

Limits To (Machine) Learning

Zhimin Chen, Bryan Kelly, and Semyon Malamud

September 2026

Table of Contents

- 1 Introduction
- 2 Limits to Learning
- 3 Deep Learning
- 4 Simulations
- 5 Empirics: 41 Return Series
- 6 Conclusions

The Billion Dollar Question

- ▶ You have X (features) and you have y (labels)
- ▶ How do you know if X predicts y ?
 - Should I (or should I not) buy this new dataset?
 - Should I have a team of PhDs spend a year trying to get some alpha from these data?

Why Is this Question So Hard To Answer? Complexity!

- ▶ Too many X s and too small T (number of observations): **statistical complexity**
- ▶ A flexible model has as many parameters as observations, or more
- ▶ Estimation error does not go away. Out-of-sample performance then understates what the data contain

The Usual Workflow

- ▶ Run a few regressions, XGBoost, deep learning, . . .
- ▶ If it works, put more effort and build a better model
- ▶ If it does not, abandon the project. . .
- ▶ **But what if we are wrong?**

A Little History: Long Plateaus, Sudden Breakthroughs

- ▶ **Computer vision (1960s–2009):** decades of hand-crafted features (edges, SIFT, HOG); performance improved marginally despite intense effort. **Breakthrough:** large labeled datasets + deep CNNs $\Rightarrow$ sudden jump in accuracy
- ▶ **Machine translation (1990s–2014):** rule-based and phrase-based systems plateaued for years; improvements incremental and expensive. **Breakthrough:** end-to-end neural models and attention
- ▶ **Speech recognition (1980s–2010s):** HMMs and GMMs dominated with slow gains. **Breakthrough:** deep learning cut word error rates dramatically
- ▶ **Return prediction (1980s–today):** decades of “nothing works” out of sample with a handful of predictors and linear models. Breakthrough still to come?

The Key Lesson

- ▶ Long periods where “nothing works” do *not* imply that no signal exists
- ▶ They often reflect a **limit to learning** given: limited sample size T ; limited model complexity (**virtue of complexity**); poor model design (bad **alignment**)
- ▶ Progress can appear suddenly once a critical threshold is crossed, or the right model is found
- ▶ **This paper:** a lower bound on how much predictability the data contain, computed from a forecast that fails out of sample
- ▶ **Message:** for U.S. returns the bound is large. The data carry far more predictability than any current forecast extracts

Preview of the Empirics

- ▶ 41 monthly U.S. return series: market, Treasury and corporate bonds, 10 size deciles, 13 factor themes, 12 industries; signals: the 14 Goyal–Welch predictors
- ▶ Realized R^2_{OOS} : at most a few percent, even with a cross-validated penalty (median 0.6%)
- ▶ 95% lower bound on population R^2 : **positive for 36 of 41 series**, median **15%**, market **17%**, seven-year zero-coupon bond **24%**, size and value themes **20% and 18%**
- ▶ The gap is the room for improvement: what better methods and richer signals can still capture

Table of Contents

- 1 Introduction
- 2 Limits to Learning**
- 3 Deep Learning
- 4 Simulations
- 5 Empirics: 41 Return Series
- 6 Conclusions

The New Paradigm

- ▶ Train a model and evaluate it Out-Of-Sample (OOS): get R_{OOS}^2
- ▶ Compute the **Limits-to-Learning Gap (LLG)** from the signal matrix alone
- ▶ Convert R_{OOS}^2 into a **lower bound on the population R^2** , with a confidence interval
- ▶ If the bound is high while R_{OOS}^2 is low: the signal is there, the method is not

Estimation Error

- Data-generating process

$$y_{t+1} = f_t + \varepsilon_{t+1} \quad (1)$$



$$MSE_{Oos}(\hat{f}) = \underbrace{E[(y_{t+1} - \hat{f}_t)^2]}_{\text{out-of-sample error}} = \underbrace{E[(f_t - \hat{f}_t)^2]}_{\text{estimation error}} + \underbrace{\sigma_\varepsilon^2}_{\text{irreducible noise}} \quad (2)$$

- Population (infeasible) R^2 :

$$R_*^2 = 1 - \frac{\sigma_\varepsilon^2}{E[f_t^2] + \sigma_\varepsilon^2} \quad (3)$$

Linear Estimators

- ▶ A linear estimator is of the form

$$\hat{f}_T = \mathcal{K}(S_T, S) y, \quad (4)$$

where $\mathcal{K}(S_T, S) \in \mathbb{R}^{1 \times T}$ collects the weights on past outcomes

- ▶ **Example 1:** ridge on P signals

$$\hat{f}_T = \underbrace{\left(S_T' (zI + S'S/T)^{-1} \frac{1}{T} S' \right)}_{\mathcal{K}(S_T, S)} y \quad (5)$$

- ▶ **Example 2:** kernel ridge with a positive definite kernel k

$$\hat{f}_T = \underbrace{k(S_T, S) (zI + k(S, S))^{-1}}_{\mathcal{K}(S_T, S)} y \quad (6)$$

Assumption

- ▶ $\mathcal{H} = \sigma\{S_t\}$: the whole signal path; $\mathcal{G}_t = \sigma(\mathcal{H}, \varepsilon_1, \dots, \varepsilon_t)$
- ▶ Innovations are a martingale difference with constant conditional variance:

$$E[\varepsilon_{t+1} \mid \mathcal{G}_t] = 0, \quad E[\varepsilon_{t+1}^2 \mid \mathcal{G}_t] = \sigma_\varepsilon^2 \quad (7)$$

- ▶ No independence across time, no stationarity of the signals, any dependence of f_t on the signals, any model misspecification

LLG i

Let $\hat{\mathcal{K}} = (\mathcal{K}(S_\tau, S))_{\tau=T}^{T+T_{OOS}-1} \in \mathbb{R}^{T_{OOS} \times T}$ be the weights placed on past outcomes over the evaluation block.

$$\hat{\mathcal{L}} = \frac{1}{T_{OOS}} \text{tr}(\hat{\mathcal{K}}' \hat{\mathcal{K}}) \quad (8)$$

is the **Limits-to-Learning Gap (LLG)**: how much training noise the estimator carries into its forecasts. It depends only on the signals, not on f .

Theorem

As $T, T_{OOS} \rightarrow \infty$,

$$\liminf \frac{MSE_{OOS}(\hat{f})}{1 + \hat{\mathcal{L}}} \geq \underbrace{\sigma_\varepsilon^2}_{\text{infeasible}} \quad \text{and} \quad R_*^2 \geq \limsup \frac{R_{OOS}^2(\hat{f}) + \hat{\mathcal{L}}}{1 + \hat{\mathcal{L}}} \quad (9)$$

For ridge, $\hat{\mathcal{L}}(z) = T^{-1} \text{tr} (\hat{\Psi}_{OOS} \hat{\Psi} (zI + \hat{\Psi})^{-2}) = O(P/T)$: large when the penalty is small and the number of signals is large.

Confidence Bound

Theorem (CLT)

Under martingale-CLT regularity, the statistic $\hat{R}_{LB}^2 = \frac{R_{OOS}^2(\hat{f}) + \hat{\mathcal{L}}}{1 + \hat{\mathcal{L}}}$ satisfies

$$\limsup \text{Prob}\left\{\sqrt{T_{OOS}}(R_*^2 - \hat{R}_{LB}^2)/\sigma_{R^2} < \eta\right\} \leq \Phi(\eta) \quad (\eta \leq 0), \quad (10)$$

so $\hat{R}_{LB}^2 - z_{1-\gamma} \hat{\sigma}_{R^2}/\sqrt{T_{OOS}}$ is a one-sided $(1 - \gamma)$ lower confidence bound for R_*^2 , with a pivotal $\hat{\sigma}_{R^2}$ computed from observables.

- Two corrections: the LLG removes the estimation loss; the standard error accounts for evaluation noise

Two Feature Sets, One Bound

- ▶ We use two random-feature maps on the same signals, **tanh and ReLU**. Each gives a statistic $\hat{R}_j^2$ and a valid lower bound on the **same** R_*^2
- ▶ Natural statistic: the larger one. Taking the larger of two noisy statistics is a selection
- ▶ **Proposition (merged bound)**. The two statistics are jointly asymptotically normal (a joint CLT on a common training block), and

$$\max_j \left\{ \hat{R}_j^2 - \frac{q_\gamma(\hat{\rho}) \hat{\sigma}_j}{\sqrt{T_{OOS}}} \right\}$$

is a valid $(1 - \gamma)$ bound, with $q_\gamma(\rho)$ the quantile of the max of two normals with correlation ρ , between $z_{1-\gamma}$ and the Bonferroni $z_{1-\gamma/2}$

- ▶ Feasible: a conservative kernel-based bound $\hat{\rho}^-$; in the data $\hat{\rho}^- = -1$ for all series, so $q = z_{1-\gamma/2} = 1.96$ (the point estimates $\rho \approx 0.1$ give the same to two decimals). **Only the merged bound is reported**

Virtue of Complexity

- ▶ As the number of random features P grows, the realized R_{OOS}^2 of the reference ridge collapses. . .
- ▶ . . . but the LLG grows with it, and the lower bound rises and then flattens out (Kelly, Malamud, Zhou 2024)
- ▶ We therefore compute every bound at a very large $P = 20,000$, where it has saturated, with a small penalty $z_{ref} = 0.01$ that makes the LLG large
- ▶ The reference forecast is deliberately a bad forecast. The bound is not

Table of Contents

- 1 Introduction
- 2 Limits to Learning
- 3 Deep Learning**
- 4 Simulations
- 5 Empirics: 41 Return Series
- 6 Conclusions

Gradients Are Learned Features

- ▶ Network $g(x; \theta)$ trained for $\mathcal{E}$ steps: $\theta_0 \rightarrow \theta_{\mathcal{E}}$. **Feature learning**: the internal representation of the signals moves with the data
- ▶ Gradient features at the trained parameters (the neural tangent kernel, NTK):

$$S_t^{(\mathcal{E})} = \nabla_{\theta} g(X_t; \theta_{\mathcal{E}}) \quad (11)$$

- ▶ One hidden layer $g = \sum_k \beta_k \sigma(\omega'_k x + b_k)$: $\partial g / \partial \beta_k$ is the **trained hidden unit** itself; $\partial g / \partial (\omega_k, b_k)$ its sensitivities to the signals
- ▶ Ridge on $S^{(\mathcal{E})}$ = kernel ridge with the trained tangent kernel: a **linear estimator on learned features**, so the LLG theory applies

The Catch, and the Fix

- ▶ If the network is trained on the same labels that enter the ridge, the features depend on the noise ε : the kernel is not exogenous, the bound fails
- ▶ Standard escape: *lazy training* (wide nets, features stay at initialization). It rules out feature learning, the reason to train a network
- ▶ **Our fix, by design:** three consecutive blocks
 - **representation block:** train the network any way you like (architecture, width, depth, optimizer, steps, amount of feature learning)
 - **ridge block:** compute the trained gradient features and fit the ridge
 - **evaluation block:** R_{OOS}^2 , LLG, bound
- ▶ Conditional on the representation block the features are fixed functions of the signals: **the same theorem and the same confidence bound apply, whatever the network**

Why the Third Block Is Necessary

Mean-field one-hidden-layer net, $T = T_{OOS} = 400$, features move substantially (movement 0.9 vs 0.12 lazy). Overshoot = $\hat{R}_{LB}^2 - \tilde{R}_*^2$ must be ≤ 0 ; averages over 20 replications (s.e.).

$\tilde{R}_*^2$	ζ	Design	LLG	Noise ratio	R_{OOS}^2	$\hat{R}_{LB}^2$	Overshoot
0.49	0.05	Two-block	0.53	0.70	0.30	0.54	+0.049 (0.008)
0.49	0.05	Three-block	0.53	1.03	0.22	0.49	-0.003 (0.008)
0.49	0.01	Two-block	0.84	0.82	0.14	0.53	+0.039 (0.010)
0.49	0.01	Three-block	0.84	1.04	0.06	0.49	-0.003 (0.010)
0.19	0.05	Two-block	0.53	0.73	-0.11	0.27	+0.077 (0.012)
0.19	0.05	Three-block	0.53	1.03	-0.24	0.19	-0.005 (0.013)
0.19	0.01	Two-block	0.86	0.82	-0.37	0.26	+0.069 (0.016)
0.19	0.01	Three-block	0.86	1.04	-0.52	0.18	-0.009 (0.016)

Two blocks: the bound overshoots by 4–6 s.e. (trained features absorb the training noise, noise ratio < 1). Three blocks: overshoot zero within sampling error.

Table of Contents

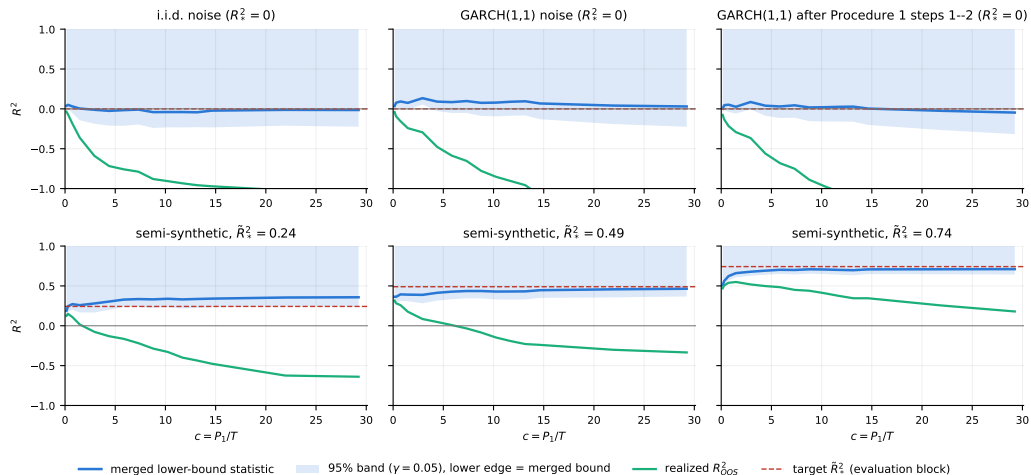
- 1 Introduction
- 2 Limits to Learning
- 3 Deep Learning
- 4 Simulations**
- 5 Empirics: 41 Return Series
- 6 Conclusions

Semi-Synthetic Design

- ▶ Real signals: the 14 Goyal–Welch predictors after Procedure 1, 1933–2024, split at 1990
- ▶ Truth: $y_{t+1} = \gamma f(X_t) + \varepsilon_{t+1}$ with a random expansion over 20,000 tanh units with their own directions (independent of the estimator's features), scaled to population R^2 of 0.25, 0.5, 0.75
- ▶ Noise benchmarks: i.i.d.; GARCH(1,1) raw and after the first two steps of Procedure 1
- ▶ Same estimator as in the empirics: two feature maps, $z_{ref} = 0.01$, merged bound at each P

Semi-Synthetic Labels: One Draw

Noise and semi-synthetic labels, merged bound, $z_{ref} = 0.01$



What the Simulation Shows

- ▶ At $P = 100$ the ridge captures only part of the signal: R^2_{OOS} of 13%, 31%, 47% against targets of 24%, 49%, 74%; then it falls with P , below zero unless the signal is strong
- ▶ The merged statistic rises with P to 36%, 46%, 71%; the 95% bound at $P = 20,000$ is 24%, 37%, 64%: **above the realized R^2 at every P**
- ▶ Under the noise benchmarks the statistic hovers around zero and the bound is negative at high P

Coverage of the 95% Bound (200 noise draws)

Coverage (%)	pivotal	s.e.		kernel-form	s.e.	
Label process	$P=10^3$	$5 \cdot 10^3$	$2 \cdot 10^4$	$P=10^3$	$5 \cdot 10^3$	$2 \cdot 10^4$
i.i.d. noise	92.0	91.0	88.5	100.0	100.0	99.5
GARCH(1,1) noise	72.5	71.0	70.5	98.5	90.0	89.0
GARCH(1,1), Procedure 1 steps 1–2	86.5	85.5	90.0	100.0	99.5	99.5
semi-synthetic, $\tilde{R}_*^2 = 0.24$	98.0	97.5	96.5	100.0	100.0	99.5
semi-synthetic, $\tilde{R}_*^2 = 0.49$	100.0	99.5	98.5	100.0	100.0	100.0
semi-synthetic, $\tilde{R}_*^2 = 0.74$	100.0	100.0	100.0	100.0	100.0	100.0

Nominal 95. The pivotal standard error is slightly liberal at high complexity for pure noise; raw GARCH violates homoskedasticity, and Procedure 1's standardization restores most of the coverage. The kernel-form standard error is conservative but 1.2–2.7 times larger.

Table of Contents

- 1 Introduction
- 2 Limits to Learning
- 3 Deep Learning
- 4 Simulations
- 5 Empirics: 41 Return Series**
- 6 Conclusions

Data

- ▶ **Signals:** the 14 Goyal–Welch predictors (dp, dy, ep, de, bm, ntis, svar, tbl, lty, ltr, tms, dfy, dfr, infl), 1930-01 to 2024-11
- ▶ **Targets:** 41 monthly return series in four blocks
 - market and yield curve: CRSP market, long-term Treasury and corporate bonds, 2-, 5-, 7-year zero-coupon Treasuries (GSW)
 - 10 size deciles (Ken French)
 - 13 Jensen–Kelly–Pedersen long-short factor themes
 - 12 Fama–French industries
- ▶ Training 1933-01 to 1989-12 ($T = 684$), evaluation 1990-01 to 2024-11 ($T_{OOS} = 419$)

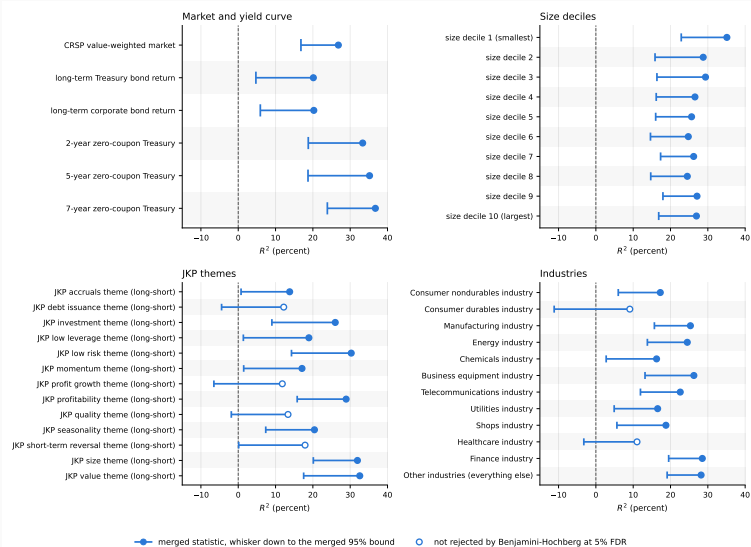
Procedure 1 and the Two Tests

- ▶ **Procedure 1** (signals and targets alike): rolling 36-month standardization lagged one month, clip at ± 3 , expanding-window AR(1) residual
- ▶ Two tests on training and evaluation blocks: $\max |\text{acf}|$ at lags 1–3 ≤ 0.2 ; training/evaluation variance ratio in $[0.5, 2]$
- ▶ Raw returns: 10 of 41 fail (9 on the variance ratio: 1970s–80s bond volatility, small caps, several themes)
- ▶ After Procedure 1: **all 41 pass**; $\max |\text{acf}|$ 0.13, variance ratios in $[0.82, 1.25]$

Inference

- ▶ Every bound at $P = 20,000$, fixed in advance; merged over the two feature maps; pivotal standard errors; $\gamma = 0.05$, quantile 1.96
- ▶ All 41 series are reported and none selected: a per-series bound needs no multiplicity adjustment
- ▶ Panel statements: under no predictability about 2 of 41 bounds would be positive by chance; Benjamini–Hochberg at 5% FDR; Bonferroni over 41 as the conservative reading
- ▶ Benchmark: R^2_{OOS} of the ridge on the same features with penalty and feature map chosen by blocked 5-fold cross-validation on the training block

Merged Bound at $P = 20,000$: All 41 Series



Headline Numbers

- ▶ Reference ridge: R_{OOS}^2 negative for **every** series (between -173% and -43%)
- ▶ Cross-validated ridge: positive for 34 of 41, median **0.6%**, max **3.8%**; penalties chosen so large the forecast is nearly zero
- ▶ Merged statistic: positive for all 41, between 9% and 37%
- ▶ **95% bound positive for 36 of 41; median 15%**. BH at 5% FDR rejects for 35; Bonferroni leaves 25
- ▶ Open dot in the figure: the six series BH does not reject

The Penalty Cuts Both Ways

- ▶ **The bound needs a small penalty.** The LLG is the estimation error of the reference forecast: it is large only when the ridge is free to overfit ($z_{ref} = 0.01$, 20,000 features on 684 months)
- ▶ **Prediction needs a large penalty.** Cross-validation picks $z_{ref} \geq 1000$ for every one of the 41 series

z_{ref}	0.01	1	10	30	100	1000	3000	10^5
median R^2_{OOS} (%)	-130	-9.3	-1.5	-0.2	0.3	0.6	0.7	0.1

At $z_{ref} = 0.01$ the forecast is negative for **all 41** series. At the CV penalty the LLG correction is small, so a bound computed there would just restate the 1% the forecast already delivers.

The two ends measure different things; neither does the other's job.

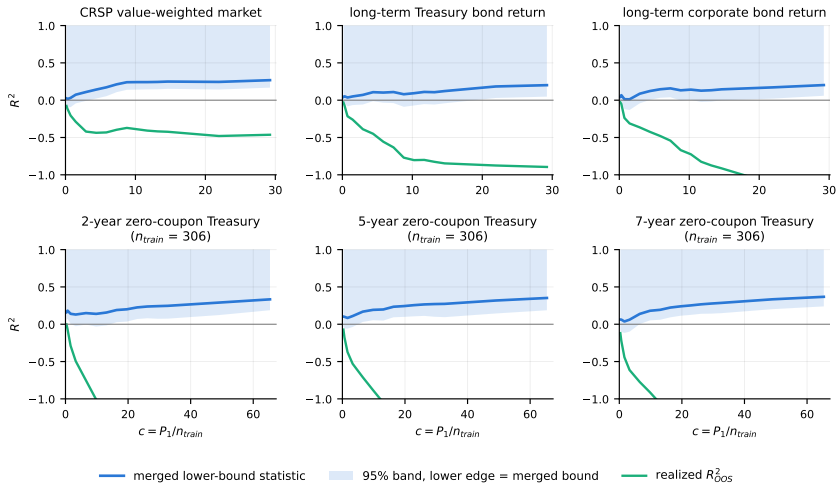
Market and Yield Curve

Series	R^2_{OOS} ref.	R^2_{OOS} CV	Statistic	95% bound
Market (CRSP value-weighted)	-46	1	27	17 ✓
Long-term Treasury bond	-89	1	20	5 ✓
Long-term corporate bond	-119	1	20	6 ✓
2-year zero-coupon Treasury	-166	2	33	19 ✓
5-year zero-coupon Treasury	-165	0	35	19 ✓
7-year zero-coupon Treasury	-173	0	37	24 ✓

Percent. ✓: Benjamini–Hochberg rejects no predictability at 5% FDR. Zero-coupon series start in 1961 ($T = 306$).

Market and Yield Curve: Against Complexity

Market and yield curve: merged bound against complexity



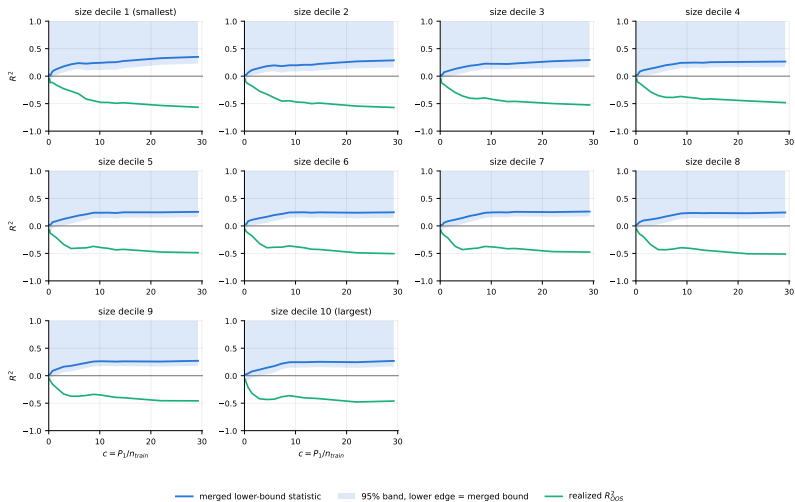
Size Deciles

Series	R^2_{OOS} ref.	R^2_{OOS} CV	Statistic	95% bound
Size decile 1 (smallest)	-57	1	35	23 ✓
Size decile 2	-57	1	29	16 ✓
Size decile 3	-52	0	29	16 ✓
Size decile 4	-48	1	27	16 ✓
Size decile 5	-49	1	26	16 ✓
Size decile 6	-50	1	25	15 ✓
Size decile 7	-48	1	26	17 ✓
Size decile 8	-51	1	25	15 ✓
Size decile 9	-46	1	27	18 ✓
Size decile 10 (largest)	-46	1	27	17 ✓

All ten deciles: bounds between 15% and 23%, largest for the smallest stocks.

Size Deciles: Against Complexity

Size deciles: merged bound against complexity



Factor Themes (JKP, long-short)

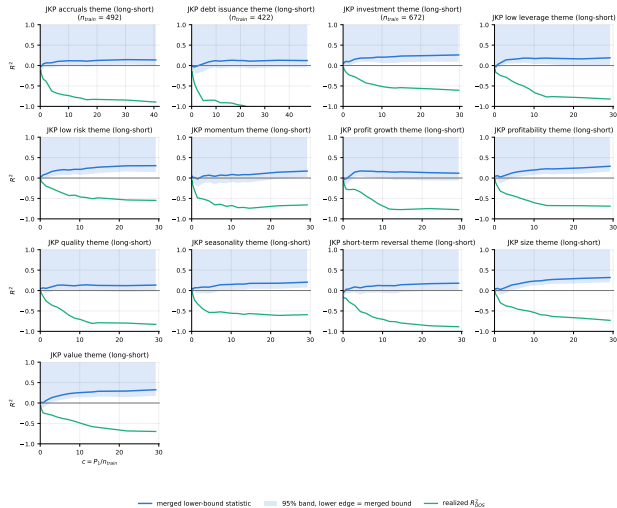
Series	R_{OOS}^2 ref.	R_{OOS}^2 CV	Statistic	95% bound
Size	−73	4	32	20 ✓
Value	−70	0	33	18 ✓
Profitability	−69	2	29	16 ✓
Low risk	−55	0	30	14 ✓
Investment	−61	−0	26	9 ✓
Seasonality	−59	1	20	7 ✓
Momentum	−66	0	17	1 ✓
Low leverage	−82	−0	19	1 ✓
Accruals	−89	−2	14	1 ✓
Short-term reversal	−89	0	18	0
Quality	−83	0	13	−2
Debt issuance	−110	−1	12	−4
Profit growth	−77	0	12	−7

Factor Timing

- ▶ The question for a long-short theme is not whether it earns a premium but whether its expected return **varies over time**
- ▶ Haddad, Kozak, Santosh (2020): anomaly-portfolio expected returns are predictable from their own valuation ratios; factor momentum (Arnott et al. 2019); principal portfolios (Kelly, Malamud, Pedersen 2023)
- ▶ Our bounds say how much there is to find: for size, value, profitability, low risk the population R^2 of the monthly theme return on 14 macro-financial signals is at least **14–20%**; a cross-validated ridge realizes at most 4%
- ▶ **Potential, not a verdict:** moving from past factor returns to the cross-section of signals, new timing variables, better methods

Factor Themes: Against Complexity

JKP themes: merged bound against complexity

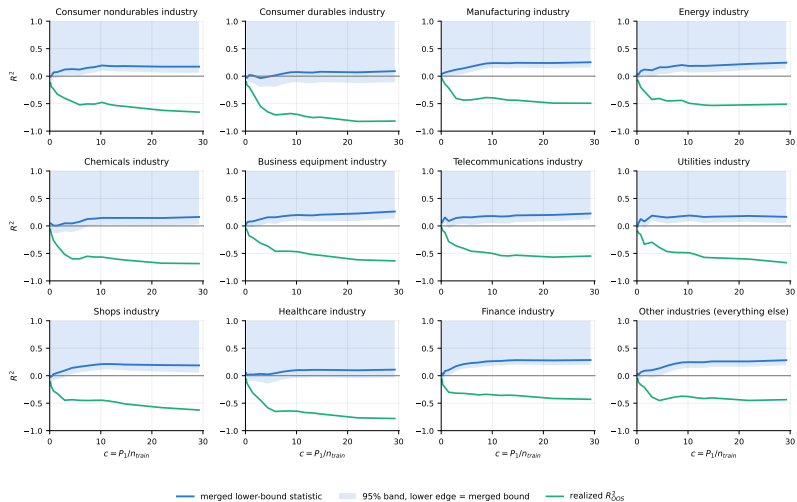


Industries (Fama–French 12)

Series	R^2_{OOS} ref.	R^2_{OOS} CV	Statistic	95% bound
Finance	−43	1	29	20 ✓
Other	−44	1	28	19 ✓
Manufacturing	−49	0	25	16 ✓
Energy	−51	−0	24	14 ✓
Business equipment	−64	1	26	13 ✓
Telecommunications	−55	2	23	12 ✓
Consumer nondurables	−65	−1	17	6 ✓
Shops	−62	1	19	6 ✓
Utilities	−67	−2	17	5 ✓
Chemicals	−69	2	16	3 ✓
Healthcare	−78	1	11	−3
Consumer durables	−82	0	9	−11

Industries: Against Complexity

Industries: merged bound against complexity



How to Read the Bounds

- ▶ Bounds are on the population R^2 of the Procedure 1 label (standardized AR(1) residual of the return) given the 14 signals: not the R^2 of raw returns, not trading profits
- ▶ They say: a nonlinear function of 14 familiar predictors explains a large share of monthly return variation, and a ridge on 20,000 random features recovers only a small part of it from 684 months
- ▶ Returns are far more predictable than realized forecasts suggest; the shortfall is a limit to learning from a short history with a given method, **not a limit on what can be learned**
- ▶ Caveats: coverage at high complexity is 86–95% in simulation; Bonferroni column (25 series) is the conservative reading

Table of Contents

- 1 Introduction
- 2 Limits to Learning
- 3 Deep Learning
- 4 Simulations
- 5 Empirics: 41 Return Series
- 6 Conclusions

Conclusions

- ▶ “Nothing works” is often a statement about *learning*, not about *reality*
- ▶ The LLG converts a failed out-of-sample forecast into a lower bound, with a confidence interval, on population predictability; it covers ridge, kernels, and ridge on the trained gradient features of any network (three-block design)
- ▶ For 36 of 41 U.S. return series the 95% bound is positive, median 15%, while every realized forecast is at or near zero
- ▶ **Where the gap is large, the data already contain predictability that current forecasts leave on the table. The bound says how much.**