AlphaQCM: Alpha Discovery in Finance with Distributional Reinforcement Learning

Zhoufan Zhu, Ke Zhu

Xiamen University, The University of Hong Kong

May 12, 2025

Background

- Extensive research has investigated the predictive power of historical stock information for forecasting future returns, resulting in the development of several well-known alphas, such as **long-term momentum** (Jegadeesh and Titman, 1993) and **short-term reversal** (Jegadeesh, 1990).
- Here, **each alpha is a function** that transforms noisy historical stock data into signals for predicting future stock returns.
- Emerging literature focuses on **how to automatically discover a set of synergistic formulaic alphas**.

Synergistic Formulaic Alphas

- The **formulaic** nature of these alphas implies that they can be expressed by a simple formula, usually making them compact, interpretable, and generalizable;
- Meanwhile, their **synergistic** nature allows them to be combined into a meta-alpha via some interpretable models (e.g., linear models).

Contributions

1. This paper contributes to the literature by introducing a novel distributional reinforcement learning method, AlphaQCM.
2. Specifically, it leverages the **IQN** algorithm (Dabney et al., 2018), to learn the quantiles, while the **quantiled conditional moments (QCM)** method is adopted to estimate variance unbiasedly.
3. The variance serves as a natural exploration bonus for the mining agent's action selection to relieve the issue of non-stationary and reward-sparsity.
4. Our work clearly generalizes and extends the AlphaGen method (Yu et al., 2023) and can be applied to other non-stationary and/or reward-sparse environments.
5. Empirical results from three real-world stock market dataset further support it.

Reverse Polish Notation (RPN)

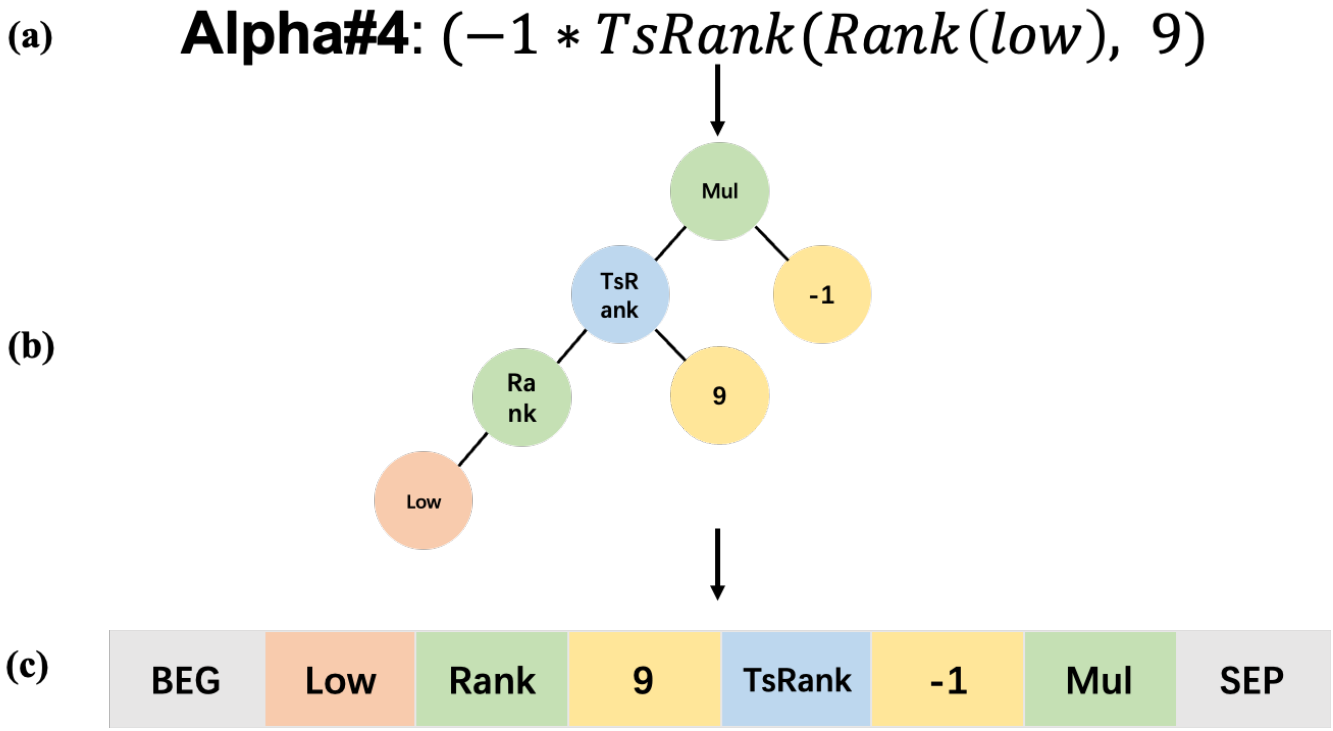

Figure 1: (a) Formulaic expression of Alpha#4 factor in Kakushadze (2016). (b) Its expression tree. (c) Its RPN representation.

MDP

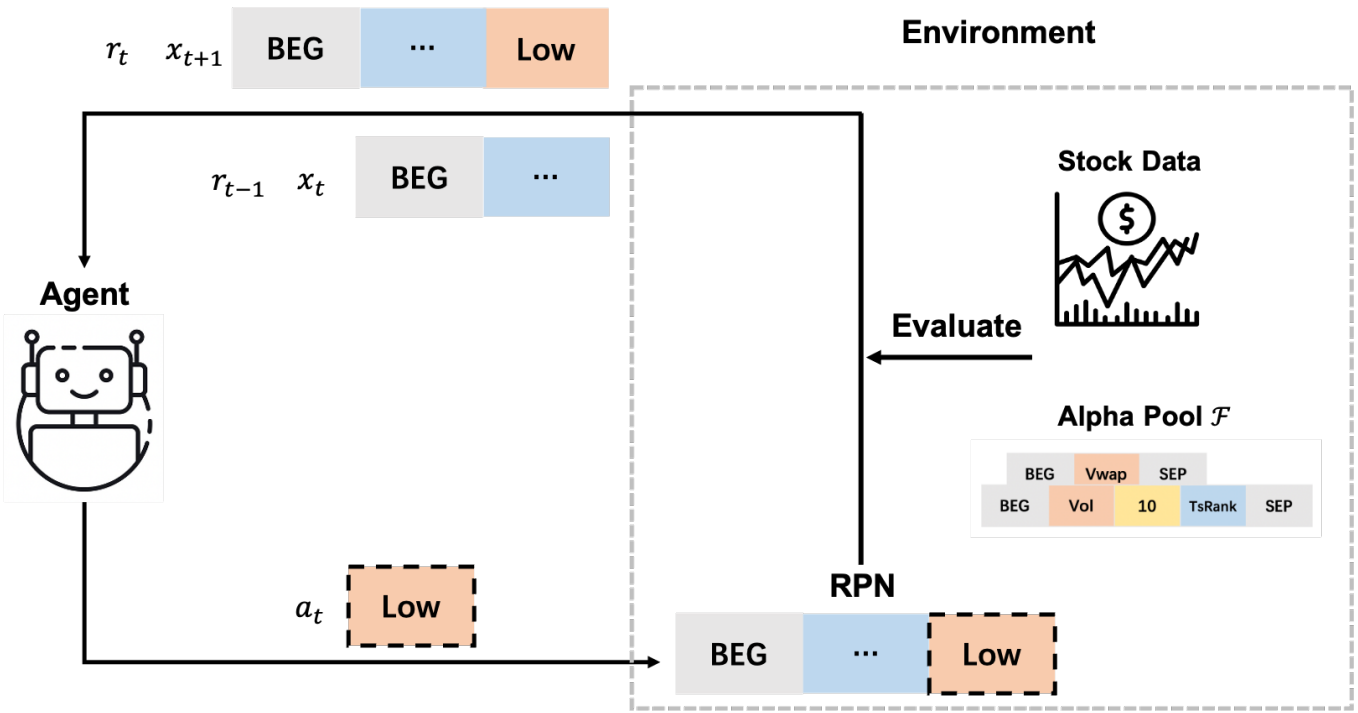

Figure 2: Agent-environment interaction diagram.

AlphaQCM

For each state-action pair (x, a), we have the linear regression model:

$$\widehat{\theta}_k(x,a) = \zeta(x,a) + Q^*(x,a) + \mathbf{\Phi}_k' \boldsymbol{\delta}(x,a) + \varepsilon_k(x,a),$$

where

1. $\widehat{\theta}_k(x,a)$ is the learned quantile,
2. $\boldsymbol{\delta}(x,a)$ is the vector built upon the variance, skewness, and kurtosis,
3. $Q^*(x,a)$ is the Q function,
4. $\mathbf{\Phi}_k$ is the vector of given quantiles of Gaussian distribution,
5. $\zeta(x,a)$ and $\varepsilon_k(x,a)$ represent the deterministic bias and stochastic residual, arising from the presence of both non-stationarity and expansion error, respectively.

AlphaQCM

- Although the MDP is non-stationary, the QCM variance estimator from solving the above linear model **remains unbiased with some mild conditions**, whereas there is no such guarantee for the other estimators, even in stationary MDPs.
- Using the QCM variance estimator as an exploration bonus, our agent tends to explore the most uncertain states, which also lead to the most informative experiences for overcoming non-stationarity and reward-sparsity.

Action Selection

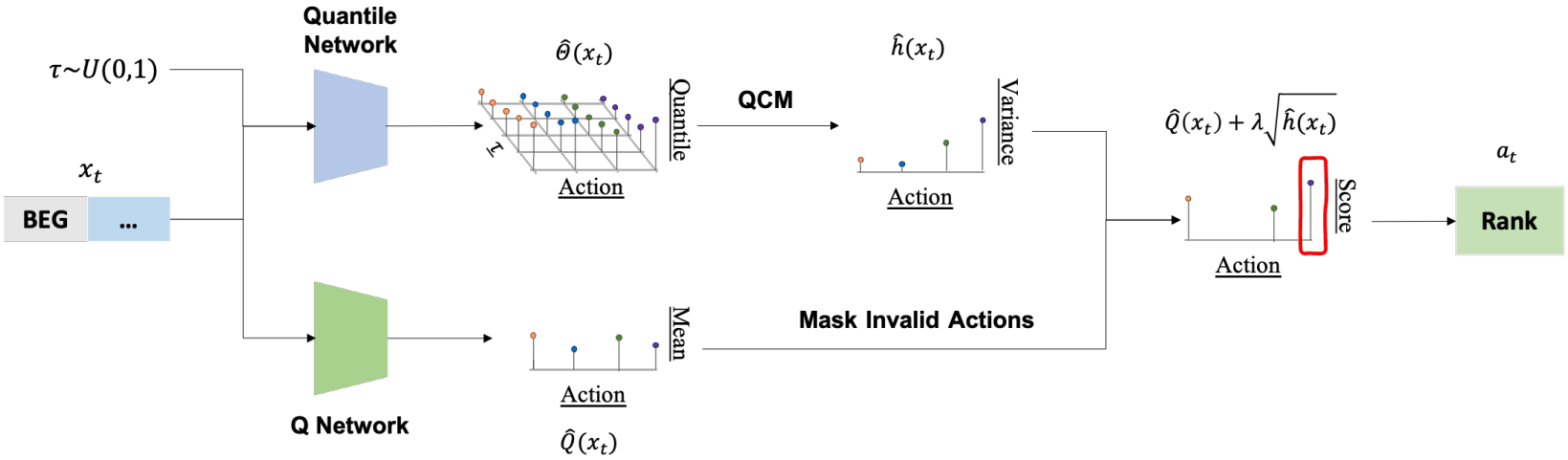

Figure 3: An illustration of action selection in our AlphaQCM framework

Settings

- Our experiments are also conducted on **Chinese A-share stock market** datasets to capture the **20-day** future stock returns.
- Three different stock pools: (1) the largest 300 stocks (**CSI 300**), (2) the largest 500 stocks (**CSI 500**), and (3) all stocks (**Market**) listed on the Shanghai and Shenzhen Stock Exchanges.
- Each dataset is split chronologically into a training set (2010/01/01 to 2019/12/31), a validation set (2020/01/01 to 2020/12/31), and a test set (2021/01/01 to 2022/12/31).

Baselines

(1) **Alpha101** (human-designed formulaic alphas): Fix the alpha pool as the formulaic alphas provided by Kakushadze (2016), and fit a linear model to form a mega-alpha.

(2) **MLP**, **XGBoost**, **LightGBM** (ML-based non-formulaic alphas): Use the MLP model, XGBoost model, or LightGBM model to form a mega-alpha.

(3) **GP w/o filter**, **GP w/ filter** (GP-based formulaic alphas): Use the GP method to generate expressions and apply top-P performing alphas without or with a mutual IC filter to form a mega-alpha.

(4) **PPO w/ filter**, **AlphaGen** (RL-based formulaic alphas): Use the PPO with a mutual IC filter or AlphaGen method to find the optimal alpha pool and then form a linear mega-alpha.

Impact of Method

Table 1: Out-of-sample IC values across different methods.

	CSI 300		CSI 500		Market	
Method	Mean	Std	Mean	Std	Mean	Std
Alpha101	3.44%	-	4.38%	-	3.15%	-
MLP	1.99%	0.24%	2.72%	0.65%	2.81%	0.72%
XGBoost	3.19%	0.81%	4.31%	0.96%	4.07%	1.22%
LightGBM	2.93%	0.76%	4.16%	0.81%	4.28%	0.93%
GP w/o filter	2.01%	1.46%	1.79%	1.62%	1.32%	2.01%
GP w/ filter	3.71%	2.01%	4.52%	1.93%	0.84%	2.27%
PPO w/ filter	1.14%	1.71%	0.98%	1.36%	2.15%	1.86%
AlphaGen	8.13%	0.94%	8.08%	1.23%	6.04%	1.78%
AlphaQCM	**8.49%**	**1.03%**	**9.55%**	**1.16%**	**9.16%**	**1.61%**

Impact of QCM Method and DRL Backbones

Table 2: Out-of-sample IC values across different action selection criteria and DRL backbones.

	CSI 300		CSI 500		Market	
Variance	Mean	Std	Mean	Std	Mean	Std
	Panel A: QRDQN as backbone					
No	6.96%	1.64%	8.54%	1.56%	7.06%	1.92%
Vanilla	6.14%	1.23%	8.80%	1.34%	7.60%	1.17%
QCM	7.59%	0.81%	9.08%	1.07%	9.12%	1.74%
	Panel B: IQN as backbone					
No	7.17%	2.40%	8.58%	1.47%	7.04%	1.82%
Vanilla	6.16%	1.73%	8.75%	1.03%	8.42%	1.59%
QCM	**8.49%**	**1.03%**	**9.55%**	**1.16%**	**9.16%**	**1.61%**

Impact of Domain Knowledge

Table 3: Out-of-sample IC values of the AlphaQCM method with or without domain knowledge.

	CSI 300		CSI 500		Market	
Domain	Mean	Std	Mean	Std	Mean	Std
	Panel A: After 10% Training					
w/	4.93%	0.71%	5.76%	0.68%	6.02%	0.92%
w/o	4.27%	1.75%	5.68%	1.51%	5.87%	1.34%
	Panel B: After 20% Training					
w/	6.32%	1.29%	7.01%	1.77%	6.85%	1.44%
w/o	5.54%	0.78%	6.43%	1.38%	6.43%	2.83%
	Panel C: After 50% Training					
w/	6.41%	1.47%	7.15%	1.22%	7.33%	1.56%
w/o	6.82%	1.35%	7.57%	2.12%	7.48%	1.84%
	Panel D: After 100% Training					
w/	8.17%	1.17%	8.96%	1.51%	8.60%	1.23%
w/o	**8.49%**	**1.03%**	**9.55%**	**1.16%**	**9.16%**	**1.61%**

Impact of Stock Market

Table 4: Out-of-sample IC values on CSI 500 and S&P 500 datasets

	CSI 500		S&P 500	
Method	Mean	Std	Mean	Std
Alpha101	4.38%	-	3.12%	-
MLP	2.72%	0.65%	2.61%	0.49%
XGBoost	4.31%	0.96%	3.08%	0.67%
LightGBM	4.16%	0.81%	3.29%	0.56%
GP w/o filter	1.79%	1.62%	1.88%	1.29%
GP w/ filter	4.52%	1.93%	4.27%	1.39%
PPO w/ filter	0.98%	1.36%	2.03%	1.88%
AlphaGen	8.08%	1.23%	7.48%	0.77%
AlphaQCM	**9.55%**	**1.16%**	8.46%	0.89%

Conclusion

1. This paper proposes a novel DRL method, AlphaQCM, for alpha discovery in the realm of big market data.
2. Unlike the existing methods in the literature, the key idea of the AlphaQCM method relies on the unbiased estimation of variance derived from potentially biased quantiles. This approach enables the efficient alpha discovery in the non-stationary and reward-sparse MDP.
3. To implement the AlphaQCM method, we employ the IQN algorithm as the backbone to obtain quantiles, while approximating the Q function using the DQN algorithm.

Thank you!

References I

Dabney, W., G. Ostrovski, D. Silver, and R. Munos (2018). Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, Volume 80, pp. 1096–1105. PMLR.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of Finance 45*, 881–898.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance 48*, 65–91.

Kakushadze, Z. (2016, March). 101 Formulaic Alphas.

References II

Yu, S., H. Xue, X. Ao, F. Pan, J. He, D. Tu, and Q. He (2023). Generating synergistic formulaic alpha collections via reinforcement learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 5476–5486. Association for Computing Machinery.