Can Machines Learn Weak Signals?

Zhouyu Shen & Dacheng Xiu

Chicago Booth working paper, 2024

Presented by John Walker

All mistakes my own, all smart sounding ideas probably Frank's

Background - High-dimensional covariate regression

Consider a standard regression setting.

$$\underset{n\times 1}{\mathrm{Y}} = \underset{n\times p}{X} \times \underset{p\times 1}{\beta_0} + \underset{n\times 1}{\varepsilon} \tag{1}$$

Suppose that X is high-dimensional, that is, p and n are of similar magnitude, or $p \gg n$.

This might occur because

- Many coviarates are available and we want to control for them
- We introduce a dictionary of transformations / technical regressors to manage non-linearoty

OLS does a bad job of prediction in these high-dimensional settings because it **overfits** in finite samples

Penalized least squares estimators

Penalized least squares estimators add a penalty to the OLS problem to encourage regularized (less complex/shrunk) solutions

Penalized least squares estimators

Penalized least squares estimators add a penalty to the OLS problem to encourage regularized (less complex/shrunk) solutions

LASSO estimator - performs ℓ_1 regularization, penalizing by the sum of the absolute values of the coefficients

$$\hat{\beta}_{\text{lasso}}(\lambda) = \arg\min_{\beta} \left(\frac{1}{n} \|y - X\beta\|^2 + \frac{p}{n} \lambda_n \|\beta\|_1 \right)$$

Penalized least squares estimators

Penalized least squares estimators add a penalty to the OLS problem to encourage regularized (less complex/shrunk) solutions

LASSO estimator - performs ℓ_1 regularization, penalizing by the sum of the absolute values of the coefficients

$$\hat{\beta}_{\text{lasso}}(\lambda) = \arg\min_{\beta} \left(\frac{1}{n} \|y - X\beta\|^2 + \frac{p}{n} \lambda_n \|\beta\|_1 \right)$$

Ridge estimator - performs ℓ_2 regularization, penalizing by the sum of least squares of the coefficients

$$\hat{\beta}_{\text{ridge}}(\lambda) = \arg\min_{\beta} \left(\frac{1}{n} \|y - X\beta\|^2 + \frac{1}{\sqrt{n}} \lambda_n \|\beta\|^2 \right)$$

Penalized least squares estimators

Penalized least squares estimators add a penalty to the OLS problem to encourage regularized (less complex/shrunk) solutions

LASSO estimator - performs ℓ_1 regularization, penalizing by the sum of the absolute values of the coefficients

$$\hat{\beta}_{\text{lasso}}(\lambda) = \arg\min_{\beta} \left(\frac{1}{n} \|y - X\beta\|^2 + \frac{p}{n} \lambda_n \|\beta\|_1 \right)$$

Ridge estimator - performs ℓ_2 regularization, penalizing by the sum of least squares of the coefficients

$$\hat{\beta}_{\text{ridge}}(\lambda) = \arg\min_{\beta} \left(\frac{1}{n} \|y - X\beta\|^2 + \frac{1}{\sqrt{n}} \lambda_n \|\beta\|^2 \right)$$

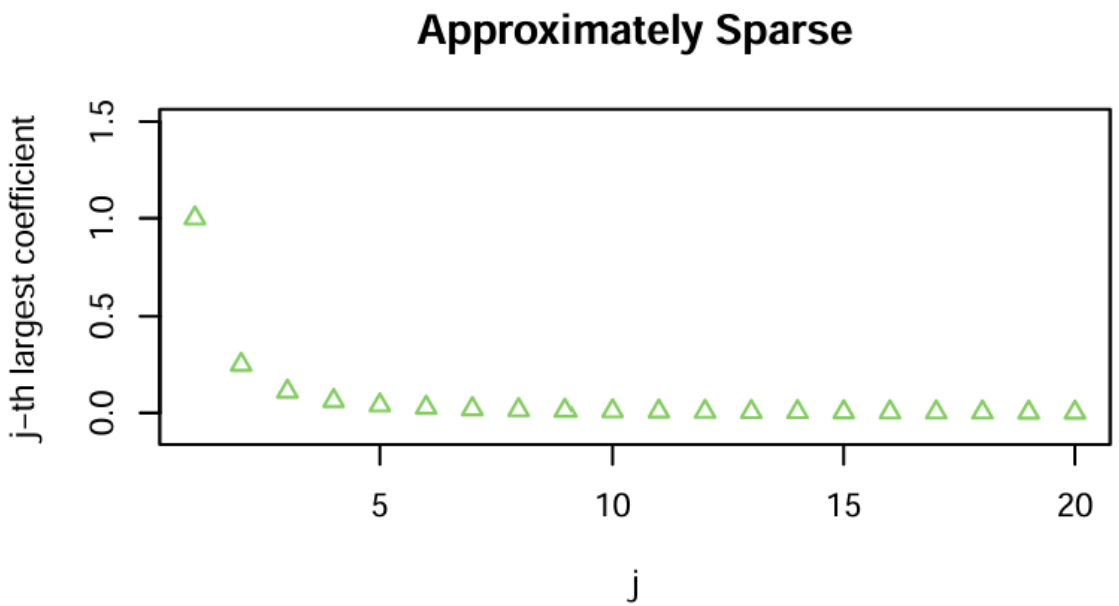
Comparison: Ridge penalizes the large values of coefficients much more aggressively and small values much less aggressively

When do they work? LASSO

Existing theory says LASSO performs well in **sparse** settings[1]

A characterization of approximate sparsity is that the sorted values of the coefficients on standardised regressors decay sufficiently fast

$$|\beta|_j \leq Aj^{-a}, \quad a > 1/2$$

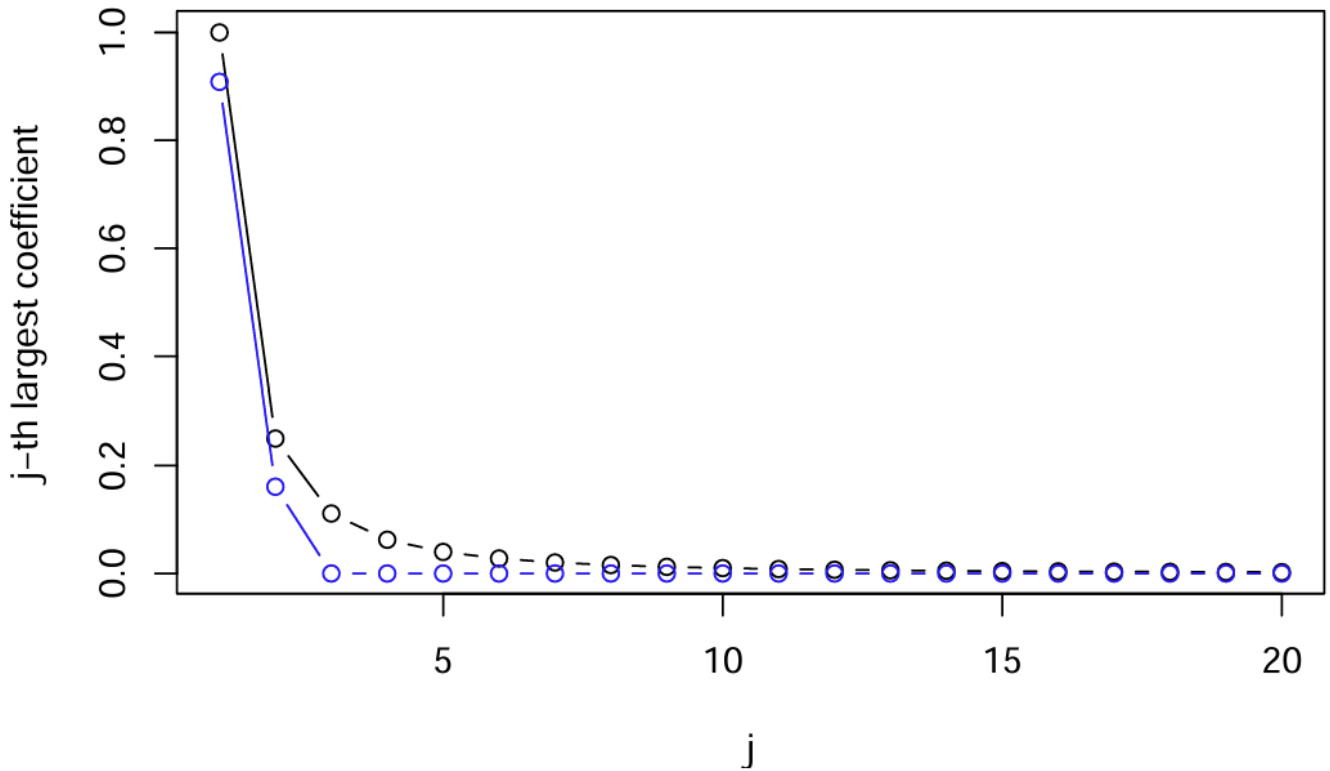


[1] see e.g. Zou (2006), Zhao and Yu (2006), Bickel et al. (2009), Bellon et al. (2013). The asymptotic theory additionally requires the restricted eigenvalue condition

What is LASSO doing?

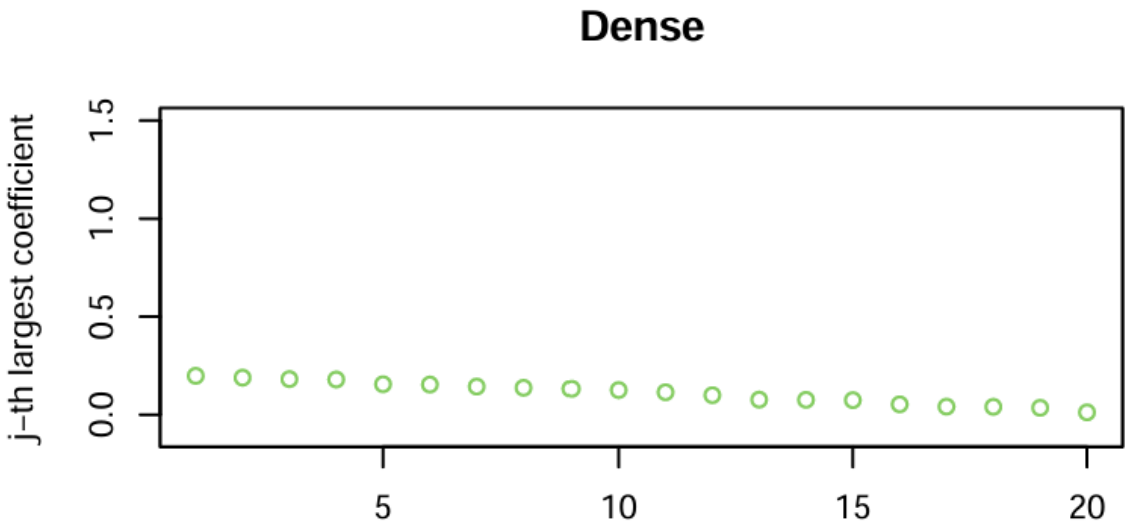
LASSO performs **variable selection**. It shrinks all coefficients, and sets the value of small coefficients to zero

The penalty function has a kink at zero, so there is a positive marginal cost of inclusion of additional regressors - it drops them if they are lower than this threshold



When do they work? Ridge

By contrast, Ridge performs well in **dense** settings. It shrinks all estimates but doesn't do variable selection.[2]



There are other PLS estimators that work well in both settings, e.g. Elastic net (ensemble of LASSO & Ridge; SCAD; LAVA)

[2]The convergence rate of Ridge's prediction error requires intricate conditions on the eigenvalue structure of the design matrix, see Tsigler and Bartlett (2023)

Motivation - Sparsity isn't sufficient

"The idea that economic data are informative enough to identify sparse predictive models might simply be an illusion"
- (Giannone et al., 2021)

Motivation - Sparsity isn't sufficient

"The idea that economic data are informative enough to identify sparse predictive models might simply be an illusion"
- (Giannone et al., 2021)

Simulation evidence indicates that LASSO variable selection performance is poor under weak signals (Wang et al., 2020)

Motivation - Sparsity isn't sufficient

"The idea that economic data are informative enough to identify sparse predictive models might simply be an illusion"
- (Giannone et al., 2021)

Simulation evidence indicates that LASSO variable selection performance is poor under weak signals (Wang et al., 2020)

Research question: Can we characterize LASSO's poor performance in these settings?

Motivation - Sparsity isn't sufficient

"The idea that economic data are informative enough to identify sparse predictive models might simply be an illusion"
- (Giannone et al., 2021)

Simulation evidence indicates that LASSO variable selection performance is poor under weak signals (Wang et al., 2020)

Research question: Can we characterize LASSO's poor performance in these settings?

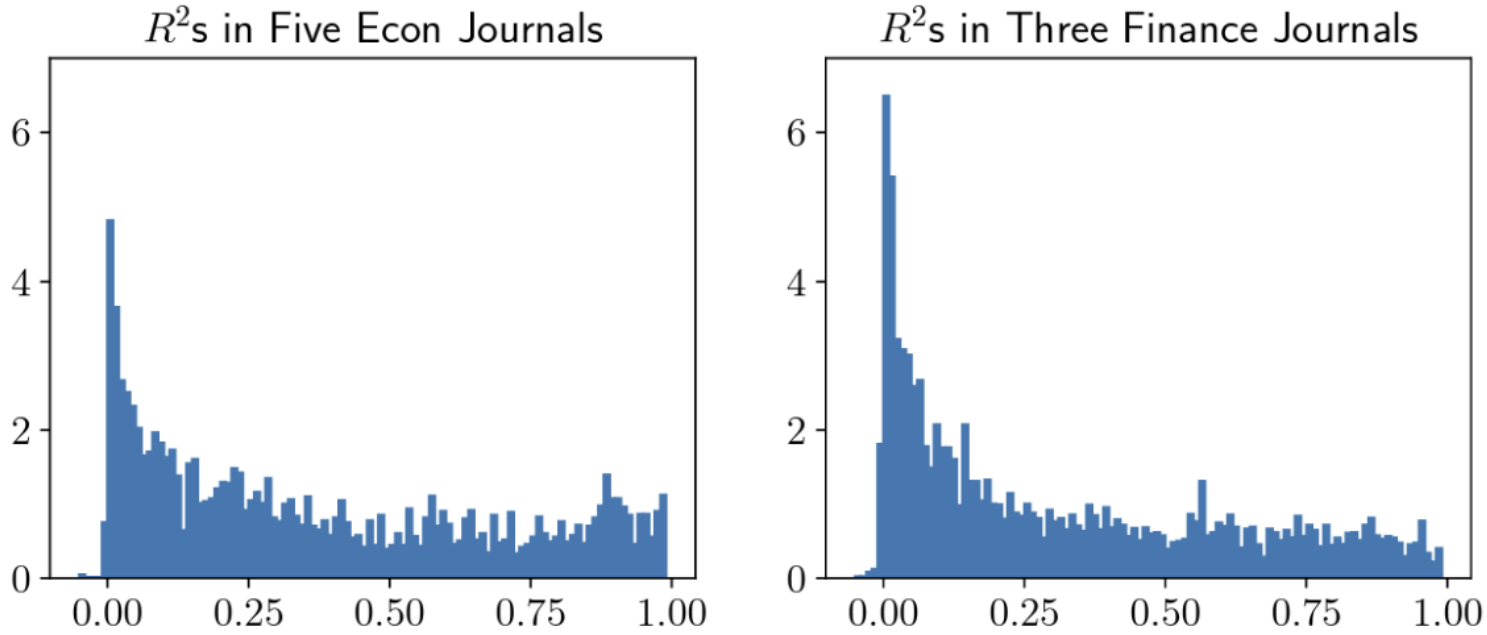
Main contributions of this paper

- Provides a new theoretical analysis of the asymptotic behaviour of LASSO and Ridge in a weak signal setting
- The results are under a unified framework, allowing comparison
- Shows that LASSO performs poorly in weak signal environment (not just non-sparse environments)
  - Verifies the results in simulations

Motivation - Econ models predict poorly

Weak signal environments are **common** in economics and finance applications, so this setting is important

Figure 1: Histograms of R^2s in Selected Economics and Finance Journals



Lots of mass near zero and the 25% quantiles of these R^2 values stand at 9.7% for economics and 5.8% for finance.

Setting and & signal strength

Back to the high-dimensional linear regression setting in which all covariates are standardised and we allow $p \gg n$

$$\underset{n\times 1}{\mathrm{Y}} = \underset{n\times p}{X} \times \underset{p\times 1}{\beta_0} + \underset{n\times 1}{\varepsilon} \tag{2}$$

Setting and & signal strength

Back to the high-dimensional linear regression setting in which all covariates are standardised and we allow $p \gg n$

$$\underset{n\times 1}{\mathrm{Y}} = \underset{n\times p}{X} \times \underset{p\times 1}{\beta_0} + \underset{n\times 1}{\varepsilon} \tag{2}$$

A **true signal** is a covariate with a non-zero coefficent in the population model, while a **false signal** has coefficient zero

Setting and & signal strength

Back to the high-dimensional linear regression setting in which all covariates are standardised and we allow $p \gg n$

$$\underset{n\times 1}{\mathrm{Y}} = \underset{n\times p}{X} \times \underset{p\times 1}{\beta_0} + \underset{n\times 1}{\varepsilon} \tag{2}$$

A true signal is a covariate with a non-zero coefficent in the population model, while a false signal has coefficient zero

In **population**, the distinction is clear, but under sampling uncertainty very small non-zero covariates are not well-separated from true zeros

Weak signals

Two potential characterizations of a **weak signal**

1. **Inference/Signal detection** - Below some **detection boundary**, we cannot detect differences from zero
   - Ingster et al. (2010), Cui et al. (2018), and Li et al. (2020) provide tests for coefficient separation from zero

Weak signals

Two potential characterizations of a weak signal

1. Inference/Signal detection - Below some detection boundary, we cannot detect differences from zero
   - Ingster et al. (2010), Cui et al. (2018), and Li et al. (2020) provide tests for coefficient separation from zero
2. **Prediction** - A definition in terms of signal strength (this paper)

Weak signals

Two potential characterizations of a weak signal

1. Inference/Signal detection - Below some detection boundary, we cannot detect differences from zero
   - Ingster et al. (2010), Cui et al. (2018), and Li et al. (2020) provide tests for coefficient separation from zero
2. Prediction - A definition in terms of signal strength (this paper)

A **weak signal** satisfies

$$\|\beta_0\|^2 \asymp_P \tau \to 0$$

In words, the ℓ_2 norm of the linear contribution of the covariates is asymptotically equivalent in probability to the signal strength, τ, which is vanishingly small.

Setting: assumptions on the DGP

Assumption 1. *The covariates $X \in \mathbb{R}^{n \times p}$ are generated as $X = \Sigma_1^{1/2} Z \Sigma_2^{1/2}$ for an $n \times p$ matrix Z with i.i.d. standard Gaussian entries, deterministic $n \times n$ and $p \times p$ positive definite matrices Σ_1 and Σ_2.[2] In addition, there exist positive constants c_1, C_1, c_2, C_2 such that $c_1 \leq \lambda_i(\Sigma_1) \leq C_1$, $i = 1, 2, \ldots, n$ and $c_2 \leq \lambda_i(\Sigma_2) \leq C_2$, $i = 1, 2, \ldots, p$.*

- The correlation structure of X is very general, with Σ_1 governing cross-sectional dependencies and Σ_2 auto-correlational dependencies
- The upper bounds on the eigenvalues rules out very strong dependencies. The lower bounds rule out multi-collinearity and linear dependence of an observation at time t on observations at times $\neg t$.
- The proof accomodates stochastic Σ_1 and Σ_2 if their entries are mutually independent and independent of Z
- The Gaussian process assumption is crucial to the proof via Gordon's inequality, but in simulation appears unnecessary

Assumptions on ϵ

Assumption 2. *Let $\varepsilon = \Sigma_\varepsilon^{1/2} z$, where z comprises i.i.d. variables with mean zero, variance one and finite fourth moment and Σ_ε is a positive semidefinite matrix satisfying $c_\varepsilon \leq \lambda_i(\Sigma_\varepsilon) \leq C_\varepsilon$, $i = 1, 2, \ldots, n$, for some fixed positive constants c_ε and C_ε.*

- These assumptions ensure that the spectral norm of Σ_ϵ is bounded
- Under A1, we have $||X\beta_0|| \asymp_P \sqrt{n}||\beta_0||$. The entries in X do not explode or vanish asymptotically
- Under A2, we have $||\epsilon|| \asymp_P \sqrt{n}$. The entries in ϵ do not explode or vanish asymptotically.
- The signal-to-noise ratio (prediction R^2) is then simply $||\beta_0||$.

Benchmarking - Bayes prediction risk

Bayes risk is an average case analysis in which we take the expected squared prediction error evaluated at a new, independent data point (x_{new}, y_{new}) and integrate it over the prior distribution F

Under linearity, we can write

$$\begin{aligned} E_F\left((y_{\text{new}} - \hat{y}_{\text{new}})^2\right) &= \sigma_\epsilon^2 + E_F\left((x_{\text{new}})^\top(\hat{\beta} - \beta_0)\right)^2 \\ &= \sigma_\epsilon^2 + E_F\left[E\left(\left((x_{\text{new}})^\top(\hat{\beta} - \beta_0)\right)^2 \middle| X, y, \beta_0\right)\right] \\ &= \sigma_\epsilon^2 + E_F\left\|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\right\|^2 \end{aligned}$$

σ_ϵ^2 is not a function of $\hat{\beta}$ we can define the Bayes prediction risk

$$R(\hat{\beta}, F) := E_F\left\|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\right\|^2$$

A class of functions (priors)

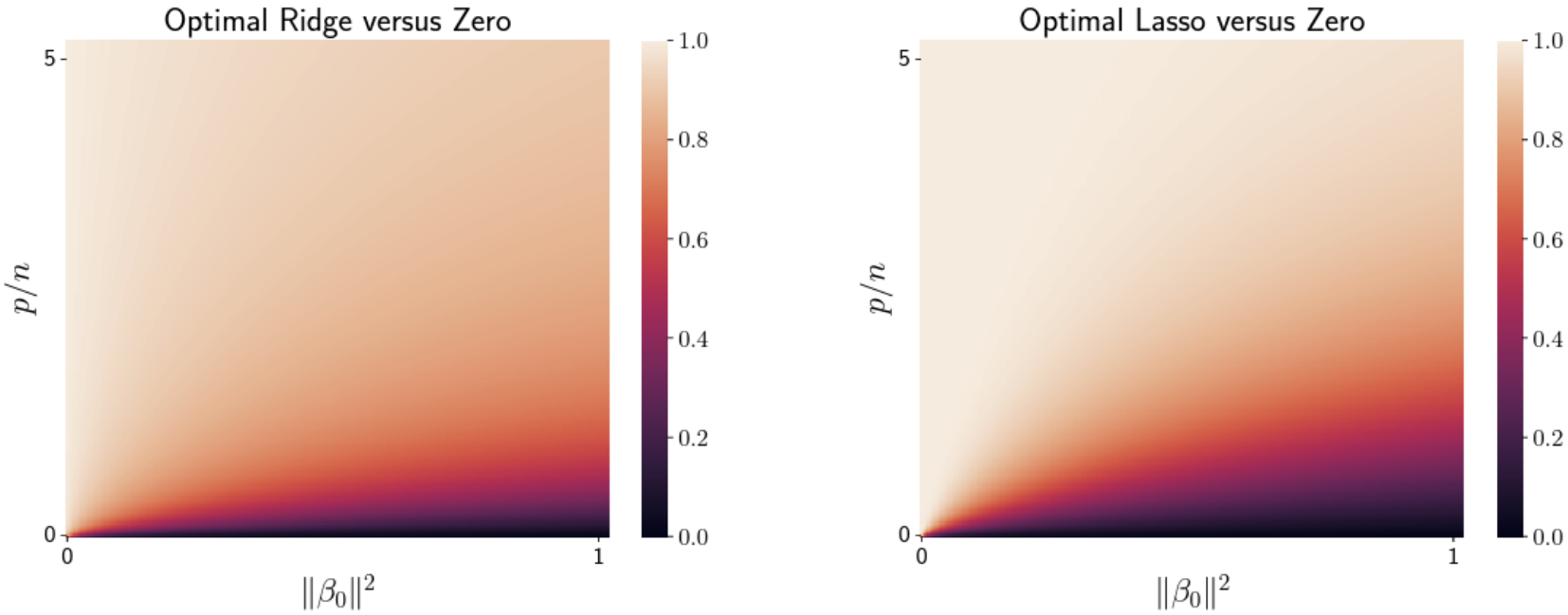
Assumption 3. *The vector* $b_0 = \sqrt{p\tau^{-1}}\beta_0$ *comprises i.i.d. random variables, each following a prior probability distribution* F *belonging to the class* $\mathcal{F}$*. The class* $\mathcal{F}$ *is defined such that any included random variable can be represented as* $q^{-1/2}b_1b_2$*, where* b_1 *and* b_2 *are independent,* b_1 *follows a binomial distribution* $B(1, q)$*, and* b_2 *is a sub-exponential random variable with a mean of zero and a variance denoted as* σ_β^2*.*

- The paper wants to avoid making strong priors on β_0, so shows a set of results for a class $\mathcal{F}$ of priors satisfying $q^{-1/2}b_1b_2$
- $q^{-1/2}b_1b_2$ accomodates a "spike-and-slab prior" as in Giannone et al. (2022) in which b_2 is modeled by $(1 - v)\psi_0 + v\psi_1$
- The constant v modulates the distribution between the spike ψ_0 and slab (ψ_1) and q dictates the degree of sparsity of β_0 - a near 0 q is sparse (favouring LASSO), but a large q benefits Ridge
- The iid assumption is **strong** - the authors claim it simplifies the math and imply it isn't necessary (without proof)
- The $\sqrt{\frac{p}{\tau}}$ re-scaling is innocuous and helps with interpretation

High signal environments

The asymptotic Bayes prediction risk of LASSO and Ridge with $p/n \to c_0 \in \mathbb{R}^+$ under strong signals ($||\beta_0||^2 \asymp_P \tau = 1$), Bayes risk minimizing tuning parameters, and $\{\Sigma_1, \Sigma_2, \Sigma_\epsilon\} = \mathbb{I}$ is known.[3]

The behaviour for $\tau \to 0$, $p/n \not\to 0$ is not well understood.



[3] Dicker (2016) and Dobriban & Wager (2018) derive asymptomptotic behaviour of the Bayes risk for Ridge, and Bayati & Montanri (2012) & Thrampoulidis et al. (2018) for LASSO

Zero's optimality

The authors then specify the asymptotics of the setting.[4] The first four constraints exclude areas near the origin and positive infinity. The final constraint imposes a **rate of decay** on signal strength τ. This is important because they show that LASSO does poorly **below** this rate

Assumption 4. $\tau \to 0$, $n^{-1}p \to c_0 \in (0, \infty]$, $n^{-1}\tau p(\log p)^4 \to 0$, $n\tau p^{-2/3}(\log p)^{-4} \to \infty$, *and* $n^{-1}pq\tau^{-1}(\log p)^{-4} \to \infty$.

[4] In **Assumption 5.** they impose additional conditions on $\Sigma_1, \Sigma_2, \Sigma_\epsilon$

Zero's optimality

The authors then specify the asymptotics of the setting.[4] The first four constraints exclude areas near the origin and positive infinity. The final constraint imposes a rate of decay on signal strength τ. This is important because they show that LASSO does poorly below this rate

Assumption 4. $\tau \to 0$, $n^{-1}p \to c_0 \in (0, \infty]$, $n^{-1}\tau p(\log p)^4 \to 0$, $n\tau p^{-2/3}(\log p)^{-4} \to \infty$, *and* $n^{-1}pq\tau^{-1}(\log p)^{-4} \to \infty$.

Definition: the estimator $\hat{\beta}$ is asymptotically optimal relative to F, if its asymptotic risk is equal to that of the Bayes risk minimizing estimator (the Bayes predicter, here $E_F(\beta_0|X, y)$)

$$\lim_{n\to\infty} \frac{\mathcal{R}(\hat{\beta}, f)}{\mathcal{R}(F)} = 1$$

Theorem 1. *Assume that Assumptions 1–4 hold. Furthermore, assume that the error term ε in Assumption 2 follows a Gaussian distribution.[7] Under these conditions, the zero estimator is asymptotically optimal relative to any distribution $F \in \mathcal{F}$.*

[4] In **Assumption 5.** they impose additional conditions on $\Sigma_1, \Sigma_2, \Sigma_\epsilon$

Benchmarking - Zero estimator

- The authors benchmark against this **zero estimator**, which just sets the entire vector $\hat{\beta}$ to zero
- Under a sufficiently high penalization, at the limit, we expect LASSO and Ridge to achieve the zero benchmark in Bayes risk

Benchmarking - Zero estimator

- The authors benchmark against this zero estimator, which just sets the entire vector $\hat{\beta}$ to zero
- Under a sufficiently high penalization, at the limit, we expect LASSO and Ridge to achieve the zero benchmark in Bayes risk
- The authors therefore introduce a different benchmark: the **relative prediction error**, $\Delta(\hat{\beta}(\lambda_n))$

$$\Delta(\hat{\beta}) = pn^{-1}\tau^{-2}\left(\|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2\right) \tag{3}$$

Benchmarking - Zero estimator

- The authors benchmark against this zero estimator, which just sets the entire vector $\hat{\beta}$ to zero
- Under a sufficiently high penalization, at the limit, we expect LASSO and Ridge to achieve the zero benchmark in Bayes risk
- The authors therefore introduce a different benchmark: the relative prediction error, $\Delta(\hat{\beta}(\lambda_n))$

$$\Delta(\hat{\beta}) = pn^{-1}\tau^{-2}\left(\|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2\right) \tag{3}$$

- This object will allow for meaningful inference at the limit
- $\Delta(\hat{\beta}) < 0$ implies that $\hat{\beta}$ outperforms zero
- As usual, we need an appropriate scaling so that the limiting distribution tells us something interesting. Here it is $\frac{p}{n} \times \frac{1}{\tau^2}$

Ridge - a precise error limit

Theorem 2. *Assuming that Assumptions 1–5 hold, and setting* $\lambda_n = \tau^{-1}\lambda$, *we establish the following convergence result:*

$$\Delta(\hat{\beta}_r(\lambda_n)) \xrightarrow{\text{P}} \alpha^* := 2\theta_2\sigma_x^4\left(\frac{\sigma_\varepsilon^2\theta_1}{2\lambda^2} - \frac{\sigma_\beta^2}{\lambda}\right).$$

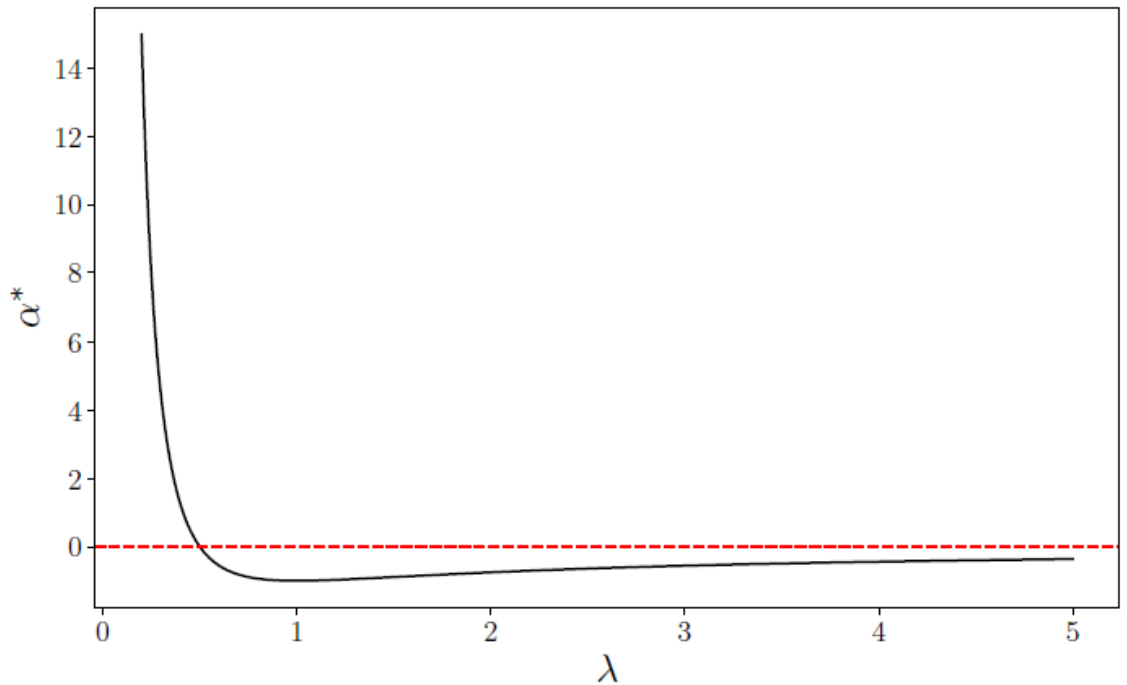
A few **consequences**

- Minimizing α^* wrt provides an optimal tuning parameter $\lambda_n^* = \tau^{-1}\sigma_\epsilon^2\theta_1/\sigma_\beta^2$
- α^* is negative for λ^*, indicating Ridge learns weak signals well
- $\alpha^* \to 0^-$ as $\lambda \to \infty$, so Ridge consistently **outperforms** the zero estimator as it approaches the limit
- As $\lambda \to 0$, $\alpha^* \to \infty$. That is, the Ridgeless estimator ($\lambda = 0$) is outperformed by the zero estimator

Ridge - Graphical performance

A graphical representation of the behaviour of $\Delta\hat{\beta}_r(\lambda_n)) \to 0^{-1}$

Figure 3: Ridge vs. Zero Estimator's Relative Precise Error



Aside: cross-validation of λ works

(K-fold) **Cross-validation** is a method of hyper-parameter selection that involves splitting the data into K folds estimating performance, then averaging across folds and choosing the value that maximises performance

- In strong signal settings, K-fold cross-validation is an effective way of choosing λ for Ridge (Hastie et al., 2022)
- Under weak signals, λ^* is often diverging in n, and the rate of divergence depends on (unknown) τ

Aside: cross-validation of λ works

(K-fold) Cross-validation is a method of hyper-parameter selection that involves splitting the data into K folds estimating performance, then averaging across folds and choosing the value that maximises performance

- In strong signal settings, K-fold cross-validation is an effective way of choosing λ for Ridge (Hastie et al., 2022)
- Under weak signals, λ^* is often diverging in n, and the rate of divergence depends on (unknown) τ

Under relatively **strong assumptions**, the authors recover optimality of cross-validated Ridge for **weak signals**

Theorem 3. *Under the same assumptions as in Theorem 2, if we also assume that* $\Sigma_1 = \mathbb{I}$, $\Sigma_\varepsilon = \sigma_\varepsilon^2 \mathbb{I}$, ε *follows a sub-exponential distribution, and that* $q^{-1}\tau^{-1}n^{-1/2}\log(p)$, $q^{-1/2}\tau^{-3/2}n^{-1/2}\log(p) \to 0$, *then we can establish that:*

$$\tau\hat{\lambda}_n^{K-CV} \xrightarrow{\text{P}} \lambda^{opt} = \sigma_\varepsilon^2/\sigma_\beta^2.$$

Main result - LASSO asymptotics

The authors provide probability bounds on the asymptotic behaviour of LASSO (not the precise error)

Theorem 4. *Assume that Assumptions 1–5 are satisfied and the tuning parameter λ_n is chosen such that the following equation holds for some $C_\lambda > 0$:*

$$pn^{-2}\tau^{-2}\mathbb{E}_{U\sim\mathcal{N}(0,\Sigma_2)}\left\|(2\sigma_\varepsilon\sqrt{\theta_1}|U| - \lambda_n)_+\right\|^2 = C_\lambda.^9 \tag{9}$$

Then, with probability approaching one, we have $c_\alpha \leq \Delta(\hat{\beta}_l(\lambda_n)) \leq C_\alpha$, where c_α and C_α are the solutions to the following equation in terms of x:

$$x - \sqrt{\frac{2C_\lambda}{c_2}x} = -\frac{C_\lambda}{100C_2}, \tag{10}$$

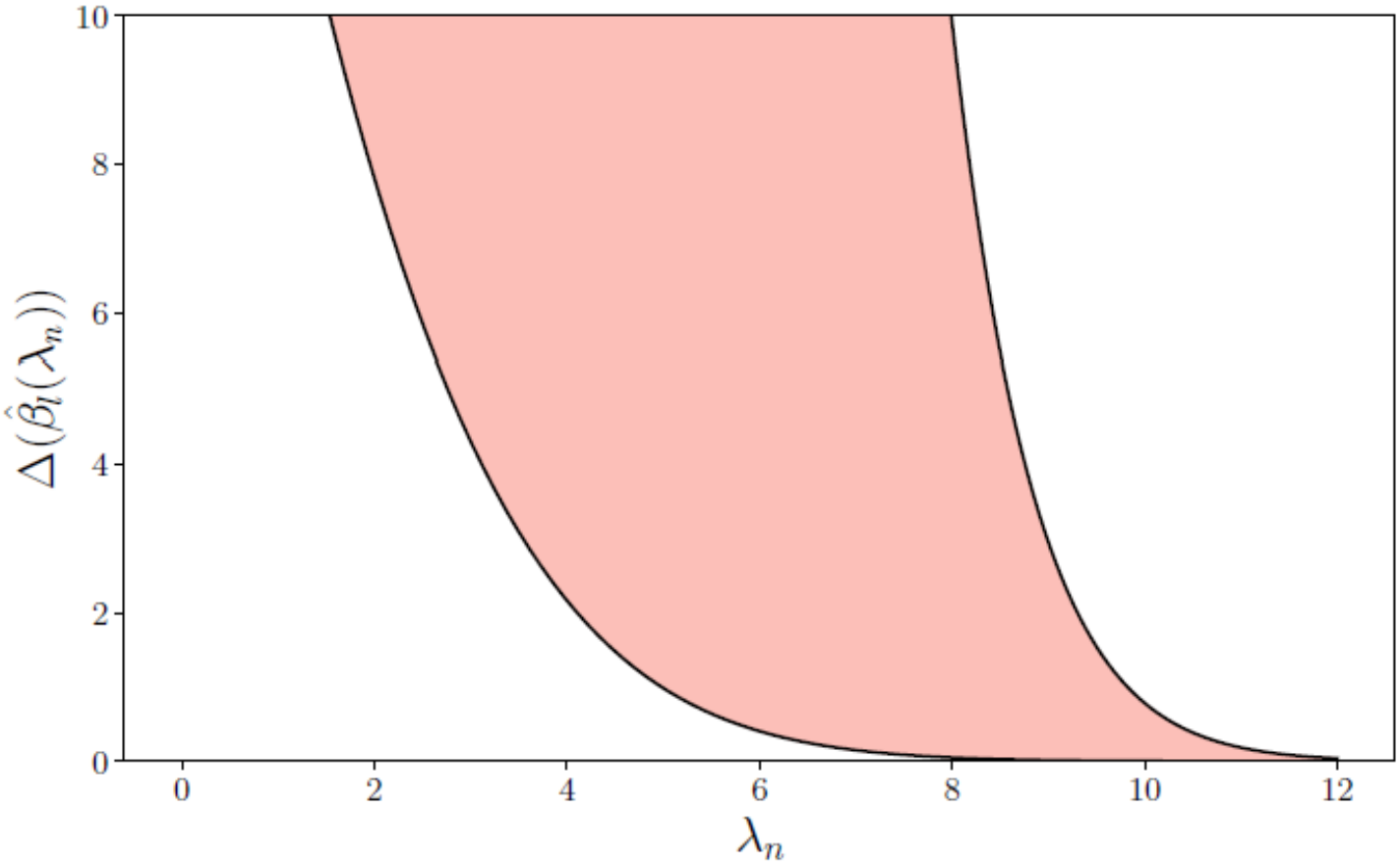
where c_2 and C_2 are constants defined in Assumption 1.

C_α and c_α are non-negative, so LASSO does not outperform the zero estimator for any tuning parameter value as $\{C_\alpha, c_\alpha\} \to 0^+$

LASSO performance - the punchline

The bounds are sufficient to show that Lasso cannot outperform zero for any value of λ in the weak signal setting

Figure 4: Lasso vs. Zero Estimator: Relative Precise Error Bounds



Implications of LASSO's breakdown

- Lasso cannot **effectively distinguish between genuine and spurious signals in the weak signal setting**
- It breaks down unless the data are extremely sparse (to a degree violating Assumption 4)
- "While a sufficiently large tuning parameter could address this problem, our theory suggests that only when the penalty is so substantial that Lasso becomes identical to the zero estimator does it enforce an appropriate penalty."
- This finding has significant implications, especially in areas like economics and finance, where large-scale regression analyses are commonplace, and signal-to-noise ratios tend to be low
- In this setting, elastic net is unlikely to outperform Ridge (as it mixes LASSO and Ridge, and LASSO is dominated)

Theorem 4 - some interpretation

Theorem 4. *Assume that Assumptions 1–5 are satisfied and the tuning parameter λ_n is chosen such that the following equation holds for some $C_\lambda > 0$:*

$$pn^{-2}\tau^{-2}\mathbb{E}_{U\sim\mathcal{N}(0,\Sigma_2)}\left\|(2\sigma_\varepsilon\sqrt{\theta_1}|U| - \lambda_n)_+\right\|^2 = C_\lambda.^9 \tag{9}$$

Then, with probability approaching one, we have $c_\alpha \leq \Delta(\hat{\beta}_l(\lambda_n)) \leq C_\alpha$, where c_α and C_α are the solutions to the following equation in terms of x:

$$x - \sqrt{\frac{2C_\lambda}{c_2}x} = -\frac{C_\lambda}{100C_2}, \tag{10}$$

where c_2 and C_2 are constants defined in Assumption 1.

- For fixed $C_\lambda > 0$, we can solve (9) for λ_n and (10) for $\{C_\alpha\ c_\alpha\}$
- As $\lambda \to 0^+$, $C_\lambda \to \infty$, causing the bounds to diverge
- As λ increases, $\{C_\alpha\ c_\alpha\} \to 0^+$, so LASSO converges towards the performance of the zero estimator

Characterising low signal environments

In-sample R^2 is prone to over-fitting, but out-of-sample R^2 is a common metric of the signal-to-noise ratio. It is consistent if $\hat{\beta}$ is consistent wrt $\hat{\beta}_0$: $||\hat{\beta} - \beta_0|| = o_P(1)$

$$R^2_{\text{oos}}\left(\hat{\beta}\right) = 1 - \frac{\sum_{i\in\text{OOS}}\left(y_i - X_i\hat{\beta}\right)^2}{\sum_{i\in\text{OOS}} y_i^2}$$

In the weak signal case, the outcome depends on the choice of estimator.

Result: Ridge R^2_{oos} does work for assessing the signal-to-noise ratio.

Proposition 1. *Under the same assumptions as Theorem 3, and assuming that the out-of-sample data follows the same DGP as the in-sample data, if $n_{oos}p^{-2}n^2\tau^2 \to \infty$, where n_{oos} is the size of the out-of-sample data, then for the optimal Ridge estimator, it holds that*

$$R^2_{oos}(\hat{\beta}_r(\lambda_n^{opt})) = p^{-1}n\theta_2(R^2)^2\left(1 + o_{\text{P}}(1)\right),$$

where R^2 denotes the population R-squared, given by $\tau\sigma_x^2\sigma_\beta^2/(\tau\sigma_x^2\sigma_\beta^2 + \sigma_\varepsilon^2)$ in this context.

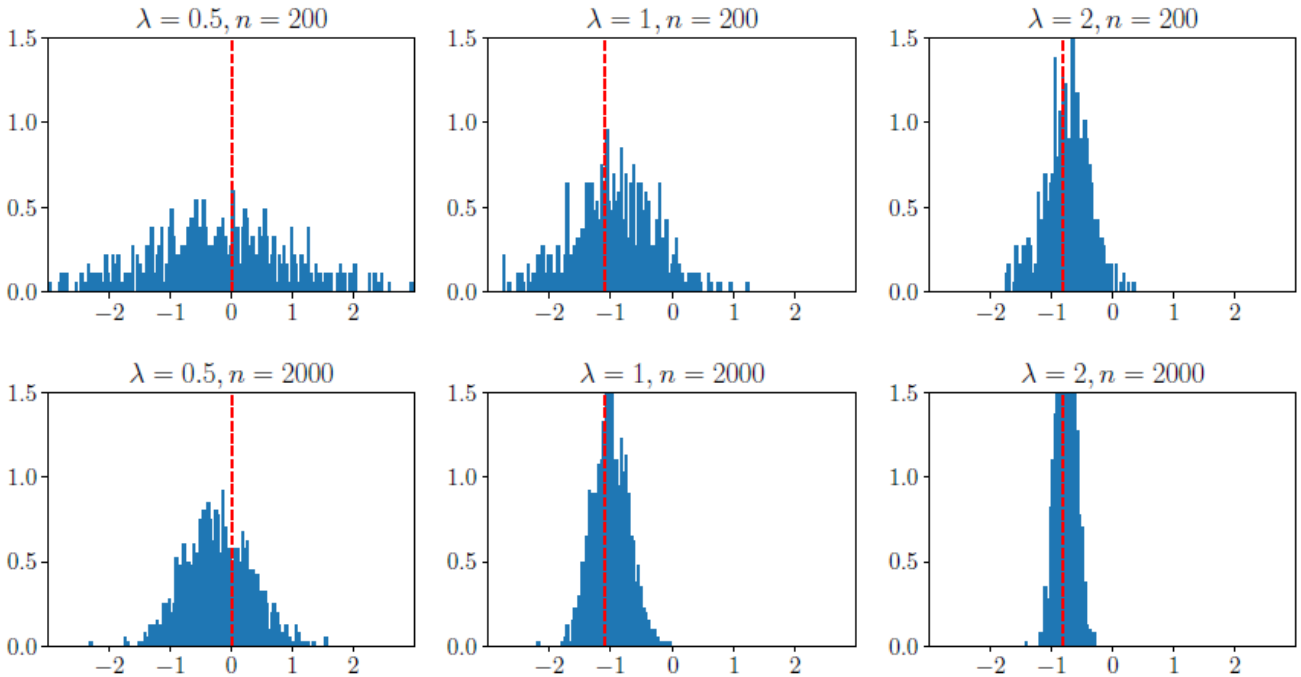
Mixed signals and the OLS benchmark

- The authors consider another setting in which both weak and strong signals are allowed
- In this setting, they find that Ridge still outperforms the zero estimator, see Theorem 5

Monte Carlo Simulations - Ridge

The theoretical part of the paper has established a set of asymptotic results. What about **finite sample** performance?

Figure 5: Simulation Results for Ridge in Linear DGPs



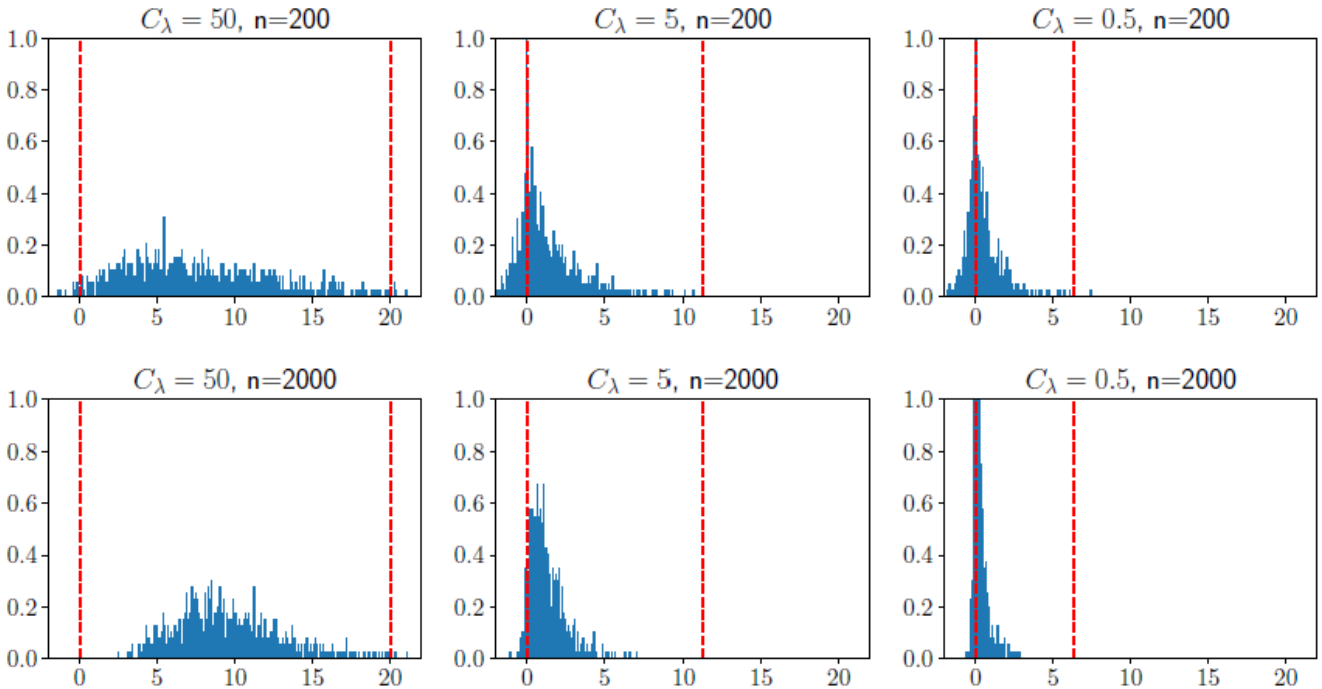
Note: The histograms depict the relative prediction error $\Delta(\hat{\beta}_r(\lambda_n))$ following equation (8) across 1,000 Monte Carlo samples. We consider two different sample sizes ($n = 200$ and $2,000$) and examine three different values of $\lambda_n = \tau^{-1}\lambda$, where $\lambda = 0.5, 1$, and 2. Notably, $\lambda = 1$ represents the optimal tuning parameter. The red dashed line indicates the values of α^*.

Monte Carlo Simulations - Ridge

- The probability mass of each histogram centres at the red vertical line, which aligns with the theoretical result
- As the sample size increases, the relative prediction error collapses to this value, in accord with the asymptotic theory
- Most of the probability mass is negative, which coincides with outperforming the zero estimator

Monte Carlo - LASSO

Figure 6: Simulation Results for Lasso in Linear DGPs



Note: The histograms depict the relative prediction error $\Delta(\hat{\beta}_l(\lambda_n))$ following equation (8) across 1,000 Monte Carlo samples. We consider two sample sizes ($n = 200$ and $2,000$) and examine three different values of C_λ. The two red dashed lines in each figure indicate the values of c_α and C_α that are solutions to (10).

Monte Carlo - LASSO

- As the sample size increases, the probability mass falls within the intervals (verifying the theoretical result)
- LASSO systematically underperforms the zero estimator
- As C_λ decreases (the tuning parameter increases), the bounds approach zero. This verifies that LASSO increases regularisation, steering the estimator closer to the zero estimator

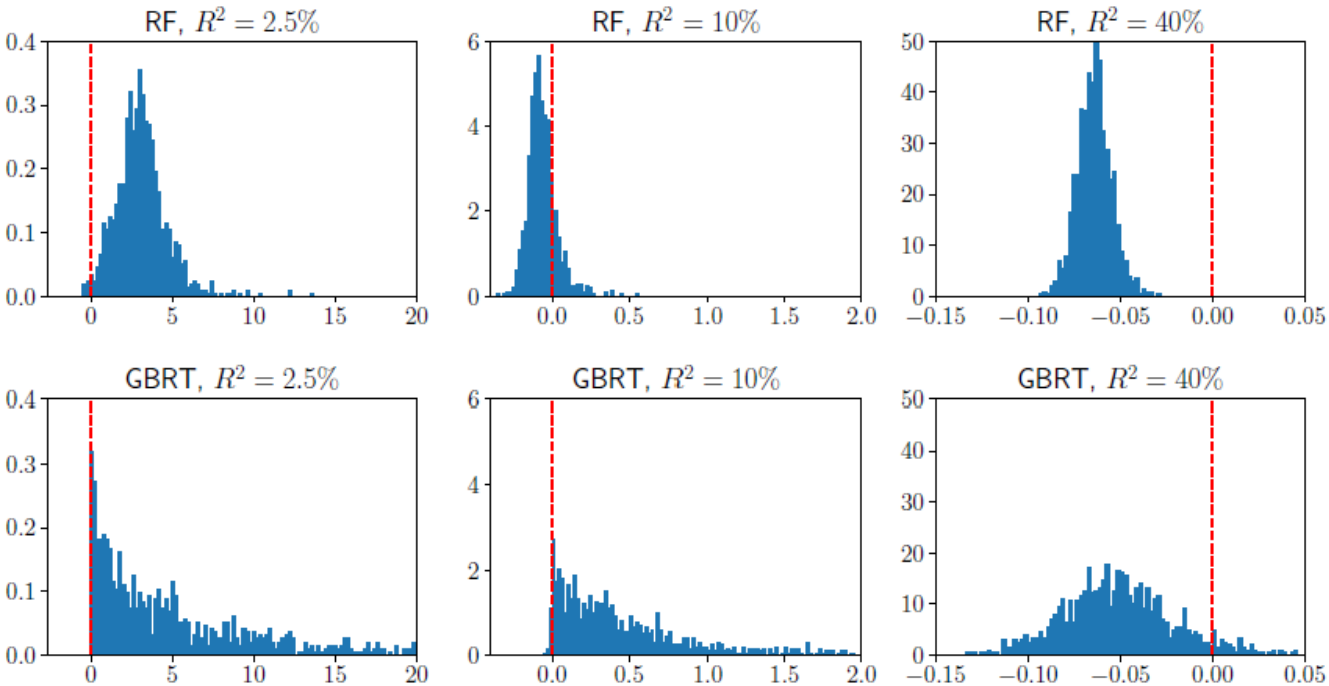
Other estimators - some headline findings

While they do not provide theoretical results, the authors simulate the behaviour of nonlinear ML estimators in the weak signal environment.

The **main takeaways** are

- Random forests perform well if there is moderate signal ($R^2 \geq 10\%$), but not for low signal ($R^2 = 2.5\%$)
- GBRT does poorly unless signal is very high ($R^2 = 40\%$)
- ℓ_2-regularized Neural Nets and early stopping NN perform very well
- ℓ_1 Neural Nets perform very poorly

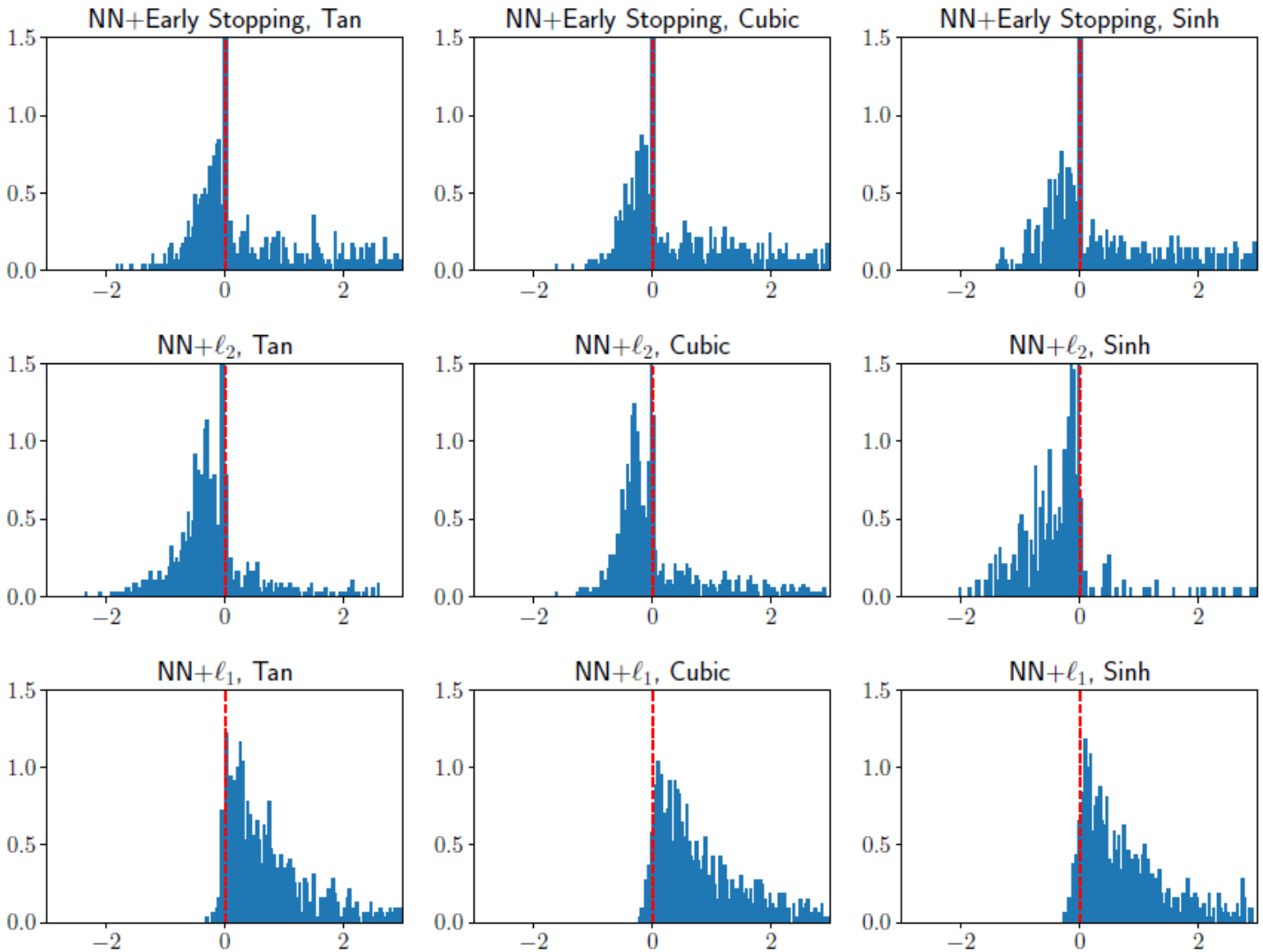
Figure 7: Simulation Results for RF and GBRT in Linear DGPs



Note: The histograms depict the relative prediction error, $pn^{-1}\tau^{-2}n_{oos}^{-1}\sum_{i\in\text{OOS}}((y_i-\hat{y}_i)^2-y_i^2)$, across 1,000 Monte Carlo samples. We consider RF and GBRT for $n = p = 2,000$, $q = 0.5$, $n_{oos} = 60,000$, and examine three different values of R^2, achieved by adjusting the value of τ. The red dashed line indicates the y axis.

Reguarlized neural nets

Figure 8: Simulation Results for NNs in Nonlinear DGPs



Taking it to the data

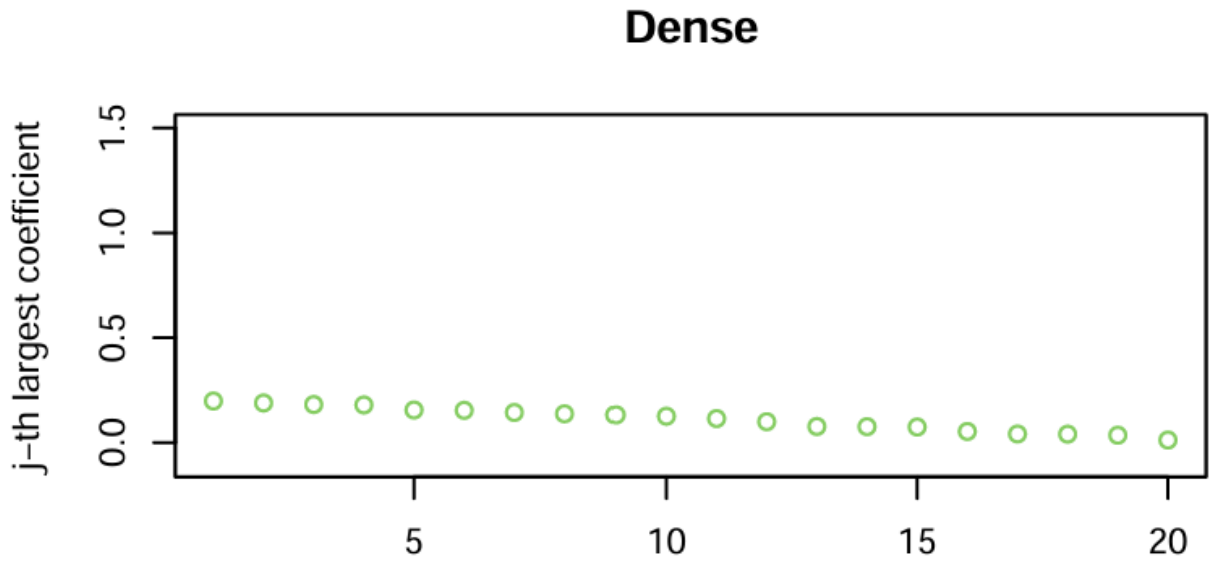
Table 1: Out-of-sample R-squared Values in Empirical Studies

	Ridge	Lasso	OLS/Ridgeless	RF	GBRT	NN(ℓ_2)	NN(ℓ_1)
Finance 1	0.80	−12.19	−81.08	−1.08	−14.21	1.41	−10.31
Finance 2	0.19	0.10	−1.25	0.08	−0.30	0.26	0.14
Macro 1a	15.29	15.40	−1375	24.39	16.44	16.94	19.09
Macro 1b	3.49	3.69	−2939	8.01	1.11	7.09	5.39
Macro 2	6.58	−14.58	−837	9.67	1.28	4.00	1.92
	(4.83)	(43.74)	(854)	(9.33)	(14.04)	(18.42)	(13.36)
Micro 1	0.48	−1.01	−13198	−10.25	−5.07	0.49	−6.77
	(0.84)	(2.01)	(12479)	(5.08)	(6.60)	(0.27)	(17.87)
Micro 2a	26.27	20.37	−12729	27.63	16.44	23.87	23.37
	(7.50)	(6.41)	(9213)	(6.10)	(3.40)	(10.07)	(10.09)
Micro 2b	1.89	-3.43	-14724	0.72	-6.45	1.11	-1.73
	(3.09)	(5.25)	(10506)	(2.41)	(6.83)	(2.20)	(5.09)

Note: This table reports R^2_{oos} values, presented in percentages, for Ridge, Lasso, OLS/Ridgeless, RF, GBRT, and NNs with respective ℓ_1 and ℓ_2 penalties, across six empirical studies spanning Finance, Macroeconomics, Microeconomics. For the first example in Macroeconomics and the second example in Microeconomics, two benchmark models are considered for comparison. Where standard deviations are applicable, they are provided in parentheses.

Some hot takes for discussions

Doesn't this paper just show what we already knew - that LASSO breaks down in **dense** environments?



The authors try to claim a contribution to contribution to a literature on **weakness** in econometrics, such as weak instrumental variables.

I get that in economics your final contribution can be speculative / a bit of a reach, but this feels tongue-in-cheek