Designing Time-Series Models With Hypernetworks

Filip Staněk
stanek.fi@gmail.com
CERGE-EI

The Future of Forecasting & The M6 Competition Conference

October 30, 2023

Outline

forecasting challenge (long form at https://bit.ly/M6_for_ch)

- MtMs model:
  general description of the model
- Application:
  details of its application to M6

investment challenge (long form at https://bit.ly/M6_inv_ch)

- description of the principles guiding portfolio decisions

Motivation

problem setup:
We observe samples from multiple similar but not necessarily identical DGPs.
How to best model them?

two extreme approaches:

- local modeling:
  parsimonious parametric models applied to individual data
- global modeling:
  non/semi-parametric models applied to pooled data

research question:
Can we use pooled data to design a suitable parametric model for given family of DGPs?

Problem setup

adapting the framework of Hospedales et al. (2021):

data:

$$\mathcal{T} = \left\{ \underbrace{\{(x_t, y_t)\}_{t=1}^{K}}_{\mathcal{D}_{train}}, \underbrace{\{(x_t, y_t)\}_{t=K+1}^{N}}_{\mathcal{D}_{val}} \right\} \sim p(\mathcal{T}) \tag{1}$$

model:

$$\{f_\omega(\cdot;\cdot),\, \kappa_\omega(\cdot)\} \tag{2}$$

- prediction function: $\hat{y}_t = f_\omega(x_t; \hat{\theta})$
- estimation function: $\hat{\theta} = \kappa_\omega(\mathcal{D}_{train})$

loss:

$$\mathcal{L}(\mathcal{D}; \hat{\theta}, \omega) = \frac{1}{|\mathcal{D}|} \sum_{(x_t, y_t) \in \mathcal{D}} \gamma\left(y_t, f_\omega(x_t; \hat{\theta})\right) \tag{3}$$

Problem setup

adapting the framework of Hospedales et al. (2021):

data:

$$\mathcal{T} = \left\{ \underbrace{\{(x_t, y_t)\}_{t=1}^{K}}_{\mathcal{D}_{train}}, \underbrace{\{(x_t, y_t)\}_{t=K+1}^{N}}_{\mathcal{D}_{val}} \right\} \sim p(\mathcal{T}) \quad (1)$$

model:

$$\{f_\omega(\cdot;\cdot),\ \kappa_\omega(\cdot)\} \quad (2)$$

- prediction function: $\hat{y}_t = f_\omega(x_t; \hat{\theta})$
- estimation function: $\hat{\theta} = \kappa_\omega(\mathcal{D}_{train}) \approx \arg\min_{\theta \in \Theta} \mathcal{L}(\mathcal{D}_{train}; \theta, \omega)$

loss:

$$\mathcal{L}(\mathcal{D}; \hat{\theta}, \omega) = \frac{1}{|\mathcal{D}|} \sum_{(x_t, y_t) \in \mathcal{D}} \gamma\left(y_t, f_\omega(x_t; \hat{\theta})\right) \quad (3)$$

Problem setup

objective:

$$\begin{aligned} \omega^* &= \arg\min_{\omega \in \Omega} \mathbb{E}_{\mathcal{T} \sim p(\mathcal{T})} \left[\mathcal{L}(\mathcal{D}_{val}\ ; \hat{\theta}, \omega) \right] \\ \text{s.t.: } \hat{\theta} &= \kappa_\omega(\mathcal{D}_{train}) \approx \arg\min_{\theta \in \Theta} \mathcal{L}(\mathcal{D}_{train}; \theta, \omega). \end{aligned} \tag{4}$$

Problem setup

objective:

$$\begin{aligned} \hat{\omega} &= \arg\min_{\omega \in \Omega} \frac{1}{M} \sum_{m=1}^{M} [\mathcal{L}(\mathcal{D}_{val}^{(m)}; \hat{\theta}, \omega)] \\ \text{s.t.: } \hat{\theta}^{(m)} &= \kappa_{\omega}(\mathcal{D}_{train}^{(m)}) \approx \arg\min_{\theta \in \Theta} \mathcal{L}(\mathcal{D}_{train}^{(m)}; \theta, \omega). \end{aligned} \tag{4}$$

MtMs - model

A1: uniqueness & vanilla optimization

$$\forall \omega \in \Omega \, \forall \mathcal{D}^{(m)}_{train} \in \left(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}\right)^K : \mathrm{card}(\kappa_\omega(\mathcal{D}^{(m)}_{train})) = 1 \tag{5}$$

$$\kappa_\omega(\mathcal{D}^{(m)}_{train}) = \arg\min_\theta \mathcal{L}(\mathcal{D}^{(m)}_{train}; \theta, \omega). \tag{6}$$

A2: explicit dependance

$$f_\omega(x_i^{(m)}; \theta^{(m)}) = f(x_i^{(m)}; \underbrace{g(\theta^{(m)}; \omega)}_{\beta^{(m)}}). \tag{7}$$

A3: train-train spit (motivated by Bai et al. (2021) and Wang et al. (2021))

$$\begin{aligned} &\hat{\omega} = \arg\min_\omega \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}(\mathcal{D}^{(m)}_{train}; \hat{\theta}^{(m)}, \omega) \\ &\text{s.t.: } \hat{\theta}^{(m)} = \kappa_\omega(\mathcal{D}^{(m)}_{train}) \end{aligned} \tag{8}$$

MtMs - model

Proposition 1:

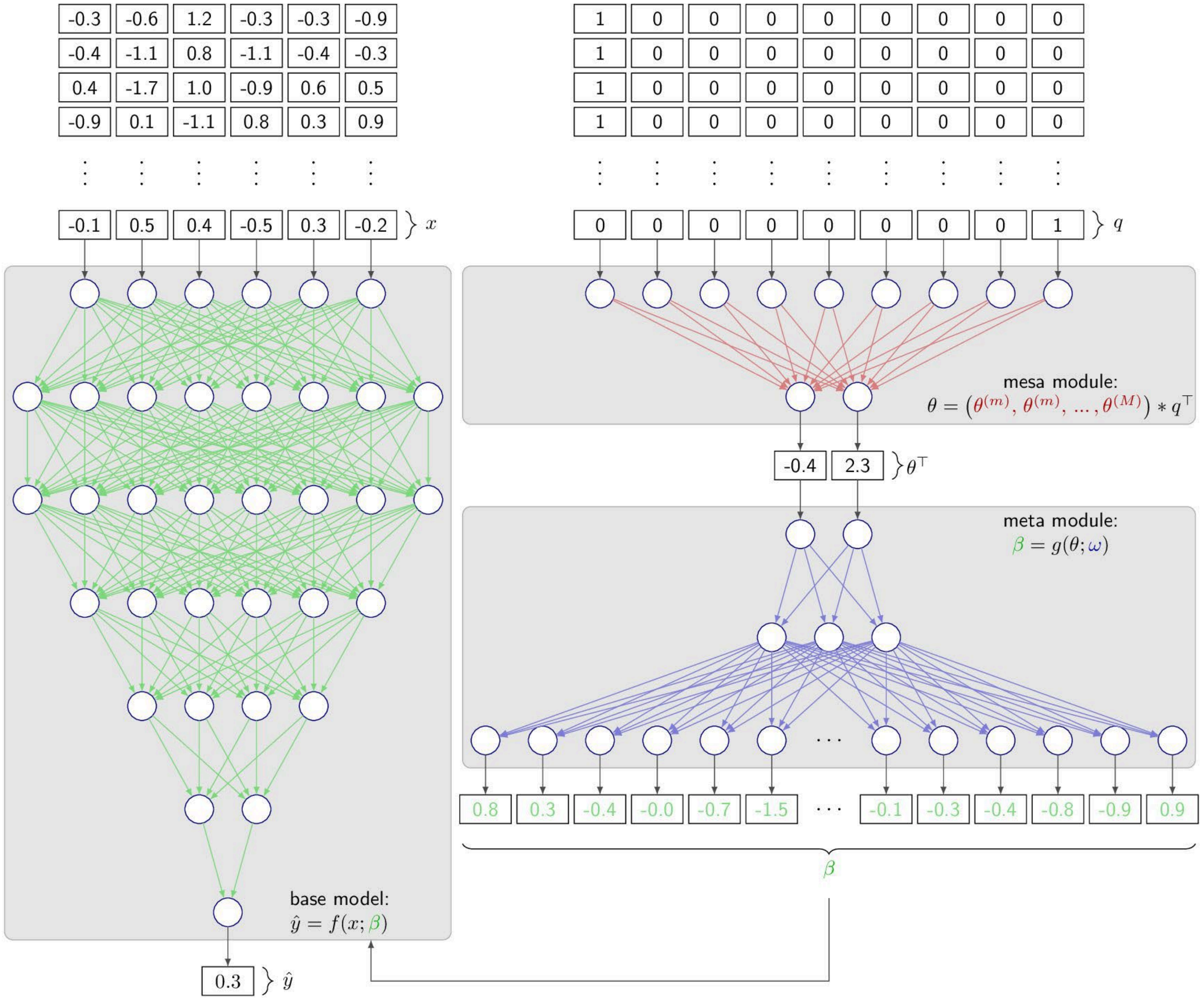
$$\hat{\omega} = \arg\min_{\omega \in \Omega} \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}(\mathcal{D}_{val}^{(m)}; \hat{\theta}^{(m)}, \omega)$$

$$\text{s.t.: } \hat{\theta}^{(m)} = \kappa_{\omega}(\mathcal{D}_{train}^{(m)}) \approx \arg\min_{\theta \in \Theta} \mathcal{L}(\mathcal{D}_{train}^{(m)}; \theta, \omega)$$

$$\Big\Updownarrow \text{A1-A3} \tag{9}$$

$$\left\{\hat{\omega}, \left\{\hat{\theta}^{(m)}\right\}_{m=1}^{M}\right\} = \arg\min_{\substack{\omega \in \Omega \\ \{\theta^{(m)}\}_{m=1}^{M} \in \Theta^{M}}} \frac{1}{M} \sum_{m=1}^{M} \frac{1}{K} \sum_{i=1}^{K} \gamma(y_i^{(m)}, f(x_i^{(m)}; g(\theta^{(m)}; \omega)))$$

MtMs - model



} x
} q
mesa module:
$\theta = \left(\theta^{(m)}, \theta^{(m)}, \dots, \theta^{(M)}\right) * q^\top$
-0.4
2.3
} $\theta^\top$
meta module:
$\beta = g(\theta; \omega)$
0.8
0.3
-0.4
-0.0
-0.7
-1.5
-0.1
-0.3
-0.4
-0.8
-0.9
0.9
β
base model:
$\hat{y} = f(x; \beta)$
0.3
} $\hat{y}$

Application to M6 - imputs

features:

- $y_t^{(m)} \in [0,1]^5$
- $x_t^{(m)}$
  - indicator of stock/ETF
  - 7 lags of 4-weeks returns
  - 7 lags of 4-weeks volatility
  - 35 technical trading indicators from TTR package (Ulrich, 2021)

data:

- time intervals
  - training 2000-2022
  - testing 2022-2023
- augmentation across assets
  - 9 additional M6-like universes (450 stocks and 450 ETFs)
- augmentation across time
  - shifting the 4-week intervals (1, 2, and 3-week offsets)

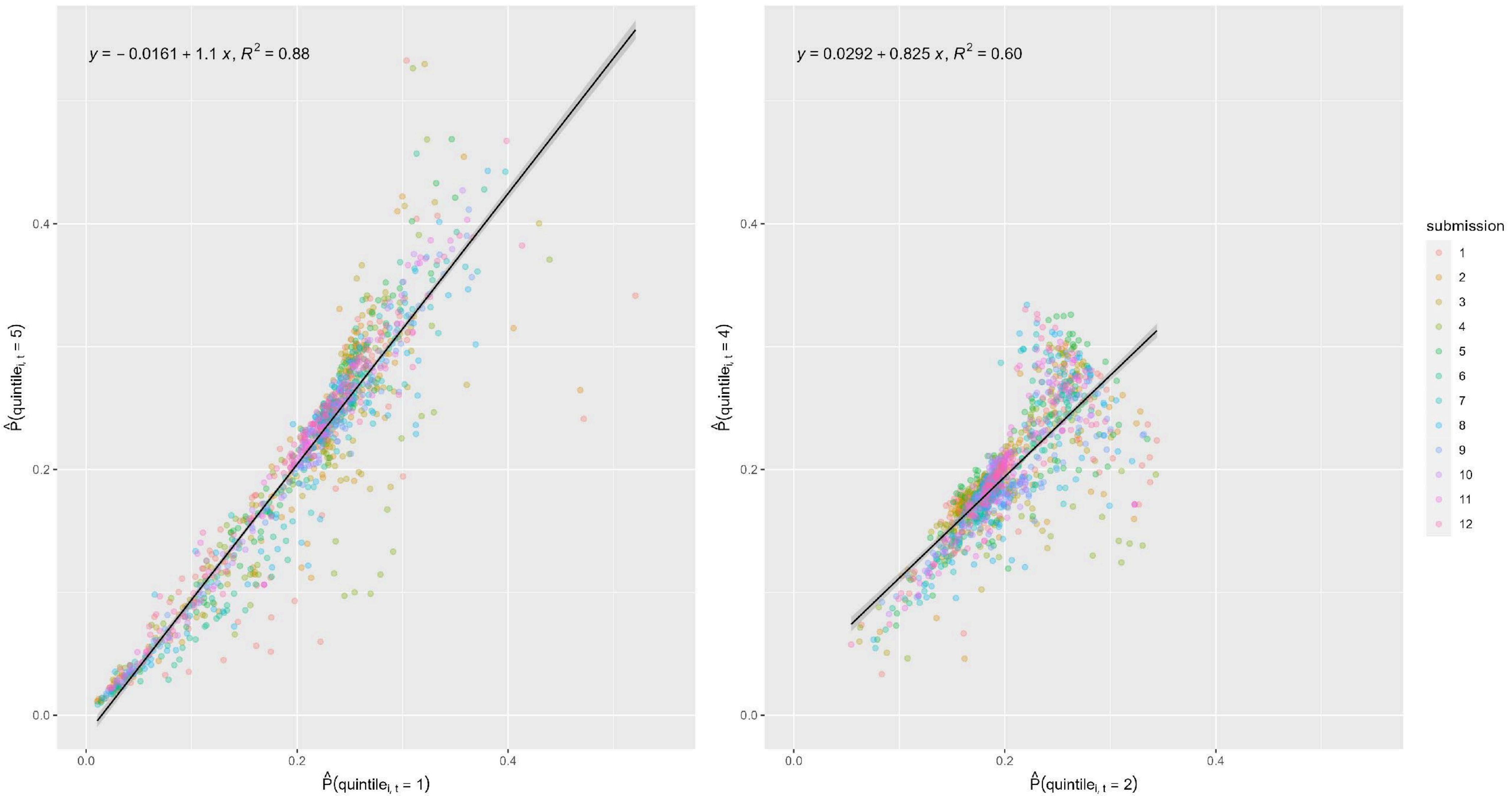
Application to M6 - model

model:

- base model $f(\cdot;\beta)$
  - FFNN with hidden layers $\{32, 8\}$
  - leaky ReLU nonlinearity
  - softmax transform on the final layer
  - dropout 0.2
- mesa module
  - one mesa parameter per asset $(\mathrm{card}(\theta) = 1)$
- meta module $g(\cdot;\omega)$
  - trivial FFNN with no hidden layer or nonlinearity
  - connected only to the last layer of $f(\cdot;\beta)$, rest orphaned constants

Application to M6 - results

RPS: 0.15689 (4-th)



Investment challenge

initial attempts:

- non-parametric approach via FFNN
- constructing joint distribution from predictions via copulas and optimizing for IR

Investment challenge

initial attempts:

- non-parametric approach via FFNN
- constructing joint distribution from predictions via copulas and optimizing for IR

$\Longrightarrow$ *none passed backtesting*

chosen strategy:

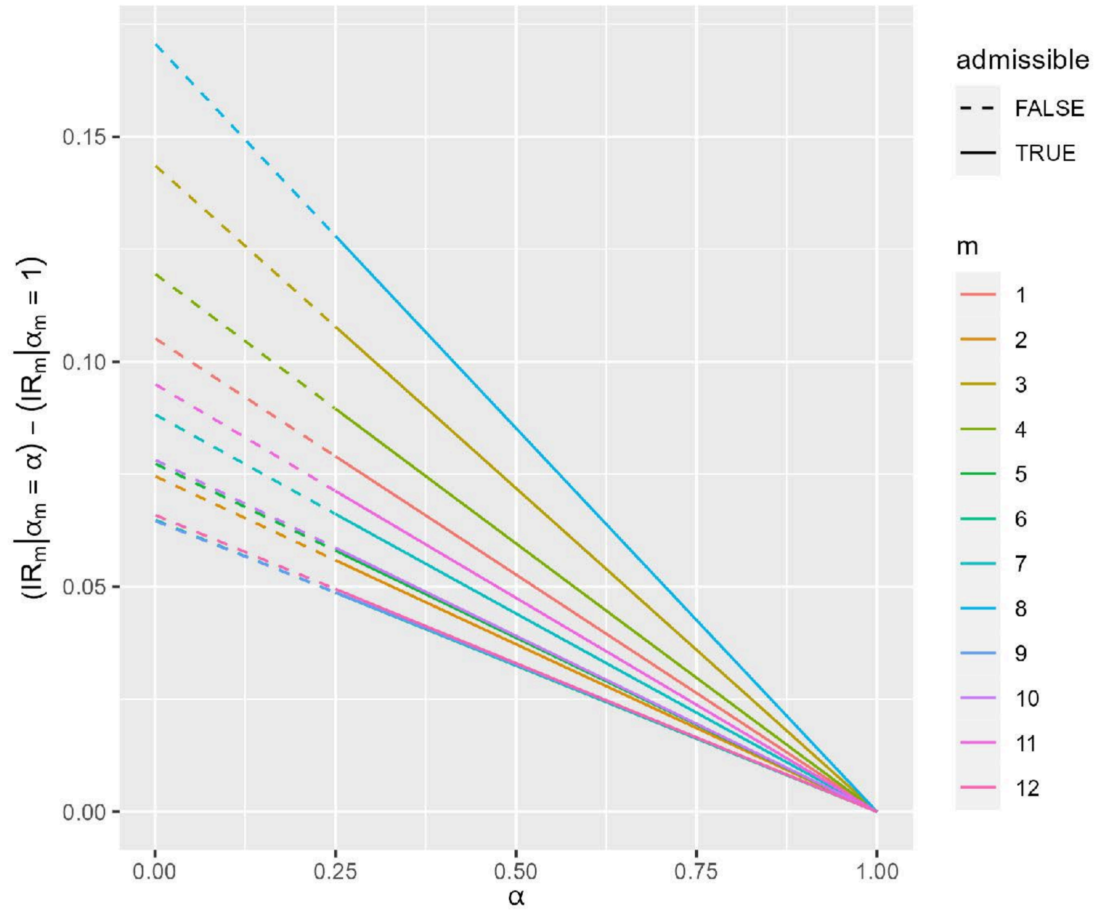
- decomposition:

$$w_{i,m} = \underbrace{\alpha_m}_{\sum_{i'=1}^{100} |w_{i',m}|} * \underbrace{\tilde{w}_{i,m}}_{\alpha_m^{-1} w_{i,m}} \quad \text{with} \quad m \in \{1, 2, ..., 12\} \tag{10}$$

- guiding principles:
  - $\alpha_m \Longrightarrow$ **optimal scaling**
  - $\tilde{w}_{:,m} \Longrightarrow$ **strategic risk taking** (discretionary)

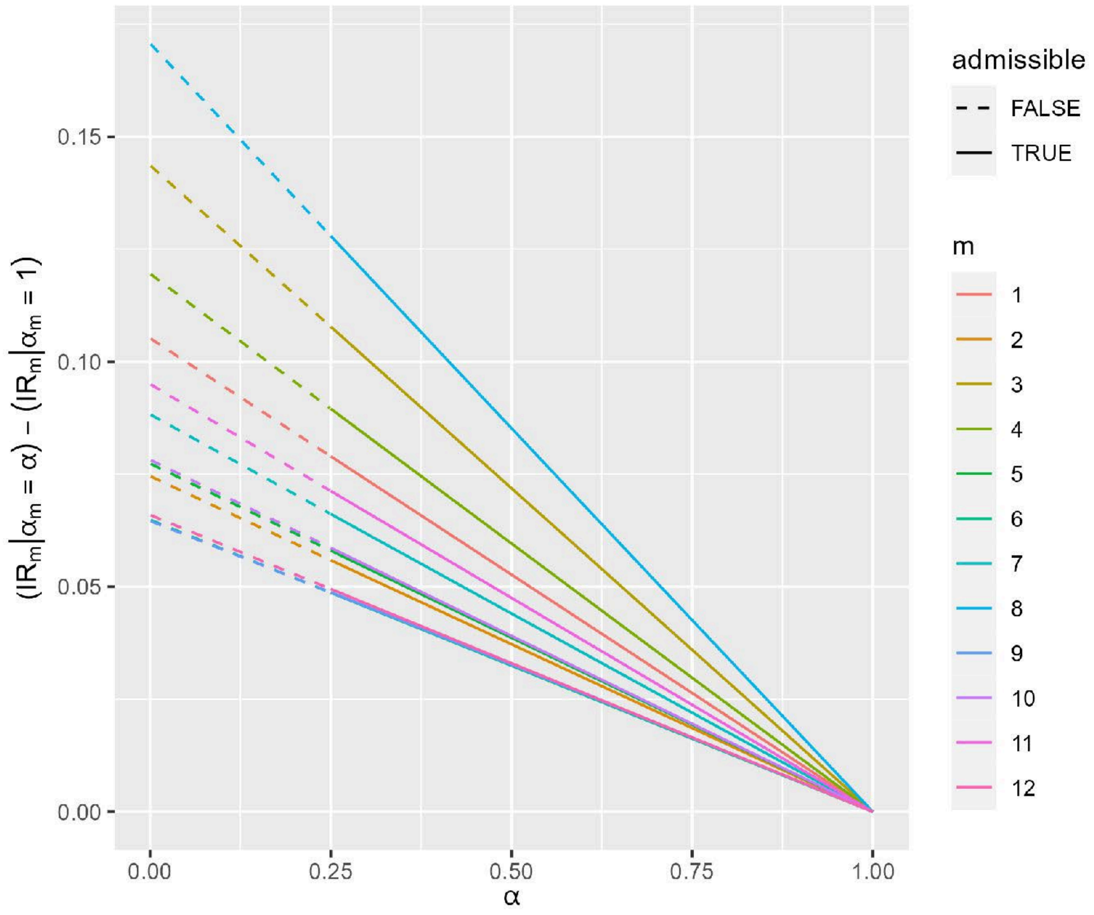
Investment challenge - optimal scaling

M6 rules: $0.25 \leq \alpha_m \leq 1$



Investment challenge - optimal scaling

M6 rules: $0.25 \leq \alpha_m \leq 1$



$\Longrightarrow$ It is *almost* always optimal to take the **smallest** exposure possible: $\alpha_m^* = 0.25$

Investment challenge - strategic risk taking

denote:

- n_+ number of long position in the portfolio
- n_- number of short position in the portfolio
- q target rank one wishes to achieve

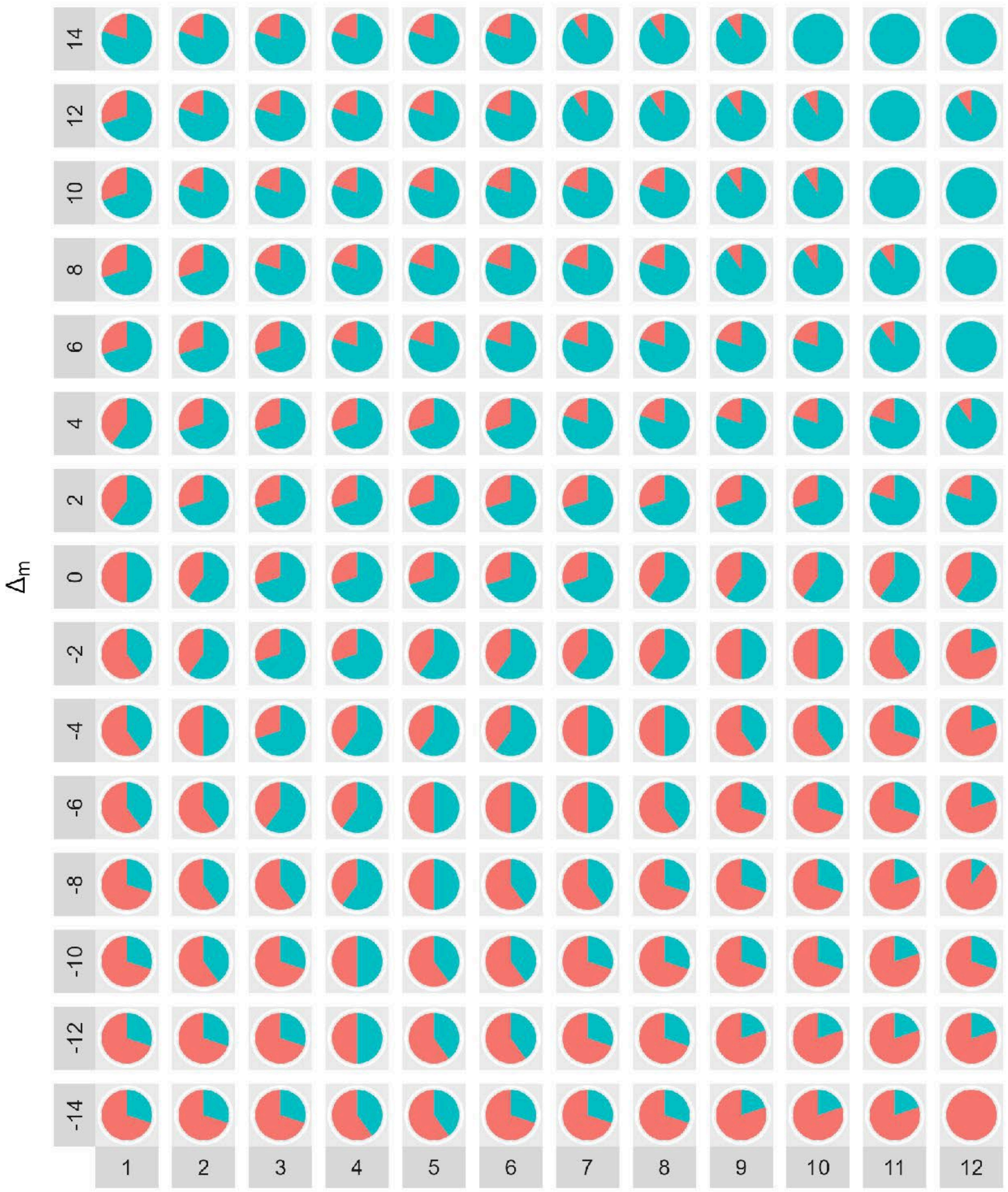
key observation:

- only relative performance matters
  $\rightarrow$ focus on $P(rank_{T_1:T_{12}} \leq q)$ rather than $\mathbb{E}\left[IR_m\right]$
- participants take predominantly long positions

strategy:

- observe your current rank $rank_{T_1:T_m}$ in the public leaderboard
  - $rank_{T_1:T_m} \leq q$
    - $\downarrow \frac{n_-}{n_+ + n_-} \rightarrow\uparrow corr(IR_{m+1,own}, IR_{m+1,cont.}) \rightarrow\uparrow P(rank_{T_1:T_{m+1}} \leq q)$
  - $rank_{T_1:T_m} \geq q$
    - $\uparrow \frac{n_-}{n_+ + n_-} \rightarrow\downarrow corr(IR_{m+1,own}, IR_{m+1,inc.}) \rightarrow\uparrow P(rank_{T_1:T_{m+1}} \leq q)$
    - more aggressively the higher the m

Investment challenge - strategic risk taking (ex-post derived)



Investment challenge - results

IR: 13.8 (6-th)

- optimal scaling
  - only very modest gains (0.07 IR per month, $\sim$ 1 IR in total)
    - 0.07 IR per month, $\sim$ 1 IR in total
- strategic risk taking
  - increases probability of securing a good rank
    - $P(rank_{T_1:T_{12}} \leq 20)$ approx. from 0.12 to 0.19
  - but worsens the $\mathbb{E}\left[IR_{T_1:T_{12}}\right]$
    - approx. by 8 IR relative to long equal weighted portfolio

Investment challenge - results

IR: 13.8 (6-th)

- optimal scaling
  - only very modest gains (0.07 IR per month, $\sim$ 1 IR in total)
    - 0.07 IR per month, $\sim$ 1 IR in total
- strategic risk taking
  - increases probability of securing a good rank
    - $P(rank_{T_1:T_{12}} \leq 20)$ approx. from 0.12 to 0.19
  - but worsens the $\mathbb{E}\left[IR_{T_1:T_{12}}\right]$
    - approx. by 8 IR relative to long equal weighted portfolio

...nonetheless, our IR rank is still **largely attributed to luck** and would likely not be repeated.

Conclusions

investment challenge (long form at https://bit.ly/M6_inv_ch)

- forecasts contained only limited directional information
  → our attempts at using them to construct portfolios failed
- strategic risk taking & optimal scaling
  - highlights the distinction between optimizing the probability of winning and optimizing the expected IR
  - limited applicability outside of the scope of M6
  - still largely influenced luck

forecasting challenge (long form at https://bit.ly/M6_for_ch)

- testing the MtMs model:
  - allows to design a suitable parametric model for a given family of DGPs
  - delivered respectable RPS in M6
  - very promising results in other applications (see Appendix C)

MtMs - application to M6 - features

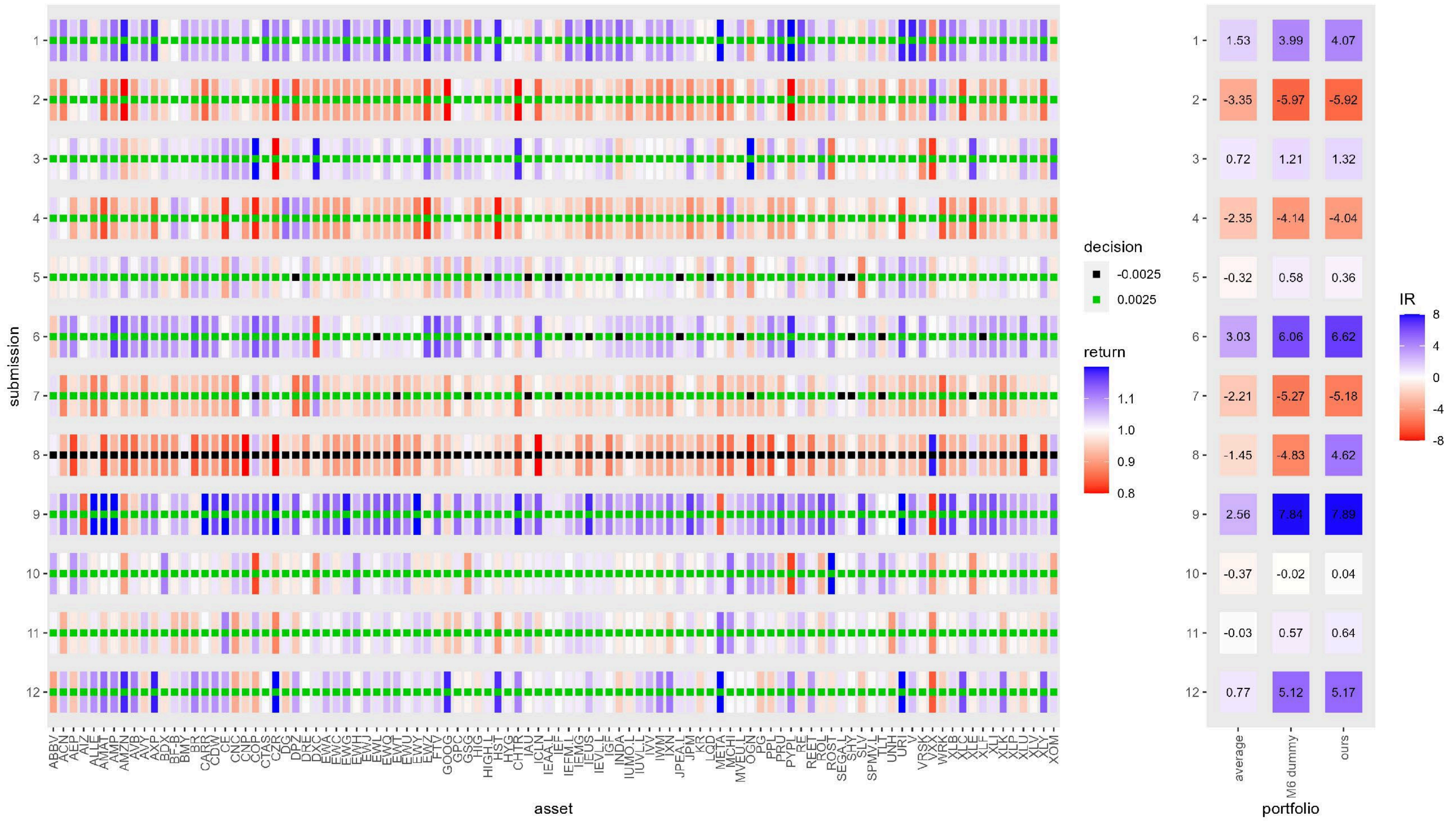
Source	Feature	Transformation
own	Volatility(lag = [1,2,3,4,5,6,7])	
own	Return(lag = [1,2,3,4,5,6,7])	
own	IsETF	
TTR	ADX	
TTR	aroon	
TTR	ATR(n=[7, 14, 28])	Norm.
TTR	BBands	Norm.
TTR	CCI	
TTR	chaikinAD	diff(1)
TTR	chaikinVolatility	
TTR	CLV	
TTR	CMF	
TTR	CMO	
TTR	CTI	
TTR	DEMA	Norm.
TTR	DonchianChannel	Norm.
TTR	EMA	Norm.
TTR	EVWMA	Norm.
TTR	GMMA(short=10, long=[30, 60])	Norm.
TTR	HMA	Norm.
TTR	KST	
TTR	MACD	
TTR	MFI	
TTR	OBV	diff(1)
TTR	PBands	Norm.
TTR	ROC	
TTR	RSI	
TTR	runPercentRank(n=100)	
TTR	SMI	
TTR	SNR(n=[20,60])	
TTR	TDI	Norm.
TTR	TRIX	
TTR	ultimateOscillator	
TTR	VHF	
TTR	volatility	
TTR	williamsAD	diff(1)
TTR	WPR	
TTR	ZLEMA	Norm.

MtMs - application to M6 - training

training:

- RPS loss
- Adam optimizer
- early stopping
- two stage training (see e.g. Beck et al. (2023))
  - 1-st stage
    - lr $\{0.01\}$
    - minibatch 200
  - 2-nd stage
    - bias of $g(\cdot;\omega)$ initialized with output of the 1-st stage
    - weights $g(\cdot;\omega)$ initialized with 0
    - $\theta^{(m)}$ initialized with $U\,[-1,1]$
    - lr $\{0.01, 0.001, 0.001, 0.0005, 0.0003, 0.0001, 0.00005\}$
    - minibatch 100 assets

Portfolio weights



IR computation

portfolio returns:

$$RET_t = \sum_{i=1}^{I} w_{i,m} r_{i,t} \quad \text{with} \quad r_{i,t} = \frac{S_{i,t}}{S_{i,t-1}} - 1 \tag{11}$$

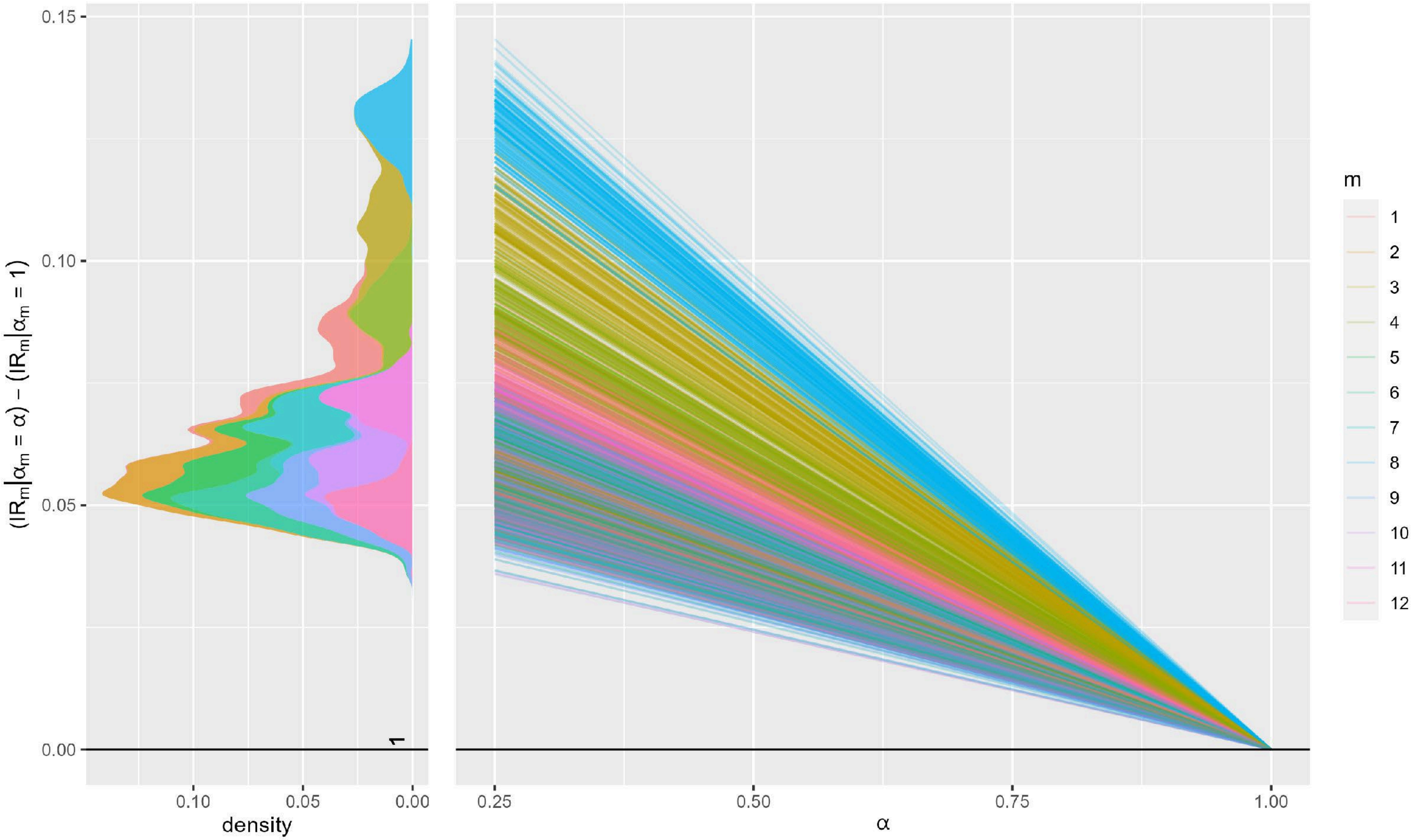
portfolio log-returns:

$$ret_t = ln(1 + RET_t) \tag{12}$$

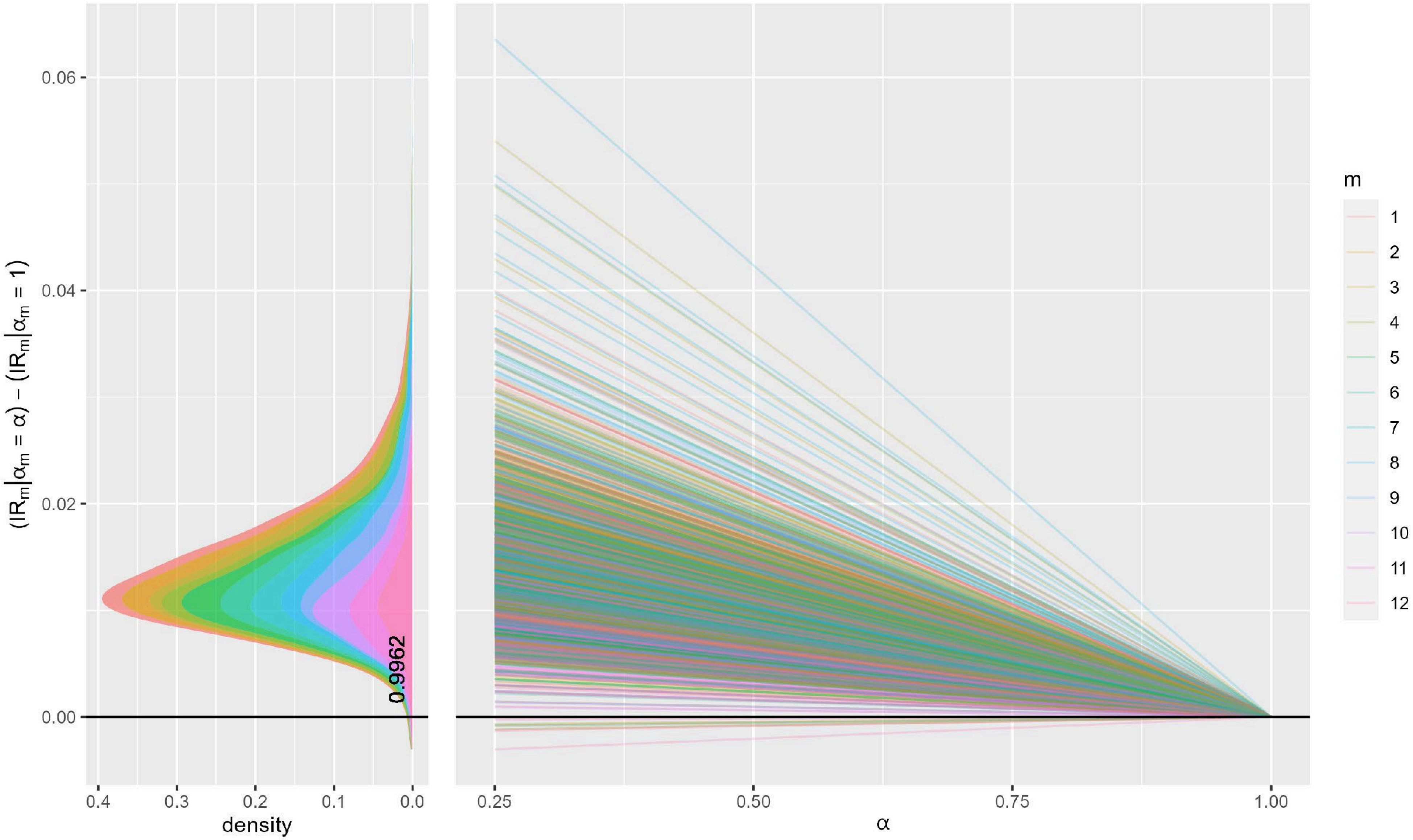
IR:

$$\begin{aligned} IR_{t_1:t_2} &= (t_2 - t_1 + 1)\frac{\widehat{\mu}_{ret_{t_1:t_2}}}{\widehat{\sigma}_{ret_{t_1:t_2}}} \\ &= \frac{\sum_{t=t_1}^{t_2} ret_t}{\sqrt{\frac{1}{t_2-t_1}\sum_{t=t_1}^{t_2}\left(ret_t - \frac{1}{t_2-t_1+1}\sum_{t=t_1}^{t_2} ret_t\right)^2}} \end{aligned} \tag{13}$$

Gamma distributed weights



Normal distributed weights



Sinusoidal regression task

following Finn et al. (2017):

$$
\begin{aligned}
y_i^{(m)} | x_i^{(m)}, A^{(m)}, b^{(m)} &= A^{(m)} * sin(x_i^{(m)} + b^{(m)}) \\
x_i^{(m)} | A^{(m)}, b^{(m)} &\sim U(-5, 5) \\
A^{(m)} &\sim U(0.1, 5) \\
b^{(m)} &\sim U(0, \pi)
\end{aligned} \tag{14}
$$

with

$$
\mathcal{D}^{(m)} = \{\mathcal{D}_{train}^{(m)}, \mathcal{D}_{val}^{(m)}\} = \{\{(x_i^{(m)}, y_i^{(m)})\}_{i=1}^{K}, \{(x_i^{(m)}, y_i^{(m)})\}_{i=K+1}^{N}\} \tag{15}
$$

$$
\begin{aligned}
\mathcal{L}(\mathcal{D}_{val}^{(m)}; \hat{\theta}^{(m)}, \omega) &= (N-K)^{-1} \sum_{i=K+1}^{N} (y_i^{(m)} - f_\omega(x_i^{(m)}; \hat{\theta}^{(m)}))^2 \\
\text{s.t.: } \hat{\theta}^{(m)} &= \kappa(\mathcal{D}_{train}^{(m)}; \omega)
\end{aligned} \tag{16}
$$

Sinusoidal regression task (MAML & Meta-SGD)

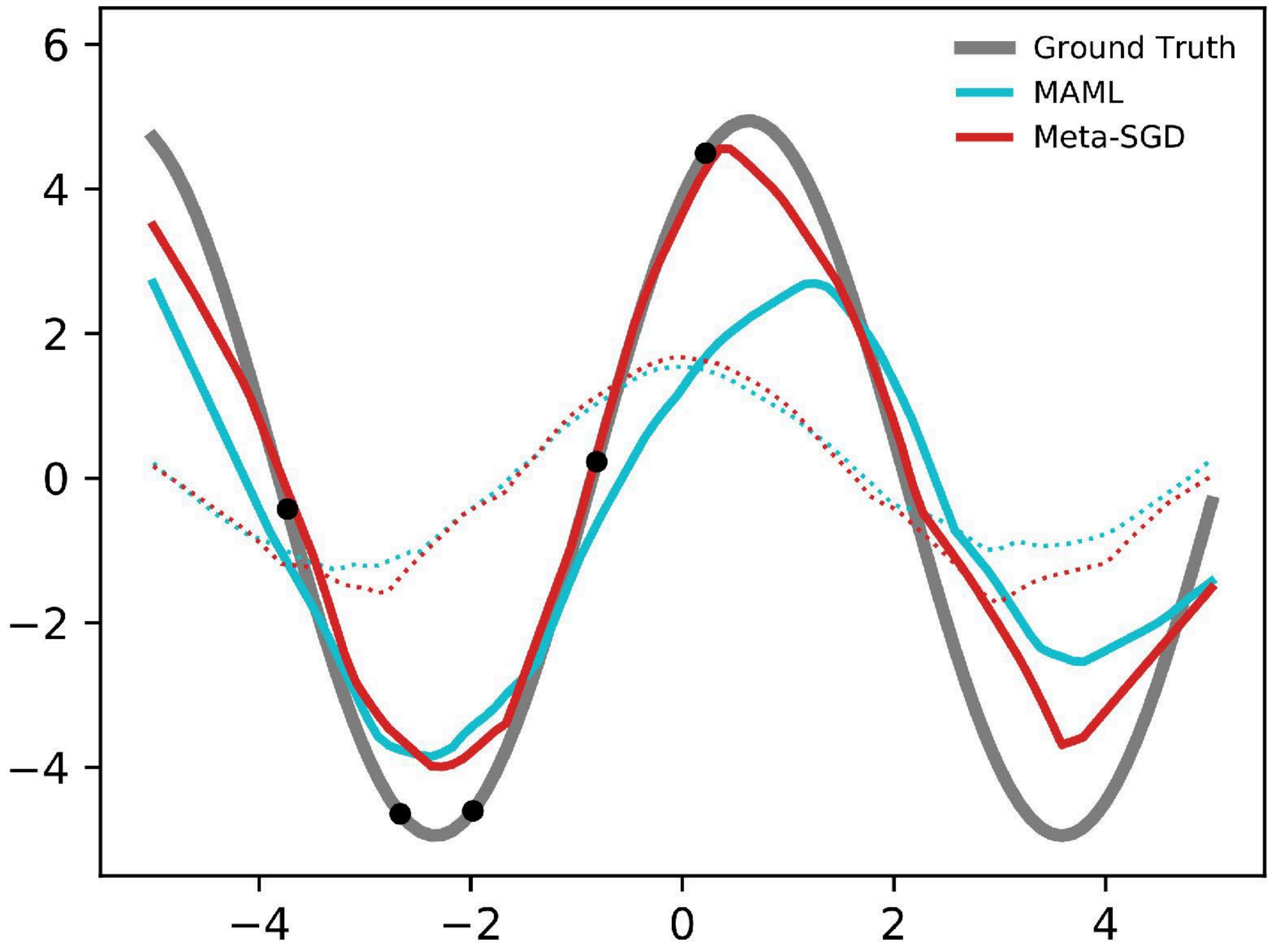
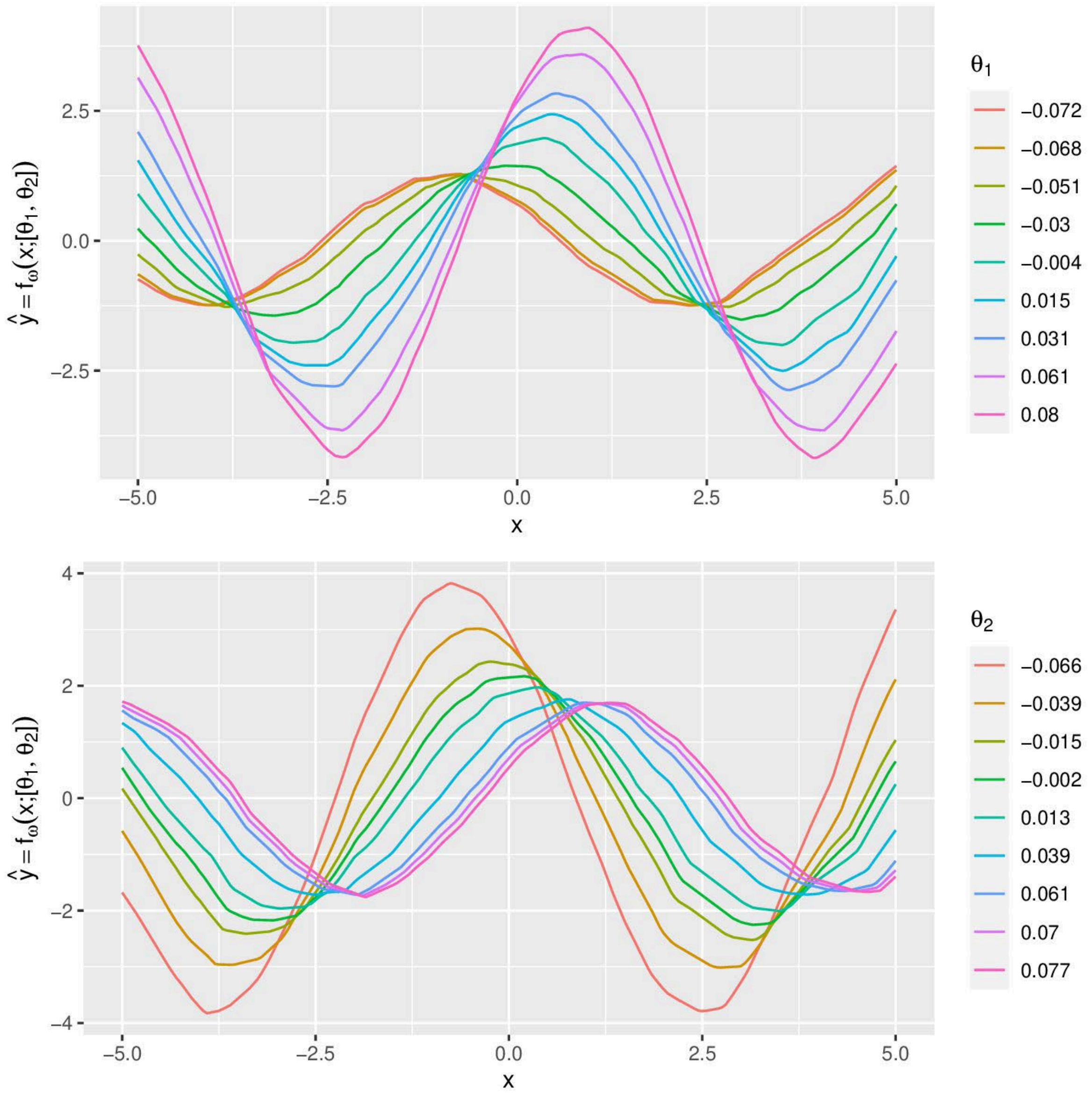


Figure 1: MAML & Meta-SGD
(source: Li et al. (2017))

Sinusoidal regression task (MtMs)



Sinusoidal regression task (comparison)

Method	$K = 5$	$K = 10$
MAML (Finn et al., 2017)	$0.686^{\pm 0.070}$	$0.435^{\pm 0.039}$
LayerLR (Park and Oliva, 2019)	$0.528^{\pm 0.068}$	$0.269^{\pm 0.027}$
Meta-SGD (Li et al., 2017)	$0.482^{\pm 0.061}$	$0.258^{\pm 0.026}$
MC1 (Park and Oliva, 2019)	$0.426^{\pm 0.054}$	$0.239^{\pm 0.025}$
MC2 (Park and Oliva, 2019)	$0.405^{\pm 0.048}$	$0.201^{\pm 0.020}$
MH (Zhao et al., 2020)	$0.501^{\pm 0.082}$	$0.281^{\pm 0.072}$
MtMs (ours)	$0.022^{\pm 0.003}$	$0.014^{\pm 0.001}$

M4

following Montero-Manso and Hyndman (2021):

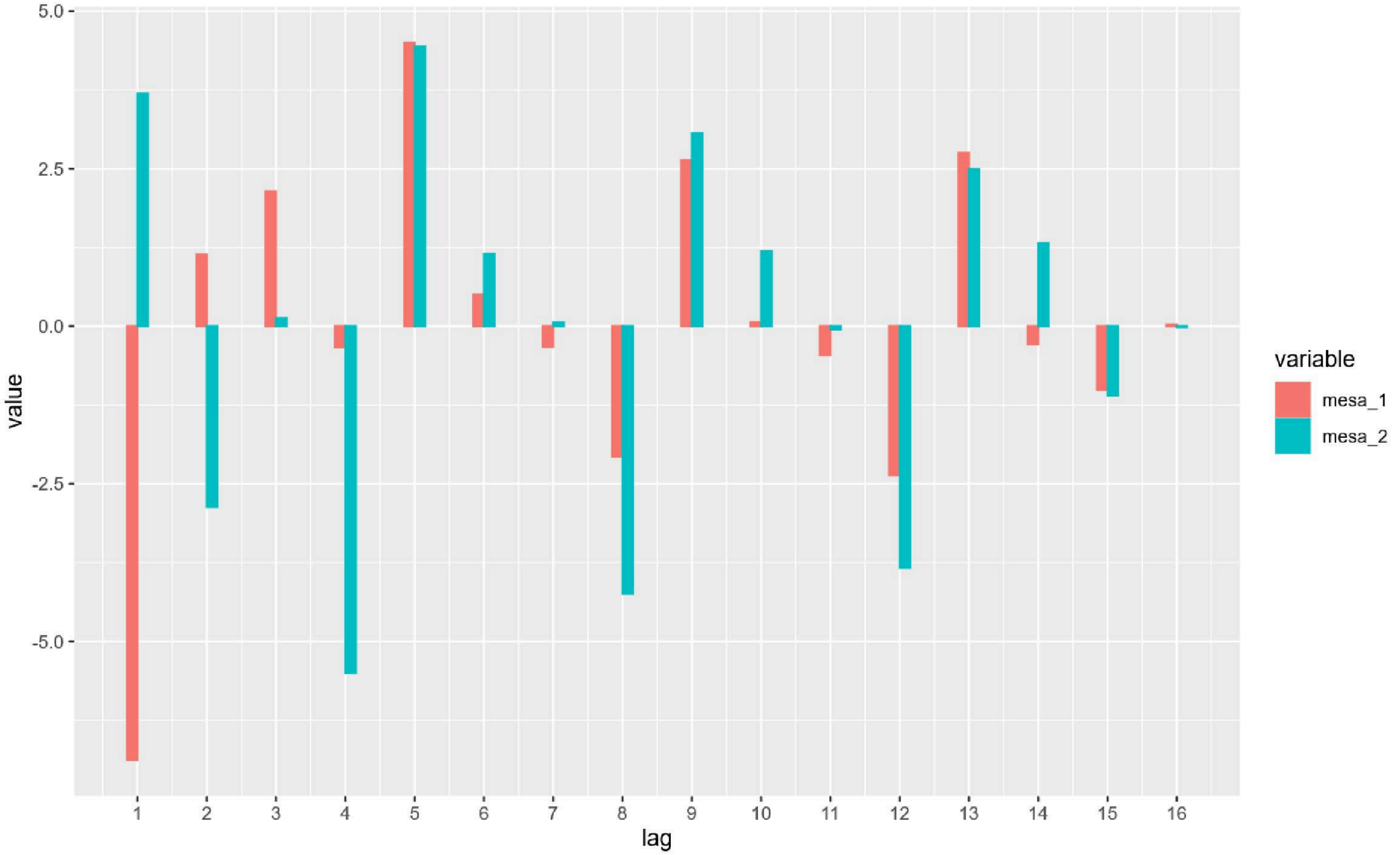
Model:

- x consists of last d_x lags
- f : 5 hidden layers 32 of nodes each with leaky ReLU nonlinearity and skip connections
- g: no hidden layers or nonlinearities
- $d_\theta = 2$
- initial values for MtMs taken from f trained on pooled data and MtMsOLS

M4

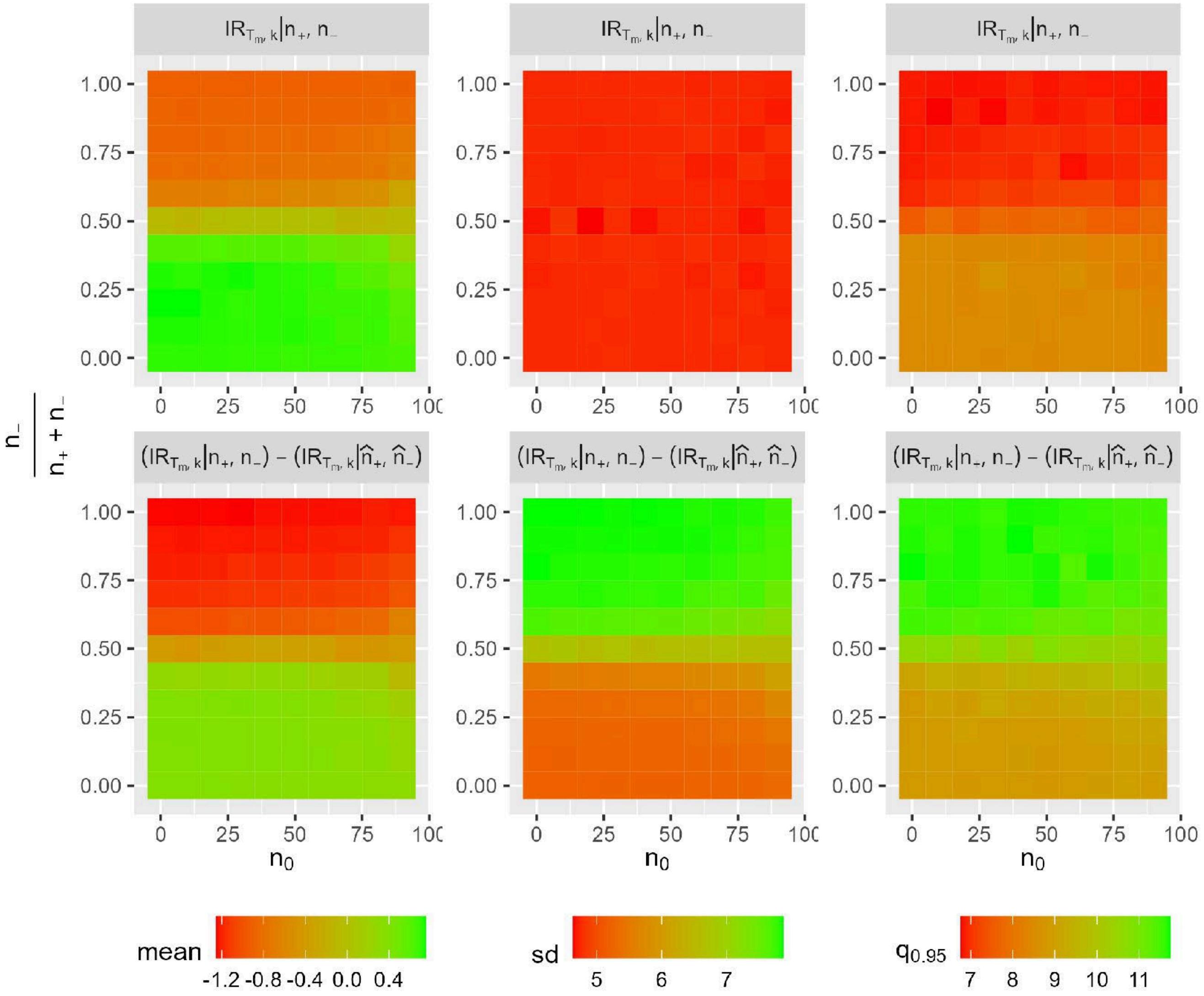
Method	Quarterly (10T subset)
OLS	$1.216^{\pm 1.400}$
NN	$1.170^{\pm 1.032}$
MtMsOLS	$1.178^{\pm 1.044}$
MtMs	$1.132^{\pm 1.010}$

M4



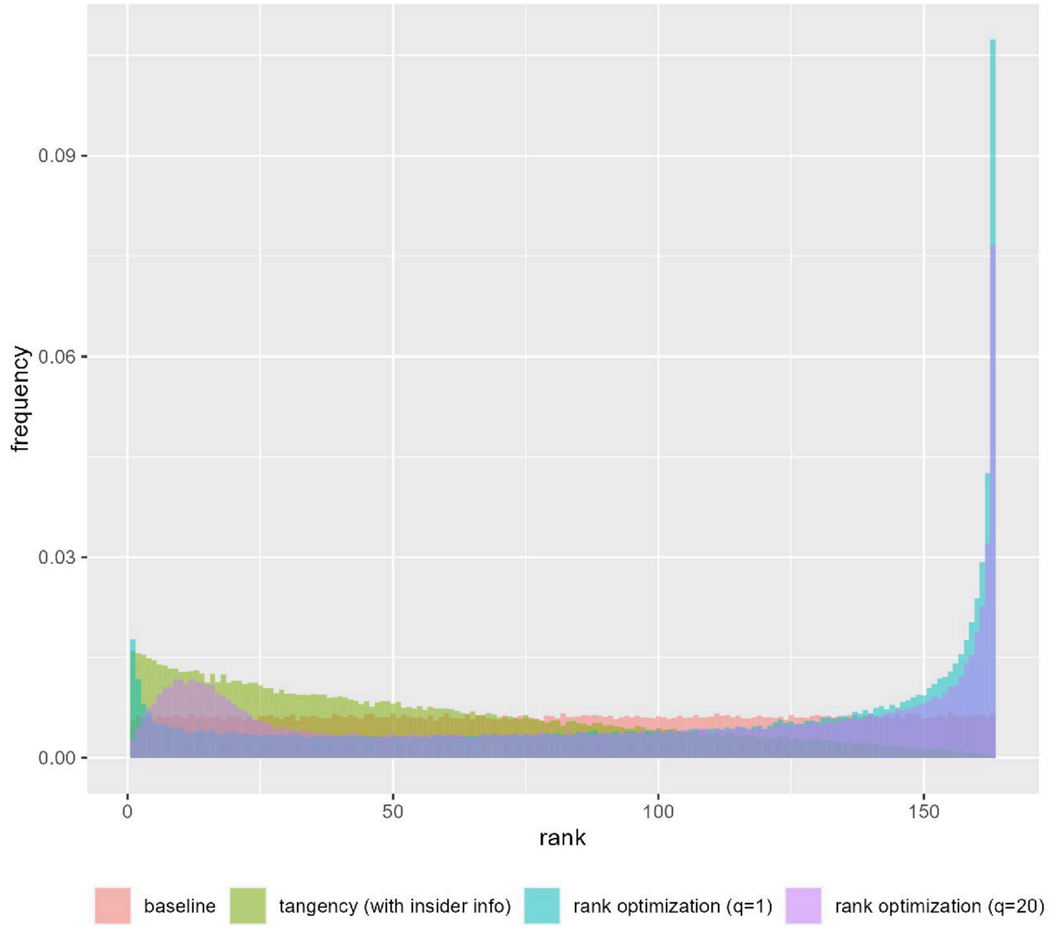
5.0
2.5
0.0
-2.5
-5.0
value
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
lag
variable
mesa_1
mesa_2

Position effects

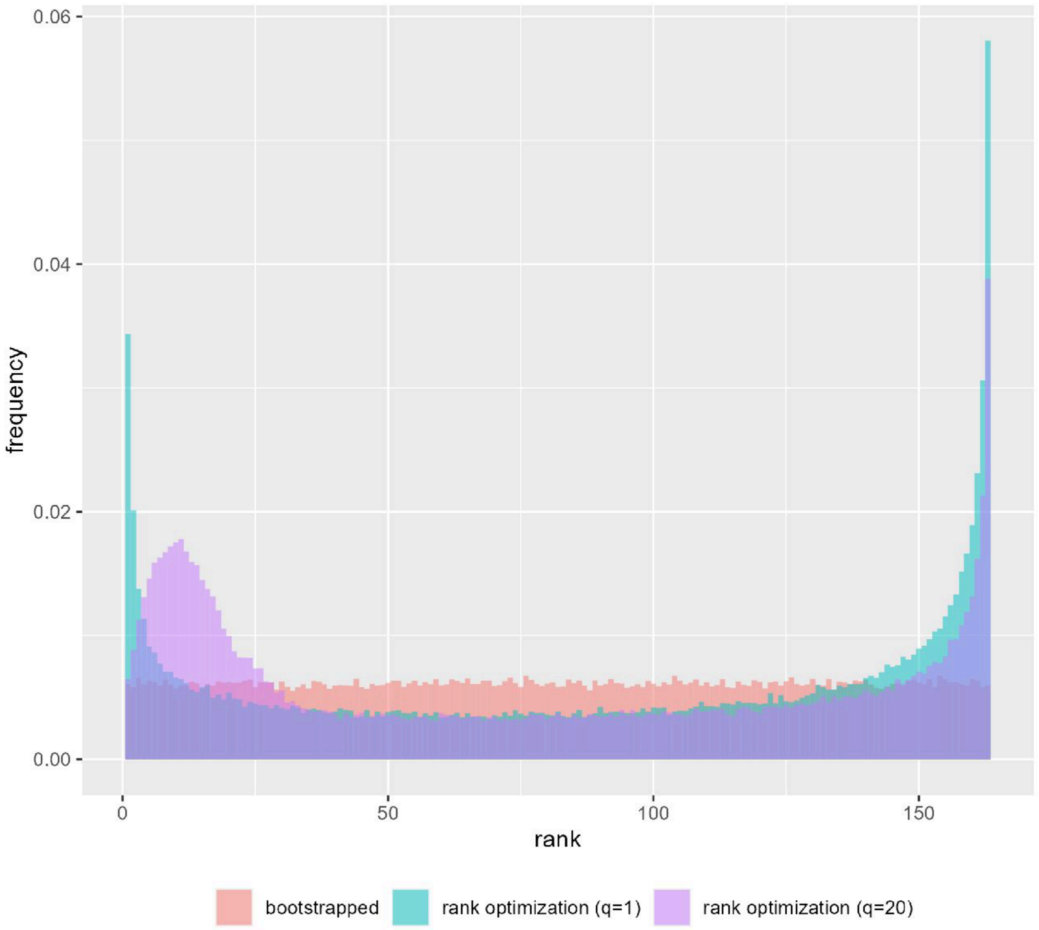


Observed & Simulated IR

month	$mean(IR_{T_m,:})$ obs.	$mean(IR_{T_m,:})$ sim.	$sd(IR_{T_m,:})$ obs.	$sd(IR_{T_m,:})$ sim.	$q_{0.01}(IR_{T_m,:})$ obs.	$q_{0.01}(IR_{T_m,:})$ sim.	$q_{0.99}(IR_{T_m,:})$ obs.	$q_{0.99}(IR_{T_m,:})$ sim.
1	1.53	1.47	3.39	3.19	-10.10	-7.05	6.74	10.03
		(0.25)		(0.21)		(1.30)		(1.42)
2	-3.23	-2.12	4.31	4.23	-11.09	-12.72	10.78	9.05
		(0.33)		(0.26)		(1.68)		(1.56)
3	0.70	0.57	3.17	3.71	-9.24	-9.27	10.57	10.15
		(0.29)		(0.23)		(1.46)		(1.41)
4	-2.37	-2.46	3.48	4.23	-11.89	-12.64	7.36	9.10
		(0.33)		(0.26)		(1.44)		(1.57)
5	-0.28	0.62	2.25	3.56	-6.39	-8.44	5.68	10.14
		(0.28)		(0.22)		(1.35)		(1.41)
6	2.97	2.02	4.42	4.04	-8.95	-8.49	10.74	12.16
		(0.31)		(0.24)		(1.43)		(1.47)
7	-2.20	-2.49	4.73	3.69	-9.81	-11.60	11.52	7.45
		(0.29)		(0.23)		(1.31)		(1.36)
8	-1.37	-2.41	4.59	3.94	-8.85	-12.11	9.11	8.45
		(0.31)		(0.24)		(1.38)		(1.61)
9	2.74	3.13	5.91	4.50	-12.61	-9.09	10.54	13.60
		(0.35)		(0.26)		(1.68)		(1.33)
10	-0.47	-0.03	3.81	4.94	-11.91	-12.51	8.56	12.80
		(0.39)		(0.30)		(1.73)		(1.83)
11	0.04	0.27	2.14	3.67	-6.91	-9.10	5.28	9.61
		(0.29)		(0.22)		(1.22)		(1.26)
12	0.50	1.29	5.66	5.53	-13.04	-12.20	9.11	12.94
		(0.43)		(0.28)		(1.44)		(1.21)
total	-1.43	-0.14	10.44	14.37	-30.35	-35.64	28.68	35.29
		(1.13)		(0.80)		(4.41)		(4.40)



portfolio	$\mathbb{E}[IR_{T_1:T_{12}}]$	$P(rank_{T_1:T_{12},k} \leq q)$ $q=1$	$q=5$	$q=10$	$q=20$
baseline	2.86	0.006	0.030	0.061	0.122
tangency (without insider info)	6.89	0.001	0.009	0.029	0.089
tangency (with insider info)	12.29	0.016	0.077	0.144	0.266
rank optimization (q=1)	−7.90	0.018	0.050	0.073	0.112
rank optimization (q=20)	−2.85	0.003	0.029	0.083	0.186



portfolio	$\mathbb{E}[IR_{T_1:T_{12}}]$	$P(rank_{T_1:T_{12},k} \leq q)$ $q = 1$	$q = 5$	$q = 10$	$q = 20$
bootstrapped	-1.47	0.006	0.031	0.061	0.122
rank optimization (q=1)	-6.00	0.034	0.089	0.126	0.182
rank optimization (q=20)	-1.38	0.007	0.054	0.138	0.278

References I

Bai, Y., M. Chen, P. Zhou, T. Zhao, J. Lee, S. Kakade, H. Wang, and C. Xiong (2021). How Important is the Train-Validation Split in Meta-Learning? In *International Conference on Machine Learning*, pp. 543–553. PMLR.

Beck, J., M. T. Jackson, R. Vuorio, and S. Whiteson (2023, March). Hypernetworks in Meta-Reinforcement Learning. In *Proceedings of The 6th Conference on Robot Learning*, pp. 1478–1487. PMLR.

Finn, C., P. Abbeel, and S. Levine (2017, July). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1126–1135. PMLR.

Hospedales, T., A. Antoniou, P. Micaelli, and A. Storkey (2021). Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence 44*(9), 5149–5169.

Li, Z., F. Zhou, F. Chen, and H. Li (2017, September). Meta-SGD: Learning to Learn Quickly for Few-Shot Learning.

References II

Montero-Manso, P. and R. J. Hyndman (2021, October). Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting 37*(4), 1632–1653.

Park, E. and J. B. Oliva (2019). Meta-curvature. *Advances in Neural Information Processing Systems 32*.

Ulrich, J. (2021, December). TTR: Technical Trading Rules.

Wang, H., H. Zhao, and B. Li (2021). Bridging multi-task learning and meta-learning: Towards efficient training and effective adaptation. In *International Conference on Machine Learning*, pp. 10991–11002. PMLR.

Zhao, D., S. Kobayashi, J. Sacramento, and J. von Oswald (2020, December). Meta-Learning via Hypernetworks. In *Zhao, Dominic; Kobayashi, Seijin; Sacramento, João; von Oswald, Johannes (2020). Meta-Learning via Hypernetworks. In: 4th Workshop on Meta-Learning at NeurIPS 2020 (MetaLearn 2020), Virtual Conference, 11 December 2020, IEEE.*, Virtual Conference. IEEE.