Can Machines Learn Weak Signals?

Zhouyu Shen[†] Dacheng Xiu[‡]

Chicago Booth[†]

Chicago Booth and NBER[‡]

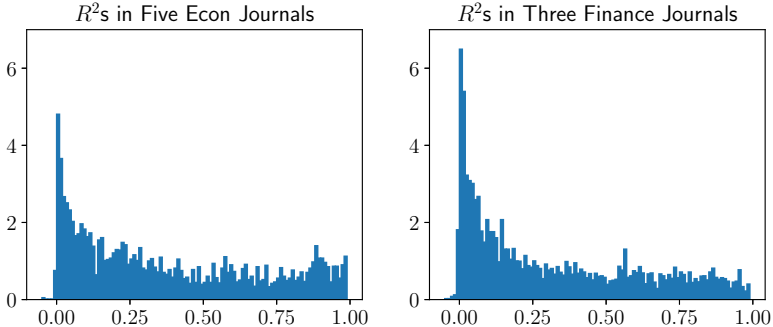
EDHEC

March 17, 2025

Weak Signals

- In the population model, covariates with non-zero coefficients are recognized as true signals, while those with zero coefficients are considered false signals, creating a clear-cut "black and white" distinction.
- However, in finite samples, the presence of minuscule non-zero coefficients introduces a "gray" area, blurring the lines between true and false signals.
- This gray area represents weak signals – covariates that, individually, exert negligible influence on the outcome variable.
- The scenario of weak signals is often encountered in economics, as evidenced by the limited explanatory power observed in empirical regression analyses.

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



- The histograms depict R^2s manually collected from published papers in a selection of Economics (AER, ECTA, JPE, QJE, RES) and Finance (JF, JFE, RFS) journals in 2022.

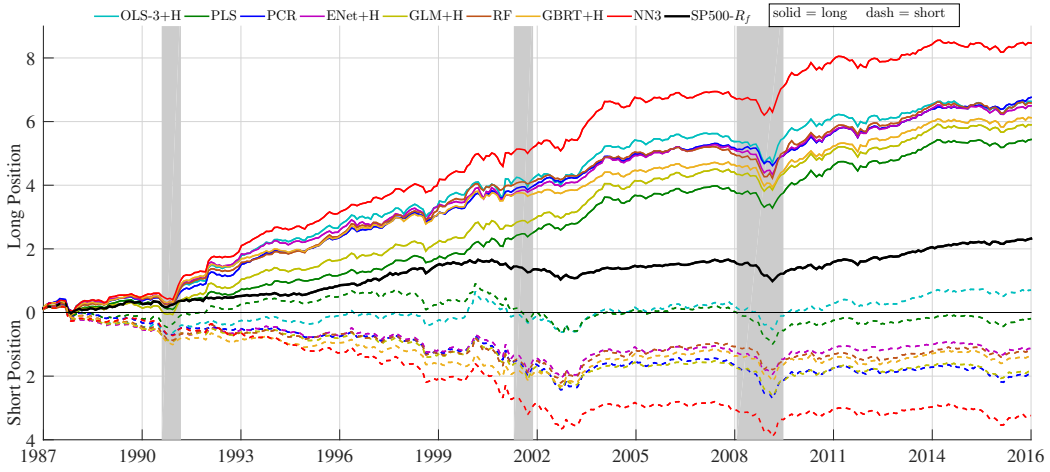
R^2s for Portfolio Returns Prediction

	OLS-3	PLS	PCR	ENet	GLM	RF	GBRT	NN1	NN2	NN3	NN4	NN5
VW-S&P500 Index	-0.11	-0.86	-2.62	-0.38	0.86	1.39	1.13	0.84	0.96	1.80	1.46	1.60
Big Growth	0.41	0.75	-0.77	-1.55	0.73	0.99	0.80	0.70	0.32	1.67	1.42	1.40
Big Value	-1.05	-1.88	-3.14	-0.03	0.70	1.41	1.04	0.78	1.20	1.57	1.17	1.42
Big Neutral	0.12	-0.81	-2.39	-0.46	0.41	1.05	1.03	1.33	0.78	1.81	1.93	1.93
Small Growth	0.35	1.54	0.72	-0.03	0.95	0.54	0.62	1.68	1.26	1.48	1.53	1.44
Small Value	-0.06	0.40	-0.12	-0.57	0.02	0.71	0.90	0.00	0.47	0.46	0.41	0.53
Small Neutral	-0.01	0.78	-0.10	-0.25	0.36	0.41	0.38	0.58	0.55	0.68	0.62	0.72
Big Conservative	-0.24	-0.17	-1.97	0.19	0.69	0.96	0.78	1.08	0.67	1.68	1.46	1.56
Big Aggressive	-0.12	-0.77	-2.00	-0.91	0.68	1.83	1.45	1.14	1.65	1.87	1.55	1.69
Big Neutral	-0.36	-1.65	-3.20	-0.11	0.76	0.99	0.73	0.54	0.62	1.62	1.44	1.60
Small Conservative	0.02	0.75	0.48	-0.46	0.55	0.59	0.60	0.94	0.91	0.93	0.99	0.88
Small Aggressive	0.14	0.97	0.06	-0.54	0.19	0.86	1.04	0.25	0.66	0.75	0.67	0.79
Small Neutral	-0.04	0.53	-0.17	0.08	0.45	0.23	0.20	0.73	0.60	0.81	0.73	0.80
Big Robust	-0.58	-0.22	-2.89	-0.27	1.54	1.41	0.70	0.60	0.84	1.14	1.05	1.21
Big Weak	-0.24	-1.47	-1.95	-0.40	-0.26	0.67	0.83	0.24	0.60	1.21	0.95	1.07
Big Neutral	-0.08	-1.02	-2.77	-0.21	0.10	1.46	1.44	0.95	1.00	1.78	1.70	1.73
Small Robust	-0.77	0.77	0.18	-0.32	0.41	0.27	-0.06	-0.06	-0.02	0.06	0.13	0.15
Small Weak	0.02	0.32	-0.28	-0.25	0.17	0.90	1.31	0.84	0.85	1.09	0.96	1.08
Small Neutral	0.22	1.05	0.09	0.03	0.48	0.76	0.97	1.08	1.04	1.19	1.12	1.18
Big Up	-1.53	-2.54	-3.93	-0.21	0.40	1.12	0.68	0.46	0.85	1.28	0.99	1.05
Big Down	-0.10	-1.20	-2.05	-0.26	0.36	1.09	0.77	0.48	0.89	1.34	1.17	1.36
Big Medium	0.24	1.38	0.57	0.01	1.32	1.56	1.37	1.60	1.76	2.28	1.83	2.01
Small Up	-0.79	0.42	-0.36	-0.33	-0.33	0.31	0.40	0.23	0.60	0.67	0.55	0.61
Small Down	0.40	1.16	0.47	-0.46	0.62	0.93	1.20	0.80	0.97	0.97	0.97	0.96
Small Medium	-0.29	0.03	-0.61	-0.56	-0.20	0.11	0.18	0.05	0.29	0.41	0.30	0.45

Source: "Empirical Asset Pricing via Machine Learning", Gu, Kelly, and Xiu, RFS (2020)

Predictions are Economically Meaningful

Market-timing Strategy's Sharpe ratio: 0.77; Stock-picking Strategy's Sharpe ratio: 1.35



Source: "Empirical Asset Pricing via Machine Learning", Gu, Kelly, and Xiu, RFS (2020)

Weak Signals and High Dimensionality

- Incorporating many weak signals into a model can result in overfitting, and, consequently, compromise predictive performance.
- Machine learning methods have proved effective in mitigating overfitting and discerning true signals from fake ones when the true signals are strong.
- These methods employ regularization techniques, such as penalizing ℓ_1 or ℓ_2 norms of model parameters, to achieve this.
- A pivotal question arises: Can machines learn weak signals?
  - Only if they can, can they outperform the (naive) baseline zero-estimator!

Caveats on the Use of Lasso

- Bayesian regression with a Spike-and-Slab prior on various economic datasets by Giannone, Lenza, and Primiceri (2022) suggests that sparsity may be an illusion, as optimal predictive models often rely on a large number of covariates.
- Kolesar, Muller, and Roelsgaard (2024) highlight limitations of sparsity-based estimators, including their lack of invariance to reparameterization and sensitivity to normalizations that are otherwise innocuous to OLS.

Model Setup

We study the following linear model:

$$\mathrm{y} = \mathrm{X}\beta_0 + \varepsilon,$$

where $\mathrm{X} \in \mathbb{R}^{n \times p}$, $\beta_0 \in \mathbb{R}^p$ and $\varepsilon \in \mathbb{R}^n$ are random variables.

- High dimension: $p/n \to c_0 \in (0, \infty]$.
- Weak signal: $\|\beta_0\|^2 \asymp_{\mathrm{P}} \tau \to 0$.

The choice of ℓ_2-norm is partially due to its close relationship with the widely-adopted R^2 metric in regression analysis, which provides a familiar and intuitive understanding of signal strength.

Assumptions on X and ε

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

- $\varepsilon = \Sigma_\varepsilon^{1/2} \mathsf{z}$, where z is composed of i.i.d. variables with mean zero, variance one and fourth moment finite. In addition, the $n \times n$ matrix Σ_ε satisfies $c_\varepsilon \leq \lambda_i(\Sigma_\varepsilon) \leq C_\varepsilon$ for $1 \leq i \leq n$.

Assumptions on β_0

- $\sqrt{p\tau^{-1}}\beta_0$ comprises i.i.d. random variables, each following a prior probability distribution F belonging to the class $\mathscr{F}$.
  - The class $\mathscr{F}$ 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.
  - This assumption allows for important classes of models, such as a spike-and-slab prior.
  - Under Gaussian noise, it is well-established (e.g., (Hastie, Tibshirani, and Friedman(2009))) that Ridge is equivalent to posterior mean under Gaussian priors, while Lasso is the same as posterior mode under Laplace priors.

- X, ε, and β_0 are asssumed to be independent of each other.

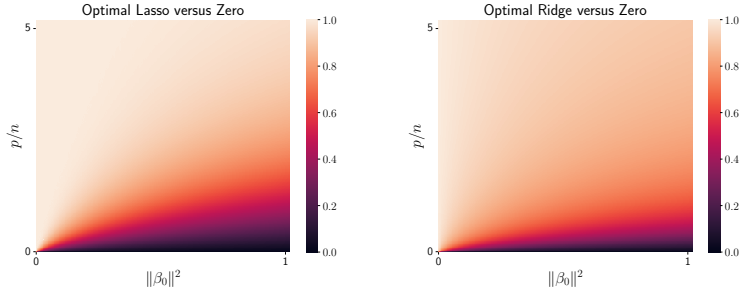
Bayes Prediction Risk

For any predictor derived from an estimator $\hat{\beta}$, the prediction error is:

$$\mathbb{E}_F(y^{\text{new}} - \hat{y}^{\text{new}})^2 = \sigma_\varepsilon^2 + \mathbb{E}_F\left[x^{\text{new}}(\hat{\beta} - \beta_0)\right]^2 = \sigma_\varepsilon^2 + \mathbb{E}_F\|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\|^2.$$

Existing Results for Strong Signals: Lasso, Ridge, vs Zero

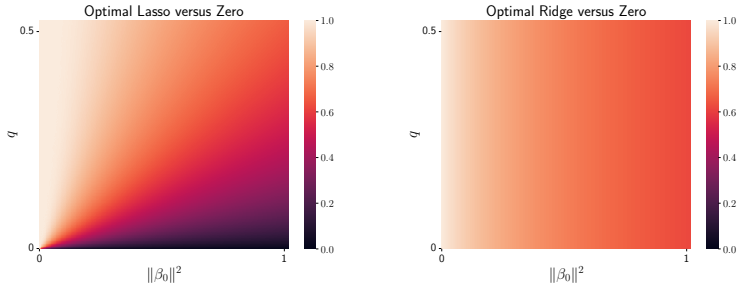
Under strong signals, i.e., $\tau = 1$, and suppose that Σ_1, Σ_2 and Σ_ε are identity matrices, $p/n \to c_0 \in \mathbb{R}^+$, significant progress has been made in understanding the asymptotic behavior of Bayes risk.



- In this figure, we fix $q = 0.2$ and vary the ratio p/n as well as $\|\beta_0\|^2$.
- Both Lasso and Ridge outperform zero.
- The disparity becomes less pronounced as $\|\beta_0\| \to 0$ and the ratio $p/n \to \infty$.

Existing Results for Strong Signals with Varying Sparsity

In this figure, we fix $p/n = 1$ and vary the sparsity parameter q as well as $\|\beta_0\|^2$.



Again, as $\|\beta_0\| \to 0$, the ratio of the Bayes risk for both Ridge and Lasso compared to zero converges to one.

Relative Prediction Error

- The zero estimator can be considered as a particular case of both Ridge and Lasso estimators when a sufficiently large tuning parameter is chosen.
- Merely comparing their Bayes risk ratios may not be an effective approach to tell any differences among these estimators.
- In light of this, for any estimator $\hat{\beta}$, define the higher-order relative error (compared with zero):

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

  - If $\Delta(\hat{\beta}) > 0$ w.p.a.1: $\hat{\beta}$ performs worse than the zero estimator.
  - If $\Delta(\hat{\beta}) < 0$ w.p.a.1: $\hat{\beta}$ surpasses the performance of the zero estimator.

Predictive Performance of Ridge

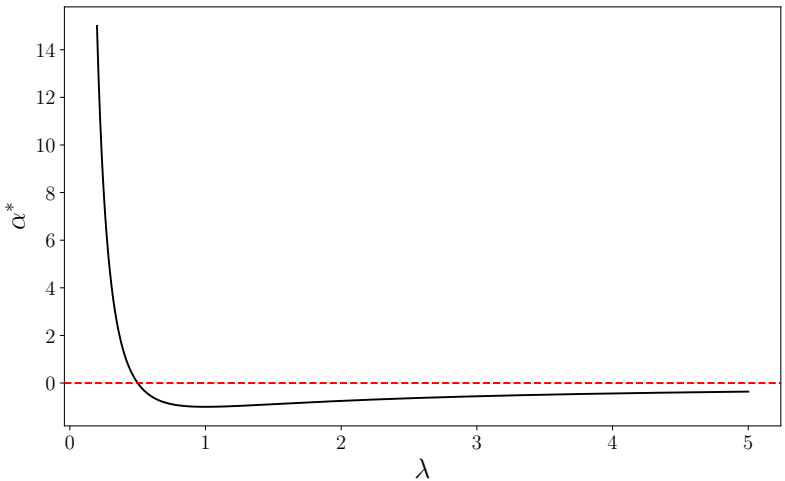
Consider the ridge estimator,

$$\hat{\beta}_r(\lambda_n) = \arg\min_{\beta} \frac{1}{n}\|y - X\beta\|^2 + \frac{p\lambda_n}{n}\|\beta\|^2.$$

- Special cases: $\lambda_n = 0$ (Ridgeless).
  - $p \leq n$: OLS
  - $p > n$: Minimum-norm interpolation, see Bartlett, Long, Lugosi, and Tsigler (2020)
- When the tuning parameter $\lambda_n = \tau^{-1}\lambda$, we have

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

The Optimal Ridge Can Beat Zero!



Predictive Performance of Lasso

Consider the Lasso estimator,

$$\hat{\beta}_l(\lambda_n) = \arg\min_{\beta} \frac{1}{n}\|y - X\beta\|^2 + \frac{\lambda_n}{\sqrt{n}}\|\beta\|_1.$$

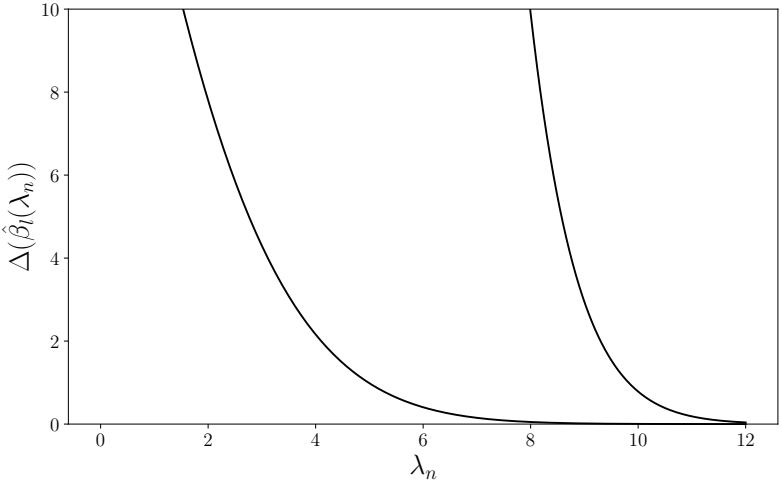
- We have, w.p.a.1,

$$c_\alpha \leq \Delta(\hat{\beta}_l(\lambda_n)) \leq C_\alpha,$$

where c_α and C_α are the solutions to the equation (in terms of x):

$$x - \sqrt{\frac{2C_\lambda}{c_2}x} = -\frac{C_\lambda}{100C_2}.$$

The Optimal Lasso is Zero!



Assessing Signal-to-noise Ratio

In practice, the out-of-sample R-squared, defined as

$$R^2_{oos} = 1 - \frac{\sum_{i \in \text{OOS}} (y_i - \hat{y}_i)^2}{\sum_{i \in \text{OOS}} y_i^2},$$

can serve as an indicator of the signal-to-noise ratio.

- Assuming that $\Sigma_1 = \mathsf{I}$, $\Sigma_\varepsilon = \sigma_\varepsilon^2 \mathsf{I}$, and 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

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

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

Mixed Signal Strengths and Alternative Benchmarks

The previous slides study the scenarios where all signals are weak and the zero estimator serves as the benchmark model. This slide expands our analysis to include models where potentially strong signals serve as benchmarks. Consider the model

$$y = W\gamma + X\beta_0 + \varepsilon$$

- $W \in \mathbb{R}^{n\times d}$ represents a predefined set of covariates. These covariates include potentially strong signals and form the basis of the benchmark model.
  - Each covariate within W is assumed to have a finite second moment. Furthermore, the eigenvalues of $n^{-1}W^\top W$ are lower bounded by some positive constant and $d = o\left(n^2 p^{-1}\tau\right)$.
- X is generated as $X = W\eta_0 + U$
  - The triplet $(U, \beta_0, \varepsilon)$ is assumed to follow the same distribution as $(X, \beta_0, \varepsilon)$ in the previous slides.
- new benchmark predictor: $\hat{y}_b^{\text{new}} = \left(w^{\text{new}}\right)^\top \hat{\gamma}$, where $\hat{\gamma} = \left(W^\top W\right)^{-1} W^\top y$

Ridge-augmented Regression

We study the Ridge-augmented regression:

$$(\hat{\beta}(\lambda_n), \hat{\gamma}) := \underset{(\beta,\gamma)}{\arg\min} \frac{1}{n}\|y - W\gamma - X\beta\|^2 + \frac{p\lambda_n}{n}\|\beta\|^2$$

and its prediction $\hat{y}^{\text{new}} = (w^{\text{new}})^\top \hat{\gamma}(\lambda_n) + (x^{\text{new}})^\top \hat{\beta}(\lambda_n)$.

- It holds that

  $$pn^{-1}\tau^2 \left(\mathbb{E}_F \left[(\hat{y}^{\text{new}} - y^{\text{new}})^2 \mid \mathscr{I} \right] - \mathbb{E}_F \left[(\hat{y}_b^{\text{new}} - y^{\text{new}})^2 \mid \mathscr{I} \right] \right) \xrightarrow{\text{P}} \alpha^* = 2\theta_2 \sigma_x^4 \left(\frac{\sigma_\varepsilon^2 \theta_1}{2\lambda^2} - \frac{\sigma_\beta^2}{\lambda} \right),$$

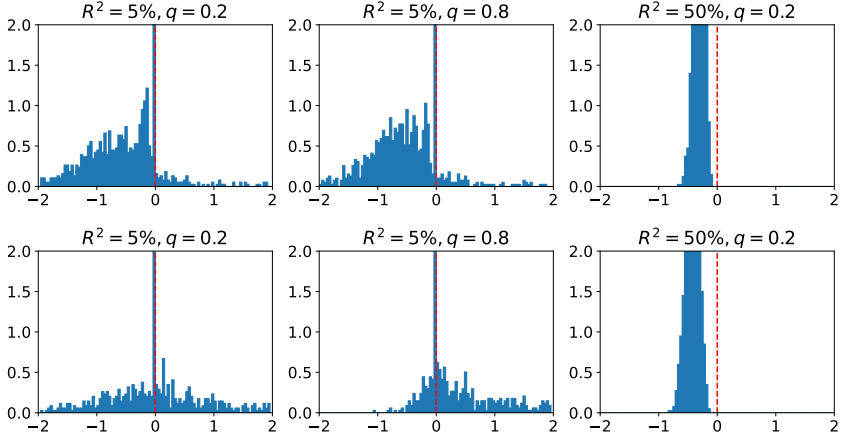
  where $\mathscr{I}$ denotes the information set generated by $(W, X, y, \gamma_0, \beta_0)$.

Strong Signals vs Weak Signals: Summary

	Strong Signals	Weak Signals
OLS/Interpolation $>$ zero ?	Yes (Hastie et al. (2022))	No
Lasso $>$ zero ?	Yes (Bayati and Montanari (2011))	No
Ridge $>$ zero ?	Yes (Dobriban and Wager (2018))	Yes
Cross-validation valid ?	Yes (Liu and Dobriban (2020))	Yes

Simulation Results for Ridge and Lasso

The histograms of $\Delta(\hat{\beta})$ for Ridge (top) and Lasso (bottom) ($n = 500,\ p = 300$):



▸ Additional simulation results

Simulations for Extremely Sparse Scenario

The simulation results below show that our conclusion—Lasso performance deteriorates when τ is small—holds even for extremely sparse cases.

		Lasso				Ridge			
q	R^2 (%)	Q1	Q2	Q3	#Zero	Q1	Q2	Q3	#Zero
0.20	5%	-0.127	0.000	0.521	360	-0.992	-0.501	-0.129	97
0.10	5%	-0.871	0.000	0.187	327	-0.981	-0.475	-0.077	113
0.10	2%	0.000	0.000	3.435	493	-0.622	0.000	0.440	237
0.05	5%	-2.688	-0.305	0.000	255	-1.037	-0.387	0.000	130
0.05	2%	0.000	0.000	2.948	473	-0.642	0.000	0.426	238
0.02	5%	-6.542	-2.050	0.000	215	-1.304	-0.230	0.000	149
0.02	2%	0.000	0.000	1.695	432	-0.605	0.000	0.625	254

Simulation Experiments for Advanced Machine Learning Methods

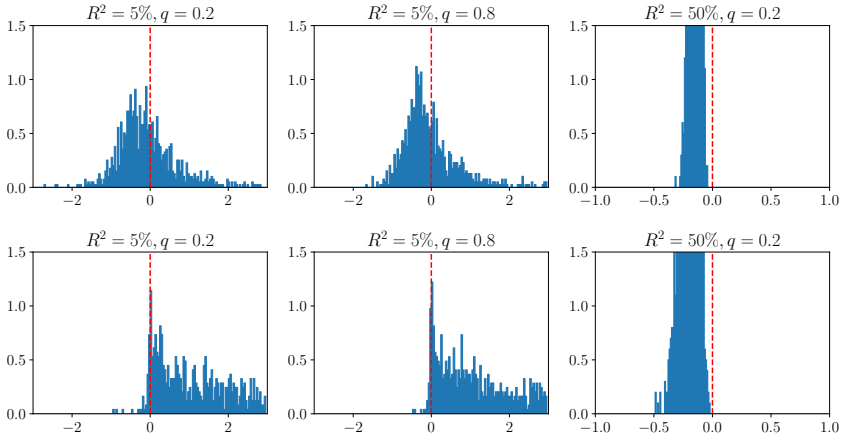
We expand our inquiry into the relevance of our theory to nonlinear machine learning methodologies such as Random Forest (RF), Gradient Boosted Regression Trees (GBRT), and Neural Networks (NN), through simulation experiments. We simulate the following DGP with $f(x) = \tan(x)$

$$y_i = \sum_{j=1}^{p} \beta_{0,j} f(Z_{ij}) + \varepsilon_i, \quad i = 1, \ldots, n.$$

- Generate Z_{ij} by applying an inverse transform to X_{ij}, which was previously simulated. Specifically, Z_{ij} is defined as $f^{-1}(X_{ij})$, where the design matrix X is constructed using the identical DGP as previously outlined.
- Both the coefficients β_0 and the error term ε_i follow the same DGPs as before.
- We analyze the benchmark scenario where $n = 500, p = 300$ and report the the relative prediction error, $pn^{-1}\tau^{-2}n_{\text{oos}}^{-1} \sum_{i \in \text{OOS}} \left((y_i - \hat{y}_i)^2 - y_i^2 \right)$.

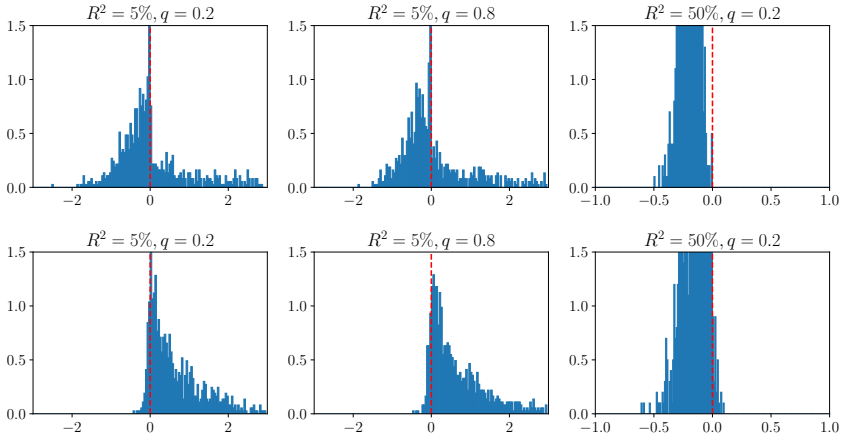
Simulations with Tree Algorithms

The histograms of higher order relative prediction error for Random Forest (top) and GBRT (bottom) ($n = 500,\ p = 300$):



Relative Error Comparisons for Neural Networks

The histograms of higher order relative prediction error for NN(ℓ_2) (top) and NN(ℓ_1) (bottom) ($n = 500,\ p = 300$):



Equity Premium

Dataset constructed by Welch and Goyal (2008).

- Dependent variable: US market (S&P 500) return
- Possible predictors: 16 lagged financial and macroeconomic indicators
- Sample: 74 annual time-series observations from 1948 to 2021
- 10−fold cross-validation
- Expanding window method, 57 exercises (following Giannone, Lenza and Primiceri (2021))

	Ridge	Lasso	OLS/Ridgeless	RF	GBRT	NN(ℓ_2)	NN(ℓ_1)
R^2_{oos}	0.8%	−12.19%	−81.08%	1.30%	−14.21%	1.41%	−10.31%

The Cross Section of Expected Returns

Dataset constructed by Gu, Kelly and Xiu (2020).

- Dependent variable: US Individual Equity returns
- Possible predictors: 920 covariates, including characteristics, macroeconomic predictors and their interactions
- Sample: Monthly returns from CRSP for all firms listed in the NYSE, AMEX, and NASDAQ from $1957-2021$, with the average number of stocks per month exceeding 6,200
- $2-$fold cross-validation
- Expanding window, 35 exercises (following Gu, Kelly and Xiu (2020))

	Ridge	Lasso	OLS/Ridgeless	RF	GBRT	NN(ℓ_2)	NN(ℓ_1)
R^2_{oos}	0.19%	0.10%	-1.25%	0.10%	-0.30%	0.26%	0.14%

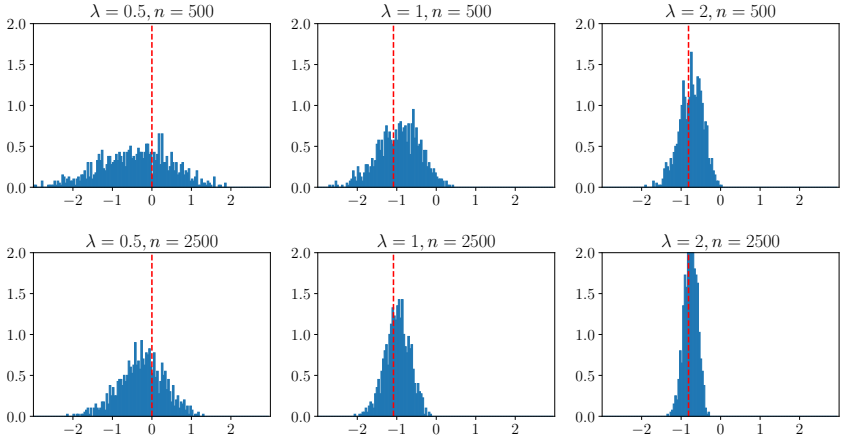
Using a stock selection portfolio strategy, NN(ℓ_2) can achieve a sharp ratio at 2.13, followed by Ridge regression (1.64) and NN(ℓ_1) (1.55).

Conclusion

- When the field of Economics and Finance adopted Machine Learning from Computer Science, there were concerns about its efficacy given the low signal-to-noise ratio. This paper offers some thoughts on this matter.
- While Lasso is often regarded as a modern equivalent to OLS, we should exercise caution when applying it in economics and financial settings, where signals are often weak.
- Giannone, Lenza and Primiceri (2021) argues that sparsity is an illusion. Our findings suggest that signal weakness is a more prevalent issue, providing a complementary explanation for the observed poor performance of Lasso.

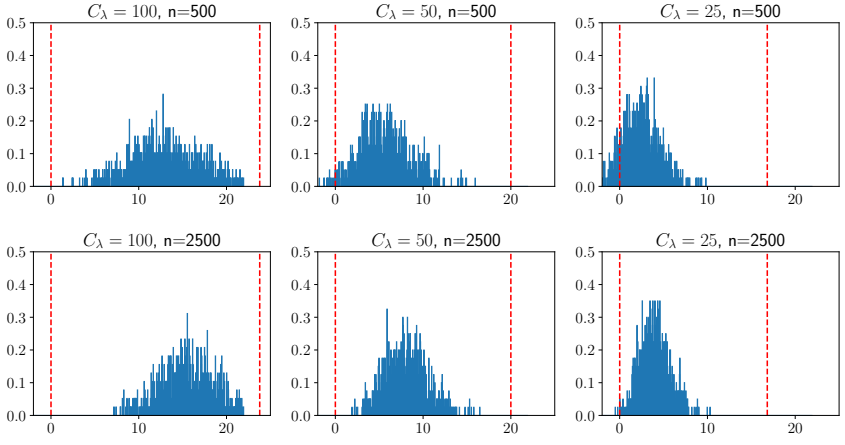
Simulation Results for Ridge with Fixed Tuning Parameters

The histograms of $\Delta(\hat{\beta})$ for Ridge ($p/n = 3/5$, $q = 0.2$, $R^2 = 5\%$, $\lambda^{opt} = 1$):



Simulation Results for Lasso with Fixed Tuning Parameters

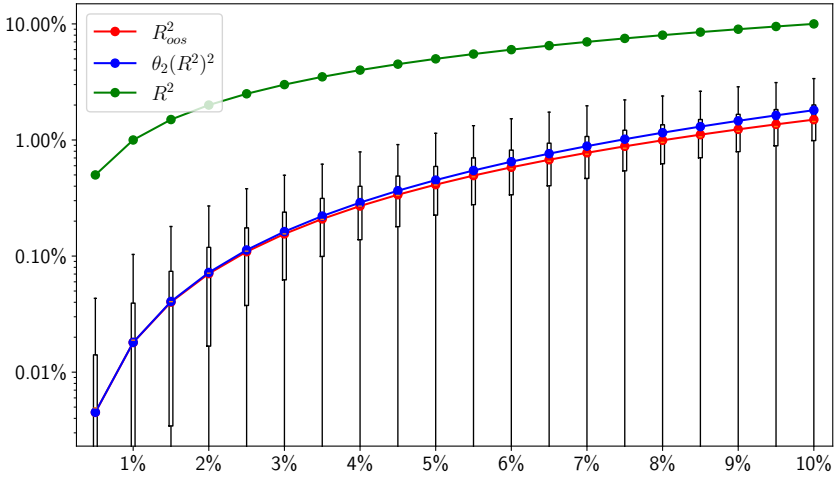
The histograms of $\Delta(\hat{\beta})$ for Lasso:



Back

Simulation Results for Optimal Ridge's R^2_{oos}

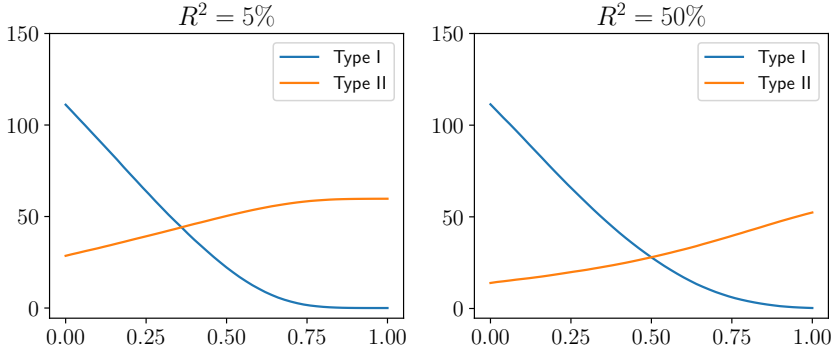
The boxplots of R^2_{oos} for optimal Ridge ($p = 300$, $n = 500$, $q = 0.2$, $R^2 = 5\%$):



Back

Why Lasso Fails? Type I Error

- The failure to identify true signals has a minor impact since zero does not utilize any true signals. The primary challenge lies in its failure to adequately filter out irrelevant signals.



Back