Maximally Machine-Learnable Portfolios

Philippe Goulet Coulombe
Université du Québec à Montréal

Maximilian Göbel
Bocconi University

June 20, 2024
SUERF | UniCredit: AI in Banking and Finance
p.gouletcoulombe@gmail.com

What Can Be Learned About the Future?

- Predicting classic portfolios' return at short horizons with modern machine learning (ML) tools and gigatons of data delivers mostly small out-of-sample R^2's (e.g., Gu et al. (2020)). And outperformance often tarnishes (or vanishes) after the mid-2000s.
- Yet, predictability is *risk-adjusted* profit.
- Maybe we don't need to predict all the future, but merely find one piece of it for which we have higher forecasting power.

Maximally Predictable Portfolios

- Lo and MacKinlay (1997) proposes MPPs made of a handful of stocks predicted by a sparse linear factor model.
- However, increased predictability may also likely be found in
    1. complex portfolios of many stocks
    2. predicted by *nonlinear* ML tools
    3. based on more than a few factors (here: nonlinear mean-reversion and the well-known Welch and Goyal (2007) dataset).

MACE

Usually:

$$\underbrace{Y_{t+h}}_{1\times 1} = f\,\underbrace{(\mathbf{X}_t)}_{1\times K} + \varepsilon_{t+h} \tag{1}$$

Alternating Conditional Expectations (ACE, Breiman and Friedman (1985))

$$g\,\underbrace{(Y_{t+h})}_{1\times 1} = f\,\underbrace{(\mathbf{X}_t)}_{1\times K} + \varepsilon_{t+h} \tag{2}$$

Multivariate Alternating Conditional Expectations (MACE)

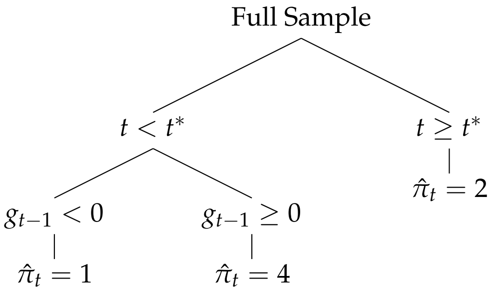
$$\underbrace{g\,\underbrace{(\mathbf{Y}_{t+h})}_{1\times N}}_{1\times 1} = f\,\underbrace{(\mathbf{X}_t)}_{1\times K} + \varepsilon_{t+h} \tag{3}$$

Detour: A Primer on Random Forest

What is a tree?

RF is a diversified ensemble of regression trees. What is a tree?

- Let π_t be inflation at time t.
- t^* is inflation targeting implementation date.
- Let g_t be some measure of output gap.



Why Random Forest

The usual

- It works tremendously well on all sorts of data
- More often than not, it's better than Neural Networks for tabular data.
- Can approximate a wide range of nonlinearities
- Tuning parameters do not alter prediction much
- Can easily deal with a very large X (no matrix operation involved)
- In general, does not overfit (Goulet Coulombe 2020)

For today's application :

- Reliable (i.e., non-overfitted) out-of-bag predictions readily available.

▸ MACE and Traditional Mean-Variance Optimization

Experiment I – Daily Returns Prediction

$\mathbf{r}_{t+1}$: Individual stock returns from Yahoo Finance for firms listed on the NASDAQ. Keep $N \in \{20,50,100\}$ of them with highest market capitalization on January 3rd 2017 (hence, no look ahead bias for the *test* sample).

Training Period: February 3rd 2000 - December 30th 2016

Out-of-Sample: January 03rd 2017 - December 7th 2022, *for now* MACE estimated once on 2016-30-12 and projected on the whole test set.

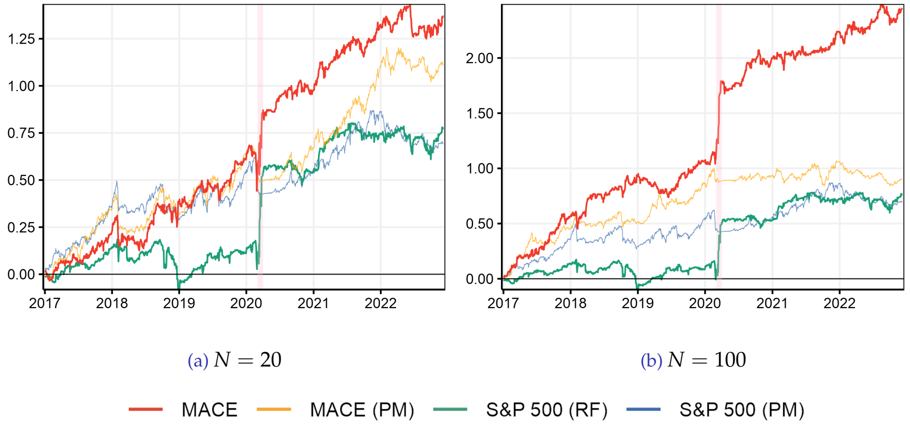
$\mathbf{X}_t$: one-sided moving-averages of past portfolio-returns

Experiment I – Results

	R^2_{OOS}	R^2_{CovidW1}	$R^2_{\neg\text{CovidW1}}$	R^2_{2022}	r^A	SR	r^A_{2022}	Ω
N = 20								
MACE	**3.42**	**7.86**	**0.56**	-0.55	**23.10**	0.99	5.03	**1.18**
MACE (PM)	0.01	-0.05	0.04	-0.12	18.71	**1.04**	0.75	1.13
EW (RF)	-4.71	-9.39	-1.21	**0.33**	9.15	0.34	**36.43**	1.02
N = 100								
MACE	**4.05**	**12.20**	**0.86**	**0.23**	**41.36**	**1.59**	**23.93**	**1.33**
MACE (PM)	-0.02	0.01	-0.03	-0.22	15.20	0.91	-16.81	1.09
EW (RF)	0.00	1.17	-0.99	-0.94	9.88	0.38	-10.48	1.03
S&P 500 (RF)	2.88	7.41	-0.14	0.09	13.29	0.64	2.80	1.09
S&P 500 (PM)	-0.01	-0.06	0.03	-0.32	11.65	0.69	-17.98	1.06

Notes: The first column-wise panel consists of out-of-sample R^2's for different test (sub-)samples. The second are economic metrics, where r^A := Annualized Returns, SR := Sharpe Ratio, r^A_{2022} := Annualized Returns for 2022, Ω := Omega Ratio.

Experiment I – Log Cumulative Returns



(a) $N = 20$ (b) $N = 100$

▸ An Implicit Statistical Test

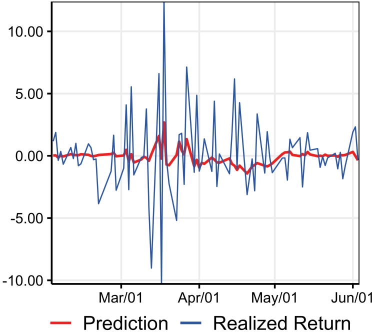
Experiment I – Understanding March 2020

Table: First Order Approximation to Nonlinear Dynamics in Returns

	CovidW1		¬CovidW1		2008	
	MACE_{100}	S&P 500 (RF)	MACE_{100}	S&P 500 (RF)	MACE_{100}	S&P 500 (RF)
Coefficient	-0.451	-0.402	-0.074	-0.036	-0.104	-0.156
Standard Error	0.097	0.100	0.027	0.027	0.063	0.062
$\text{Corr}(\hat{r}_t^{\text{RF}}, r_{t-1})$	-0.616	-0.577	0.102	0.014	-0.243	-0.173

Notes: This table reports the AR(1) coefficient and its standard error for two different return series on two non-overlapping subsamples of the test set spanning from 2016 to 2022 as well as 2008 from the training sample. $\text{Corr}(\hat{r}_t^{\text{RF}}, r_{t-1})$ is the correlation between Random Forest's prediction of the portfolio's return and the realized return on the previous business day.

$\rightarrow$ MACE_{100} hits a local R^2 of 20% and its sign prediction accuracy is 78 % by leveraging strong day-to-day mean reversion during episodes of high volatility.



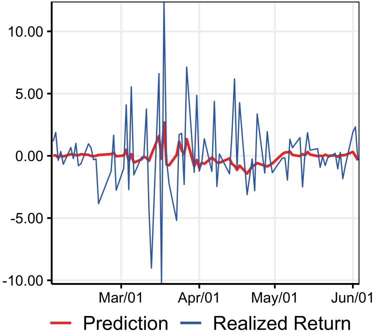
Experiment I – Understanding March 2020

Table: First Order Approximation to Nonlinear Dynamics in Returns

	CovidW1		¬CovidW1		2008	
	MACE_{100}	S&P 500 (RF)	MACE_{100}	S&P 500 (RF)	MACE_{100}	S&P 500 (RF)
Coefficient	$\boxed{-0.451}$	-0.402	-0.074	-0.036	-0.104	-0.156
Standard Error	0.097	0.100	0.027	0.027	0.063	0.062
$\text{Corr}(\hat{r}_t^{\text{RF}}, r_{t-1})$	-0.616	-0.577	0.102	0.014	-0.243	-0.173

Notes: This table reports the AR(1) coefficient and its standard error for two different return series on two non-overlapping subsamples of the test set spanning from 2016 to 2022 as well as 2008 from the training sample. $\text{Corr}(\hat{r}_t^{\text{RF}}, r_{t-1})$ is the correlation between Random Forest's prediction of the portfolio's return and the realized return on the previous business day.

→ MACE_{100} hits a local R^2 of 20% and its sign prediction accuracy is 78 % by leveraging strong day-to-day mean reversion during episodes of high volatility.



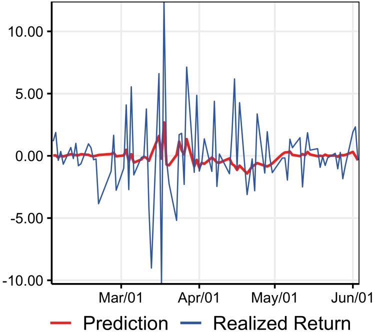
Experiment I – Understanding March 2020

Table: First Order Approximation to Nonlinear Dynamics in Returns

	CovidW1		¬CovidW1		2008	
	MACE_{100}	S&P 500 (RF)	MACE_{100}	S&P 500 (RF)	MACE_{100}	S&P 500 (RF)
Coefficient	$\boxed{-0.451}$	-0.402	-0.074	-0.036	-0.104	-0.156
Standard Error	0.097	0.100	0.027	0.027	0.063	0.062
$\text{Corr}(\hat{r}_t^{\text{RF}}, r_{t-1})$	$\boxed{-0.616}$	-0.577	0.102	0.014	-0.243	-0.173

Notes: This table reports the AR(1) coefficient and its standard error for two different return series on two non-overlapping subsamples of the test set spanning from 2016 to 2022 as well as 2008 from the training sample. $\text{Corr}(\hat{r}_t^{\text{RF}}, r_{t-1})$ is the correlation between Random Forest's prediction of the portfolio's return and the realized return on the previous business day.

$\rightarrow$ MACE_{100} hits a local R^2 of 20% and its sign prediction accuracy is 78 % by leveraging strong day-to-day mean reversion during episodes of high volatility.

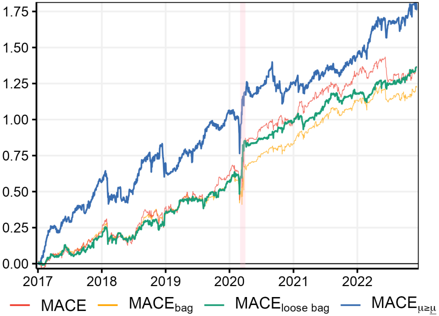


Experiment I – Some Refinements

Table: Benefits of Refinements for MACE_{20}

	r^A	SR	Ω
MACE	23.10	0.99	1.18
MACE_{bag}	20.60	1.07	1.20
$\text{MACE}_{\text{loose bag}}$	20.50	**1.21**	**1.21**
$\text{MACE}_{\mu \geq \underline{\mu}}$	**29.76**	1.17	1.19

Notes: Economic metrics are r^A := Annualized Returns, SR := Sharpe Ratio, Ω := Omega Ratio. All statistics but SR and Ω are in percentage points. Returns and risk-reward ratios are based on trading each portfolio using a simple mean-variance scheme with risk aversion parameter $\gamma = 5$. Numbers in **bold** are the best statistic of the column.



▸ Factor-Based Explainability

Conclusion and (ongoing) Excursions

- MACE: creates maximally machine-learnable linear combinations of $\boldsymbol{Y}_t$.
- In the era of abundant alternative data, it offers an algorithmic solution to an increasingly common question: what is this dataset useful for?
- Maximally predictable macroeconomic aggregates
  - **Core inflation** as the combination (or trimming) of CPI components that is maximally predictable – or is maximally predictable based on the subset of $\boldsymbol{X}_t$ that the central bank actually influences.
  - **Synthetic USA** as the combination (or trimming) of county- or state-level unemployment rates that is maximally predictable.

Appendix

MACE and Traditional Mean-Variance Optimization

- We build a portfolio return $z_{t+1}(\boldsymbol{w})$, characterized by a relative weight vector $\boldsymbol{w}$ combining single security returns.
- Our overall trading position is determined by $\omega_{t+1} \in [-1, 2]$, which is the absolute position over the portfolio.
- Relative weights are fixed (unless re-estimated), but absolute weights are changing every period based on forecasts for $z_{t+1}(\boldsymbol{w})$ and its volatility.
- The problem looks like

$$\max_{\omega_{t+1},\, \boldsymbol{w}} \quad E_t\left[\omega_{t+1} z_{t+1}(\boldsymbol{w}) \; - \; 0.5\,\gamma\,\omega_{t+1}^2\,\sigma_{t+1}^2(\boldsymbol{w})\right]$$

where γ is the risk aversion parameter. Plugging in the solution for ω_{t+1} (i.e., $\omega_{t+1} = \frac{1}{\gamma}\,\frac{\hat{z}_{t+1}(\boldsymbol{w})}{\hat{\sigma}_{t+1}^2(\boldsymbol{w})}$) conditional on $\boldsymbol{w}$, re-arranging, etc, we get

$$\max_{\boldsymbol{w}} \quad \frac{1}{2\gamma} \times \frac{\hat{z}_{t+1}^2(\boldsymbol{w})}{\hat{\sigma}_{t+1}^2(\boldsymbol{w})} \quad \Leftrightarrow \quad \max_{\boldsymbol{w}} \quad R_{t+1}^2(\boldsymbol{w})$$

back

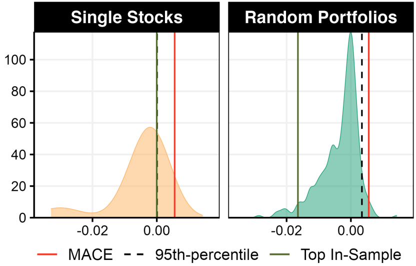
Hyperparameters

Table: **Summary of Tuning Parameters and Their Values in Applications**

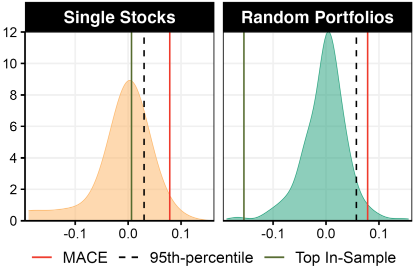
	Monthly Data	Daily Data ($N \in \{20{,}50\}$)	Daily Data ($N = 100$)
η	0.05	0.01	0.05
$s_{\max}$	150	250	500
`stopping.rule`	$s = s_{\max}$	early stopping	early stopping
`mtry`	1/3	1/10	1/10
`minimal.node.size`	20	200	200
`block.size`	24 months	2 months	2 months
`subsampling.rate`	80%	80%	80%
`number.of.trees`	500	1500	1500
λ	$R^2_{s,\text{train}}(\lambda) = 0.05$	$R^2_{s,\text{train}}(\lambda) = 0.01$	$R^2_{s,\text{train}}(\lambda) = 0.01$

Experiment I – An Implicit Statistical Test

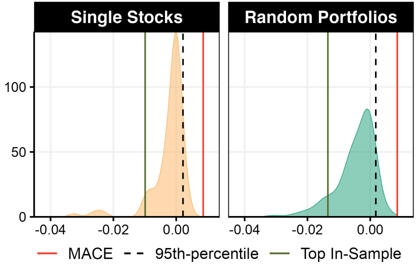
back



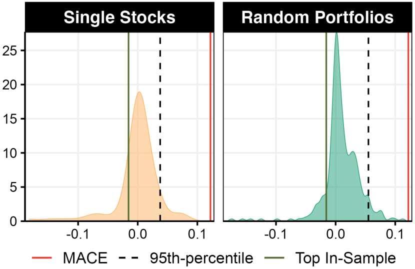
(a) $N = 20$, $R^2_{\neg\text{CovidW1}}$



(b) $N = 20$, R^2_{CovidW1}



(c) $N = 100$, $R^2_{\neg\text{CovidW1}}$



(d) $N = 100$, R^2_{CovidW1}

Experiment I – Factors-Based Explainability

Table: Factor Regressions Results

	$N = 20$			$N = 100$	
	MACE	MACE (PM)	$\text{MACE}_{\text{loose bag}}$	MACE	MACE (PM)
	Sample Period: 2016/01/01 - 2022/12/07				
α	0.05	0.01	0.05^{**}	0.13^{***}	0.00
MKT	0.64^{***}	0.59^{***}	0.45^{***}	0.35^{***}	0.39^{***}
SMB	-0.31^{***}	-0.07	-0.22^{***}	-0.07	-0.05
HML	0.01	-0.28^{***}	0.00	-0.11	-0.48^{***}
RMW	-0.02	0.17^{***}	0.00	0.02	0.41^{***}
CMA	0.55^{***}	0.75^{***}	0.44^{***}	0.15	0.24^{***}
MOM	0.01	-0.01	0.00	0.03	-0.02^{*}
R^2	0.29	0.43	0.29	0.08	0.41

back

Experiment I – Transaction Costs

Table: Daily Stock Returns after Transaction Costs

	0.01%			0.015%			0.03%		
	r^A	SR	Ω	r^A	SR	Ω	r^A	SR	Ω
MACE_{20}	19.72	0.84	1.14	18.07	0.77	1.12	13.14	0.56	1.07
$\text{MACE}_{\text{loose bag}}$	20.23	1.19	1.20	18.82	1.11	1.18	14.58	0.86	1.11
MACE_{100}	32.83	1.26	1.24	28.58	1.1	1.20	15.81	0.61	1.08

Notes: This table reports annualized returns (r^A), Sharpe ratio (SR) and the Omega Ratio (Ω) for various MACE portfolios after accounting for transaction costs with $\mathfrak{C} \in \{0.01\%, 0.015\%, 0.03\%\}$.

Experiment II – Monthly Returns Prediction

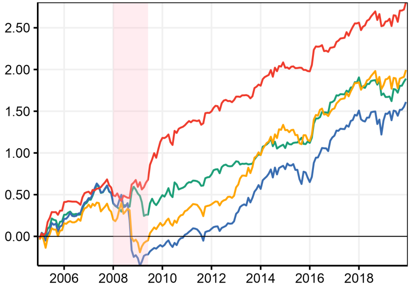
$\mathbf{r}_{t+1}$: Individual stock returns from CRSP for firms listed on the NYSE, AMEX, and NASDAQ (Gu et al., 2020)

$\mathbf{X}_t$: The well-known 16 macroeconomic indicators of Welch and Goyal (2007) (we include 12 lags)

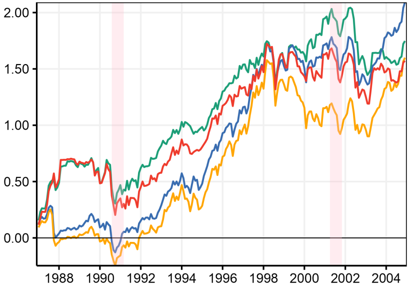
Training Period: 1957m3 - 1986m12

Out-of-Sample: 1987m1- 2019m12, expanding window, with MACE re-estimated every 3 months.

Experiment II – Log Cumulative Returns



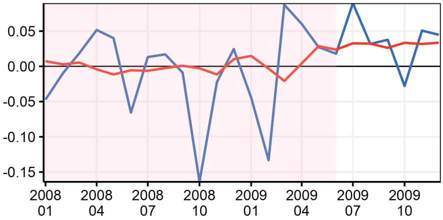
(a) (01/2005-12/2019)



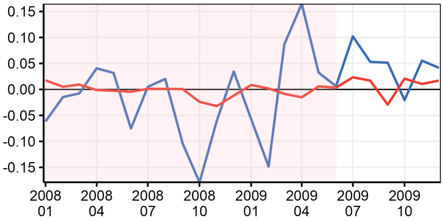
(b) (01/1987-12/2004)

	01/2005 - 12/2019				01/1987 - 12/2004			
	R^2_{OOS}	r^A	SR	DD^{MAX}	R^2_{OOS}	r^A	SR	DD^{MAX}
MACE	**4.13**	**18.57**	**1.04**	**27.02**	-7.99	6.40	0.29	71.80
MACE (PM)	-0.30	11.51	0.54	70.84	-0.43	6.88	0.33	84.57
EW (RF)	1.88	11.60	0.65	42.73	-8.24	7.82	0.38	66.24
EW (PM)	-0.24	8.39	0.38	113.95	**-0.39**	**10.00**	**0.50**	**53.16**

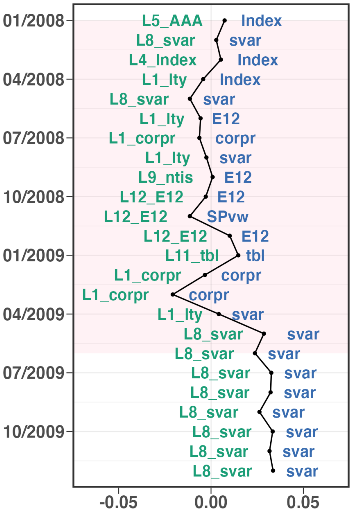
Experiment II – Some Interpretability



(a) MACE



(b) EW (RF)



(c) MACE: Most Important Predictors

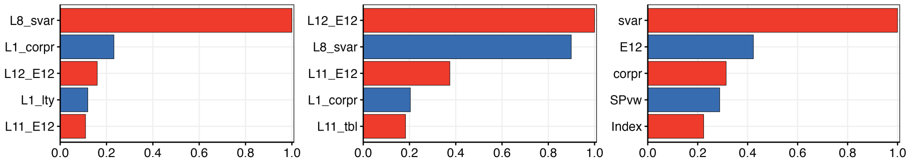
Experiment II – Results

Table: **Summary Statistics for Monthly Stock Returns Prediction**

	01/2005 - 12/2019				01/1987 - 12/2004			
	R^2_{OOS}	r^A	SR	DD^{MAX}	R^2_{OOS}	r^A	SR	DD^{MAX}
Main Results								
MACE	**4.13**	**18.57**	**1.04**	**27.02**	-7.99	6.40	0.29	71.80
MACE (PM)	-0.30	11.51	0.54	70.84	-0.43	6.88	0.33	84.57
Benchmarks								
EW (RF)	1.88	11.60	0.65	42.73	-8.24	7.82	0.38	66.24
EW (PM)	-0.24	8.39	0.38	113.95	**-0.39**	**10.00**	**0.50**	**53.16**
S&P 500 (RF)	-3.53	11.07	0.65	45.57	-13.77	2.63	0.12	125.03
S&P 500 (PM)	-0.52	6.15	0.34	101.70	-0.48	8.25	0.47	78.27
Refinements								
MACE_{bag}	**4.84**	16.56	0.95	**23.43**	-7.61	7.19	0.34	65.71
$\text{MACE}_{\mu \geq \underline{\mu}}$	4.27	**19.04**	0.99	38.32	-4.04	**11.46**	**0.53**	59.30

Notes: The first column-wise panel consists of out-of-sample R^2's for different test (sub-)samples. The second are economic metrics, where r^A := Annualized Returns, SR := Sharpe Ratio, and DD^{MAX} = Maximum drawdown. All statistics but SR are in percentage points. Returns and risk-reward ratios are based on trading each portfolio using a simple mean-variance scheme with risk aversion parameter $\gamma = 3$. PM means the prediction is based on the respective prevailing mean with a lookback period of twenty years, while RF means using that of a Random Forest. Numbers in **bold** are the best statistic within the first two panels (that is, excluding MACE refinements). Numbers in **green** are the best statistic of whole column.

Experiment II – Variable Importance

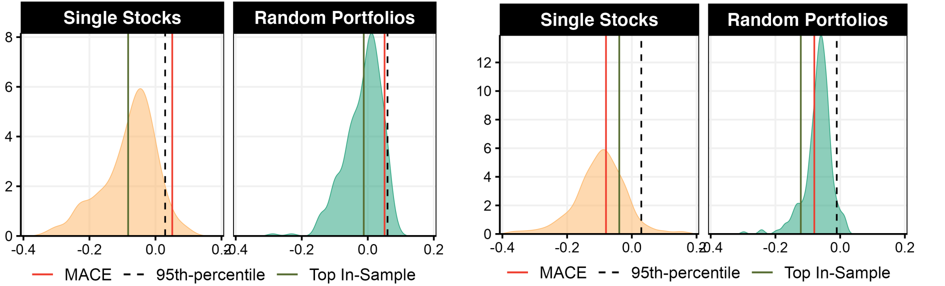


(a) MACE: VI_i^{oos} (b) EW (RF): VI_i^{oos} (c) MACE: VI_g^{oos}

Notes: The bars represent the VI of predictor i and grouped versions (VI_g, i.e. , summing the Shapley Values across all lags of variable i), scaled by the corresponding maximum value.

Figure: Shapley Value Importance: 01/2008 - 12/2009

Experiment II – An Implicit Statistical Test



(a) (01/2005-12/2019)

(b) (01/1987-12/2004)

Figure: R^2 Comparison of MACE to Random Alternatives

Experiment II – Transaction Costs

Table: Monthly Stock Returns after Transaction Costs

	0.05%			0.01%			0.1%		
	r^A	SR	Ω	r^A	SR	Ω	r^A	SR	Ω
				01/2005 - 12/2019					
MACE	18.43	1.03	1.87	17.90	1.00	1.82	8.37	0.46	1.16
MACE_{bag}	16.64	0.96	1.71	16.13	0.93	1.67	6.94	0.40	1.09
$\text{MACE}_{\mu \geq \underline{\mu}}$	19.33	1.00	1.77	19.08	0.99	1.75	14.60	0.76	1.46
				01/1987 - 12/2004					
MACE	7.57	0.34	1.10	7.02	0.32	1.07	-2.90	-0.13	0.75
MACE_{bag}	7.72	0.37	1.11	7.22	0.35	1.08	-1.73	-0.08	0.78
$\text{MACE}_{\mu \geq \underline{\mu}}$	11.79	0.56	1.28	11.49	0.54	1.26	6.03	0.28	1.04

Notes: This table reports annualized returns (r^A), Sharpe ratio (SR) and the Omega Ratio (Ω) for various MACE portfolios after accounting for transaction costs with $\mathfrak{C} \in \{0.05\%, 0.01\%, 0.1\%\}$.

Experiment II – Factor-Based Explainability

Table: Factor Regression Results

	01/1987-12/2004		01/2005-12/2019	
	MACE	MACE (PM)	MACE	MACE (PM)
α	-0.33	-0.54^{**}	0.99^{***}	0.11
MKT	1.09^{***}	1.26^{***}	0.58^{***}	1.20^{***}
SMB	0.22	0.26^{***}	0.42^{***}	0.33^{**}
HML	0.69^{***}	0.82^{***}	-0.07	-0.05
RMW	0.16	0.31^{**}	0.51^{***}	0.46^{**}
CMA	0.00	-0.03	0.69^{**}	0.43^{*}
MOM	-0.14	-0.04	0.09	0.01
R^2	0.50	0.72	0.26	0.67