Appendix for "Automatic De-Biased Temporal-Relational Modeling for Stock Investment Recommendation"

Weijun Chen[1,4,5] , **Shun Li**[2*] , **Xipu Yu**[3] , **Heyuan Wang**[1,4,5] , **Wei Chen**[1,4,5] and **Tengjiao Wang**[1,4,5]

[1]Key Lab of High Confidence Software Technologies (MOE), School of Computer Science, Peking University
[2]University of International Relations
[3]New York University
[4]Research Center for Computational Social Science, Peking University
[5]Institute of Computational Social Science, Peking University (Qingdao)
oncecwj@stu.pku.edu.cn, lishunmail@foxmail.com, xy2253@nyu.edu,
{wangheyuan,pekingchenwei,tjwang}@pku.edu.cn

A The Selection of MI

A.1 Measurement Algorithms

Mutual information (MI) is a fundamental concept used to quantify the degree of dependence or relationship between random variables. For two variables X and Y, it is typically formulated as:

$$I(X,Y) = \int_{\mathcal{X}\times\mathcal{Y}} log\frac{d\mathbb{P}_{XY}}{d\mathbb{P}_X \otimes d\mathbb{P}_Y} d\mathbb{P}_{XY} \qquad (1)$$

where $\mathbb{P}_{XY}$ is the joint probability distribution, and $\mathbb{P}_X = \int_Y \mathbb{P}_{XY}$ and $\mathbb{P}_Y = \int_X \mathbb{P}_{XY}$ are the marginals. MI differs from correlation information with the ability to comprehensively measure non-linear statistics beyond linear relationships [Kinney and Atwal, 2014]. In *ADB-TRM*, the selection of MI is closely tied to the profitability and robustness of the model. To explore the detailed impact, we conduct the selection experiment by exploring various MI estimation methods. These methods can be categorized into two main groups: those that rely on similarity or dependency estimation, and those that employ distance functions for estimation. For functions based on similarity or dependency measures, they can be directly used as MI. However, for functions based on distance, we utilize the **negative** value of the distance as the measure of MI. In this case, smaller distance values indicate greater MI between variables or distributions. Please note that in *RAT*, the automatically generated graph samples may not always meet the criteria for adversarial graph samples as described in the original paper, and relational adversarial training is not performed at this stage. Consequently, in the training curve shown below, the number of training batches for $\mathcal{L}_\mathcal{O}$ and $\mathcal{L}_\mathcal{G}$ is the same, but it is less than that for $\mathcal{L}_\mathcal{T}$ and $\mathcal{L}_{meta}$.

Cosine Similarity. Given the variable X and Y, cosine similarity is defined as:

$$cos(X,Y) = \frac{<X,Y>}{\|X\| \cdot \|Y\|}. \qquad (2)$$

where $< \cdot,\cdot >$ denotes the inner product between vectors and $\|\cdot\|$ denotes the L_2 normalization.

Deep Correlation Alignment (CORAL) The CORAL [Sun and Saenko, 2016] similarity is defined as the distance between the second-order statistics (covariance) of samples drawn from two distributions:

$$CORAL(X,Y) = \frac{1}{4N^2}\|C_\mathcal{X} - C_\mathcal{Y}\|_F^2. \qquad (3)$$

where N is the dimension of the variables, and $C_\mathcal{X}$ and $C_\mathcal{Y}$ are the covariance matrices of the two distributions.

Wasserstein Distance The Wasserstein distance of order p between two probability distributions $\mathcal{X}$ and $\mathcal{Y}$ on a given metric space S is defined as follows:

$$W_p(\mathcal{X},\mathcal{Y}) = (\inf_{\gamma\in\Gamma(\mathcal{X},\mathcal{Y})} \int_{S\times S} d(x,y)^p d\gamma(x,y))^{1/p}. \qquad (4)$$

where γ encompasses all conceivable joint distributions (couplings) that maintain the specified marginal distributions $\mathcal{X}$ and $\mathcal{Y}$, $d(x,y)$ is the distance between points x and y in the metric space. W_p is the p-th root of the infimum (the greatest lower bound) of the expected value of the distance $d(x,y)$ raised to the power of p. In our experiments, we choose to set $p = 1$. Given the difficulties of directly calculating the exact Wasserstein distance, we opt for approximating it using the differentiable Sinkhorn algorithm [Cuturi, 2013].

Mutual Information Neural Estimation MINE [Belghazi *et al.*, 2018] is a versatile neural estimator that leverages dual representations of the Kullback-Leibler (KL) divergence [Nguyen *et al.*, 2010]. It is designed to create a statistics network $T_\theta : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ for neural information measurement $I_\theta(\mathcal{X},\mathcal{Y})$:

$$I_\theta(\mathcal{X},\mathcal{Y}) = \underset{\theta\in\Theta}{sup}\, \mathbb{E}_{\mathbb{P}_{XY}}[T_\theta] - log(\mathbb{E}_{\mathbb{P}_X\otimes\mathbb{P}_Y}[e^{T_\theta}]). \qquad (5)$$

The expectations in the equation above are estimated using empirical samples from both $\mathbb{P}_{XY}$ and $\mathbb{P}_X \otimes \mathbb{P}_Y$, or by

[*]Corresponding Author.

Table 1: Profitability comparison with different MI metrics.

	Methods	NASDAQ		NYSE		TSE	
		SR	IRR	SR	IRR	SR	IRR
ADB-TRM	Cosine	<u>1.61</u>	<u>0.60</u>	<u>1.36</u>	<u>0.52</u>	<u>1.37</u>	<u>0.91</u>
	CORAL	1.36	0.39	1.14	0.37	1.21	0.68
	Wasserstein	**1.66**	**0.66**	**1.42**	**0.58**	**1.38**	**0.93**
	MINE	1.46	0.47	1.23	0.39	1.31	0.80
	SMILE	1.55	0.58	1.27	0.43	1.35	0.86

shuffling the samples from the joint distribution along the batch axis. The optimization objective can be maximized using gradient ascent.

Smoothed Mutual Information ”Lower-bound” Estimator SMILE [Song and Ermon, 2019] aim to address the high-variance issue in some of the neural estimators (e.g., MINE) and propose to clip the density ratios when estimating the partition function:

$$clip(v, l, u) = max(min(v, u), l),$$

$$I_{SMILE}(T_\theta, \tau) := \mathbb{E}_{\mathbb{P}_{XY}}[T_\theta] - log(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Y}[clip(e^{T_\theta}, e^{-\tau}, e^{\tau})]) \tag{6}$$

where $\tau \geq 0$ is a hyperparameter and T_θ is a neural network that estimates the log-density ratio; this is equivalent to clipping the log density ratio estimator between $-\tau$ and τ.

It is important to emphasize that the aforementioned two methods estimating MI with statistical network T_θ require simultaneous optimization of T_θ. Therefore, we concurrently train the statistical network T_θ alongside the training of *ADB-TRM*. This can lead to considerable additional computational complexity, particularly when dealing with large-scale stock datasets (e.g., NASDAQ and NYSE).

A.2 Experiments

In the subsequent experiments, the profit performance of applying different MI functions is summarized in Table 1. We present the optimization curves of the loss functions $\mathcal{L}_{meta}$, $\mathcal{L}_{\mathcal{T}}$, $\mathcal{L}_{\mathcal{G}}$, and $\mathcal{L}_{\mathcal{O}}$ as the training batch progresses the TSE dataset. Similar performance can be observed on NASDAQ and NYSE datasets.

In-depth Analyses Analyzing the above optimization curves, adopting cosine similarity as the MI metric tends to yield the best convergence effects for the model. The cosine similarity has distinctive numerical properties and well-defined boundaries, hence can support robust optimization. By broadly calculating the angle between vectors, it effectively quantifies the dependency information between variables, providing a straightforward method for assessing mutual information. However, this simplicity of cosine similarity might limit its effectiveness in accurately estimating the degree of non-linear mutual information between variables. Adopting Wasserstein distance as the MI metric also appears to have a positive impact on the model's convergence. While its convergence effect on $\mathcal{L}_{\mathcal{O}}$ might not be as smooth as cosine similarity, it generally performs better than other measures. Furthermore, Wasserstein distance provides a more detailed and comprehensive description of the non-linear correlation and depen-

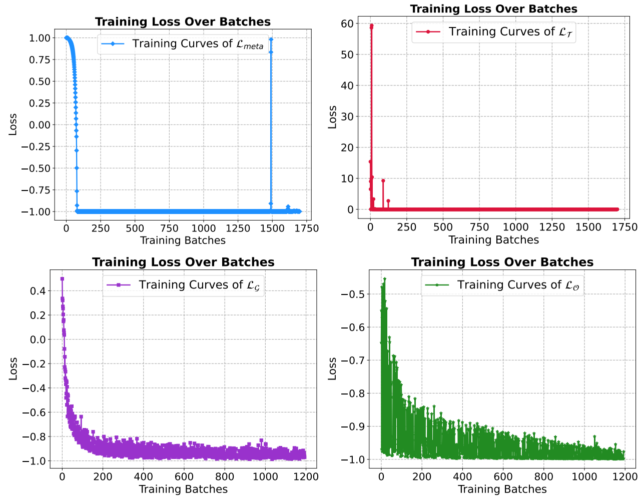

Figure 1: Training Curves of loss functions $\mathcal{L}_{meta}$, $\mathcal{L}_{\mathcal{T}}$, $\mathcal{L}_{\mathcal{G}}$, and $\mathcal{L}_{\mathcal{O}}$ using **cosine similarity** as MI.

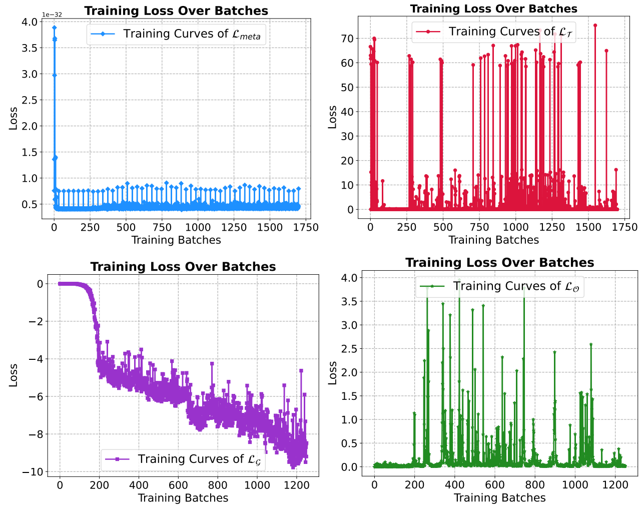

Figure 2: Training Curves of loss functions $\mathcal{L}_{meta}$, $\mathcal{L}_{\mathcal{T}}$, $\mathcal{L}_{\mathcal{G}}$, and $\mathcal{L}_{\mathcal{O}}$ using **CORAL** as MI.

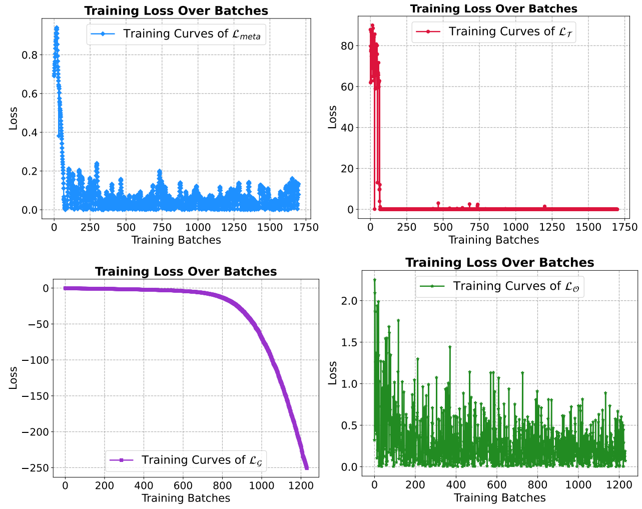

Figure 3: Training Curves of loss functions $\mathcal{L}_{meta}$, $\mathcal{L}_{\mathcal{T}}$, $\mathcal{L}_{\mathcal{G}}$, and $\mathcal{L}_{\mathcal{O}}$ using **Wasserstein distance** as MI.

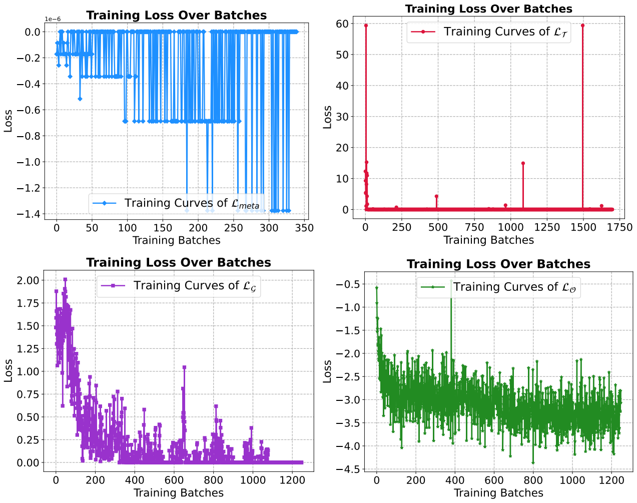

Figure 4: Training Curves of loss functions $\mathcal{L}_{meta}$, $\mathcal{L}_{\mathcal{T}}$, $\mathcal{L}_{\mathcal{G}}$, and $\mathcal{L}_{\mathcal{O}}$ using **MINE** as MI.

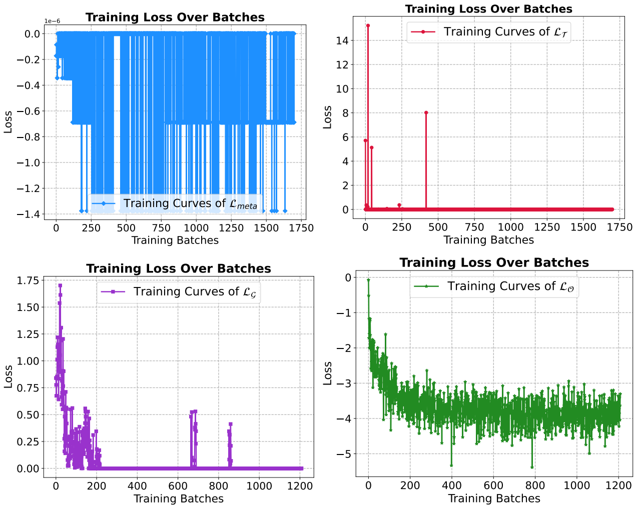

Figure 5: Training Curves of loss functions $\mathcal{L}_{meta}$, $\mathcal{L}_{\mathcal{T}}$, $\mathcal{L}_{\mathcal{G}}$, and $\mathcal{L}_{\mathcal{O}}$ using **SMILE** as MI.

dence between distributions in contrast to cosine similarity. Table 1 shows that the use of Wasserstein distance as the MI metric in *ADB-TRM* yields the highest revenue, underscoring its effectiveness as the optimal MI metric in our framework.

Utilizing CORAL as the MI metric has a detrimental effect on the model's convergence, as all four loss functions exhibit suboptimal convergence. The optimization curve for $\mathcal{L}_{meta}$ reveals that the initial measure value of CORAL is extremely low (around 10^{-32}), suggesting its unsuitability as an MI metric in the overall framework due to its ineffectiveness in differentiating or calculating mutual dependence between distributions. Table 1 further underscores that using CORAL as the MI metric results in the lowest revenue, highlighting the importance of selecting an appropriate MI measure for *ADB-TRM*.

The MINE and SMILE methods that utilize statistical network T_θ to formulate MI between distributions also show promising results in Table 1. However, when examining the optimization curves, it's clear that both MINE and SMILE

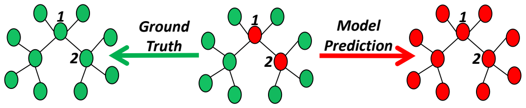

Figure 6: An illustration of the adversarial attack.

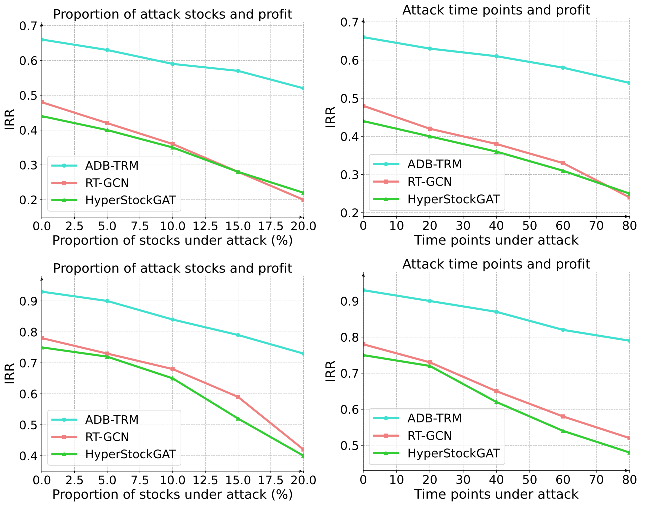

Figure 7: Comparison of revenue curves on NASDAQ (upper two figures) and TSE (lower two figures) for *ADB-TRM*, *RT-GCN*, and *HyperStockGAT* with the proportion of stock nodes and the number of time points under attack.

lead to suboptimal convergence effects for the model, particularly in the case of measurement $\mathcal{L}_{meta}$. Similar to CORAL, MINE and SMILE also face challenges in effectively measuring mutual information between variable distributions, leading to noticeable oscillations and small numerical estimates (around 10^{-6}) in the $\mathcal{L}_{meta}$. This issue is partly due to the fact that the statistical networks T_θ in both MINE and SMILE require training to adapt to the varying distributions before measurement. However, the distributions that require measurement often change during the model's training, meaning that the data distribution encountered by the statistical network T_θ is constantly shifting. This continuing change in distribution introduces inconsistencies and incoherence in the measurement functions of MINE and SMILE, leading to the above fluctuations when estimating MI.

Based on the optimization curves for the five mutual information measurement methods, the $\mathcal{L}_{meta}$ loss is extremely difficult to optimize. Most of these methods struggle to effectively estimate mutual information between *global* and *local* distributions, often leading to very small initial values or significant optimization fluctuations. These reveal the importance and challenges in identifying effective invariant temporal representations.

B Adversarial Attacks

Numerous studies highlight the efficacy of *TRM* in generating excess returns. These models primarily rely on traditional graph neural networks to aggregate relational information and lack specific defenses against potential attacks. However, the information propagation and aggregation traits of conventional graph neural networks make them susceptible to attacks. As depicted in Figure 6, green nodes signify declining stocks, while red nodes represent rising stocks. Attackers could exploit the message propagation mechanism of current *TRM* to skew the overall predictions by, for example, altering stock prices at pivotal nodes 1 and 2, as shown. This manipulation, using the model's propagation mechanism, could lead to inaccurate forecasts and great financial losses. To emphasize the vulnerability of the current TRM and demonstrate the superiority of *ADB-TRM*, we simulate such attacks in our following experiments.

In general, our objective is to further assess the generalization ability and robustness of *ADB-TRM* in a volatile and competitive stock market. We conduct manual adversarial attacks, which involve introducing sudden increases or decreases in the stock price features of selected stock nodes within the test set. During the attack, we randomly choose 5% to 20% of the stock nodes to simulate sudden changes in stock prices, akin to real-world scenarios (possibly due to human manipulation). Specifically, within the test set, we randomly select between 20 to 80 time points and make targeted modifications to the original closing price. These modifications involve increasing or decreasing the closing price by 5%, based on the closing price of the day. Importantly, the final closing price is constrained to remain within the bounds of the highest and lowest stock prices recorded on that day, regardless of whether it increased or decreased. During the time points of the attack, the model refrains from executing any buying or selling operations. When conducting attacks in the temporal domain (time points), the proportion of stocks under attack remains fixed at 5%. In contrast, for attacks in the relational domain (proportion of stock nodes), the number of time points under attack is fixed at 20. We conduct each temporal and relational attack independently five times and report the average results. We compare the final experimental results on NASDAQ and TSE with two state-of-the-art temporal-relational baselines: *RT-GCN* and *HyperStockGAT*.

The results from the above four figures in Figure 7 clearly show the effectiveness of adversarial attacks on these two representative *TRM*. Increasing the proportion of attacking stock nodes and the number of attacking temporal points significantly impairs the model's ability to make profitable judgments and decisions. On the TSE dataset, when exposed to the most extensive temporal attack (with 80 time points under attack), *RT-GCN* and *HyperStockGAT* experience profit losses of 33.3% and 36.0%, respectively, while *ADB-TRM* incurs only a 15.1% profit reduction compared to its original profit. On the NASDAQ dataset, when subjected to the most extensive relational-domain attack (with 20% of stock nodes under attack), *RT-GCN* and *HyperStockGAT* experience profit losses of 57.42% and 49.12%, respectively. In contrast, *ADB-TRM* incurs a 21.21% profit reduction compared to its original profit.

It is evident from the above figures that the relative resilience of *ADB-TRM* to malicious attacks compared to *RT-GCN* and *HyperStockGAT*, aided by the well-designed *RAT*, *TAT*, and *Outer Meta-Learner* modules. Nevertheless, despite exhibiting better robustness compared to other models, *ADB-TRM* still suffers from the negative impact of attacks on overall revenue to some extent. Mitigating noise and attacks for *TRM* remains a valuable research direction in de-biased stock modeling, which also plays a crucial role in the practical implementation of stock investment recommendations.

References

[Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[Kinney and Atwal, 2014] Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014.

[Nguyen *et al.*, 2010] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[Song and Ermon, 2019] Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2019.

[Sun and Saenko, 2016] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016.