

DARE

Disagreement-Adaptive Risk-Aware Ensemble

Learned model trust, joint covariance risk control, and cost-aware trading

Cindy Wang, Lilian Li, Megan Wang · Yale School of Management · MGT 924, Problem Set 3

Over 1990–2023, DARE earns a Sharpe ratio of **4.16** before costs and **3.54** after an assumed 10 bp per dollar traded. It extends our previous model, CRE, which placed third on the competition leaderboard with an official Sharpe ratio of 3.9997 (3.98 before and 3.37 after costs in our backtest), with a disagreement-aware layer: instead of trusting the tree and ridge forecasts uniformly, DARE learns from past out-of-sample errors how much each deserves, and it pairs this with risk control under two covariance models at once and with cost-aware trading. Every new setting was fixed in advance or chosen each year from past data only, and the model was run once on the test period. The improvement holds at every cost level but is only marginally significant ($z = 1.77$, two-sided $p \approx 0.08$).

4.16

Sharpe ratio (s.e. 0.23)

3.54

Sharpe after
10bp costs

2.58

Sharpe after
25bp costs

-6.9%

maximum drawdown

90%

months with a gain

0.00

beta to the market

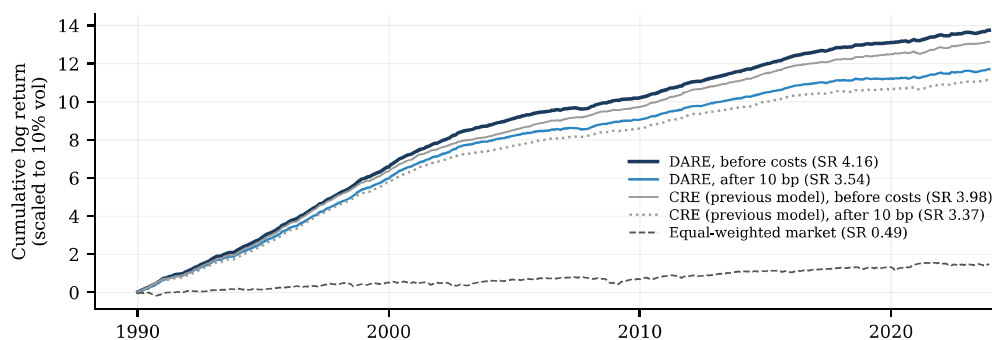


Figure 1: Cumulative log returns over the test period for DARE and our previous model, CRE, before and after an assumed 10 bp per dollar traded; each series is scaled to 10% volatility (Sharpe ratios are unchanged).

1 Why we added disagreement

Our previous model, CRE, blended gradient-boosted trees and a ridge regression and let a yearly out-of-sample vote choose among 25 configurations. Submitted to the competition as “CRE: GBM + ridge, factor and sample-covariance books”, it earned an official Sharpe ratio of 3.9997 with a maximum drawdown of -10.91% , third on the leaderboard; our own backtest gives 3.98, the small gap reflecting the leaderboard’s own return computation. It had a blind spot: it trusted each forecast equally for every stock, whether the two models agreed or contradicted each other. Disagreement between a linear and a non-linear model is a natural signal of forecast uncertainty, so we set out to use it. Our first attempt, submitted as “Disagreement-Adaptive Risk-Aware Ensemble”, shrank forecasts by a fixed function of the disagreement and lowered the official Sharpe ratio to 3.96 (maximum drawdown -10.88% ; 3.92 in our backtest): the models often disagree precisely where the trees detect effects the ridge cannot, so blanket shrinkage discarded real signal. The lesson was that disagreement must be *learned*, not assumed. We then conducted research on how to learn it, rebuilt the allocation around that idea, and fixed three implementation issues along the way (Table 1). The model described here, DARE, keeps that entry’s name but replaces its fixed rule with learned trust.

Protocol. New settings were either fixed in advance or chosen each January by a predefined rule using past data only; nothing was tuned on 1990–2023, and the full model was run once. One later fix (letting short samples that end before 1975 run the walk-forward) leaves the full-sample weights unchanged bit for bit.

Table 1: What we changed in our previous model.

Return-realisation dates	A month's labels are used only after every return in it is realised (<code>ecom_ret</code>), for training, validation, selection and the learned weights.
Short samples	No setting is tuned on fewer than 36 validation or 60 fitting months; selection needs 36 months of history; a sample ending before 1975 runs the full walk-forward.
Clustering	Risk clusters use the characteristics of every stock known before the year, whether or not its next return is known.
Learned disagreement	Weights on the two forecasts learned from past out-of-sample errors, by disagreement level (Section 2).
Joint risk control	One portfolio with risk at most the target under both covariance models and one common beta constraint (Section 3).
Smooth allocation	Weights on 12 configurations proportional to $\exp(\eta \cdot \text{trailing Sharpe})$ instead of the top three.
Trading costs	Holdings move a fraction τ towards the target; (η, τ) chosen yearly by trailing Sharpe net of 10 bp.

2 Forecasts and learned trust

Base forecasts. As in our previous model, a pooled ridge regression on monthly characteristic ranks and scikit-learn gradient-boosted trees on raw characteristics (127 leaves, minimum leaf 2,000 firm-months) are refit every January; both forecasts are standardised across stocks, giving z_i^{tree} and z_i^{ridge} .

Learning when disagreement matters. Each month, stocks are split into terciles of disagreement $|z_i^{\text{tree}} - z_i^{\text{ridge}}|$. Each January, using all matured out-of-sample forecasts from the previous ten years, the target y_i (next-month return standardised within each month and clipped at ± 3 , the same target the forecasters are trained on) is regressed on the two forecasts $z_i = (z_i^{\text{tree}}, z_i^{\text{ridge}})$ separately within each tercile g , and negative coefficients are then set to zero:

$$\tilde{b}_g = \arg \min_b \sum_{i \in g} (y_i - b' z_i)^2 + n_g \|b - \bar{b}\|^2, \quad b_g = \max(\tilde{b}_g, 0), \quad \mu_i = b_{g(i)}^{\text{tree}} z_i^{\text{tree}} + b_{g(i)}^{\text{ridge}} z_i^{\text{ridge}}.$$

Here $\bar{b}$ is the pooled least-squares fit across all terciles and n_g the tercile's number of observations; because the two forecasts are correlated, this penalty yields a precision-weighted average of the tercile and pooled estimates rather than an exact half-way point. The maximum is taken element by element, so a tercile in which neither model has predicted returns receives a small forecast: "both unreliable". Until 36 months of matured forecasts exist, both weights are 0.5. Unlike our first, fixed rule, the data now decide whether the trees, the ridge or neither deserve trust where the models disagree.

Why disagreement should matter. When a linear and a non-linear model agree, the signal is robust to the modelling choice; when they disagree, at least one is wrong. Forecast combination weights forecasts by their reliability (Bates and Granger, 1969), disagreement across an ensemble is a standard measure of predictive uncertainty in machine learning (Lakshminarayanan et al., 2017), and in finance differences of opinion carry information about returns (Diether et al., 2002).

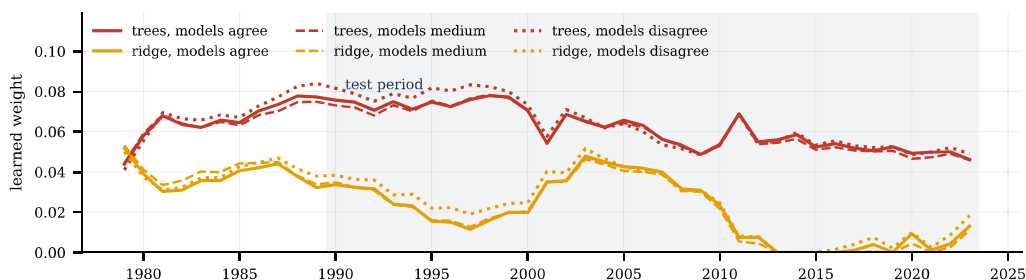


Figure 2: Learned weights on the tree (red) and ridge (orange) forecasts in the three disagreement terciles (line styles), re-estimated each January; shaded: test period. The near-overlap of the three line styles shows that the level of disagreement changed the weights little, while trust moved steadily from the ridge to the trees.

3 Portfolio construction

Two covariance models. Risk model A is a Barra-style factor model (2-digit SIC industries plus 13 characteristic clusters, exponentially weighted covariances; Jensen et al. 2024); risk model B is the Ledoit and Wolf (2004) shrunk

sample covariance of the last 756 daily returns.

Joint risk control. Our previous model optimised one portfolio under each model and averaged the two. DARE builds a single portfolio whose risk stays within the target under both, with one common beta-neutrality constraint, in the spirit of multi-prior portfolio choice (Garlappi et al., 2007):

$$\max_w \mu'w \quad \text{s.t.} \quad w'\Sigma_A w \leq \sigma^2, \quad w'\Sigma_B w \leq \sigma^2, \quad \mathbf{1}'w = 0, \quad \tilde{\beta}'w = 0.$$

By Lagrangian duality, the exact solution is a mean-variance portfolio under some mix $\omega\Sigma_A + (1 - \omega)\Sigma_B$. We solve a finite-grid approximation: for each $\omega \in \{0, 0.25, 0.5, 0.75, 1\}$ we compute the constrained mean-variance portfolio under that mix plus a ridge κ (a regularisation not shown in the displayed problem), scale it so that the riskier of the two predicted volatilities equals $\sigma = 10\%$, and keep the ω with the highest expected return (average chosen $\omega = 0.60$). $\tilde{\beta}$ averages the two models' betas on the equal-weighted market.

Smooth allocation and trading. Three forecasts (learned, 50/50 blend, trees) times four ridges $\kappa \in \{0, 0.25, 1, \infty\}$ give 12 configurations, weighted by $\exp(\eta \cdot \text{SR}_c)$ on their trailing ten-year Sharpe ratios. Holdings then move a fraction τ of the way towards this target each month (Gârleanu and Pedersen, 2013). Each January the pair (η, τ) , with $\eta \in \{1, 3, 10\}$ and $\tau \in \{1, 0.5, 0.25\}$, is the one with the best trailing Sharpe ratio net of an assumed 10 bp per dollar traded.

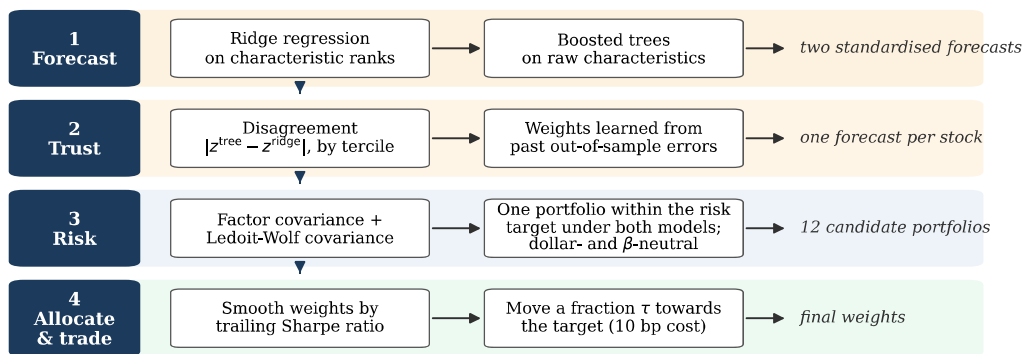


Figure 3: How DARE turns forecasts into weights: four decision layers, each re-estimated every January from realised data only.

4 Results

Table 2 reports the final model, its single configurations and ablations; standard errors follow Lo (2002).

Finding 1: DARE improves on CRE at every cost level. The Sharpe ratio rises from 3.98 to 4.16 before costs, from 3.37 to 3.54 after 10 bp and from 2.44 to 2.58 after 25 bp, and the maximum drawdown falls from -12.1% to -6.9% . The two models' returns are 0.92 correlated; a paired Sharpe-ratio test (Jobson and Korkie, 1981; Memmel, 2003) gives $z = 1.77$ (two-sided $p \approx 0.08$): consistent, but short of conventional significance.

Finding 2: joint risk control is the clearest gain. With the learned forecast and $\kappa = 1$, optimising under both covariance models jointly earns 4.07, against 3.88 for averaging two separately optimised books; ignoring risk altogether gives 2.84.

Finding 3: learned trust succeeds where fixed shrinkage failed. Learning from past errors beats fixed blending: the learned forecast earns 4.07 against 3.86 for the fixed 50/50 blend, whereas fixed shrinkage had lowered every configuration. It does not beat the trees alone (4.16), and the reason is instructive. Its weights hardly differ across disagreement terciles (trees 0.062, 0.061, 0.064; ridge 0.020, 0.019, 0.023 on average): trust shifted over time rather than with disagreement, a steady move from the ridge to the trees, the ridge weight reaching zero after 2012 (Figure 2), rather than regimes in which disagreement signals that both models fail.

Finding 4: the cost-aware rule rarely slowed trading. The past-only rule chose partial trading ($\tau < 1$) in only 17% of test years, so turnover fell modestly, from 99% to 92% of gross per month, and the after-cost gain mostly carries over the gross gain (Figure 4).

Finding 5: with hindsight, simpler choices would have done as well or better. Taking the top three configurations instead of smooth weights would have earned 4.34; always trading at full speed 4.32 before and 3.60 after 10 bp costs. These comparisons use the test period, so they are reported, not adopted: switching now would be tuning on the scored window.

Table 2: Test-period performance (1990–2023). SR: annualised Sharpe ratio (standard error). MDD: maximum drawdown with returns scaled to 10% volatility.

Portfolio	Configuration	SR	(s.e.)	<i>t</i> -stat	MDD (%)	Skew
DARE	final model, before costs	4.16	(0.23)	24.3	-6.9	0.20
DARE, 10 bp	after 10 bp per dollar traded	3.54	(0.21)	20.6	-8.0	0.22
DARE, 25 bp	after 25 bp per dollar traded	2.58	(0.19)	15.1	-22.9	0.24
Learned forecast	single configuration, $\kappa = 1$	4.07	(0.22)	23.7	-10.6	0.17
50/50 blend	single configuration, $\kappa = 1$	3.86	(0.22)	22.5	-7.5	0.16
Trees only	single configuration, $\kappa = 1$	4.16	(0.22)	24.2	-10.7	0.24
Two books averaged	learned, $\kappa = 1$, as in previous model	3.88	(0.22)	22.6	-12.5	0.17
Top-3 allocation	instead of smooth weights	4.34	(0.23)	25.3	-7.1	0.22
No turnover control	$\eta = 3$, $\tau = 1$, before costs	4.32	(0.23)	25.2	-6.4	0.24
No turnover control, 10 bp	$\eta = 3$, $\tau = 1$	3.60	(0.21)	21.0	-7.7	0.24
No risk model	learned, $\kappa = \infty$	2.84	(0.20)	16.6	-9.9	0.39
CRE (previous model)	before costs	3.98	(0.22)	23.2	-12.1	0.31
CRE (previous model), 10 bp	after 10 bp per dollar traded	3.37	(0.21)	19.6	-13.1	0.32

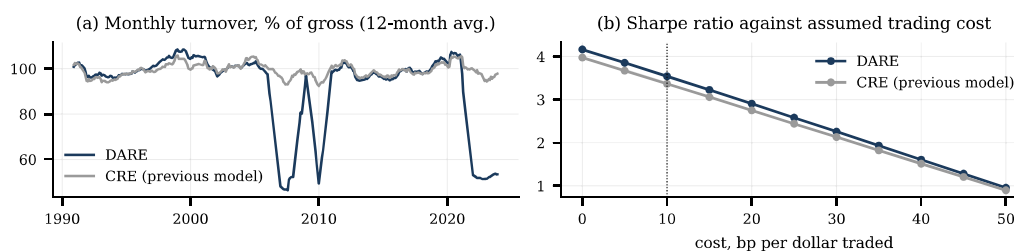


Figure 4: (a) Turnover, DARE against CRE. (b) Sharpe ratio as a function of the assumed trading cost; the dotted line marks the 10 bp used for selection.

Risk and neutrality. Realised volatility is 1.99 times the 10% target (CRE: 1.72): scaling to the riskier of two estimates still understates risk, because both models are fitted to the same past and the optimiser concentrates where measured risk is lowest. Every configuration is exactly dollar- and beta-neutral; after partial rebalancing the final holdings are rescaled but not re-projected, so neutrality holds approximately (net exposure 0.05% of gross on average, at most 1.56%; realised market beta 0.00).

5 Investor outlook

Expected performance. The 95% confidence interval for the Sharpe ratio is [3.72, 4.60], and the test period probably overstates the future: our previous model's Sharpe ratio fell from 6.9 in the 1990s to 2.95 in 2010–2023, consistent with the decay of published anomalies (McLean and Pontiff, 2016). The unscored 1975–1989 walk-forward earned 6.60, but most characteristics were discovered on data from that era. **Costs and risks.** After 25 bp per dollar traded the Sharpe ratio is 2.58. The book is long-short and leveraged; it is exposed to crowding, factor crashes, short-sale constraints and limited capacity in small stocks, and an investor seeking 10% realised volatility should target about half that ex ante. **Limitations.** Disagreement terciles turned out to be uninformative; conditioning trust on market states rather than disagreement levels is a natural next step. The tuned number of trees still varies between 50 and 600, the largest risk cluster still holds about 351 of 402 characteristics, and the improvement over our previous model is only marginally significant.

6 Reproducibility and disclosures

Development log. ① Studied the three leading leaderboard entries. ② Validated the pipeline on simulated data with the competition's schema. ③ CRE on the leaderboard: Sharpe ratio 3.9997, maximum drawdown -10.91% , ranked third (1.7 hours locally on a MacBook). ④ Fixed disagreement shrinkage on the leaderboard: Sharpe ratio 3.96, maximum drawdown -10.88% . ⑤ Conducted research on using disagreement: learning trust from past errors, plus three fixes and three further changes. ⑥ DARE with learned trust, run once: 4.16 before costs, 3.54 after 10 bp; all competition checks passed on the real data.

The code is deterministic, uses only pre-installed packages pinned to exact versions, and needs no GPU; two runs give identical weights, a run on truncated data reproduces all earlier weights exactly, and every test-period observation receives a weight. *Credits.* The idea of using the disagreement between our two models came from our group, and we developed it through further research into learning trust from past errors, joint risk control and cost-aware trading (Section 1). The factor model follows Jensen et al. (2024), and the cluster count follows Jensen et al. (2023). Leaderboard entries inspired our use of Ledoit–Wolf shrinkage (P. Geertsema) and two risk models with beta neutrality (A. Keshri). *AI use.* We used Claude (Anthropic) as a coding, graph-generation assistant under our direction; we downloaded the data, ran the analysis and all checks, and are responsible for all content.

References

- Bates, J. M., Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly* 20, 451–468.
- Diether, K. B., Malloy, C. J., Scherbina, A. (2002). Differences of opinion and the cross section of stock returns. *Journal of Finance* 57, 2113–2141.
- Gârleanu, N., Pedersen, L. H. (2013). Dynamic trading with predictable returns and transaction costs. *Journal of Finance* 68, 2309–2340.
- Garlappi, L., Uppal, R., Wang, T. (2007). Portfolio selection with parameter and model uncertainty: a multi-prior approach. *Review of Financial Studies* 20, 41–81.
- Hellum, O., Jensen, T. I., Kelly, B., Pedersen, L. H. (2025). The power of the common task framework. SSRN 5242901.
- Jensen, T. I., Kelly, B., Pedersen, L. H. (2023). Is there a replication crisis in finance? *Journal of Finance* 78, 2465–2518.
- Jensen, T. I., Kelly, B., Malamud, S., Pedersen, L. H. (2024). Machine learning and the implementable efficient frontier. *Review of Financial Studies*, forthcoming.
- Jobson, J. D., Korkie, B. M. (1981). Performance hypothesis testing with the Sharpe and Treynor measures. *Journal of Finance* 36, 889–908.
- Lakshminarayanan, B., Pritzel, A., Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems* 30.
- Ledoit, O., Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* 88, 365–411.
- Lo, A. W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal* 58, 36–52.
- McLean, R. D., Pontiff, J. (2016). Does academic research destroy stock return predictability? *Journal of Finance* 71, 5–32.
- Mommel, C. (2003). Performance hypothesis testing with the Sharpe ratio. *Finance Letters* 1, 21–23.