How do investors learn as data becomes bigger? Evidence from a FinTech platform

Ahmed Guecioueur

INSEAD

Contributions

- Identify the effect of access to additional predictive signals on investors' ability to attain their objectives, disentangling from experience effects
- Experienced investors are able to exploit "wider" data availability
- Surprisingly, less experienced investors do not similarly benefit
- Rationalize these empirical findings by investors fearing model uncertainty when using historical data to predict the future

Institutional setting for identification

- Typically, learning dynamics are difficult to identify:
  1. Investor information sets are unknown
  2. Confounding effects: different preferences, horizons, etc
  3. Must proxy for experience
- Deal with all these issues by using a unique institutional setting as a laboratory: a FinTech platform (Quantiacs) that runs fixed-horizon trading contests for investors to systematically trade futures contracts on a daily basis using real market data on a simulation platform
- Identify learning dynamics by studying investor outcomes:
  1. Investors can only use a common set of predictive variables that the platform makes available to all; cannot upload their own
  2. Common objective: investors are incentivized to maximize their out-of-sample Sharpe Ratio over a common, fixed horizon – the out-of-sample "Live period" of each contest
  3. Panel dataset since investors can (and do) take part in multiple contests
- Data became bigger: Quantiacs suddenly expanded the set of common predictive variables in between the 7th & 8th trading contests

Learning with experience

- Investors better attain their (known) objective of maximizing their Live-period Sharpe Ratios as they gain in experience
- Consistent with prior work using brokerage or exchange data

Table 1. OLS & panel regressions of in-sample ("backtest") & out-of-sample ("live") performance outcomes against experience.

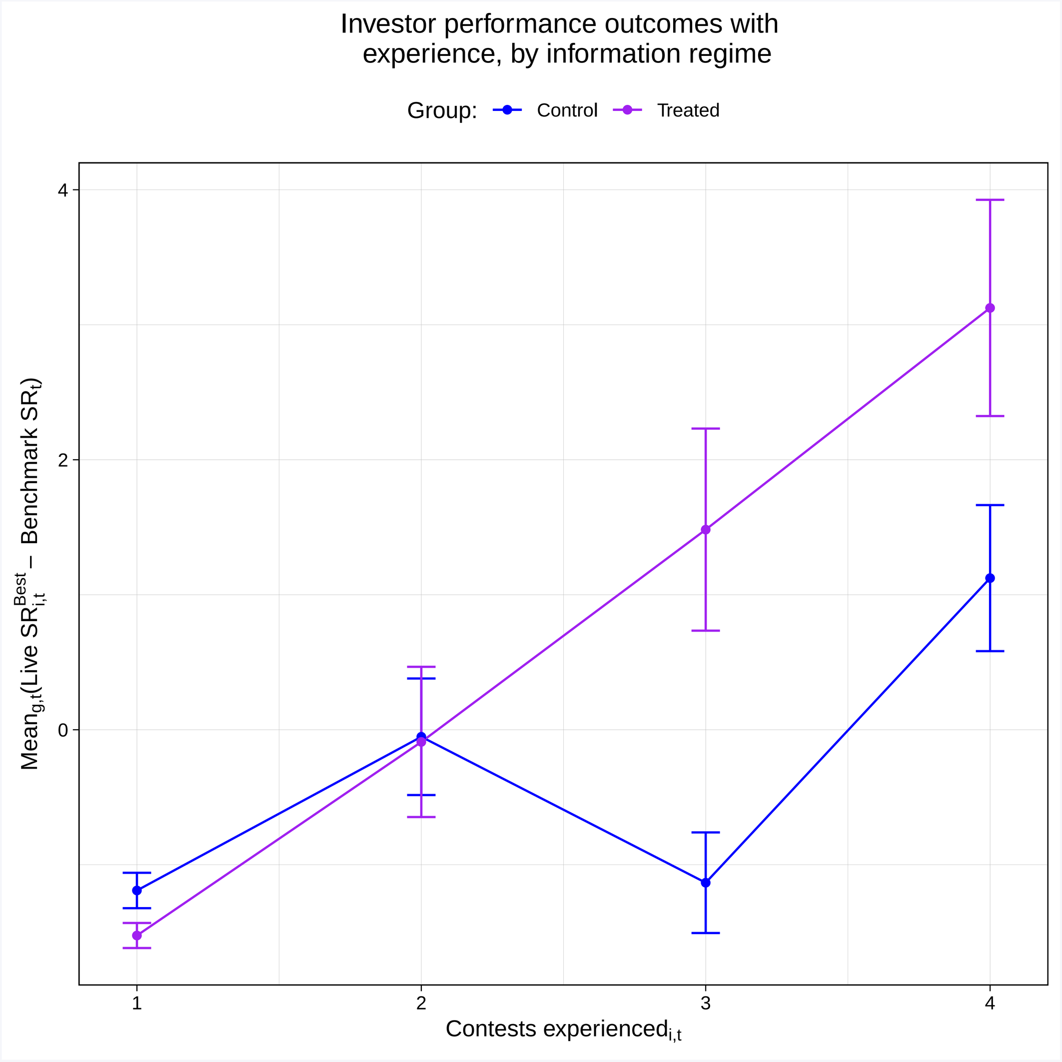
	Dependent variable:			
	Backtest $SR^{Best}_{i,t}$		Live $SR^{Best}_{i,t}$	
	OLS	*panel linear*	*OLS*	*panel linear*
	(1)	(2)	(3)	(4)
Contests experienced$_{i,t}$	1.161***	1.338***	0.445**	1.261***
	(0.055)	(0.505)	(0.178)	(0.456)
Intercept	✓		✓	
Contest FEs		✓		✓
Contestant FEs		✓		✓
Observations	874	874	874	874
R^2	0.156	0.024	0.035	0.040

Note: std. errs. (in parentheses) are double-clustered by contest & contestant. *p<0.1; **p<0.05; ***p<0.01

Learning from data interacts with experience

- Main empirical finding: availability of additional predictive variables is associated with a steepening of the learning dynamics…
- …with the beneficial effects of additional data availability found only for the more experienced investors
- Surprising because all investors should rationally be making use of all available signals to attain their (common) objectives

Figure 1. Learning dynamics in our main sample, split into the Treatment & Control groups. Bars represent standard errors.



- Treated investors are those who only trade from contest 8 onwards
- Time trends are controlled for by focussing on results in excess of the benchmark portfolio (EW daily rebalanced) presented to contestants as the default
- See paper for regression versions of Figure 1 with similar results

Learning from data: within-investor dynamics

- Similar steepening within-investor for those who traded both before & after the introduction of the new predictive variables

Ruling out potential explanations

- Selection effects: using "Heckit" regressions with exogenous first-stage covariates to correct for selection effects implies an *increased* magnitude of the learning effect, in agreement with the intuition of Linnainmaa (2011)
- Competition effects interacting with data abundance, as in Dugast and Foucault (2021): no significant interaction detected in this setting

Model uncertainty as explanation for results

- Experienced investors appear to benefit from wider data
- Why don't inexperienced investors also take advantage?
- Explanation rooted in model uncertainty:
  - Inexperienced investors fear model uncertainty more, leading them to discard some predictive signals that are available to them
  - As they gain in experience, investors shed some model uncertainty
- This mechanism is captured by the following model of investor learning

Investor learning under model uncertainty

- Follow Martin and Nagel (2021) in modeling each investor as behaving like an econometrician when using historical data
- Recall Quantiacs investors are incentivized to maximize out-of-sample (i.e. future) Sharpe Ratios over a fixed horizon,

$$\max_{\boldsymbol{w}} \frac{\boldsymbol{\mu}^T \boldsymbol{w}}{\sqrt{\boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w}}}, \tag{1}$$

- Assume the variance is known (Merton 1980) and that the expected return is a linear combination of the given predictive signal values, $\mu = \sum_{i=1}^m b_i s_i = \boldsymbol{sb}$. Then the investor must learn $\boldsymbol{b}$ based on historical expected returns from (similar but not identical) futures contracts that expired in the past $\boldsymbol{v}$ and corresponding historical signals $\boldsymbol{S}$.
- Fearing worst-case model uncertainty, her learning problem is thus to

$$\min_{\boldsymbol{b} \in \mathbb{R}^m} \max_{\boldsymbol{U} \in \mathcal{U}} ||\boldsymbol{v} - (\boldsymbol{S} + \boldsymbol{U})\boldsymbol{b}||_2, \tag{2}$$

where the model uncertainty can be represented as a matrix of signal-wise perturbations $\boldsymbol{U}$ that maximizes the ℓ_2 norm-based error for any choice of $\boldsymbol{b}$ and is constrained by an uncertainty set

$$\mathcal{U} := \{[\boldsymbol{u}_1\ \boldsymbol{u}_2\ \ldots\ \boldsymbol{u}_m] : ||\boldsymbol{u}_i||_2 \leq \delta_i\ \forall\ i = 1, \ldots, m\} \tag{3}$$

that is characterized by a set of upper bounds $\delta_i \geq 0$ on the ℓ_2 norm of each possible signal-wise disturbance $\boldsymbol{u}_i$.

- Assuming orthonormal $\boldsymbol{S}$, it follows from results by Xu, Caramanis, and Mannor (2010) and Tibshirani (1996) that the investor should use

$$\widehat{\mu} = \boldsymbol{s}\widehat{\boldsymbol{b}}, \tag{4}$$

in her portfolio choice problem, with elements of $\widehat{\boldsymbol{b}}$ being

$$\widehat{b}_k = \text{sign}(\boldsymbol{s}_k^T \boldsymbol{v}) \max\{|\boldsymbol{s}_k^T \boldsymbol{v}| - \lambda, 0\}, \tag{5}$$

where $\lambda \geq 0$ is a scaling of $\delta := \max_i \delta_i$ in (3).

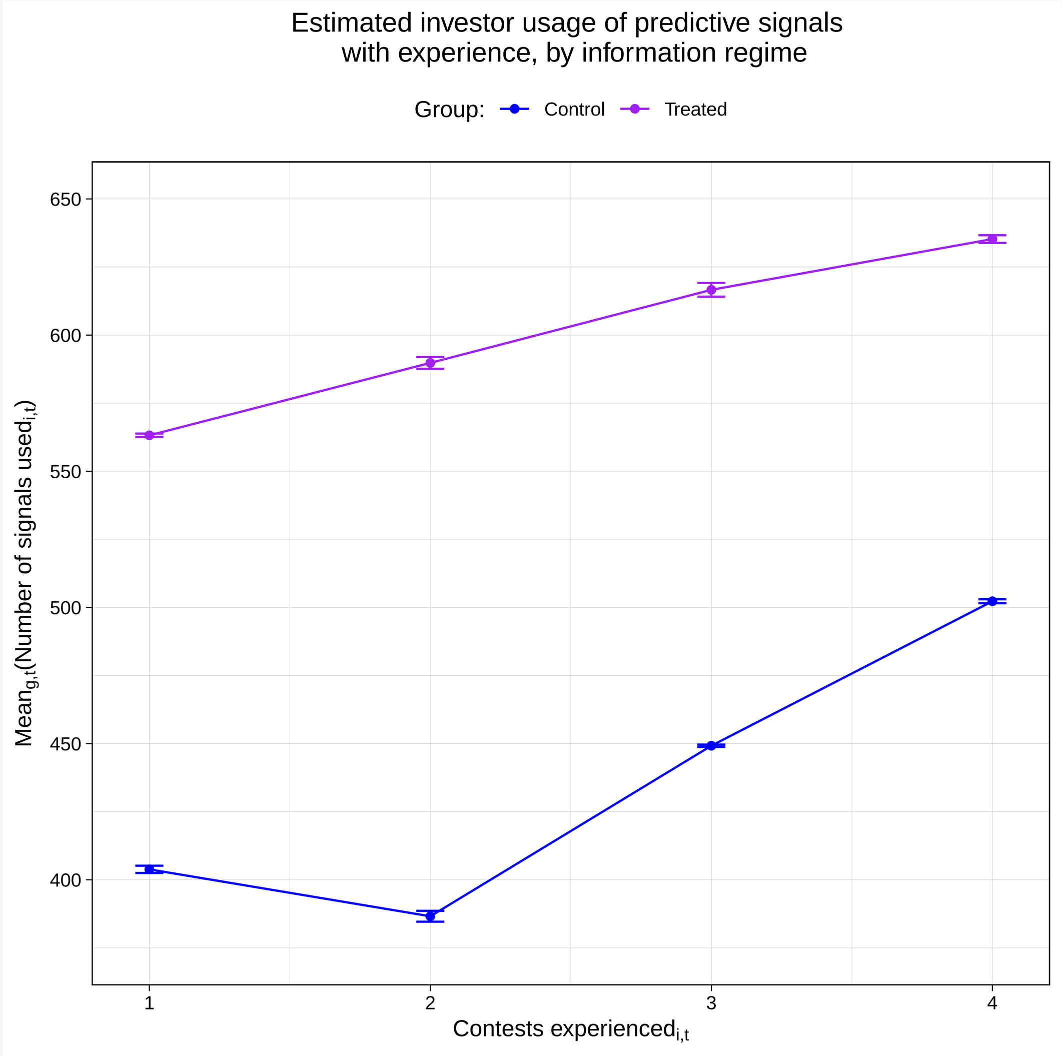
Implications of Eqn. (5)

- The investor should ignore signals whose historical predictive contribution is less than her subjective model uncertainty threshold λ
- The higher her fear of model uncertainty λ, the fewer predictive signals she should use (informal statement)
- Conjecture: investor's fear of model uncertainty λ falls with experience
- Therefore, the number of predictive variables she uses should increase with her experience

Estimating investors' usage of predictive variables

- Investors use more predictive variables as they gain in experience
- Once again, highlights the interaction between the complementary channels of learning with experience & learning from data

Figure 2. The dynamics of the estimated number of predictive variables used by investors to solve their portfolio choice problem. Bars represent standard errors.



- Set of hundreds of lagged predictive variables based on daily market data and (for contest 8 onwards) the values of the additional predictive variables
- For realism, the orthonormality assumption is dropped, so investor-portfolio-level estimates of $\widehat{\boldsymbol{b}}$ are performed using Friedman et al. (2007)'s lasso estimation procedure

More results in the paper

- Identification by exploiting the fact that all the new predictive variables happen to be lower-frequency macroeconomic variables
- Secondary results on: realized ex-post moments of returns, dispersions (within-investor & across-investor), overconfidence

References

Dugast, Jérôme, and Thierry Foucault. 2021. "Equilibrium data mining and data abundance." *HEC Paris Research Paper No. FIN-2020-1393, Université Paris-Dauphine Research Paper* 3710495.

Friedman, Jerome, Trevor Hastie, Holger Höfling, and Robert Tibshirani. 2007. "Pathwise coordinate optimization." *Annals of Applied Statistics* 1 (2): 302–332.

Linnainmaa, Juhani T. 2011. "Why do (some) households trade so much?" *Review of Financial Studies* 24 (5): 1630–1666.

Martin, Ian WR, and Stefan Nagel. 2021. "Market efficiency in the age of big data." *Journal of Financial Economics* (in print).

Merton, Robert C. 1980. "On estimating the expected return on the market: An exploratory investigation." *Journal of Financial Economics* 8 (4): 323–361.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–288.

Xu, Huan, Constantine Caramanis, and Shie Mannor. 2010. "Robust regression and Lasso." *IEEE Transactions on Information Theory* 56 (7): 3561–3574.