Handout 4
Sparsity. High-dimensional models. Lasso estimator

Instructor: Vira Semenova Note author: Vira Semenova

1 Definitions and Notation. Ideal Noise Model

In this handout, we will introduce the basics of **high-dimensional** regime, where the parameter of interest $\theta_0 \in \mathrm{R}^p$ can have dimension exceeding sample size:

$$\dim(\theta) = p \gg n.$$

Of course, additional assumptions should be placed on θ to make informative learning possible. An example of this assumption is sparsity, that is, the number of non-zero coordinates of

$$\|\theta_0\|_0 = s = \sum_{j=1}^{p} 1\{\theta_{0,j} \neq 0\}$$

is small relative to sample size n:

$$s \ll n \ll p.$$

Most of the coordinates of θ_0 are actually zero, but which ones are non-zero is unknown. Section 3 gives examples of sparse structures in economic data.

Notation.

Let me introduce notation. Define

$$(a)_+ = \max\{a, 0\}, \quad a \vee b = \max\{a, b\}, \quad a \wedge b = \min\{a, b\}.$$

We also use the notation $a \lesssim b$ to denote $a \leqslant cb$ for some constant $c > 0$ that does not depend on n; and $a \lesssim_P b$ to denote $a = O_P(b)$. The l_2-norm and l_1-norm are denoted by

$$\|\theta\| := (\sum_{j=1}^{p} \theta_j^2)^{1/2}, \quad \|\theta\|_1 = \sum_{j=1}^{p} |\theta_j|$$

and the l_0-norm and l_∞-infty norms are

$$\|\theta\|_0 = \sum_{j=1}^{p} 1\{\theta_j \neq 0\}, \quad \|\theta\|_\infty = \max_{1 \leq j \leq p} |\theta_j|.$$

Finally, denote the sample average

$$\mathbb{E}_n[f] = n^{-1} \sum_{i=1}^{n} f(X_i).$$

Given a vector $\theta \in \mathbb{R}^p$, denote the support set

$$T := \{j : \quad 1 \leq j \leq p, \theta_{0,j} \neq 0\}$$

which consists of the non-zero elements of T. Denote the **sparsified** θ_T as

$$\theta_{Tj} := \begin{cases} \theta_j & j \in T \\ 0 & j \notin T \end{cases}, \quad j = 1, 2, \ldots, p.$$

By construction,

$$\|\theta_T\|_0 \leq |T| = s.$$

Ideal Noise Model.

Here we consider a high-dimensional parameter θ_0 in $\mathbb{R}^p$, where p is high-dimensional in the sense that it is not (necessarily) small compared to the sample size. Specifically we allow for cases where p/n is large, albeit within reasons. Consider the following model:

$$Y_i = X_i'\theta_0 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \tag{1.1}$$

where X_i's are fixed such that the *orthonormal condition holds*:

$$\mathbb{E}_n X_i X_i' = I_p. \tag{ORT}$$

and σ^2 is *known* parameter. The Maximum Likelihood Estimator (MLE), which coincides with OLS, takes the form:

$$\hat{\theta}^{\text{MLE}} = (\mathbb{E}_n X_i X_i')^{-1} \mathbb{E}_n X_i Y_i = \mathbb{E}_n X_i Y_i. \tag{1.2}$$

The squared estimation error of MLE is

$$\mathbb{E}_n (X_i'(\hat{\theta}^{\text{MLE}} - \theta_0))^2 = (\hat{\theta}^{\text{MLE}} - \theta_0)' \mathbb{E}_n X_i X_i' (\hat{\theta}^{\text{MLE}} - \theta_0) = \|\hat{\theta}^{\text{MLE}} - \theta\|_2^2.$$

Hence to estimate the regression function well, we need to have the norm of the error, $\|\hat{\theta}^{\text{MLE}} - \theta_0\|_2$, small. □

In the ideal noise model (1.1), we have

$$\|\hat{\theta}^{\text{MLE}} - \theta_0\|_2^2 \sim \chi^2(p)/n = (p/n)(1 + O_P(1/\sqrt{p})),$$

which diverges to infinity if p increases more rapidly than n. The maximal risk of this estimator is

$$\sup_{\theta_0 \in \mathbb{R}^p} \mathbb{E}\left[\|\hat{\theta}^{\text{MLE}} - \theta_0\|_2^2\right] = p/n,$$

and there are no estimators having smaller maximal risk. This is a rather pessimistic conclusion. On a more optimistic side, there is a fundamental result due to Stein that shows that there are regularized estimators whose risk is no higher than that of $\hat{\theta}^{\text{MLE}}$ for each θ_0 and can be strictly lower for some values of θ_0. Unfortunately the maximal risk of Stein's estimators is still no better than p/n.

Sparsity. A simple structure that provides useful intuition is the *exact sparsity structure*. Under this structure, θ_0 has exactly s non-zero components, while the rest $p - s$ are zero. If the support set

$$T = \{j \in \{1, 2, \ldots, p\} : \theta_j \neq 0\}$$

was known, the ideal (oracle) estimator

$$\theta^{\text{HRD}o}_j = \begin{cases} \hat{\theta}^{\text{MLE}}_j, & \text{if } j \in T, \\ 0, & \text{if } j \notin T \end{cases}$$

attains the rate

$$\|\tilde{\theta} - \theta_0\|_2 \leq \sqrt{\sum_{j \in T} |\hat{\theta}_j - \theta_{0j}|^2} \leq \sqrt{\|N(0, I_s/n)\|^2} = \sqrt{\chi^2(s)/n} \lesssim_P \sqrt{s/n}\,, \tag{1.3}$$

and, therefore, is small. However, T is unknown and needs to be estimated. The *Hard Thresholded Estimator* below attains a *near*-oracle convergence rate up to $\sqrt{\log p}$.

Given the threshold ρ to be chosen later, the *Hard Thresholded Estimator* is

$$\theta_j^{\text{HRD}} = \begin{cases} \hat{\theta}_j^{\text{MLE}}, & \text{if } |\hat{\theta}_j^{\text{MLE}}| > 2\rho, \\ 0, & \text{if } |\hat{\theta}_j^{\text{MLE}}| \leq 2\rho, \end{cases}, \quad j = 1, 2, \ldots, p$$

and

$$\theta^{\text{HRD}} := (\theta_1^{\text{HRD}}, \theta_2^{\text{HRD}}, \ldots, \theta_p^{\text{HRD}}).$$

Lemma 1 is the first main Lemma. It bounds the mean square disk

Lemma 1 (Hard thresholding estimator)**.** *1. (a) If*

$$\rho \geq \|\hat{\theta}^{MLE} - \theta_0\|_\infty, \tag{1.4}$$

the mean squared risk of $\hat{\theta}^{HRD}$ is

$$\|\hat{\theta}^{HRD} - \theta_0\|_2^2 = \sum_{j=1}^{p} (\hat{\theta}_j^{HRD} - \theta_{0j})^2 \leq 16 s\rho^2. \tag{1.5}$$

2. (b) If $\min_{j \in T} |\theta_{0j}| > 3\rho$, then

$$supp(\widehat{\theta}^{HRD}) = supp(\theta).$$

Proof. See appendix. □

Lemma 2 ($\hat{\theta}^{\text{HRD}}((2\rho)) = \hat{\theta}(\lambda)$ if $\lambda = 4\rho^2$)**.** *One can show that, under the ORT condition, $\hat{\theta}^{HRD}$ coincides with ℓ_0-penalized estimator (see HW2)*

$$\hat{\theta}(\lambda) = \arg\min_{\theta \in \mathrm{R}^p} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i'\theta)^2 + \lambda \|\theta\|_0, \tag{1.6}$$

where $\lambda = 4\rho^2$.

Decompose the least squares criterion as

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} (Y_i - X_i'\theta)^2 &= n^{-1} \sum_{i=1}^{n} Y_i^2 - 2n^{-1} \sum_{i=1}^{n} Y_i X_i'\theta + \theta' n^{-1} \sum_{i=1}^{n} X_i X_i'\theta \\ &= n^{-1} \sum_{i=1}^{n} Y_i^2 - 2\widehat{\theta}_{\text{ML}}'\theta + \theta'\theta. \end{aligned}$$

Since the first term does not depend on θ, it suffices to consider

$$-2\widehat{\theta}_{\text{ML}}'\theta + \theta'\theta + \lambda\|\theta\|_0 := Q(\theta),$$

Note that $Q(\theta_1, \theta_2, \ldots, \theta_p) = \sum_{j=1}^{p} Q_j(\theta_j)$ is an additively separable function of θ. Then,

$$\min_{\theta} Q(\theta_1, \theta_2, \ldots, \theta_p) = \min_{\theta_j} Q_j(\theta_j) \quad \forall j = 1, 2, \ldots, p.$$

Using the hint gives

$$\begin{aligned} \min_{\theta} Q(\theta) &\Leftrightarrow \min_{\theta} -2\widehat{\theta}_{\text{ML}}'\theta + \theta'\theta + \lambda\|\theta\|_0 \\ &\Leftrightarrow \min_{\theta_j} -2\widehat{\theta}_{\text{ML},j}\theta_j + \theta_j^2 + \lambda 1\{\theta_j \neq 0\} \quad \forall j = 1, 2, \ldots, p. \end{aligned}$$

The indicator function $1\{\theta_j \neq 0\}$ can take two values depending on θ_j. If one chooses $\theta = 0$, it results in $Q_j(0) = 0$. The second option is to solve the unrestricted problem

$$\min_{\theta_j} -2\widehat{\theta}_{\mathrm{ML},j}\theta_j + \theta_j^2 + \lambda \cdot 1$$

whose optimal value is

$$Q_j(\theta_j^*) = -\widehat{\theta}_{\mathrm{ML},j}^2 + \lambda.$$

Comparing $Q_j(\theta_j^*)$ with $Q_j(0) = 0$ gives an optimal estimator

$$\theta_j^* = \begin{cases} \theta_{\mathrm{ML},j}, & |\theta_{\mathrm{ML},j}| \geq \sqrt{\lambda} \\ 0, & |\theta_{\mathrm{ML},j}| < \sqrt{\lambda} \end{cases}$$

Under the orthonormal condition ORT, one can show that

$$\hat{\theta}(\lambda) = \hat{\theta}^{\mathrm{HRD}}(2\rho), \quad \lambda = 4\rho^2.$$

The objective of BIC estimator is not convex in θ, and solving for (1.6) is an NP hard problem. A convex relaxation of $\hat{\theta}^{\mathrm{HRD}}$ is LASSO estimator

$$\hat{\theta}_L = \underset{\theta \in \mathrm{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i'\theta)^2 + \lambda \|\theta\|_1,$$

which replaces $\|\theta\|_0$ by $\|\theta\|_1$ to obtain a convex objective function. The resulting estimator has some statistical guarantees.

Lemma 3 (Lasso estimator)**.** *If $\rho \geq \|\hat{\theta} - \theta_0\|_\infty$, the mean square risk of $\hat{\theta}^{SFT}$ is*

$$\|\hat{\theta}^{SFT} - \theta_0\|_2^2 \leq 36 s \rho^2.$$

Proof. See appendix. □

Lemma 4 (LASSO = Soft thresholded.)**.** *The LASSO estimator coincides with the soft-thresholding estimator*

$$\hat{\theta}_j^{SFT} = \begin{cases} \hat{\theta}_j - 2\rho, & \text{if } \hat{\theta}_j > 2\rho \\ 0, & \text{if } |\hat{\theta}_j| < 2\rho, \\ \hat{\theta}_j + 2\rho & \text{if } \hat{\theta}_j < -2\rho, \end{cases},$$

where $j = 1, 2, \ldots, d$. That is,

$$\hat{\theta}^{SFT} = \hat{\theta}_L \ .$$

Lemmas 1 and 3 are stated as if-then statements. If a random event $\rho \geq \|\hat{\theta} - \theta_0\|_\infty$ holds, then we have a rate bound in terms of ρ. We choose ρ such that the event

$$\rho \geq \|\hat{\theta}^{\mathrm{MLE}} - \theta_0\|_\infty$$

holds with probability $1 - \delta$ for a given $\delta \in (0, 1)$. Then, the desired statement will also hold with a given probability. Section 2 discusses the choice of ρ.

2 Choice of penalty

A mean zero random variable X is σ^2-subGaussian if

$$\mathbb{E}\left[\exp^{\lambda X}\right] \leq \exp^{\lambda^2 \sigma^2 / 2} \quad \forall \lambda \in \mathrm{R}.$$

A r.v. $X \sim N(0, \sigma^2)$ is, by definition, σ^2-subGaussian. Any a.s. bounded random variable bounded by a, such as $U[-a, a]$ is a^2-subGaussian (Hoeffding lemma). Basic facts about subGaussian random variables:

1. X is σ^2-subGaussian, therefore cX is $c^2\sigma^2$-subGaussian. Proof: definition.
2. $\cup(X_i)_{i=1}^n$ is σ^2-subGaussian and i.i.d, therefore $\sum_{i=1}^n X_i$ is $n\sigma^2$-subGaussian. Proof:

$$\mathbb{E}\left[\exp^{\lambda\sum_{i=1}^n X_i}\right] = \prod_{i=1}^n \mathbb{E}\left[\exp^{\lambda X_i}\right] \leq \prod_{i=1}^n \exp^{\lambda^2\sigma^2/2} = \exp^{n\lambda^2\sigma^2/2}.$$

Gaussian errors. Suppose $(\epsilon_i)_{i=1}^n$ are $N(0,\sigma^2)$ errors. Then,

$$\widehat{\theta} - \theta_0 = \frac{1}{n}\sum_{i=1}^n X_i(Y_i - X_i'\theta_0) = \frac{1}{n}\sum_{i=1}^n X_i\epsilon_i$$

Given fixed $(X_i)_{i=1}^n$, the sample average is

$$\frac{1}{n}\textstyle\sum_{i=1}^n X_i\epsilon_i \text{ is a } N(0,\sigma^2 I_p) \text{ vector}$$

- $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\epsilon_i\right] = 0$
- $\text{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\epsilon_i\right) = \frac{1}{n}\sum_{i=1}^n X_iX_i' = I_p$

If $\left\|\frac{1}{n}\sum_{i=1}^n X_i\epsilon_i\right\|_\infty > t$ holds, there must exist $j \in \{1,2,\ldots,p\}$ such that $|\frac{1}{n}\sum_{i=1}^n X_{ij}\epsilon_i| > t$. Union bound implies

$$\begin{aligned}
\Pr\left(\left\|\frac{1}{n}\sum_{i=1}^n X_i\epsilon_i\right\|_\infty > t\right) &\leq \sum_{j=1}^p \Pr\left(\left|\frac{1}{n}\sum_{i=1}^n (X_i)_j\epsilon_i\right| > t\right) \\
&\leq 2p\Pr(\xi > t/\sigma) && \text{(for some r.v. } \xi \sim N(0,1)) \\
&\leq 2p\exp^{-\frac{t^2}{2\sigma^2}} = \delta
\end{aligned}$$

Solving for t in terms of δ, p, σ gives

$$t(\delta) = 2\sigma\sqrt{\log\left(\frac{2p}{\delta}\right)}$$

Therefore, w.p. $1-\delta$,

$$\|\widehat{\theta} - \theta_0\|_\infty \leq 2\sigma\sqrt{\log\left(\frac{2p}{\delta}\right)}$$

3 Examples of Sparsity (Optional)

Example 1: Sparse Models for Earning Regressions. In this example we consider a model for the conditional expectation of log-wage Y_i given education Z_i, measured in years of schooling. We can expand the conditional expectation of wage Y_i given education Z_i:

$$E[Y_i|Z_i] = \sum_{j=1}^p \beta_{0j}P_j(Z_i), \tag{3.1}$$

Sparse Approximation	L_2 error	L_∞ error
Conventional	0.12	0.29
Lasso	0.08	0.12
Post-Lasso	0.04	0.08

Table 1: Errors of Conventional and the Lasso-based Sparse Approximations of the Earning Function. The Lasso method minimizes the least squares criterion plus the ℓ_1-norm of the coefficients scaled by a penalty parameter λ. The nature of the penalty forces many coefficients to zero, producing a sparse fit. The Post-Lasso minimizes the least squares criterion over the non-zero components selected by the Lasso estimator. This example deals with a pure approximation problem, in which there is no noise.

using some dictionary of approximating functions $P(Z_i) = (P_1(Z_i), \ldots, P_p(Z_i))'$, such as polynomial or spline transformations in Z_i and/or indicator variables for levels of Z_i. In fact, since we can consider an overcomplete dictionary, the representation of the function using $P_1(Z_i), \ldots, P_p(Z_i)$ may not be unique, but this is not important for our purposes.

A conventional sparse approximation employed in econometrics is, for example,

$$f(Z_i) := E[Y_i|Z_i] = \tilde{\beta}_1 P_1(Z_i) + \cdots + \tilde{\beta}_s P_s(Z_i) + \tilde{r}_i, \tag{3.2}$$

where the P_j's are low-order polynomials or splines, with typically one or two (linear or linear and quadratic) terms. Of course, there is no guarantee that the approximation error $\tilde{r}_i$ in this case is small or that these particular polynomials form the best possible s-dimensional approximation. Indeed, we might expect the function $E[Y_i|Z_i]$ to change rapidly near the schooling levels associated with advanced degrees, such as MBAs or MDs. Low-degree polynomials may not be able to capture this behavior very well, resulting in large approximation errors $\tilde{r}_i$.

A sensible question is then, "Can we find a better approximation that uses the same number of parameters?" More formally, can we construct a much better approximation of the sparse form

$$f(Z_i) := E[Y_i|Z_i] = \beta_{k_1} P_{k_1}(Z_i) + \cdots + \beta_{k_s} P_{k_s}(Z_i) + r_i, \tag{3.3}$$

for some regressor indices $k_1, \ldots, k_s$ selected from $\{1, \ldots, p\}$? Since we can always include (3.2) as a special case, we can in principle do no worse than the conventional approximation; and, in fact, we can construct (3.3) that is much better, if there are some important higher-order terms in (3.1) that are completely missed by the conventional approximation. Thus, the answer to the question depends strongly on the empirical context.

Consider for example the earnings of prime age white males in the 2000 U.S. Census see, e.g., [**?**]. Treating this data as the population data, we can compute $f(Z_i) = E[Y_i|Z_i]$ without error. Figure 3.1 plots this function. We then construct two sparse approximations and also plot them in Figure 3.1. The first is the conventional approximation of the form (3.2) with $P_1, \ldots, P_s$ representing polynomials of degree zero to $s-1$ ($s = 5$ in this example). The second is an approximation of the form (3.3), with $P_{k_1}, \ldots, P_{k_s}$ consisting of a constant, a linear term, and three linear splines terms with knots located at 16, 17, and 19 years of schooling. We find the latter approximation automatically using the ℓ_1-penalization or Lasso methods discussed below,[1] although in this special case we could construct such an approximation just by eye-balling Figure 3.1 and noting that most of the function is described by a linear function with a few abrupt changes that can be captured by linear spline terms that induce large changes in slope near 17 and 19 years of schooling. Note that an exhaustive search for a low-dimensional approximation in principle requires looking at a very large set of models. Methods for HDS models, such as ℓ_1-penalized least squares (Lasso), which we employed in this example, are designed to avoid this search. □

Example 2: Approximate Sparsity through Smoothness of Target Functions. Approximate sparsity incorporates both substantial generalizations and improvements over the conventional series approximation of regression functions in [**?**]. In order to explain this consider the set $\{P_j(z), j \geq 1\}$ of orthonormal basis functions on $[0,1]^d$, e.g. orthopolynomials, with respect to the Lebesgue measure. Suppose Z_i have a uniform distribution on $[0,1]^d$ for simplicity.[2] Assuming $\mathbb{E}\left[f^2(Z_i)\right] < \infty$, we can represent f via a Fourier expansion, $f(z) = \sum_{j=1}^{\infty} \delta_j P_j(z)$, where $\{\delta_j, j \geq 1\}$ are Fourier coefficients that satisfy $\sum_{j=1}^{\infty} \delta_j^2 < \infty$.

[1] The set of functions considered consisted of 12 linear splines with various knots and monomials of degree zero to four. Note that there were only 12 different levels of schooling.

[2] The discussion in this example continues to apply when Z_i has a density that is bounded from above and away from zero on $[0,1]^d$.

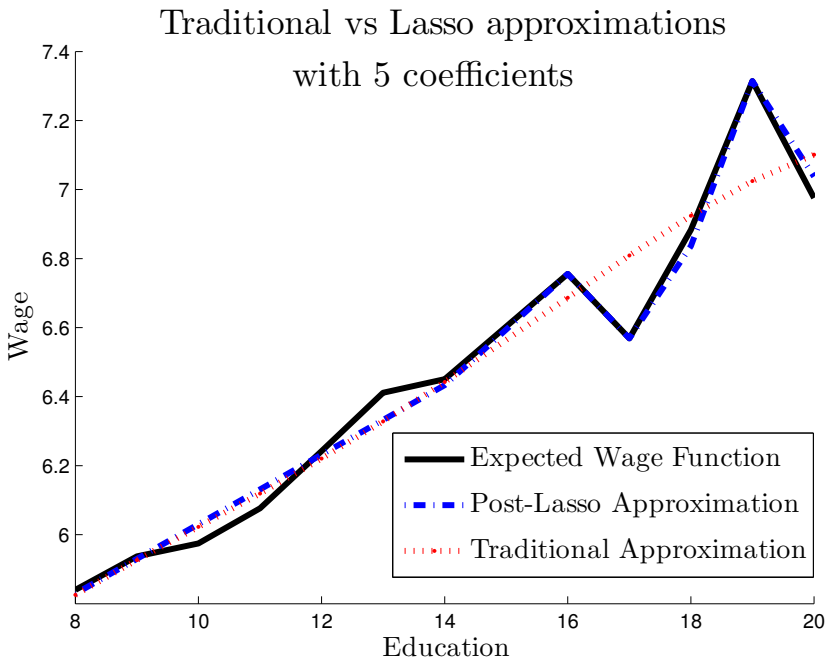

Figure 3.1: The figures illustrates the Post-Lasso sparse approximation and the fourth order polynomial approximation of the wage function.

Let us consider the case that f is a smooth function so that Fourier coefficients feature a polynomial decay $\delta_j \propto j^{-\nu}$, where ν is a measure of smoothness of f. Consider the conventional series expansion that uses the first K terms for approximation, $f(z) = \sum_{j=1}^{K} \beta_{0j} P_j(z) + a_c(z)$, with $\beta_{0j} = \delta_j$. Here $a_c(Z_i)$ is the approximation error which obeys $\sqrt{\mathbb{E}_n[a_c^2(Z_i)]} \lesssim_P \sqrt{\mathbb{E}\left[a_c^2(Z_i)\right]} \lesssim K^{\frac{-2\nu+1}{2}}$. Balancing the order $K^{\frac{-2\nu+1}{2}}$ of approximation error with the order $\sqrt{K/n}$ of the estimation error gives the oracle-rate-optimal number of series terms $s = K \propto n^{1/2\nu}$, and the resulting oracle series estimator, which knows s, will estimate f at the oracle rate of $n^{\frac{1-2\nu}{4\nu}}$. This also gives us the identity of the most important series terms $T = \{1, ..., s\}$, which are simply the first s terms. We conclude that Condition ASM holds for the sparse approximation $f(z) = \sum_{j=1}^{p} \beta_{0j} P_j(z) + a(z)$, with $\beta_{0j} = \delta_j$ for $j \leq s$ and $\beta_{0j} = 0$ for $s+1 \leq j \leq p$, and $a(Z_i) = a_c(Z_i)$, which coincides with the conventional series approximation above, so that $\sqrt{\mathbb{E}_n[a^2(Z_i)]} \lesssim_P \sqrt{s/n}$ and $\|\beta_0\|_0 \leq s$.

Next suppose that Fourier coefficients feature the following pattern $\delta_j = 0$ for $j \leq M$ and $\delta_j \propto (j-M)^{-\nu}$ for $j > M$. Clearly in this case the standard series approximation based on the first $K \leq M$ terms, $\sum_{j=1}^{K} \delta_j f_j(z)$, has no predictive power for $f(z)$, and the corresponding standard series estimator based on the first K terms therefore fails completely.[3] In contrast, Condition AS is easily satisfied in this case, and the Lasso-based estimators will perform at a near-oracle level in this case. Indeed, we can use the first p series terms to form the approximation $f(z) = \sum_{j=1}^{p} \beta_{0j} P_j(z) + a(z)$, where $\beta_{0j} = 0$ for $j \leq M$ and $j > M + s$, $\beta_{0j} = \delta_j$ for $M + 1 \leq j \leq M + s$ with $s \propto n^{1/2\nu}$, and p such that $M + n^{1/2\nu} = o(p)$. Hence $\|\beta_0\|_0 = s$, and we have that $\sqrt{\mathbb{E}_n[a^2(Z_i)]} \lesssim_P \sqrt{\mathbb{E}\left[a^2(Z_i)\right]} \lesssim \sqrt{s/n} \lesssim n^{\frac{1-2\nu}{4\nu}}$. □

Example 3. Individual Effects are not Sparse Consider individual effects a_i in the linear panel models:

$$Y_{it} = a_i + X_{it}'\beta + \epsilon_{it}.$$

We can think of $a_i's$ as random variables generated according to some non-degenerate distribution, so the approximate sparsity does not hold (the mass of a_i's close to zero is too big to follow a polynomial decay). One approach is to simply avoid penalizing fixed effects and estimate them directly (this is equivalent to still doing de-meaning or taking the first differences and then applying lasso). Another approach is to think of other ways of building models that

[3] This is not merely a finite sample phenomenon but is also accommodated in the asymptotics since we expressly allow for array asymptotics; i.e. the underlying true model could change with n. Recall that we omit the indexing by n for ease of notation.

have some sparsity relative to a sensible benchmark. For example, if we follow Mundlack's correlated random effects approach, then we can model

$$a_i = \bar{X}_i'\lambda + v_i,$$

where v_i is treated as a part of the regression approach, and λ is treated as approximately sparse. This generates a plausible a_i that are dense, yet having a parsimonious form relative to a benchmark.

A Appendix

Proof of Lemma 1. For each coordinate $j = 1, 2, \ldots, p$

$$|\hat{\theta}_j| > 2\rho \quad \Rightarrow |\theta_{0,j}| > |\hat{\theta}_j| - \rho > \rho \tag{A.1}$$

$$|\hat{\theta}_j| \leq 2\rho \quad \Rightarrow |\theta_{0,j}| \leq |\hat{\theta}_j| + \rho \leq 3\rho \tag{A.2}$$

Therefore, for every coordinate j

$$\begin{aligned}
|\hat{\theta}_j^{\mathrm{HRD}} - \theta_{0,j}| &= |\hat{\theta}_j - \theta_{0,j}|1_{\{|\hat{\theta}_j|>2\rho\}} + |\theta_{0,j}|1_{\{|\hat{\theta}_j|\leq 2\rho\}} \\
&\leq \rho 1_{\{|\hat{\theta}_j|>2\rho\}} + |\theta_{0,j}|1_{\{|\hat{\theta}_j|\leq 2\rho\}} && \text{(by choice of } \rho) \\
&\leq \rho 1_{\{|\theta_{0,j}|>\rho\}} + |\theta_{0,j}|1_{\{|\theta_{0,j}|<3\rho\}} && \text{(from (1) and (2))} \\
&\leq 4\min(|\theta_{0,j}|, \rho).
\end{aligned}$$

Summing over $j = 1, 2, \ldots, j$ gives

$$\|\hat{\theta}^{\mathrm{HRD}} - \theta_0\|_2^2 = \sum_{j=1}^{p}(\hat{\theta}_j^{\mathrm{HRD}} - \theta_{0,j})^2 \leq \sum_{j\in T} 16\rho^2 + \sum_{j\in T^c} 0 = 16s\rho^2.$$

Therefore, thresholding at ρ overcomes the curse of dimensionality if s is small.
To see (b), note that if $\theta_{0j} \neq 0$, then $|\theta_{0j}| > 3\rho$ so that

$$|\hat{\theta}_j^{\mathrm{HRD}}| \geq |\theta_{0j}| - \|\hat{\theta} - \theta_0\| > 3\rho - \rho \geq 2\rho.$$

Therefore, $\hat{\theta}_j^{\mathrm{HRD}} \neq 0$, and $\mathrm{supp}(\theta) \subseteq \mathrm{supp}(\hat{\theta}^{\mathrm{HRD}})$. On the other hand, if $\hat{\theta}_j^{\mathrm{HRD}} \neq 0$, it must be that $|\hat{\theta}_j| > 2\rho$, and $|\theta_{0j}| > \rho > 0$. Therefore, $\mathrm{supp}(\hat{\theta}^{\mathrm{HRD}}) \subseteq \mathrm{supp}(\hat{\theta})$.

□

Proof of Lemma 3. The first step is the same as in Lemma 1. For each coordinate $j = 1, 2, \ldots, p$

$$|\hat{\theta}_j| > 2\rho \quad \Rightarrow |\theta_{0,j}| > |\hat{\theta}_j| - \rho > \rho \tag{A.3}$$

$$|\hat{\theta}_j| \leq 2\rho \quad \Rightarrow |\theta_{0,j}| \leq |\hat{\theta}_j| + \rho \leq 3\rho \tag{A.4}$$

Therefore, for every coordinate j

$$\begin{aligned}
|\hat{\theta}_j^{\mathrm{SFT}} - \theta_{0,j}| &= |\hat{\theta}_j - \theta_{0,j} - 2\rho|1_{\{\hat{\theta}_j>2\rho\}} + |\theta_{0,j}|1_{\{|\hat{\theta}_j|\leq 2\rho\}} \\
&\quad + |\hat{\theta}_j - \theta_{0,j} + 2\rho|1_{\{\hat{\theta}_j<-2\rho\}} \\
&\leq 3\rho 1_{\{|\hat{\theta}_j|>2\rho\}} + |\theta_{0,j}|1_{\{|\hat{\theta}_j|\leq 2\rho\}} && \text{(by choice of } \rho) \\
&\leq 3\rho 1_{\{|\theta_{0,j}|>\rho\}} + |\theta_{0,j}|1_{\{|\theta_{0,j}|<3\rho\}} && \text{(from (1) and (2))} \\
&\leq 6\min(|\theta_{0,j}|, \rho).
\end{aligned}$$

Summing over $j = 1, 2, \ldots, j$ gives

$$\|\hat{\theta}^{\mathrm{SFT}} - \theta_0\|_2^2 = \sum_{j=1}^{p}(\hat{\theta}_j^{\mathrm{SFT}} - \theta_{0,j})^2 \leq \sum_{j\in T} 36\rho^2 + \sum_{j\in T^c} 0 = 36s\rho^2.$$

□