduction

We recall the predictive effect question:[1]

- How does the predicted value of Y change if a regressor D increases by a unit, while other regressors W remain unchanged?

As before, we denote the set of regressors as $X = (D, W)$. In Chapter 1, we discussed how we could use the population regression coefficient corresponding to the variable D, denoted α, to answer this question. We also discussed how to estimate this effect and construct confidence intervals with regression. Now we turn to estimation and construction of confidence intervals for α in the high-dimensional setting, using the tools we developed in Chapter 3.

Here we focus on using Lasso methods. We can use other penalized methods with the caveat that theoretical guarantees are not available unless we perform additional data splitting. We will discuss the use of data splitting and more general machine learning methods in detail when we introduce "double machine learning" or "debiased machine learning" in Chapter 9.

1: We discuss assumptions and modeling frameworks under which the predictive effect question has a causal interpretation in detail in Chapter 5 thrChapter 11. Under the framework developed in those chapters, the tools in this chapter offer one approach to performing statistical inference for causal effects. Here, we simply note that we may be interested in providing statistical inference for predictive effects regardless of whether they have a causal interpretation.

4.2 Inference with Double Lasso

Inference on One Coefficient

The key to inference will be the application of Frisch-Waugh-Lovell partialling-out. Consider the simple predictive model:

$$Y = \alpha D + \beta' W + \epsilon, \tag{4.2.1}$$

where D is the target regressor and W consists of p controls. After partialling-out W,

$$\tilde{Y} = \alpha \tilde{D} + \epsilon, \quad \mathrm{E}[\epsilon \tilde{D}] = 0, \tag{4.2.2}$$

where the variables with tildes are residuals retrieved from taking out the linear effect of W (practically, via linear regression):

$$\tilde{Y} = Y - \gamma_{YW}' W, \quad \gamma_{YW} \in \arg\min_{\gamma \in \mathbb{R}^p} \mathrm{E}[(Y - \gamma' W)^2],$$

$$\tilde{D} = D - \gamma_{DW}' W, \quad \gamma_{DW} \in \arg\min_{\gamma \in \mathbb{R}^p} \mathrm{E}[(D - \gamma' W)^2].$$

α can then be recovered from population linear regression of $\tilde{Y}$ on $\tilde{D}$:

$$\alpha = \arg\min_{a \in \mathbb{R}} \mathrm{E}[(\tilde{Y} - a\tilde{D})^2] = (\mathrm{E}[\tilde{D}^2])^{-1}\mathrm{E}[\tilde{D}\tilde{Y}].$$

Note also that $a = \alpha$ solves the moment equation:

$$\mathrm{E}[(\tilde{Y} - a\tilde{D})\tilde{D}] = 0.$$

We now consider estimation of α in a high-dimensional setting. For estimation purposes, we maintain that we have a random sample $\{(Y_i, X_i)\}_{i=1}^n$ where $X_i = (D_i, W_i)$.

To estimate α, we will mimic the partialling-out procedure in the population in the sample. In Chapter 1, where p/n was small, we employed ordinary least squares as the prediction method in the partialling-out steps. We are now considering cases where p/n is not small, and we instead employ Lasso-based methods in the partialling-out steps.

The estimation procedure for a target parameter α in a high-dimensional linear model setting can be summarized as follows:

The Double Lasso procedure:

1. We run Lasso regressions of Y_i on W_i and D_i on W_i

$$\hat{\gamma}_{YW} = \arg\min_{\gamma \in \mathbb{R}^p} \quad \sum_i (Y_i - \gamma' W_i)^2 + \lambda_1 \sum_j \hat{\psi}_j^Y |\gamma_j|,$$

$$\hat{\gamma}_{DW} = \arg\min_{\gamma \in \mathbb{R}^p} \quad \sum_i (D_i - \gamma' W_i)^2 + \lambda_2 \sum_j \hat{\psi}_j^D |\gamma_j|,$$

   and obtain the resulting residuals:

$$\check{Y}_i = Y_i - \hat{\gamma}'_{YW} W_i,$$

$$\check{D}_i = D_i - \hat{\gamma}'_{DW} W_i.$$

   In place of Lasso, we can use Post-Lasso or other Lasso relatives (the Dantzig selector, square-root Lasso, and others).

2. We run the least squares regression of $\check{Y}_i$ on $\check{D}_i$ to

obtain the estimator $\hat{\alpha}$:

$$\begin{aligned}\hat{\alpha} &= \arg\min_{a\in\mathbb{R}} \mathbb{E}_n[(\check{Y} - a\check{D})^2] \\ &= (\mathbb{E}_n[\check{D}^2])^{-1}\mathbb{E}_n[\check{D}\check{Y}].\end{aligned} \tag{4.2.3}$$

We can use standard results from this regression, ignoring that the input variables were previously estimated, to perform inference about the predictive effect, α.

Good performance of the Double Lasso procedure relies on approximate sparsity of the population regression coefficients γ_{YW} and γ_{DW}, with a sufficiently high speed of decrease in the sorted coefficients and on careful choice of the Lasso tuning parameters. For approximate sparsity, we will impose that the sorted coefficients satisfy

$$|\gamma_{YW}|_{(j)} \leq Aj^{-a} \text{ and } \quad |\gamma_{DW}|_{(j)} \leq Aj^{-a}$$

for $a > 1$ and $j = 1, \ldots, p$.[2] Under these sparsity conditions, we can use the plug-in rule outlined in Chapter 3 for choosing λ_1 and λ_2. Importantly, using these tuning parameters theoretically guarantees that we produce high quality prediction rules for D and Y while simultaneously avoiding overfitting under approximate sparsity. Absent these guarantees, we cannot theoretically ensure that first step estimation of $\check{D}$ and $\check{Y}$ does not have first-order impacts on the final estimator $\hat{\alpha}$. Practically, we have found that Lasso with penalty parameter selected via cross-validation can perform poorly in simulations in moderately sized samples. We return to this issue in Chapter 9 where we discuss a method that allows the use of complex machine learners, including Lasso and other regularized estimators, and data-driven tuning (e.g. cross-validation).

2: Note that in this case the effective dimension s of the problem is $s \approx A^{1/a}n^{1/2a} \ll n^{1/2}$. Intuitively, the effective number of non-zero coefficients grows slower than $\sqrt{n}$.

The following theorem can be shown for the Double Lasso procedure:

Theorem 4.2.1 (Adaptive Inference with Double Lasso in High-Dimensional Regression) *Under the stated approximate sparsity, the conditions required for Theorem 3.2.1 (e.g. restricted isometry), and additional regularity conditions, the estimation error in $\check{D}_i$ and $\check{Y}_i$ has no first order effect on $\hat{\alpha}$, and*

$$\sqrt{n}(\hat{\alpha} - \alpha) \approx \sqrt{n}\mathbb{E}_n[\tilde{D}\epsilon]/\mathbb{E}_n[\tilde{D}^2] \overset{a}{\sim} N(0, \mathsf{V}),$$

where

$$\mathsf{V} = (\mathrm{E}[\tilde{D}^2])^{-1}\mathrm{E}[\tilde{D}^2\epsilon^2](\mathrm{E}[\tilde{D}^2])^{-1}.$$

The above statement means that $\hat{\alpha}$ concentrates in a $\sqrt{\mathsf{V}/n}$-neighborhood of α, with deviations controlled by the normal law. Observe that the approximate behavior of the Double Lasso estimator is the same as the approximate behavior of the least squares estimator in low-dimensional models; see Theorem 1.3.2 in Chapter 1.

Just like in the low-dimensional case, we can use these results to construct a confidence interval for α. The standard error of $\hat{\alpha}$ is

$$\sqrt{\hat{\mathsf{V}}/n},$$

where $\hat{\mathsf{V}}$ is a plug-in estimator of V. The result implies, for example, that the interval

$$[\hat{\alpha} \pm 1.96\sqrt{\hat{\mathsf{V}}/n}]$$

covers α about 95% of the time.

Application to Testing the Convergence Hypothesis

R Notebook on Double Lasso for Growth Convergence and Python Notebook on Double Lasso for Growth Convergence provides code for the convergence hypothesis example.

We provide an empirical example of partialling-out with Lasso to estimate the regression coefficient α in the high-dimensional linear regression model:

$$Y = \alpha D + \beta' W + \epsilon.$$

In this example, we are interested in how economic growth rates (Y) are related to the initial wealth levels in each country (D) controlling for a country's institutional, educational, and other similar characteristics (W).

The relationship is captured by α, the "speed of convergence/-divergence," which predicts the speed at which poor countries catch up ($\alpha < 0$) or fall behind ($\alpha > 0$) rich countries, after controlling for W. Here, we are interested in understanding if poor countries grow faster than rich countries, controlling for educational and other characteristics. In other words, is the speed of convergence negative: Is $\alpha < 0$?

$\alpha < 0$ corresponds to the Convergence Hypothesis predicted by the Solow growth model. Robert M. Solow is a world-renowned MIT economist who won the Nobel Prize in Economics in 1987.

In our data, the outcome (Y) is the realized annual growth rate of a country's wealth (Gross Domestic Product per capita). The target regressor (D) is the initial level of the country's

wealth. The controls (W) include measures of education levels, quality of institutions, trade openness, and political stability in the country. The sample, which is based on the Barro-Lee data set [2], contains 90 countries and about 60 controls. Thus $p \approx 60, n = 90$ and p/n is not small. We expect the least squares method to provide a poor/ noisy estimate of α. We expect the method based on partialling-out with Lasso to provide a high-quality estimate of α.

	Estimate	Std. Error	95% CI
OLS	-0.009	0.032	[-0.073, 0.054]
Double Lasso	-0.045	0.018	[-0.080, -0.010]

Table 4.1: Estimates for the convergence coefficient. We report specification robust standard errors with finite sample correction, i.e., "HC1."

Least squares provides a rather noisy estimate of convergence speed, which does not allow drawing strong conclusions about the convergence hypothesis. For example, the 95% confidence interval is wide and includes both positive and negative values. Given that p/n is not small in this example, we should also be highly skeptical of the OLS results and especially the standard error. For example, [3] show that conventional robust standard errors are not even consistent in linear models when p/n is not small. In sharp contrast, Double Lasso provides a precise estimate for which we can obtain theoretically justified inferential statements even though p/n is not close to 0. The Lasso-based point estimate is -4.5% and the 95% confidence interval for the (annual) convergence rate is -8% to -1%. This empirical evidence is consistent with the conditional convergence hypothesis.

4.3 Why Partialling-out Works: Neyman Orthogonality

Neyman Orthogonality

In the Double Lasso approach, α is the target parameter and η are *nuisance* projection *parameters* with true value

$$\eta^o = (\gamma'_{DW}, \gamma'_{YW})'.$$

As the learned value $\hat{\alpha}$ of α depends on the values of the nuisance parameters, it is useful to explicitly consider the dependence of $\hat{\alpha}$ on the nuisance parameters:

$$\hat{\alpha}(\eta).$$

For the majority of the estimation processes we will describe in this book, we can construct a population analogue

$$\alpha(\eta)$$

of the estimator $\hat{\alpha}(\eta)$, such that the in-sample estimation procedure converges to it, in a formal sense.

For instance, the Double Lasso process constructs the residuals

$$\check{Y}_i(\eta) = Y_i - \eta_1' W_i, \quad \check{D}_i(\eta) = D_i - \eta_2' W_i$$

and then obtains $\hat{\alpha}(\eta)$ as the solution to the empirical estimating equation

$$\widehat{\mathsf{M}}(a, \eta) := \mathbb{E}_n[(\check{Y}(\eta) - a\check{D}(\eta))\check{D}(\eta)] = 0.$$

This process implicitly defines the function $\hat{\alpha}(\eta)$. We can think of the population analog of this process, where we construct the residuals

$$\tilde{Y}(\eta) = Y - \eta_1' W, \quad \tilde{D}(\eta) = D - \eta_2' W$$

and solve the population moment equation

$$\mathsf{M}(a, \eta) := \mathrm{E}[(\tilde{Y}(\eta) - a\tilde{D}(\eta))\tilde{D}(\eta)] = 0, \tag{4.3.1}$$

which again implicitly defines the function $\alpha(\eta)$.

The main idea of the Double Lasso approach is that, in the population limit, it corresponds to a procedure for learning the target parameter α that is first-order insensitive to local perturbations of the nuisance parameters around their true values, η^o:

$$\partial_\eta \alpha(\eta^o) = 0. \tag{4.3.2}$$

Formally, we use ∂_η to denote the Gateaux derivative. See Remark 9.4.2 in Chapter 9 for more details.

We will call the local insensitivity of target parameters to nuisance parameters as in (4.3.2) Neyman orthogonality of the estimation process.

Neyman orthogonality is important for providing high-quality estimation and inference, especially in high-dimensional settings. In high-dimensional settings, we use regularization procedures to estimate the nuisance parameters as solutions to suitable prediction problems. The use of regularization generally results in bias, and we may heuristically view using regularized estimates of nuisance parameters as plugging in estimates of these parameters that are close to, but not exactly equal to, the true values of the nuisance parameters η^o. Neyman

orthogonality, which guarantees that the target parameter is locally insensitive to perturbations of the nuisance parameters around their true values, then ensures that this bias does not transmit to the estimation of the target parameter, at least to the first order.

Let us prove the claim $\partial_\eta \alpha(\eta^o) = 0$ for the Double Lasso process. Since the function $\alpha(\eta)$ is implicitly defined as the solution to the equation $\mathsf{M}(a, \eta) = 0$, by the implicit function theorem and letting $\alpha = \alpha(\eta^o)$:

$$\partial_\eta \alpha(\eta^o) = -\partial_a \mathsf{M}(\alpha, \eta^o)^{-1} \partial_\eta \mathsf{M}(\alpha, \eta^o).$$

Here

$$\partial_\eta \mathsf{M}(\alpha, \eta^o)$$

consists of two components

$$\partial_{\eta_1} \mathsf{M}(\alpha, \eta^o) = \mathrm{E}[W\tilde{D}(\eta^o)] = \mathrm{E}[W(D - \gamma'_{DW} W)] = 0$$

and

$$\begin{aligned}\partial_{\eta_2} \mathsf{M}(\alpha, \eta^o) = & -\mathrm{E}[W\tilde{Y}(\eta^o)] + 2\mathrm{E}[\alpha W\tilde{D}(\eta^o)] \\ = & -\mathrm{E}[W(Y - \gamma'_{YW} W)] + 2\mathrm{E}[\alpha W(D - \gamma'_{DW} W)] = 0.\end{aligned}$$

We summarize the discussion as follows:

Neyman Orthogonality. The parameter of interest α that depends on nuisance parameters η with true value η^o is Neyman orthogonal with respect to these parameters if

$$\partial_\eta \alpha(\eta^o) = 0.$$

If the parameter α is defined as a root in a of the equation $\mathsf{M}(a, \eta) = 0$, which depends on the nuisance parameters η with true value η^o, then the equation is Neyman orthogonal if

$$\partial_\eta \mathsf{M}(\alpha, \eta^o) = 0.$$

The principle is applicable to problems outside the high-dimensional linear model problem considered in this chapter.

What Happens if We Don't Have Neyman Orthogonality?

If we don't have Neyman orthogonality, we should not expect to get high-quality estimates of the target parameters. For example, a seemingly sensible approach that one might consider for statistical inference in the high-dimensional linear model context is as follows:

> **(Invalid) Single Selection/Naive Method.**
> In this invalid method, one applies Lasso regression of Y on D and W to select relevant covariates W_Y, in addition to the covariate of interest, then refits the model by least squares of Y on D and W_Y. Inference for the target parameter is then carried out using conventional inference based on the latter regression.

Despite its simplicity and seeming intuitive appeal, the approach outlined above is not a valid approach if the goal is to perform inference on α. It is a fine approach if the goal is solely the prediction of the outcome, but it can result in very misleading conclusions about the parameter of interest α, as we demonstrate in Example 4.3.1 below.

The naive approach outlined above relies on the moment condition

$$\mathsf{M}(a, b) = \mathrm{E}[(Y - aD - b'W)D] = 0.$$

When $b = \beta$, this moment condition is satisfied by the true value, $a = \alpha$. In this case, it coincides with the classical moment condition for α underlying low-dimensional ordinary least squares which sets prediction errors to be orthogonal to each predictor variable.

However, this moment condition does not exhibit Neyman orthogonality since

$$\partial_b \mathsf{M}(\alpha, \beta) = \mathrm{E}[DW] \neq 0$$

unless D is orthogonal to W.[3] Because $\mathsf{M}(a, b)$ is not Neyman orthogonal, the bias and the slower than parametric rate of convergence,

$$\sqrt{s \log(p \vee n)/n},$$

of our estimate of $\beta'W$ will transmit to bias and slower than $\sqrt{n}$ convergence in estimates of α provided by solving the empirical analog of $\mathsf{M}(a, b)$. The "Single Selection" procedure outlined

3: In "pure" RCTs where treatment is assigned independently of everything, D's are orthogonal to W, after de-meaning D, so Neyman orthogonality automatically holds in this setting.

above exactly provides the solution to this moment condition. Consequently, while this naive procedure provides an estimator of α that will approach the true value in large samples (at a slower than $\sqrt{n}$-rate), the bias of the estimator converges too slowly for standard inference methods to provide reliable inference.

We can set up a simulation experiment to verify that this naive approach provides low-quality estimates for α.

Example 4.3.1 In R Notebook with Experiment on Orthogonal vs Non-Orthogonal Learning and Python Notebook with Experiment on Orthogonal vs Non-Orthogonal Learning, we compare the performance of the naive and orthogonal methods in a computational experiment where $p = n = 100$, $\beta_j = 1/j^2$, $(\gamma_{DW})_j = 1/j^2$, and

$$Y = 1 \cdot D + \beta' W + \varepsilon_Y, \quad W \sim N(0, I), \quad \varepsilon_Y \sim N(0, 1)$$

$$D = \gamma'_{DW} W + \tilde{D}, \quad \tilde{D} \sim N(0, 1)/4.$$

From the histograms shown in Figure 4.1, we see that the naive estimator is heavily biased, as expected from the lack of Neyman orthogonality in its estimation strategy. We also see that the Double Lasso estimator, which is based on principled partialling-out such that Neyman orthogonality is satisfied, is approximately unbiased and Gaussian.

The reason that the naive estimator does not perform well is that it only selects controls that are strong predictors of the outcome, thereby omitting weak predictors of the outcome. However, weak predictors of the outcome could still be strong predictors of D, in which case dropping these controls results in a strong omitted variable bias. In contrast, the orthogonal approach solves two prediction problems – one to predict Y and another to predict D – and finds controls that are relevant for either. The resulting residuals are therefore approximately "de-confounded."

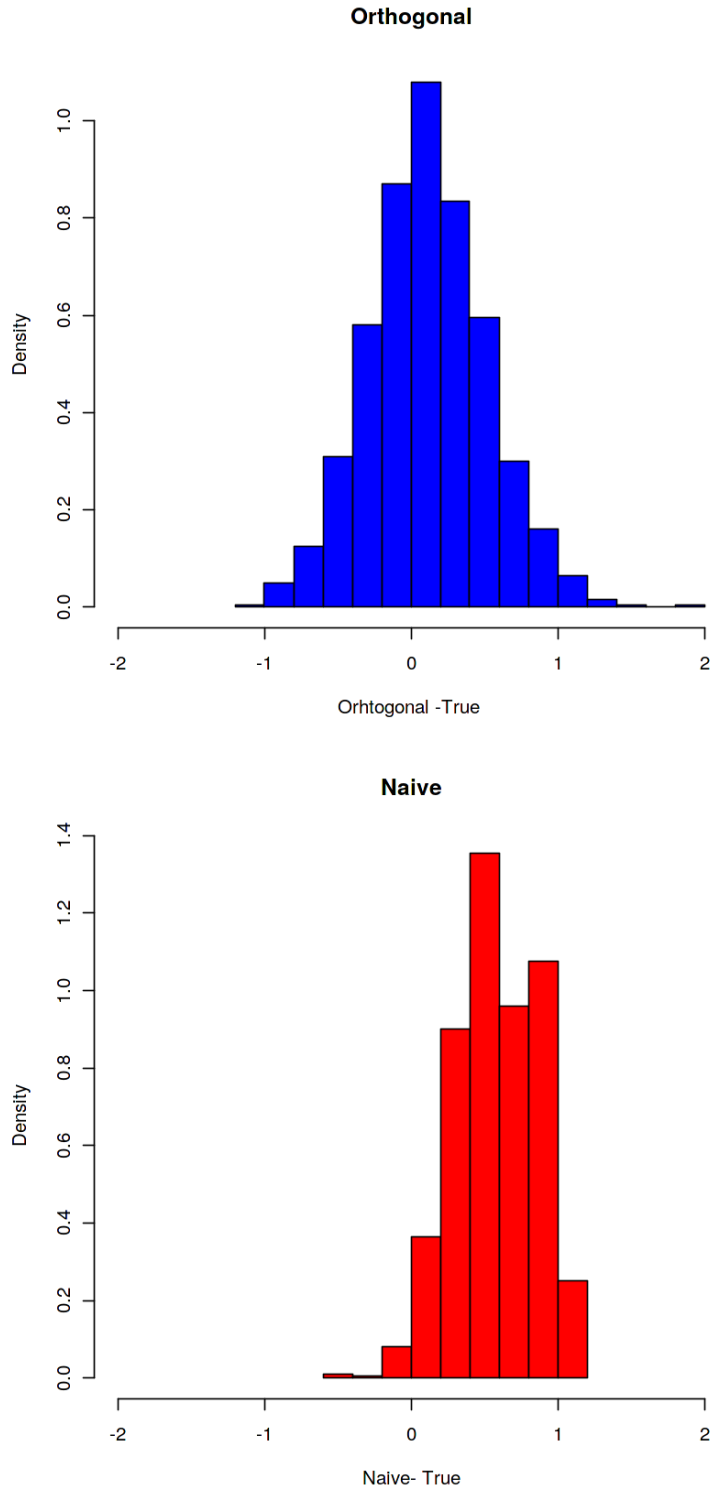

Figure 4.1: Top Panel: Simulated distribution of the orthogonal estimator centered around the true value. **Bottom Panel:** Simulated distribution of the naive (single-selection) non-orthogonal estimator centered around the true value.

4.4 Inference on Many Coefficients

If we are interested in more than one coefficient, we can repeat the one-by-one Double Lasso procedure for each of the coefficients of interest and obtain valid estimation and inference on each component under regularity conditions.

We consider the model

$$\underbrace{Y}_{\text{Outcome}} = \underbrace{\sum_{\ell=1}^{p_1} \alpha_\ell D_\ell}_{\text{Target Predictors}} + \underbrace{\sum_{j=1}^{p_2} \beta_j \bar{W}_j}_{\text{Controls}} + \epsilon,$$

where we use D_ℓ for $\ell = 1, ..., p_1$ to denote the predictors of interest and $\bar{W}_j$ for $j = 1, ..., p_2$ to denote other predictors in the model. Here, both the number of predictors of interest, p_1, and the number of additional variables, p_2, can both be very large.

There are at least three motivations for considering many coefficients of interest:

- there can be multiple policies whose predictive effect we would like to infer;
- we can be interested in heterogeneous predictive effects across pre-specified groups;
- we can be interested in nonlinear effects of policies.

This setting encompasses examples where we are interested in *heterogeneous effects*, where $D_\ell's$ are generated as

$$D_\ell = D_0 \bar{X}_\ell, \quad \ell = 1, ..., p_1,$$

where D_0 is a base variable of interest – for example, a treatment indicator, a price, or a group indicator – and $(\bar{X}_\ell)_{\ell=1}^{p_1}$ are known transformations of controls $\bar{W}$ – for example, various subgroup indicators.

The setting also encompasses cases where *nonlinear effects* are of interest. For example, we could consider D_ℓ's generated as polynomial transformations of a multi-valued base variable, such as a price:

$$D_\ell = D_0^\ell, \quad \ell = 1, ..., p_1.$$

We could further interact these transformations with other variables to study nonlinear heterogeneous effects.

One by One Double Lasso for Many Target Parameters. For each $\ell = 1, \ldots, p_1$, we apply the one-by-one Double Lasso procedure for estimation and inference on the coefficient α_ℓ in the model

$$Y = \alpha_l D_\ell + \gamma_\ell' W_\ell + \epsilon, \quad W_\ell = ((D_k)_{k \neq \ell}', \bar{W}')'.$$

Under approximate sparsity conditions, the Double Lasso method provides a high-quality estimate $\hat{\alpha} = (\hat{\alpha}_\ell)_{\ell=1}^{p_1}$ of $\alpha = (\alpha_\ell)_{\ell=1}^{p_1}$ that is approximately Gaussian. We can thus easily construct individual confidence intervals or even joint confidence bands. Under regularity conditions, these results allow for simultaneous inference on $p_1 > n$ coefficients.

Theorem 4.4.1 (Double Lasso for Many Coefficients) *Under regularity conditions including approximate sparsity as in Definition 3.1.1 with parameters (A, a) with $a > 1$ in all partialling out steps and provided $(\log p_1)^5/n$ is small, we have the adaptivity property,*

$$\sqrt{\log p_1} \max_{\ell \leq p_1} \left| \sqrt{n}(\hat{\alpha}_\ell - \alpha_\ell) - (\mathbb{E}_n[\tilde{D}_\ell^2])^{-1} \sqrt{n} \mathbb{E}_n[\tilde{D}_\ell \epsilon] \right| \approx 0,$$

and, consequently, the Gaussian approximation

$$\sqrt{n}(\hat{\alpha} - \alpha) \overset{a}{\sim} N(0, \mathsf{V}),$$

where

$$\mathsf{V}_{\ell k} = (\mathrm{E}[\tilde{D}_\ell^2])^{-1} \mathrm{E}[\tilde{D}_\ell \tilde{D}_k \epsilon^2] (\mathrm{E}[\tilde{D}_k^2])^{-1}.$$

Recall that the above distributional approximation formally means that

$$\sup_{R \in \mathscr{R}} \left| \mathrm{P}\left(\sqrt{n}(\hat{\alpha} - \alpha) \in R \right) - \mathrm{P}\left(N(0, \mathsf{V}) \in R \right) \right| \to 0,$$

where $\mathscr{R}$ is a collection of all (hyper) rectangles. The latter result allows the construction of *simultaneous confidence bands* on all target parameters α_ℓ's of the form:

$$\widehat{CR} = \times_{\ell=1}^{p_1} \left[\hat{\alpha}_\ell \pm c \sqrt{\hat{\mathsf{V}}_{\ell\ell}/n} \right],$$

The critical value c in the simultaneous confidence band is

chosen so that

$$\begin{aligned} \mathrm{P}(\alpha \in \widehat{CR}) &= \mathrm{P}\left(\sqrt{n}(\alpha - \hat{\alpha}) \in \sqrt{n}(\hat{C}R - \hat{\alpha})\right) \\ &= \mathrm{P}\left(\sqrt{n}(\alpha_\ell - \hat{\alpha}_\ell) \in [\pm c\hat{\mathsf{V}}_{\ell\ell}^{1/2}] \; \forall \; \ell \in \{1, \ldots, p_1\}\right) \\ &\approx 1 - \mathsf{a} \end{aligned}$$

where $1 - \mathsf{a}$ denotes the confidence level.

Remark 4.4.1 (Details on critical values) It can be shown that an "ideal" choice of c is

$$c = (1 - \mathsf{a}) - \text{ quantile of } \left\| N\left(0, \mathsf{D}^{-1/2}\mathsf{V}\mathsf{D}^{-1/2}\right)\right\|_\infty,$$

where $\mathsf{D} = \mathrm{diag}(\mathsf{V})$ is a matrix with variances $(\mathsf{V}_{\ell\ell})_{\ell=1}^{p_1}$ on the diagonal and zeroes off the diagonal. The critical value c can therefore be approximated by simulation plugging in $\mathsf{V} = \hat{\mathsf{V}}$. Please see [4], for example, for more details. Note that c is generally no smaller than the $(1 - \mathsf{a}/2)$-quantile of a $N(0, 1)$, so the simultaneous confidence bands are always no smaller than the component-wise confidence bands.

Remark 4.4.2 (Simultaneous vs. Marginal Confidence Intervals) A *simultaneous confidence band* guarantees that in repeated experiments the entire set of coefficients is covered by their respective intervals with a specified probability. For example, a 95% simultaneous confidence band means that if the experiment were repeated many times, every coefficient would lie within its interval in 95% of those repetitions.

In contrast, *marginal confidence intervals* ensure that each individual coefficient is covered 95% of the time when considered separately. However, when examining multiple coefficients simultaneously, these individual guarantees do not translate into 95% overall coverage. If there are p_1 coefficients, the probability that all are correctly covered is

$$(0.95)^{p_1}.$$

Thus, the probability of at least one interval failing is

$$1 - (0.95)^{p_1},$$

which can be significantly larger than 5% as p_1 increases.

This distinction is especially important when reporting *discoveries* by noting coefficients whose intervals exclude 0. With

marginal intervals, the chance that such a finding is a false discovery is $1 - (0.95)^{p_1}$, rather than the nominal 5%. In contrast, a 95% simultaneous confidence band controls the overall error probability at or below 5%.

While simultaneous bands provide robust family-wise error rate control, alternative procedures aimed at controlling the false discovery rate (FDR) may be less conservative. Such methods can be used in conjunction with the marginal confidence interval and p-value constructions we discuss in this book (see, e.g., [5]; [6]).

Discovering Heterogeneity in the Wage Gap Analysis

We apply the Double Lasso method of the preceding section to analyze heterogeneity of wage gaps using CPS 2015 data. As in Chapter 1, we use the log hourly wage as the outcome variable. To explore heterogeneity, we interact the female indicator with group indicators capturing education groups (Some High School (shs), High School Graduate (hsg), Some College (scl), College Graduate (clg), Advanced Degree (ad)), region indicators – Midwest (mw), South (so), West (we)) and a fourth degree polynomial in experience (exp1= Experience, exp2= Experience^2/100, exp3= Experience^3/1000, exp4= Experience^4/10000). In total these are 12 target parameters corresponding to the 11 interactive variables and the non-interactive variable that corresponds to the female indicator. All engineered variables used for heterogeneity were de-meaned prior to taking the interaction with sex, while the sex variable was not de-meaned. Hence, the interaction coefficients can be interpreted as "predictive effect modifiers," and the coefficient associated with the non-interactive variable sex as the average predictive effect. As additional variables, we also include all pairwise interactions of the aforementioned variables (excluding sex), as well as one-hot-encodings for occupation and industry sector, providing 990 engineered features. All engineered variables used as controls were also de-meaned prior to estimation.

Table 4.2 provides estimated coefficients, standard errors, pointwise p-values, and the 95% simultaneous confidence band for the coefficients on sex and its interactions with the schooling (shs, hsg, scl, clg, and ad), region (mw, so, and we), and experience (exp1, exp2, exp3, and exp4) variables described above. Rows give variable names with "*" indicating interaction; e.g.

	Estimate	Std. Error	p-value
sex	-0.07	0.01	0.00
sex:shs	-0.32	0.19	0.08
sex:hsg	0.05	0.05	0.29
sex:scl	0.03	0.05	0.49
sex:clg	0.06	0.05	0.19
sex:mw	-0.12	0.04	0.01
sex:so	-0.08	0.04	0.06
sex:we	-0.03	0.04	0.50
sex:exp1	0.02	0.01	0.12
sex:exp2	-0.04	0.07	0.59
sex:exp3	-0.05	0.03	0.18
sex:exp4	-0.00	0.00	0.98

Table 4.2: Estimates of Heterogeneous Predictive Effects in the CPS 2015 data

the row sex*shs provides results for the interaction between sex and shs.

Among other things, we see that having a college degree increases the predictive effect, i.e. decreases the wage gap, while the largest increase in wage gap occurs for un-educated workers. However, these heterogeneities are not statistically significant. Moreover, the wage gap is predicted to be larger in the Midwest region. Having a high potential experience also predicts larger drops in wages for female workers. Of course, the simultaneous confidence regions include 0 for all coefficients except for the main effect on sex suggesting that it may be difficult to draw any strong conclusions about heterogeneity of predictive effects in this example.

	Estimate	CI lower	CI upper
sex	-0.07	-0.11	-0.03
sex:shs	-0.32	-0.85	0.20
sex:hsg	0.05	-0.09	0.20
sex:scl	0.03	-0.11	0.17
sex:clg	0.06	-0.07	0.19
sex:mw	-0.12	-0.25	0.00
sex:so	-0.08	-0.20	0.04
sex:we	-0.03	-0.16	0.10
sex:exp1	0.02	-0.01	0.04
sex:exp2	-0.04	-0.23	0.16
sex:exp3	-0.05	-0.14	0.05
sex:exp4	-0.00	-0.01	0.01

Table 4.3: Simultaneous 95% Confidence Intervals for the Estimates of Heterogeneous Predictive Effects in the CPS 2015 data.

4.5 Other Approaches That Have the Neyman Orthogonality Property

Double Selection

One way to fix the naive "single selection" approach outlined in Section 4.3 would be to have "double selection":

Double Selection

- find controls W_Y that predict Y as judged by lasso;
- find controls W_D that predict D as judged by lasso;
- regress Y on D and the union of controls $W_Y \cup W_D$; proceed with standard inference.

This procedure is approximately equivalent to the partialling out approach, and therefore inherits the orthogonality property. This approach is more conservative compared to single selection, as it makes sure that we have not omitted controls that are strong confounders for D. It therefore guards against large omitted variable biases.

Debiased Lasso

Yet another procedure that has the orthogonality property and is approximately equivalent to the partialling out approach under suitable conditions is the debiased (also called desparsified) Lasso.

This approach uses the fact that $a = \alpha$ solves the equation,

$$\mathsf{M}(a, \eta) = \mathrm{E}[(Y - aD - b'W)\tilde{D}(\gamma)] = 0,$$

when $\eta = (b', \gamma')' = \eta^o := (\beta', \gamma'_{DW})'$ for γ_{DW} the best linear predictor coefficient from regressing D onto W and

$$\tilde{D}(\gamma) = D - \gamma' W.$$

One can verify that

$$\alpha(\eta) = \left(\mathrm{E}[D\tilde{D}(\gamma)]\right)^{-1} \mathrm{E}\left[(Y - b'W)\tilde{D}(\gamma)\right],$$

and that

$$\alpha = \alpha(\eta^o).$$

Further, the moment condition is Neyman orthogonal – verification of which is left to the reader – which implies that

$$\partial_\eta \alpha(\eta^o) = 0,$$

similarly to the argument for Double Lasso.

Debiased Lasso

- Run a Lasso estimator with suitable choice of λ as discussed in Chapter 3 of Y on D and W, and save the coefficient estimate $\hat{\beta}$.
- Run a Lasso estimator with suitable choice of λ as discussed in Chapter 3 of D on W and save the coefficient estimate $\hat{\gamma}$.
- The estimator $\hat{\alpha}$ is then the solution of the empirical analog of the moment condition above:

  $$\mathbb{E}_n[(Y - \hat{\alpha}D - \hat{\beta}'W)\tilde{D}(\hat{\gamma})] = 0,$$

  which has the explicit form

  $$\hat{\alpha} = \left(\mathbb{E}_n[D\tilde{D}(\hat{\gamma})]\right)^{-1} \mathbb{E}_n\left[(Y - \hat{\beta}'W)\tilde{D}(\hat{\gamma})\right],$$

  where $\hat{\beta}$ and $\hat{\gamma}$ are Lasso estimators.

Estimators of this form are referred to in econometrics as "instrumental variable estimators." In purely technical terms, we are using residualized $\tilde{D}$ to "instrument" for D.

4.6 Notes

We mainly follow the Double Lasso approach developed in [7] and [8], because it is nicely connected to the classical partialling out. Desparsified Lasso was developed by [9] and [10]; a closely related approach is the debiased Lasso proposed by [11]. All of these approaches could be called "debiased" Lasso and will generalize later to the approach called Debiased Machine Learning. The Double Selection method was developed by [12] and [13]. Inference on many coefficients using Double Lasso was first developed by [14] and [15]. [16] provide results for Double Lasso with clustered dependence. The Double Lasso and desparsified Lasso approaches have also been extended to time series and many time series by [17]. Both [16] and [17] take into account the temporal dependencies in the data when fitting Lasso and performing inference on the coefficients of interest.

Failure of single selection even when p is small is discussed in simple terms in [13], but the problem was first systematically examined by [18]. A recent paper [19] develops debiasing methods for shape constrained high-dimensional linear regression models.

[4] provide a recent survey on methods for simultaneous inference in high-dimensional settings.

For an in-depth analysis of heterogeneity in the wage gap based on Lasso, we refer to [20].

4.7 Notebooks

Notebook 4.7.1 (Orthogonal vs Non-Orthogonal Learning) R Notebook with Experiment on Orthogonal vs Non-Orthogonal Learning and Python Notebook with Experiment on Orthogonal vs Non-Orthogonal Learning presents the simulation experiment comparing orthogonal (partialling-out) with non-orthogonal learning (naive method).

Notebook 4.7.2 (Hard Sparsity on Orthogonal vs Non-Orthogonal) R Notebook with Hard Sparsity on Orthogonal vs Non-Orthogonal Learning and Python Notebook with Hard Sparsity on Orthogonal vs Non-Orthogonal Learning presents an alternative simulation to that shown in the main text

comparing orthogonal (partialling-out) with non-orthogonal learning. In this simulation, we consider orthogonal and non-orthogonal learning in a stylized treatment effects simulation.

Notebook 4.7.3 (Double Lasso for Growth Convergence) R Notebook on Double Lasso for Growth Convergence and Python Notebook on Double Lasso for Growth Convergence presents a Double Lasso analysis of the conditional convergence hypothesis in growth economics.

Notebook 4.7.4 (Double Lasso for the Heterogeneous Wage Gap) R Notebook on Double Lasso for the Heterogeneous Wage Gap and Python Notebook on Double Lasso for the Heterogeneous Wage Gap presents a Double Lasso analysis of the heterogeneous wage gap.

4.8 Exercises

Exercise 4.8.1 Experiment with the first Notebooks 4.7.1. Try different models. For example, try different coefficient structures for β and γ_{DW} and/or different covariance structures for W. Provide an explanation to a friend for what each step in the Double Lasso procedure is doing.

Exercise 4.8.2 (Double Lasso for Growth Convergence) Explore the Notebooks 4.7.3. Provide an explanation to a friend for what each step in the Double Lasso procedure is doing. Explain the empirical results to a friend. Experiment with making the set of controls more flexible and higher-dimensional by adding nonlinear and/or interaction terms that seem potentially interesting. Comment on how the results differ from the baseline results.

Exercise 4.8.3 (Double Lasso for the Heterogeneous Wage Gap) Explore the Notebooks 4.7.4. Provide an explanation to a friend for what each step in the inference procedure is doing. Explain the empirical results to a friend.

Exercise 4.8.4 (Neyman Orthogonality) Verify that Neyman orthogonality holds for the "de-sparsified" Lasso strategy.

4.A High-Dimensional Central Limit Theorems*

Let $X_1, \ldots, X_n$ be independent (but not necessarily identically distributed) random vectors with dimension p. Assume that X_i's have mean zero (otherwise, work with $X_i - \mathrm{E}[X_i]$ instead of X_i). Consider the scaled sample mean

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i.$$

Let $\bar{\sigma}, \underline{\sigma}$ be given positive constants such that $\underline{\sigma} \leqslant \bar{\sigma}$, and let $B_n \geqslant 1$ be a sequence of constants that may diverge as $n \to \infty$. Let $\Sigma_n = \mathrm{E}\left[S_n S_n^T\right] = n^{-1} \sum_{i=1}^{n} \mathrm{E}\left[X_i X_i^T\right]$. Also, let $\mathcal{R}$ denote the collection of closed rectangles in $\mathbb{R}^p$.

We first present a high-dimensional CLT over the rectangles under a sub-exponential condition on the coordinates. Suppose that the coordinates of X_i are sub-exponential with scale B_n, then

$$\sup_{R \in \mathcal{R}} |\mathrm{P}(S_n \in R) - \mathrm{P}(N(0, \Sigma_n) \in R)| \approx 0, \tag{4.A.1}$$

provided that $B_n^2 \log^5(pn)/n \approx 0$. Note that this allows p to be much larger than n. It turns out that a similar result applies without sub-exponential conditions, as stated formally below.

To state the results in a finite-sample form, let

$$\delta_{1,n} := \left(\frac{B_n^2 \log^5(pn)}{n}\right)^{1/4} \qquad \delta_{2,n}^{[q]} := \sqrt{\frac{B_n^2 (\log(pn))^{3-2/q}}{n^{1-2/q}}},$$

for $q > 2$.

Theorem 4.A.1 (High-Dimensional CLT, [21]) *Suppose second moments are non-degenerate,* $\min_{j \leqslant p} n^{-1} \sum_{i=1}^{n} \mathrm{E}\left[X_{ji}^2\right] \geqslant \underline{\sigma}^2$, *and fourth moments obey* $\max_{j \leqslant p} n^{-1} \sum_{i=1}^{n} \mathrm{E}\left[X_{ji}^4\right] \leqslant B_n^2 \bar{\sigma}^2$.

(A) If coordinates are subexponential, i.e. $\max_{i \leqslant n; j \leqslant p} \mathrm{E}\left[e^{|X_{ji}|/B_n}\right] \leqslant 2$, *then*

$$\sup_{R \in \mathcal{R}} |\mathrm{P}(S_n \in R) - \mathrm{P}(N(0, \Sigma_n) \in R)| \leq C \delta_{1,n},$$

where C is a constant that depends only on $\underline{\sigma}$ and $\bar{\sigma}$.

(B) If the envelope of the coordinates admits a moment bound $\max_{i \leqslant n} \mathrm{E}\left[\|X_i\|_\infty^q\right] \leq B_n^q$ *for some* $q > 2$*, then*

$$\sup_{R \in \mathcal{R}} |\mathrm{P}(S_n \in R) - \mathrm{P}(N(0, \Sigma_n) \in R)| \leq C\left(\delta_{1,n} \vee \delta_{2,n}^{[q]}\right)$$

where C *is a constant that depends only on* $q, \underline{\sigma}$ *and* $\bar{\sigma}$*.*

Notably, the above theorem does not impose any restrictions on the correlation structure between the coordinates of the random vectors, so Σ_n is permitted to be singular.

As discussed in [22], the assumption of Part (A) is satisfied if, for example, $\left|X_{ji}\right| \leqslant B_n$ for all (i, j), but also allows for unbounded coordinates. Part (B) covers the following scenario relevant to regression applications: $X_i = \epsilon_i z_i$ where ϵ_i is a univariate "error" term while $z_i \in \mathbb{R}^p$ is a vector of fixed "covariates." In this case, $\mathrm{E}\left[\|X_i\|_\infty^q\right] \leqslant \|z_i\|_\infty^q \mathrm{E}\left[|\epsilon_i|^q\right]$, so if the covariates are uniformly bounded and the q-th moments of the error terms are bounded, then $B_n = O(1)$. Notably this only requires ϵ_i to have $q = 2 + \delta$ bounded moments.

Often, statistics of interest are not exactly sample means, but can be well approximated by sample means. For example, the Double Lasso estimator, $\widehat{\alpha} = (\mathbb{E}_n[\check{D}^2])^{-1}\mathbb{E}_n[\check{D}\check{Y}] \approx (\mathrm{E}[\tilde{D}^2])^{-1}\mathbb{E}_n[\tilde{D}\tilde{Y}]$, takes this form. In order to claim a High-Dimensional CLT for such statistics, we need the approximation error to vanish at the rate faster than $1/\sqrt{\log p}$.[4]

Lemma 4.A.2 (High-dimensional CLT for approximate sample mean) . Suppose that S_n obeys (4.A.1), but S_n is not directly available. Suppose instead that we have access to $\widehat{S}_n$ that approximates S_n such that $\widehat{S}_n = S_n + \mathrm{R}_n$ with $\sqrt{\log p}\,\|\mathrm{R}_n\|_\infty \approx 0$. Assume $\min_{j \leqslant p} \Sigma_{jj} \geqslant \underline{\sigma}^2$. Then the same conclusion holds with S_n replaced by $\widehat{S}_n$.

The lemma follows from Nazarov's anticoncentration inequality for Gaussian vectors over rectangles; see [22] for the proof.

4: The requirement that approximation error, denoted R_n, vanishes faster than $1/\sqrt{\log p}$ arises from the fact that the maximum of a Gaussian random vector $N(0, \Sigma)$ concentrates in (i.e., places a probability mass of near 1 to) a $1/\sqrt{\log p}$- neighborhood of its expected value, but not in smaller neighborhoods (anti-concentration). The approximation error R_n needs to be much smaller than the size of the neighborhood. Otherwise, the probabilistic errors incurred by Gaussian approximation to the distribution of $\hat{S}$ can be as large as 1, meaning that the Gaussian approximation fails.

Bibliography

[1] Ragnar Frisch and Frederick V Waugh. 'Partial time regressions as compared with individual trends'. In: *Econometrica* (1933), pp. 387–401 (cited on page 99).

[2] Robert Barro and Jong-Wha Lee. 'A new data set of educational attainment in the world, 1950–2010'. In: *Journal of Development Economics* 104.C (2013), pp. 184–198 (cited on page 104).

[3] Matias D. Cattaneo, Michael Jansson, and Whitney K. Newey. 'Inference in linear regression models with many covariates and heteroscedasticity'. In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1350–1361 (cited on page 104).

[4] Philipp Bach, Victor Chernozhukov, and Martin Spindler. *Valid Simultaneous Inference in High-Dimensional Settings (with the hdm package for R)*. 2018. DOI: 10.48550/ARXIV.1809.04951. URL: https://arxiv.org/abs/1809.04951 (cited on pages 112, 117).

[5] Yoav Benjamini and Daniel Yekutieli. 'False Discovery Rate–Adjusted Multiple Confidence Intervals for Selected Parameters'. In: *Journal of the American Statistical Association* 100.469 (2005), pp. 71–81. DOI: 10.1198/016214504000001907 (cited on page 113).

[6] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, Christian Hansen, and Kengo Kato. 'High-dimensional econometrics and regularized GMM'. In: *arXiv preprint arXiv:1806.01888* (2018) (cited on page 113).

[7] Victor Chernozhukov, Christian Hansen, and Martin Spindler. 'Valid post-selection and post-regularization inference: An elementary, general approach'. In: *Annual Review of Economics* 7.1 (2015), pp. 649–688 (cited on page 117).

[8] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. 'Pivotal estimation via square-root lasso in nonparametric regression'. In: *Annals of Statistics* 42.2 (2014), pp. 757–788 (cited on page 117).

[9] Cun-Hui Zhang and Stephanie S. Zhang. 'Confidence intervals for low dimensional parameters in high dimensional linear models'. In: *Journal of the Royal Statistical Society: Series B* 76.1 (2014), pp. 217–242 (cited on page 117).

[10] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. 'On asymptotically optimal confidence regions and tests for high-dimensional models'. In: *Annals of Statistics* 42.3 (2014), pp. 1166–1202 (cited on page 117).

[11] Adel Javanmard and Andrea Montanari. 'Confidence intervals and hypothesis testing for high-dimensional regression'. In: *Journal of Machine Learning Research* 15.1 (2014), pp. 2869–2909 (cited on page 117).

[12] Alexandre Belloni, Victor Chernozhukov, and Christian B. Hansen. 'Inference for High-Dimensional Sparse Econometric Models'. In: *Advances in Economics and Econometrics: Tenth World Congress*. Ed. by Daron Acemoglu, Manuel Arellano, and Eddie Dekel. Vol. 3. Econometric Society Monographs. Cambridge University Press, 2013, pp. 245–295. DOI: 10.1017/CBO9781139060035.008 (cited on page 117).

[13] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. 'Inference on Treatment Effects After Selection Amongst High-Dimensional Controls'. In: *Review of Economic Studies* 81.2 (2014), pp. 608–650 (cited on page 117).

[14] Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. 'Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems'. In: *Biometrika* 102.1 (2015), pp. 77–94 (cited on page 117).

[15] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Ying Wei. 'Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework'. In: *Annals of statistics* 46.6B (2018), p. 3643 (cited on page 117).

[16] Alexandre Belloni, Victor Chernozhukov, Christian Hansen, and Damian Kozbur. 'Inference in High-Dimensional Panel Models With an Application to Gun Control'. In: *Journal of Business & Economic Statistics* 34.4 (2016), pp. 590–605 (cited on page 117).

[17] Victor Chernozhukov, Wolfgang Karl Härdle, Chen Huang, and Weining Wang. 'Lasso-driven inference in time and space'. In: *Annals of Statistics* 49.3 (2021), pp. 1702–1735 (cited on page 117).

[18] Hannes Leeb and Benedikt M. Pötscher. 'Model selection and inference: Facts and fiction'. In: *Econometric Theory* 21.1 (2005), pp. 21–59 (cited on page 117).

[19] Yufei Yi and Matey Neykov. 'A New Perspective on Debiasing Linear Regressions'. In: *arXiv preprint arXiv:2104.03464* (2021) (cited on page 117).

[20] Philipp Bach, Victor Chernozhukov, and Martin Spindler. 'Heterogeneity in the US gender wage gap'. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 187.1 (2024), pp. 209–230 (cited on page 117).

[21] Victor Chernozhukov, Denis Chetverikov, Kengo Kato, and Yuta Koike. 'Improved central limit theorem and bootstrap approximations in high dimensions'. In: *Annals of Statistics* 50.5 (2022), pp. 2562–2586 (cited on page 119).

[22] Victor Chernozhukov, Denis Chetverikov, Kengo Kato, and Yuta Koike. 'High-dimensional Data Bootstrap'. In: *Annual Review of Statistics and Applications; arXiv preprint arXiv:2205.09691* (2023) (cited on page 120).