1 Lasso and Related Methods

Consider a model:

$$Y = X'\beta + \varepsilon; \qquad \mathbb{E}[\varepsilon|X] = 0$$

where $\dim(\beta) = p$ is large, relative to n. Penalization methods rely on the *sparsity assumption*:[1]

$$s = \sum_{j=1}^{p} \mathbf{1}(\beta_j \neq 0) \text{ is small relative to } n.$$

That is, even though p may be similar to or even larger than n, we assume that only a small number s of the coefficients matter. The issue is that we don't know which ones. This assumption may be more or less reasonable depending on the application. If we believe in sparsity, we should impose it in estimation. There are many different ways to do so, as we discuss in detail below.

1.1 Best Subset Selection

Consider the following objective function:

$$\tilde{\beta}(\lambda) = \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i'b)^2 + \lambda \left|\left|b\right|\right|_0 \right),$$

where $\left|\left|b\right|\right|_0 = \sum_{j=1}^{p} \mathbf{1}(b_j \neq 0)$ and λ is a penalty parameter. This is a highly non-convex minimization problem. The only way to ensure that we solved it properly is to consider all possible non-zero subsets of b. That is, with p covariates one would have to estimate 2^p regressions, which becomes computationally infeasible very quickly.

This idea is applied in a paper by Sala-i-Martin (1997) titled "I just Ran Two Million Regressions" published in the *American Economic Review*. The goal of that paper is to understand what drives economic growth. The unit of observation is a country and the outcome variable is the long-run GDP growth. There are 62 explanatory variables available for each country. Sala-i-Martin runs all possible regressions with 7 explanatory variables and reports variables that are significant more often then others. The implicit assumption here is $s = 7$.

[1] This requirement is known as *exact sparsity*. In many cases, one can get away with a weaker condition of *approximate sparsity:* most coefficients are close to zero, and only a small number of coefficients are far away from zero.

1.2 Lasso

The Lasso objective function is:

$$\hat{\beta}(\lambda) = \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2 + \lambda \left|\left| b \right|\right|_1 \right),$$

where $\left|\left| b \right|\right|_1 = \sum_{j=1}^{p} |\beta_j|$ and λ is a penalty parameter. When $\lambda = 0$, Lasso coincides with OLS, and if $\lambda \to \infty$, all components of $\hat{\beta}_n(\lambda) \to 0$. Lasso stands for "Least Absolute Shrinkage and Selection Operator," and was proposed by Tibshirani (1996).

1.2.1 Mechanics

The Lasso objective function is convex, as a sum of two convex functions. In fact, it can be seen as a "convex relaxation" of the best subset selection problem. To appreciate the connection, consider an equivalent formulation of Lasso:

$$\min_{b \in \mathbb{R}^p} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2 \right) \quad \text{s.t. } \left|\left| \beta \right|\right|_1 \leqslant C.$$

There is a one-to-one mapping between λ and C, and λ may be interpreted as a Largange multiplier associated with the constraint $\left|\left| \beta \right|\right|_1 \leqslant C$. Notice that the level curves of the objective functions are ellipsoids, while the feasible set is diamond-shaped. This naturally produces corner solutions, so that many of the coefficients $\hat{\beta}(\lambda)$ will be exactly equal to zero. Therefore, Lasso also performs subset selection.

It is also useful to compare Lasso with OLS. This is most easily done when the regressors are orthonormal, $\frac{1}{n}\mathbb{X}'\mathbb{X} = I$. In this case,

$$\hat{\beta}_n^{OLS} = \left(\frac{1}{n} \mathbb{X}'\mathbb{X} \right)^{-1} \left(\frac{1}{n} \mathbb{X}'\mathbb{Y} \right) = \frac{1}{n} \mathbb{X}'\mathbb{Y}.$$

Writing the Lasso objective in vector form, note that:

$$\frac{1}{n}(\mathbb{Y} - \mathbb{X}'b)'(\mathbb{Y} - \mathbb{X}'b) + \lambda \left|\left| b \right|\right|_1 = \frac{1}{n}\mathbb{Y}'\mathbb{Y} - 2b' \underbrace{\left(\frac{1}{n} \mathbb{X}'\mathbb{Y} \right)}_{\hat{\beta}_n^{OLS}} + b' \underbrace{\left(\frac{1}{n} \mathbb{X}'\mathbb{X} \right)}_{I} b + \lambda \left|\left| b \right|\right|_1 .$$

Therefore,

$$\hat{\beta}_n(\lambda) = \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{j=1}^{p} \left(b_j^2 - 2 b_j \hat{\beta}_{n,j}^{OLS} + \lambda |b_j| \right),$$

so that we can solve the problem for each j separately:

$$\hat{\beta}_{n,j}(\lambda) = \underset{b_j \in \mathbb{R}}{\operatorname{argmin}} \left(b_j^2 - 2 b_j \hat{\beta}_{n,j}^{OLS} + \lambda |b_j| \right).$$

By taking first-order conditions and considering two cases $b_j \geqslant 0$ and $b_j < 0$, one can verify that the solution is:

$$\hat{\beta}_{n,j}(\lambda) = \operatorname{sign}(\hat{\beta}_{n,j}^{OLS}) \max \left(|\hat{\beta}_{n,j}^{OLS}| - \frac{\lambda}{2}, 0 \right),$$

where $\operatorname{sign}(a) = 1$ if $a \geqslant 0$ and $\operatorname{sign}(a) = -1$ if $a < 0$.

So, we make three observations: (i) Lasso and OLS coefficients have the same sign; (ii) $|\hat{\beta}_{n,j}(\lambda)| < |\hat{\beta}_{n,j}^{OLS}|$; (iii) and for $\lambda > 2 \max_j(|\hat{\beta}_{n,j}^{OLS}|)$, all $\hat{\beta}_{n,j}(\lambda) = 0$. Note that (i) and (ii) imply that in addition to selecting covariates, Lasso *shrinks* the coefficients towards zero. This is a purely mechanic effect that may not be desirable.

In practice, it is important to take the scale of covariates into account. Consider a model:

$$Y = X_1' \beta_1 + X_2' \beta_2 + \varepsilon,$$

and let $\hat{\beta}_{n,1}^{OLS}$ denote the OLS coefficient. Rescaling $\tilde{X}_1 = \gamma X_1$, we can re-write:

$$Y = \tilde{X}_1{}' \theta_1 + X_2' \beta_2 + \varepsilon,$$

where $\tilde{\theta}_1 = \beta_1 / \gamma$. Estimating this model, we naturally obtain $\hat{\theta}_{n,1}^{OLS} = \hat{\beta}_{n,1}^{OLS} / \gamma$.

Unlike OLS, Lasso estimator is sensitive to the scaling of covariates. Rescaling the variable X_1 with a large γ will make the corresponding coefficient very small and more likely to be "killed" by Lasso (i.e., set to exactly zero). For this reason, a common approach is to standartize all covariates so that $\sum_{i=1}^n X_{ij} = 0$ and $\sum_{i=1}^n X_{ij}^2 = 1$ before estimating the model.

1.2.2 Lasso as a Robust OLS

As we have seen above, one motivation for Lasso is the need of variable selection. Another one can be given via its interpretation as a version of robust OLS.

The OLS coefficients are known to be sensitive to outliers in the data. One way to deal with this issue is to consider a robust version of OLS:

$$\min_{b\in\mathbb{R}^p}\left\{\max_{\Delta\in\mathcal{D}}||\mathbb{Y}-(\mathbb{X}+\Delta)b||_2\right\},$$

where $\Delta=(\delta_1,\ldots,\delta_p)$ is a $n\times p$ matrix representing perturbation of the data and $\mathcal{D}$ is the set of all allowed perturbations, called an uncertainty set. The idea is to minimize the sum of squared residuals under the worst-case perturbation of the data within the set $\mathcal{D}$. Consider, specifically,

$$\mathcal{D}=\{(\delta_1,\ldots,\delta_p):||\delta_j||_2\leqslant\lambda \text{ for all } j\}.$$

Then, with some algebra one can show that for any fixed b:

$$\max_{\Delta\in\mathcal{D}}||\mathbb{Y}-(\mathbb{X}+\Delta)b||_2=||\mathbb{Y}-\mathbb{X}b||_2+\lambda\sum_{j=1}^{p}|b_j|.$$

This means that the above optimization problem is equivalent to:

$$\min_{b\in\mathbb{R}^p}\left\{||\mathbb{Y}-\mathbb{X}b||_2+\lambda\sum_{j=1}^{p}|b_j|\right\},$$

which corresponds to Lasso. Moreover, one can show that $\hat{\beta}_j^{LASSO}=0$ if there is an allowable perturbation $\Delta\in\mathcal{D}$ which makes the feature x_j irrelevant. See Caramanis, Mannor, and Xu (2010) for the details.

1.2.3 Computation

Generally speaking, convex functions are easy to minimize using variations of the gradient descent algorithm. Note however that the Lasso objective function is non-differentiable: it has a kink at zero. Two most popular algorithms for solving the Lasso problem are the *Cyclic Coordinate Descent (CCD)* and *Least Angle Regression (LARS)* algorithms. Both are great, but in the interest of time we will focus on the

former.

The idea of CCD algorithm is to update one coordinate at a time leveraging simplicity of univariate Lasso. Importantly, we assume that the variables are standartized so that $\frac{1}{n}\sum_{i=1}^{n} X_{ij}^2 = 1$ for all j. Suppose the values of all b_l for $l \neq j$ are fixed, and consider the problem:

$$\min_{b_j \in \mathbb{R}} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_{ij} b_j - \sum_{l \neq j} X_{il} b_j)^2 + \lambda |b|_j \right).$$

This is a univariate Lasso problem that has a closed-form solution:

$$\hat{b}_j = \text{sign}(\hat{a}_{-j}) \max \left(|\hat{a}_{-j}| - \frac{\lambda}{2}, 0 \right),$$

where

$$\hat{a}_{-j} = \frac{1}{n} \sum_{i=1}^{n} X_{ij} \left(Y_i - \sum_{l \neq j} X_l b_l \right).$$

Then, the algorithm is as follows:

1. Initiate some value $b = (b_1, \ldots, b_p)$.
2. Loop over $j = 1, \ldots, p, 1, \ldots, p, \ldots$ solving the problem above until numerical convergence.

1.2.4 Choosing λ

Choice of the penalty parameter λ in Lasso plays a familiar role of balancing bias and variance. When λ is "too small", $\hat{\beta}_n(\lambda)$ will contain "too many" non-zero coefficients and, as a result, have a large variance. On the other hand, when λ is "too large", the variance is reduced but $\hat{\beta}_n(\lambda)$ is biased towards zero.

There are four methods to choose λ: sample splitting, cross-validation, BRT rule, and BCCH rule. Typically, the first two methods are *very good* at out-of-sample prediction but *very bad* at variable selection (they tend to select too many variables). In turn, the last two methods are *reasonably good at both* out-of-sample prediction and variable selection.

1. **Sample splitting**

   (1) Randomly split the sample

   $$\underbrace{(Y_1, X_1), \ldots, (Y_m, X_m)}_{\text{training sample}}, \underbrace{(Y_{m+1}, X_{m+1}), \ldots, (Y_n, X_n)}_{\text{test sample}},$$

   where $m \approx \frac{2}{3}n$.

   (2) For each λ, estimate $\hat{\beta}_m(\lambda)$ using the training sample and calculate out-of-sample prediction error using the test sample:

   $$\mathcal{F}(\lambda) = \frac{1}{n-m} \sum_{i=m+1}^{n} (Y_i - X_i' \hat{\beta}_m(\lambda))^2.$$

   (3) Choose λ to minimize

   $$\hat{\lambda}_n^{SS} = \underset{\lambda \geqslant 0}{\operatorname{argmin}}\, \mathcal{F}(\lambda).$$

2. **Cross-Validation**

   (1) Split the sample into K subsamples of similar size (typically $K = 5, 10$).

   (2) For each subsample k train the model using all other subsamples and test it on subsample k to obtain prediction error $\mathcal{F}_k(\lambda)$.

   (3) Choose λ to minimize:

   $$\hat{\lambda}_n^{CV} = \underset{\lambda \geqslant 0}{\operatorname{argmin}} \sum_{k=1}^{K} \mathcal{F}_k(\lambda).$$

3. **Bickel-Ritov-Tsybakov rule**. This rule requires homoskedastic noise with known variance $\mathbb{V}ar(\varepsilon) = \sigma^2$.

   (1) Choose $\alpha \in (0, 1)$, typically $\alpha = 0.05$, and $c > 1$, typically $c = 1.1$.

   (2) Set

   $$\hat{\lambda}_n^{BRT} = \frac{2c\sigma}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) \max_{1 \leqslant j \leqslant p} \sqrt{\frac{1}{n} \sum_{i=1}^{n} X_{ij}^2},$$

   where Φ is the CDF of the Standard Normal distribution.

4. **Belloni-Chen-Chernozhukov-Hansen rule**. This rule allows for heteroskedastic noise and requires no prior knowledge.

   (1) Choose α and c as in Bickel-Ritov-Tsybakov rule.

   (2) Run a pilot Lasso to obtain $\hat{\beta}_n(\lambda^{\text{pilot}})$ with

$$\lambda^{\text{pilot}} = \frac{2c}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2p}\right)\max_{1\leqslant j\leqslant p}\sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^2X_{ij}^2}$$

   (3) Calculate the redisuals:

$$\hat{\varepsilon}_i = Y_i - X_i'\hat{\beta}_n(\lambda^{\text{pilot}}).$$

   (4) Set:

$$\hat{\lambda}_n^{BCCH} = \frac{2c}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2p}\right)\max_{1\leqslant j\leqslant p}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2X_{ij}^2}.$$

1.2.5 Some Theoretical Guarantees

Recall the model:

$$Y = X'\beta + \varepsilon; \quad \mathbb{E}[\varepsilon|X] = 0,$$

and the sparsity assumption:

$$s = \sum_{j=1}^{p}\mathbf{1}(\beta_j \neq 0) \quad \text{is small relative to } n.$$

For a vector $\delta \in \mathbb{R}^p$, we denote:

$$||\delta||_2 = \sqrt{\sum_{j=1}^{p}\delta_j^2} \qquad ||\delta||_\infty = \max_{1\leqslant j\leqslant p}|\delta_j|,$$

and define the so-called gradient:

$$S = \frac{2}{n}\sum_{i=1}^{n}X_i\varepsilon_i.$$

The following result is from Bickel, Ritov, and Tsybakov (2009).

Theorem 1 (Rate of convergence for Lasso)**.** *Let $c > 1$ be some constant. Then, under some technical conditions, the event*

$$\lambda > c\,||S||_\infty$$

implies

$$\left|\left|\hat{\beta}_n(\lambda) - \beta\right|\right|_2 \leqslant C\lambda s,$$

where C is a constant independent of n and p. In particular, if $\varepsilon|X \sim N(0, \sigma^2)$, then setting $\lambda = \hat{\lambda}_n^{BRT}$ gives

$$\left|\left|\hat{\beta}_n(\lambda) - \beta\right|\right|_2 \leqslant \sqrt{\frac{Cs\log p}{n}}$$

with probability at least $1 - \alpha$ for large n.

Several comments are in order. First, the result shows the familiar bias-variance trade-off: it is desired to choose λ as small as possible such that $\lambda > c\,||S||_\infty$ holds with high probability. With a small λ, it is hard to guarantee that $\hat{\beta}_n(\lambda)$ will be close to β with high probability due to the high variance of $\hat{\beta}_n(\lambda)$ in this case. With a large λ, too many Lasso coefficients are set to zero and the bias kicks in. Second, the result implies that $\hat{\beta}_n(\lambda)$ is consistent even if p is larger than n, as long as s is small. This is great, but does it imply that Lasso is going to correctly guess which coefficients β_j are zero? Not necessarily. The literature provides conditions under which Lasso consistently estimates *the set* of non-zero coefficients, but these conditions are much stronger than what is needed for the above consistency in l_2 norm. This observation has implications on interpretation of the coefficients, alternative penalization schemes, and inference. Finally, it is instructive to compare the rate result with OLS. Recalling the rate for OLS with large p from Lecture 2, we can find a constant $\tilde{C}$ such that

$$\left|\left|\hat{\beta}_n^{OLS} - \beta\right|\right|_2 \leqslant \sqrt{\frac{\tilde{C}p}{n}}$$

with probability at least $1 - \alpha$. Therefore, Lasso works much better if the true model is sparse.

1.3 Variations of Lasso

1.3.1 Post-Lasso

Recall that Lasso performs two things: variable selection and shrinkage. If the true DGP is sparse and we know the key s covariates, the OLS estimator is unbiased. Therefore, even if Lasso selects those key covariates perfectly, it will be biased towards zero. Therefore, one might prefer the following procedure, called *Post-Lasso*:

1. Run Lasso of Y on X to obtain $\hat{\beta}_n(\lambda)$.
2. Select variables X_j corresponding to $\hat{\mathcal{J}}_n = \{j : \hat{\beta}_{n,j}(\lambda) \neq 0\}$.
3. Run OLS of Y on selected X_j, $j \in \hat{\mathcal{J}}_n$.

This procedure retains the variables selected by Lasso, but avoids shrinkage. Belloni and Chernozhukov (2009) study statistical properties of a large class of "post-selection OLS" estimators and show that such estimators perform at least as good as Lasso in terms of rate of convergence but have the advantage of a smaller bias.

1.3.2 Partial Penalization

In many applications, it may not be desirable to penalize all covariates. For example, some key regressors (education in the wage regression, price in the demand regression, etc.) should always be included and penalizing them makes little sense. Also, penalizing fixed effects in panel data models defeats the purpose of including them. Luckily, Lasso provides a great deal of flexibility in this regard.

Suppose that $X = (D, W)$ where $D \in \mathbb{R}^d$ are key regressors and $W \in \mathbb{R}^p$ are possibly high-dimensional controls. Consider the following problem:

$$\min_{(b_1,b_2)\in\mathbb{R}^d\times\mathbb{R}^p} \left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - D_i'b_1 - W_i'b_2)^2 + \lambda \sum_{j=1}^{p} |b_{2j}| \right).$$

To solve this, we can first partial out D using the Frisch-Waugh-Lowell theorem. Denoting $\mathbb{Y} = (Y_1, \ldots, Y_n)$, $\mathbb{D}$ a $n \times d$ matrix with rows D_i', $\mathbb{W}$ a $n \times p$ matrix with rows W_i', and $P_D = \mathbb{D}(\mathbb{D}'\mathbb{D})^{-1}\mathbb{D}'$ the projection matrix, we obtain:

$$\tilde{\mathbb{Y}} = (I - P_D)\mathbb{Y}; \qquad \tilde{\mathbb{W}} = (I - P_D)\mathbb{W}.$$

Then, we can estimate$\hat{\beta}_{2,n}(\lambda)$ via

$$\hat{\beta}_{2,n}(\lambda) = \underset{b_2 \in \mathbb{R}^p}{\operatorname{argmin}} \left(\frac{1}{n} \sum_{i=1}^{n} (\tilde{Y}_i - \tilde{W}_i' b_2)^2 + \lambda \sum_{j=1}^{p} |b_{2j}| \right),$$

and then solve:

$$\hat{\beta}_{1,n}(\lambda) = \min_{b_1 \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - D_i' b_1 - W_i' \hat{\beta}_{2,n}(\lambda))^2.$$

1.3.3 Double Lasso

Now, consider a model:

$$Y = D\beta_1 + W'\beta_2 + \varepsilon; \quad \mathbb{E}[\varepsilon | D, W] = 0,$$

where $D \in \mathbb{R}$ is the target covariate and W is a vector of controls. Since it makes little sense to penalize β_1 and Post-Lasso works well, one may consider the following procedure: (1) run Lasso excluding b_1 from the penalty to select controls; (2) run OLS on D and the controls selected in the first step. While this approach may seem natural, it is very problematic.

Lasso, as well as other high-dimensional estimation methods, targets *prediction* rather then learning about specific model parameters Intuitively, any variable that is highly correlated with D is more likely to be excluded in the first stage since it does not improve prediction accuracy too much. But exclusion of such variables is precisely what leads to omitted variable bias (recall the discussion in Lecture 2).

Another problem is that existing inference results for Lasso and Post-Lasso estimators rely on very strong assumptions that guarantee *perfect model selection*. Outside of the small class of data generating processes in which these assumptions are believable, the asymptotic distribution of Lasso and Post-Lasso is unknown, so we cannot construct confidence intervals or test hypotheses.

A better estimator is the so-called *Double-Lasso*, proposed by Belloni, Chernozhukov, and Hansen (2014). To define it, in addition to the main model above consider an auxiliary model:

$$D = \gamma' W + \nu; \quad \mathbb{E}[\nu | W] = 0.$$

Note that we're not modeling the first stage to deal with endogeneity. Instead, we use it to construct a suitable moment condition for estimating β_1. By the Frisch-Waugh-Lowell theorem (or the Law of Iterated Expectations),

$$\mathbb{E}[(Y - D\beta_1 - W'\beta_2)(D - W'\gamma)] = 0 \quad \Longrightarrow \quad \beta_1 = \frac{\mathbb{E}[(Y - W'\beta_2)(D - W'\gamma)]}{\mathbb{E}[D(D - W'\gamma)]}. \quad (1)$$

Then, we can estimate β_1 as follows:

1. Run Lasso of D on W to obtain $\tilde{\gamma}_n$.
2. Run Lasso of Y on D and W to obtain $\tilde{\beta}_{1,n}, \tilde{\beta}_{2,n}$.
3. Use the analogy principle to estimate:

$$\hat{\beta}_{1,n}^{DL} = \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - W_i'\tilde{\beta}_{2,n})(D_i - W_i'\tilde{\gamma}_n)}{\frac{1}{n}\sum_{i=1}^n D_i(D_i - W_i'\tilde{\gamma}_n)}.$$

Then, under regularity conditions, one can show that:

$$\sqrt{n}(\hat{\beta}_{1,n}^{DL} - \beta_1) \quad \to_d \quad N(0, \sigma^2),$$

where $\sigma^2 = \mathbb{E}[\varepsilon^2\nu^2]/(\mathbb{E}[\nu^2])^2$. This is why Belloni, Chernozhukov, and Hansen (2014) changed the whole field.

Several comments are in order. First, it is important to perform variable selection (i.e., run Lasso instead of OLS) at each stage of the procedure. Intuitively, in the first stage, the procedure will select variables most relevant for predicting D and in the second stage, the procedure will select variables most relevant for predicting Y. Then, the third stage uses all of the selected variables to estimate β_1, which guards against omitted variable bias.

Second, recall that the problem with inference using Lasso or Post-Lasso was the requirement of perfect model selection. This is a requirement on the quality of estimation of *nuisance parameters* β_2 and γ. Note that the moment condition in (1) is, in some sense, "immunized" against imperfect estimation of the nuisance parameters. Specifically, denoting

$$\psi(Y, D, W, \beta_1; \beta_2, \gamma) = (Y - D\beta_1 - W'\beta_2)(D - W'\gamma),$$

we have

$$\frac{\partial}{\partial \beta_2}\mathbb{E}[\psi(Y, D, W, \beta_1; \beta_2, \gamma)] = 0; \qquad \frac{\partial}{\partial \gamma}\mathbb{E}[\psi(Y, D, W, \beta_1; \beta_2, \gamma)] = 0.$$

This property is known as *Neyman Orthogonality*. Intuitively, if the moment condition used to identify β_1 is not too sensitive to nuisance parameters, we can accommodate larger estimation errors and still get a good estimator for β_1. Formally, Neyman Orthogonality weakens the rate requirements on $\tilde{\beta}_{2,n}$ and $\tilde{\gamma}_n$. In contrast, the standard OLS moment condition:

$$\mathbb{E}\left[(Y - D\beta_1 - W'\beta_2)\begin{pmatrix} D \\ W \end{pmatrix}\right] = 0$$

is not Neyman orthogonal.

Finally, linearity of the conditional expectations in the model above is not necessary for Double-Lasso to work. Specifically, we can consider a more flexible model:

$$Y = D\beta_1 + g_1(W) + \varepsilon; \quad \mathbb{E}[\varepsilon|D, W] = 0$$

$$D = g_2(W) + \nu; \quad \mathbb{E}[\nu|W] = 0$$

and estimate functions $g_1(W)$ and $g_2(W)$ non-parametrically. We will revisit this model later in the course.