Alleviating Non-identifiability: a High-fidelity Calibration Objective for Financial Market Simulation with Multivariate Time Series Data

Chenkai Wang, Junji Ren, and Peng Yang, *Senior Member, IEEE*

Abstract—The non-identifiability issue has been frequently reported in social simulation works, where different parameters of an agent-based simulation model yield indistinguishable simulated time series data under certain discrepancy metrics. This issue largely undermines the simulation fidelity yet lacks dedicated investigations. This paper theoretically demonstrates that incorporating multiple time series data features during the model calibration phase can exponentially alleviate non-identifiability as the number of features increases. To implement this theoretical finding, a maximization-based aggregation function is proposed based on existing discrepancy metrics to form a new calibration objective function. For verification, the task of calibrating the Financial Market Simulation (FMS), a typical yet complex social simulation, is considered. Empirical studies confirm the significant improvements in alleviating the non-identifiability of calibration tasks. Furthermore, as a model-agnostic method, it achieves much higher simulation fidelity of the chosen FMS model on both synthetic and real market data. Moreover, it is both theoretically and empirically analyzed that as long as the features are selected not linearly correlated, they can contribute to the alleviation, which demonstrates the robustness of the proposed objective. Hence, this work is expected to provide not only a rigorous understanding of non-identifiability in social simulation but also an off-the-shelf high-fidelity calibration objective function for FMS.

Index Terms—Financial market simulation, Agent-based modeling, Non-identifiability, High-fidelity calibration objective

I. INTRODUCTION

Social simulation plays a pivotal role in understanding complex social systems in a way of revealing their endogenous complexity [1]. By properly modeling the endogenous components of the social system as multiple interacting agents, i.e., using the Agent-based Models (ABMs) [2], various what-if tests can be conducted by intervening the agents of interest and analyzing the evolution process of the simulators [3], [4]. These what-if tests involving simulating hypothetical scenarios to assess how changes in agent behaviors or interactions may influence the overall system dynamics.

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received XXX,XXX. (Corresponding author: Peng Yang).

Chenkai Wang is with Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen 518055, China (email: wangck2022@mail.sustech.edu.cn).

Junji Ren and Peng Yang are with the Guangdong Provincial Key Laboratory of Brain-Inspired Intelligent Computation and the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China (emails: 12432695@mail.sustech.edu.cn; yangp@sustech.edu.cn). Peng Yang is also with the Department of Statistics and Data Science at the same university.

The success of a social simulation largely lies in the model's fidelity. Note that the what-if tests are often counterfactual, i.e., they do not happen in real social systems, and thus, there are no grounds for validation. To what extent should we believe the results of what-if tests? Besides, due to the uncertainty of human activities, it is often impractical to derive mathematically trustworthy social simulation models. Usually, this issue is addressed in a data-driven manner via model calibration [5].

Specifically, a simulation model $M(\omega)$ is essentially a data-generating process that can produce sequences of arbitrary length, i.e., $M(\omega) = \mathbf{X}_{T'}$, where $T' = [1, ..., \tau]$ and $\tau \in \mathbb{N}^+$. If the simulated data $\mathbf{X}_T$ resembles the observed data $\hat{\mathbf{X}}_T$ of the targeted real system within any discrete time interval of interest $T = [t_s, t_s + 1, ..., t_e]$, we consider that $M(\omega)$ has learned the underlying data-generating probabilistic distribution of the real system for generating $\hat{\mathbf{X}}_T$. Therefore, when using $M(\omega)$ for what-if tests, by keeping the non-intervened agents of $M(\omega)$ unchanged, the simulation results of how the intervened agent influences the system are considered trustworthy. To obtain such $M(\omega)$, the discrepancy between $\mathbf{X}_T$ and $\hat{\mathbf{X}}_T$ is measured by various metrics D [6]. The smaller D is, the closer $M(\omega)$ approximates the data-generating distribution of the real social system. By fixing the ABM-based model structure M, calibrating the parameter ω to fit a given observed data sequence $\hat{\mathbf{X}}_T$ is often modeled as an optimization problem [6]:

$$\min_{\omega \in \Omega^D} D\left(\hat{\mathbf{X}}_T, M(\omega) = \mathbf{X}_T\right). \tag{1}$$

In practice, the calibration methods often suffer from the so-called non-identifiability issue [7], [8]. That is, there are multiple parameters in the parameter space Ω^D share the same discrepancy value to the given $\hat{\mathbf{X}}_T$. However, those parameters highly likely lead to different distributions of data generation due to the nonlinear model structure $M(\omega)$. In other words, the "ground-truth" parameter of deciding the "ground-truth" data generating distribution cannot be effectively identified from the other parameters in those sets. Consequently, the what-if tests on the resultant model with a randomly picked non-identifiable parameter from there are less trustworthy.

Taking the simulation of the well-known flash crush as an example. Two cases of non-identifiability are illustrated in Fig. 1. In the case 1, the two calibrated parameters represent different trading behaviors while both two corresponding simulated data enjoys the identical objective value. By analyzing the flash
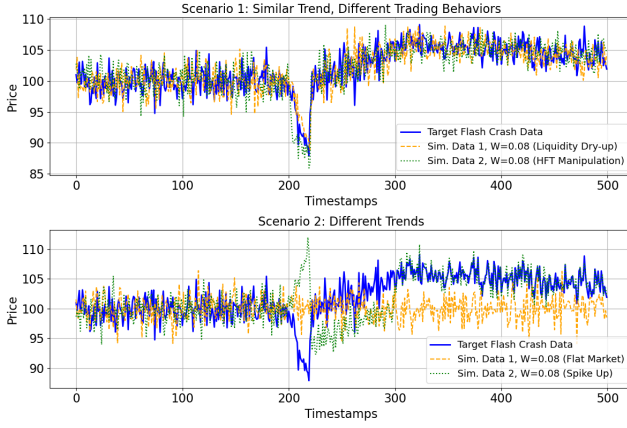
Fig. 1. Two cases of non-identifiable simulation.

Fig. 2. An illustration of the ABM-based FMS.

crush with the two parameters, one may get different causes (i.e., high liquidity taking with large volume of market orders v.s. potential manipulation with large volume of canceling orders). Also, the non-identifiability can lead to the case 2, where two simulated data has the same discrepancy to the target data but appears to be total different trends. In both cases, the calibration algorithm cannot identify the one with higher-fidelity and thus can deteriorate the downstream analysis unexpectedly.

Even though non-identifiability is frequently mentioned in the literature, there is neither a formal definition to describe it nor a rigorous way to alleviate it. This paper first defines non-identifiability as the probability of a randomly uniformly sampled parameter falling into the non-identifiable set. With this definition, this paper mathematically proves that the non-identifiability can be alleviated by calibrating with more distinct features, each of which represents an observed univariate time series. More specifically, we mathematically show that the upper bound of the non-identifiability can be reduced exponentially with the number of features if the selected features satisfy certain conditions.

This explains well that the previous works mostly calibrate their models to merely one data feature and are thus less sufficient to describe the uniqueness of a social system. For clarification and applicability, this work restricts the scenarios to the FMS, as it is a representative social simulation and has been active for over 30 years [9]. Like other complex social systems, the real-world financial markets will output multivariate state data over time, e.g., price, volume, bid/ask directions, order arrival time, and other hand-crafted features [10]. However, existing works merely consider the observed mid-price data for calibration, resulting in wide criticism for their low simulation fidelity [11], especially in the high-frequency intraday simulations [7], [12].

Note that, following our theory, the task of calibrating to a multivariate time series is naturally divided into multiple individual tasks, each of which calibrates the model to one distinct univariate time series. In practice, how to jointly execute those individual tasks so that the obtained parameter is more identifiable to the original calibration tasks with multivariate 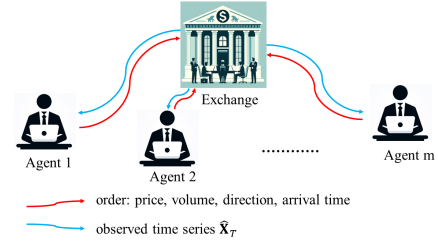time series? This challenges the implementation of the above theory. To address it, we propose a novel calibration objective by aggregating all those individual calibration tasks using the maximization function. This new objective function mathematically asks to search in the intersection of all those individual spaces. Thus, the obtained parameter can be highly identifiable across all individual tasks and thus fit our theory accurately.

Extensive empirical studies have successfully shown that the proposed method can significantly alleviate the non-identifiability issue of FMS. Such advantages are verified robust over 6 commonly seen features of financial market data. Besides, this paper also presents the high-fidelity simulation results by using the proposed new calibration objective function on 10 synthetic data and 1 real data from the Shenzhen Stock Exchange of China.

The remainder of this paper is as follows. Section II describes the background of FMS and related works. Section III theoretically discusses the alleviation of non-identifiability. Section IV reports the empirical studies in detail. Section V concludes this work.

II. BACKGROUND

A. Calibration of Social Simulators

To ensure the simulation fidelity, the simulator often needs careful calibration of the agent parameters so that the simulated data resembles the observed time series from the real system. The general form of calibration refers to (1).

In recent years, the study of the discrepancy metrics D has attracted increasing research attention, as the distance measure between time series remains unsolved and impacts significantly on the calibration performance. Methods of Simulated Moments [13] is a typical method that first transforms the time series into several statistical moments and then calculates the weighted average between the moments of observed data and simulated data. The information criterion based objective functions [14], [15] abstract the temporal information of the original time series via various techniques like histograms and measure the distances between the abstracted temporal vectors. The non-parametric Kolmogorov-Smirnov test is also introduced as D by statistically testing between the probabilistic distributions of the observed data and the simulated data [8]. Other methods employ the Bayesian theory to estimate the likelihood [16] or posterior distribution [17] of the observed data also require a well-defined D for selecting effective samples and updating the estimated distributions.
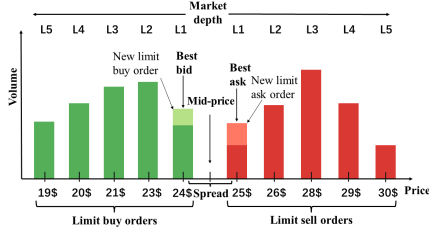
Fig. 3. LOB contains much information about the market.

Despite the progress of exploiting various metrics for calibrating social simulators, existing works frequently encounter the so-called non-identifiability issue, resulting in poor simulation fidelity [18]. From our viewpoint, one major reason is that existing works merely use one single feature to calibrate the model, ignoring that the real systems continuously output multivariate data. Although the features used might be a major factor in describing the social system, other features can help capture more details of the system. Unfortunately, existing calibration objectives are mostly designed for univariate time series data. Consequently, existing works are ill-equipped to calibrate with multivariate time series data. Yet, it is theoretically unclear how exactly it benefits from calibrating more features.

B. ABM-based FMS and Data Features

In social simulations, ABMs are dominantly preferred over the black-box deep learning models due to their interpretability of the data-generating process [19], [20]. ABM constructs a collection of agents mimicking various social activities and interacting together to simulate the whole social system [21]. In FMS, the agents are designed to model various trading strategies [22] (illustrated in Fig. 2), e.g., fundamentalists, momentum traders, high-frequency traders, market makers, and arbitrage traders.

Regardless of the differences among agents, the ABM-based FMSs follow the same workflow. That is, at the beginning of the simulation, the parameterized agents are initialized with the calibrated parameters, which determine how exactly those agents submit orders under their own simulated strategies. Each submitted order usually consists of the price, volume, direction, and arrival time. The Limited Order Book (LOB), a particular data structure for organizing the existing untraded orders, is maintained by the exchange and can be observed by all trading agents. The simulated exchange generates the next LOB by matchmaking the incoming orders submitted by the agents and the current LOB [23].

Those untraded orders are categorized into two directions (bid or ask) and sorted by their prices. Orders with the same price are again sorted by arrival time. Each agent continuously observes the current LOB and submits its order at will. If a new order matches a price in the opposite direction of the LOB, it will be traded immediately, and the matched orders will be removed from the LOB. Undoubtedly, many data features can be extracted from LOB [10]. However, existing FMSs mostly considered the mid-price as the only feature for calibration,

which represents the average between the best ask price and best bid prices (see Fig. 3) and implies the potential price movement of the market.

Recent advances in financial time series modeling prioritize two objectives: capturing nonlinear dynamics while maintaining interpretability. The RHINE model [24] addresses regime-switching through kernel-based spectral clustering and sliding-window eigengap analysis. Parallel innovations in concept drift detection, exemplified by WormKAN [25], employ Kolmogorov-Arnold Networks (KANs) with self-representation matrices to track abrupt transitions. These are unified through MT-KAN's [26] lightweight architectures for multivariate interaction modeling. Besides that, recent works like [8] and [20] tried to calibrate simulators with multi-variate time series, but did not investigate the non-identifiability issue from the perspective of using more features.

III. A HIGH-FIDELITY CALIBRATION OBJECTIVE

This section first formally defines the non-identifiability, then theoretically analyzes how it benefits from calibrating with multivariate data, and lastly proposes a new calibration objective using multiple features to implement our theory.

A. The Definition of Non-identifiability

As existing discrepancy metrics D are designed for univariate data, we re-denote the target data as $\hat{\mathbf{X}}_T^k$ with a superscript k to emphasize that it is a 1-dimensional time series describing some k-th feature of the real system. In financial markets, this feature can be either price, volume, bid/ask directions, order arrival time, or any other hand-crafted features [10]. Correspondingly, the k-th feature of the simulated data is re-denoted as $\mathbf{X}_T^k$. Let $\Omega^{D,k}$ be the parameter space of this calibration task. Then the original calibration task can be re-written as

$$\min_{\omega \in \Omega^{D,k}} D\left(\hat{\mathbf{X}}_T^k, M(\omega) = \mathbf{X}_T^k\right). \tag{2}$$

Without loss of generality, it is assumed that the optimal parameter ω^* lies in the defined parameter space, i.e., $\omega^* \in \Omega^{D,k}$, where $D(\hat{\mathbf{X}}_T^k, M(\omega^*)) < \epsilon$. Note that the optimal parameter may not be unique by definition. The cardinality of the set containing all optimal parameters depends on ϵ, which filters the observation noise in the data. This set, referred to as the non-identifiable set, is formally defined as follows:

*Definition 1 (**The Non-identifiable Set**):* There exists some region $\mathcal{S}^{D,k,\epsilon} \subseteq \Omega^{D,k}$ that for all $\omega_1, \omega_2 \in \mathcal{S}^{D,k,\epsilon}$ $D(\hat{\mathbf{X}}_T^k, M(\omega_1)) < \epsilon$ and $D(\hat{\mathbf{X}}_T^k, M(\omega_2)) < \epsilon$. Such region $\mathcal{S}^{D,k,\epsilon}$ is called the non-identifiable set.

The Lebesgue measure (intuitively the hypervolume) of $\mathcal{S}^{D,k,\epsilon}$, denoted as $\mu\left(\mathcal{S}^{D,k,\epsilon}\right)$, is influenced by ϵ. By choosing the parameter space to be finite and based on Definition 1, we know that $0 < \mu\left(\mathcal{S}^{D,k,\epsilon}\right) \leq \mu\left(\Omega^{D,k,\epsilon}\right) < +\infty$. Then the non-identifiability can be defined as follows:

*Definition 2 (**The Univariate Non-identifiability**):* Let $P\left(\omega \in \mathcal{S}^{D,k,\epsilon} \middle| \hat{\mathbf{X}}_T^k\right)$ be the non-identifiability to the k-th univariate time series data. It is defined as the probability of

uniformly randomly sampling a parameter ω from $\Omega^{D,k}$ while it falls into $\mathcal{S}^{D,k,\epsilon}$. And we have

$$P\left(\omega \in \mathcal{S}^{D,k,\epsilon}\middle|\hat{\mathbf{X}}_T^k\right) = \frac{\mu\left(\mathcal{S}^{D,k,\epsilon}\right)}{\mu\left(\Omega^{D,k}\right)}. \tag{3}$$

Therefore, the rank of any $\omega_1, \omega_2 \in \mathcal{S}^{D,k,\epsilon}$ cannot be effectively identified by the objective D given the data $\hat{\mathbf{X}}_T^k$, leading to that the comparison-based calibration process easily gets stuck in $\mathcal{S}^{D,k,\epsilon}$. Note that, $\mathcal{S}^{D,k,\epsilon}$ can involve arbitrary parameters in $\Omega^{D,k}$, including the "ground-truth" parameter ω^*. Hence, if the calibration process converges to $\mathcal{S}^{D,k,\epsilon}$, one has to randomly pick one $\omega \in \mathcal{S}^{D,k,\epsilon}$ for further what-if simulation. Unfortunately, the underlying data-generating distribution decided by any randomly picked $\omega \in \mathcal{S}^{D,k,\epsilon}$ is less likely to be the same with ω^*, given the nonlinear nature of the simulation model $M(\omega)$. Consequently, the what-if test on the randomly picked ω is less trustworthy unless the cardinality of $\mathcal{S}^{D,k,\epsilon}$ is sufficiently small.

B. Alleviating the Non-identifiability with Multiple Features

In this section, it is shown that the non-identifiability issue can be alleviated exponentially with $K > 1$ features of the observed time series data.

Suppose the observed time series generated by the real social system is multivariate containing K features, i.e., $\hat{\mathbf{X}}_T = \{\hat{\mathbf{X}}_T^k\}_{k=1}^K$, which is true for the financial markets. And the simulation model $M(\omega)$ can also generate a multivariate simulated data $\mathbf{X}_T = \{\mathbf{X}_T^k\}_{k=1}^K$ with the same K features. Given that existing discrepancy metrics D only compute for two univariate time series, there are naturally in total K individual calibration tasks: $\min_w D(\hat{\mathbf{X}}_T^k, M(\omega) = \mathbf{X}_T^k), k = 1, \cdots, K$.

Note that, given any two parameters $\omega_1, \omega_2 \in \Omega^{D,k}$, only they are simultaneously non-identifiable in all K individual calibration tasks, i.e., $\omega_1, \omega_2 \in \{\mathcal{S}^{D,k,\epsilon}\}_{k=1}^K$, are they non-identifiable in the general calibration task with respect to multivariate $\hat{\mathbf{X}}_T$. Otherwise, their discrepancy to $\hat{\mathbf{X}}_T$ can be distinguished in at least one feature and thus are identifiable. The search space $\Omega^{D,k}$ is identical across all calibration tasks and is uniformly represented as Ω^D. That is to say, only the intersection of the K individual non-identifiable set $\{\mathcal{S}^{D,k,\epsilon}\}_{k=1}^K$ can be regarded as the non-identifiable set in the multivariate time series setting, which is defined as follows:

*Definition 3 (**The Multivariate Non-identifiability**):* Let $P\left(\omega \in \{\mathcal{S}^{D,k,\epsilon}\}_{k=1}^K \middle| \hat{\mathbf{X}}_T\right)$ be the non-identifiability to the multivariate time series with K features. It is defined as the probability of uniformly randomly sampling a parameter ω from Ω^D while it falls into all K individual non-identifiable set $\{\mathcal{S}^{D,k,\epsilon}\}_{k=1}^K$. And we have

$$P\left(\omega \in \{\mathcal{S}^{D,k,\epsilon}\}_{k=1}^K \middle| \hat{\mathbf{X}}_T\right) = \frac{\mu\left(\cap_{k=1}^K \mathcal{S}^{D,k,\epsilon}\right)}{\mu\left(\Omega^D\right)}. \tag{4}$$

Since

$$\frac{\mu\left(\cap_{k=1}^K \mathcal{S}^{D,k,\epsilon}\right)}{\mu\left(\Omega^D\right)} \leq \min_k \frac{\mu\left(\mathcal{S}^{D,k,\epsilon}\right)}{\mu\left(\Omega^{D,k}\right)} \tag{5}$$

we have

$$P\left(\omega \in \{\mathcal{S}^{D,k,\epsilon}\}_{k=1}^K \middle| \hat{\mathbf{X}}_T\right) \leq \min_k P\left(\omega \in \mathcal{S}^{D,k,\epsilon}\middle| \hat{\mathbf{X}}_T^k\right). \tag{6}$$

This equality holds only if the K non-identifiable sets are fully overlapped. This can be caused by the fact that the K features are fully dependent and that time series data of one feature can be deduced from other features. This suggests that the selected features should be as diverse as possible. Therefore, with proper choice of the features and D, the non-identifiability of calibrating the multivariate time series $\hat{\mathbf{X}}_T$ is smaller than that of the univariate $\hat{\mathbf{X}}_T^k$. In other words, calibrating multiple features can alleviate the non-identifiability issue. Next, we estimate how fast it decreases with K as follows.

For clarity, let A_k denote the event of a uniformly randomly sampled parameter falling into $\mathcal{S}^{D,k,\epsilon}$. The probability of happening A_k is $P(A_k) = P(\omega \in \mathcal{S}^{D,k,\epsilon}|\hat{\mathbf{X}}_T^k)$. Then we have:

$$P\left(\omega \in \{\mathcal{S}^{D,k,\epsilon}\}_{k=1}^K \middle| \hat{\mathbf{X}}_T\right)$$
$$= P(A_1 A_2 \cdots A_K)$$
$$= P(A_1) \cdot P(A_2|A_1) \cdots P(A_K|A_{K-1}\cdots A_1)$$
$$= \frac{\mu\left(\mathcal{S}^{D,1,\epsilon}\right)}{\mu\left(\Omega^D\right)} \cdot \frac{\mu\left(\cap_{i=1}^2 \mathcal{S}^{D,i,\epsilon}\right)}{\mu\left(\mathcal{S}^{D,1,\epsilon}\right)} \cdots \frac{\mu\left(\cap_{i=1}^K \mathcal{S}^{D,i,\epsilon}\right)}{\mu\left(\cap_{i=1}^{K-1} \mathcal{S}^{D,i,\epsilon}\right)} \tag{7a}$$
$$= \beta_1 \cdot \beta_2 \cdot \beta_3 \cdots \beta_K \tag{7b}$$
$$\leq \left(\frac{\sum_{i=1}^K \beta_i^2}{K}\right)^{\frac{K}{2}}. \tag{7c}$$

The proof of (7c) is provided in appendix A. In (7b), β_1 represents the non-identifiability of feature 1, i.e., the probability that any $\omega \in \mathcal{S}^{D,1,\epsilon}$ falls into Ω^D. $\beta_i, i = 2, ..., K$ denotes the probability that any parameter $\omega \in \cap_{k=1}^i \mathcal{S}^{D,k,\epsilon}$ falls into $\cap_{k=1}^{i-1} \mathcal{S}^{D,k,\epsilon}$. In other words, β_i means the overlapping ratio between $\mathcal{S}^{D,i,\epsilon}$ and $\cap_{k=1}^{i-1} \mathcal{S}^{D,k,\epsilon}$, and $\beta_i \leq 1$. Especially, $\beta_i = 1$ only if $\cap_{k=1}^{i-1} \mathcal{S}^{D,k,\epsilon} \subseteq \mathcal{S}^{D,i,\epsilon}$, indicating that the selected i-th feature is fully dependent on the first $i-1$ features.

The order of features in (7b) can be arbitrary but it influences the upper bound of the multivariate non-identifiability, which is defined as (7c). That is,

*Theorem 1 (**The Exponential Alleviation**):* If each K-th feature is selected such that $\beta_K^2 < \frac{\sum_{i=1}^{K-1}\beta_i^2}{K-1}$, the upper bound of the multivariate non-identifiability is reduced exponentially with the increasing number of features K.

This is intuitive that if $\beta_K^2 < \frac{\sum_{i=1}^{K-1}\beta_i^2}{K-1}$, the base of (7c) will not increase. This theorem not only provides an exponential decreasing upper bound for the multivariate non-identifiability but theoretically suggests a rule to select the features. The proof of Theorem 1 is in appendix B.

C. Aggregating Multiple Features via Maximization

Note that the above discussions require the multivariate time series $\{\hat{\mathbf{X}}_T^k\}_{k=1}^K$ being calibrated separately as K univariate time series with existing discrepancy metrics D. This results

in K individual calibration tasks of $D(\hat{\mathbf{X}}_T^k, M(\omega) = \mathbf{X}_T^k), k = 1, \cdots, K$. How to jointly calibrate the parameter to satisfy the K individual tasks simultaneously? This section proposes a new objective function based on the K individual objectives so that the practical alleviation of the non-identifiability follows the above theory.

Let F be the proposed new objective function used for calibrating the multivariate time series $\hat{\mathbf{X}}_T$, and let $\mathcal{S}^{F,\epsilon}$ denote the associated non-identifiable set defined by this objective function. To achieve $\mathcal{S}^{F,\epsilon} = \cap_{k=1}^{K} \mathcal{S}^{D,k,\epsilon}$, the proposed way is to ensure every location in S^F to be the worst value among all K non-identifiable sets $\{\mathcal{S}^{D,k,\epsilon}\}_{k=1}^{K}$ (see Fig. 4 for illustration). In this case, any parameter ω that is not simultaneously in all K non-identifiable sets will make it fall outside of the $\mathcal{S}^{F,\epsilon}$. This is because the parameters outside the non-identifiable set are even worse than those that are non-identifiable from the ground-truth parameter ω^*.

Given that the calibration problem is a minimization problem, the new objective function F can be implemented with a maximization aggregation among K individual calibration tasks with D. Formally, we have the following theorem.

Theorem 2: (**The utility and uniqueness of new objective function** F) Let F aggregate the K individual calibration tasks via maximization, i.e.,

$$F\left(\hat{\mathbf{X}}_T, M(\omega) = \mathbf{X}_T\right)$$
$$= \max_{k \in \{1, \cdots, K\}} D\left(\hat{\mathbf{X}}_T^k, M(\omega) = \mathbf{X}_T^k\right), \tag{8}$$

then minimizing F uniquely achieves $\mathcal{S}^{F,\epsilon} = \cap_{k=1}^{K} \mathcal{S}^{D,k,\epsilon}$.

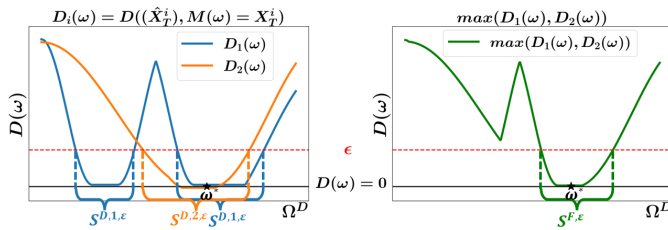The proof of Theorem 2 is provided in appendix C.

Fig. 4. Optimizing the maximization of two functions equals searching in their intersections.

IV. EMPIRICAL STUDIES

This section empirically verifies two major claims.

- With the increasing number of features applied in the calibration, the non-identifiable set shrinks exponentially.
- The fidelity of the simulation data, in terms of various performance metrics, is improved significantly by calibrating more features.

A. Experimental Settings

To conduct a calibration process, 4 terms should be specified according to (8), i.e., the discrepancy metric D, the simulation

model $M(\omega)$, the optimization algorithm $\min_{\omega}$, and the K features $\{\hat{\mathbf{X}}_T^k\}_{k=1}^{K}$.

The discrepancy metric. The Wasserstein distance is employed as D since it provides geometric interpretability and tail sensitivity [27], [28]. Given two univariate time series $\mathbf{X}_T^k$ and $\hat{\mathbf{X}}_T^k$, the Wasserstein distance is defined as

$$D(\hat{\mathbf{X}}_T^k, \mathbf{X}_T^k)$$
$$= \int_{-\infty}^{+\infty} \left| \frac{1}{T} \sum_{t=t_s}^{t_e} \mathcal{I}(\hat{\mathbf{X}}_t^k \leq x) - \frac{1}{T} \sum_{t=t_s}^{t_e} \mathcal{I}(\mathbf{X}_t^k \leq x) \right| dx, \tag{9}$$

where $\mathcal{I}(\cdot)$ is an indicator function that returns 1 if the input event is true, otherwise it returns 0.

The simulation model. This work considers the Preis-Golke-Paul-Schneid (PGPS) model as the simulation model, due to its popularity in recent FMS works [7], [12]. The PGPS model simulates the market dynamics through interactions between two types of agents, i.e., 125 liquidity providers and 125 liquidity takers. The PGPS model has 6 hyper-parameters shared by all agents that need to be calibrated, i.e., $\omega = [\delta, \lambda_0, C_\lambda, \Delta_s, \alpha, \mu]$. At the t-th time-step, each liquidity provider submits a limited order at a fixed probability α with the default volume equals 1. The probability of a limited order being bid side or ask side equals to 0.5. Each liquidity taker submits a market order at a fixed probability μ with the default volume equals 1. The probability of a market order being bid side or ask side is $q_{taker}(t)$ or $1 - q_{taker}(t)$, respectively. The probability $q_{taker}(t)$ is specified by a mean-reverting random walk with mean equals 0.5, mean reversion probability equals $0.5 + |q_{taker}(t) - 0.5|$, and the increment sizes towards mean equals $\pm \Delta_s$. Moreover, the liquidity taker has a probability δ to cancel its untraded limited order. Let $p_a(t)$ and $p_b(t)$ represents the best ask and bid price, respectively. The price of a market order is determined by the market, i.e., the best price at the opposite side of the LOB. The price of a limit order is determined as $p = p_s(t) + \lambda(t) \log(u) + s$ where $p_s(t) = p_a(t)$, $s = 1$ for an ask order, and $p_s(t) = p_b(t)$, $s = -1$ otherwise, with $u \sim U(0, 1)$. Here $\lambda(t)$ is a time-variant order placement depth parameter and is calculated as

$$\lambda(t) = \lambda_0 \left(1 + \frac{\left| q_{taker}(t) - \frac{1}{2} \right|}{\sqrt{< q_{taker}(t) - \frac{1}{2} >^2}} C_\lambda \right) \tag{10}$$

$\sqrt{< q_{taker}(t) - \frac{1}{2} >^2}$ is a pre-computed value. Before initiating the simulation, it is computed in advance through 10^5 Monte Carlo iterations, ensuring its convergence to the accurate standard deviation.

The optimization method. Since the order matchmaking process needs to align with the real market, the PGPS model should be run on the executable trading programs. As a result, the calibration objective function is non-differentiable and the calibration process often has to resort to non-derivative optimization methods. In such contexts, both Bayesian optimization methods and evolutionary algorithms are particularly suitable. This work employs the well-known Particle Swarm Optimization (PSO) algorithm [29], which has been frequently adopted to train the financial simulation or prediction models

TABLE I
THE RANGES FOR RANDOMLY SAMPLING PARAMETERS ω TO GENERATE THE SYNTHETIC DATA

Para.	Range	Remarks
δ	$[0.00, 0.050]$	The probability for each liquidity taker to cancel an untraded order.
λ_0	$[50.00, 300.000]$	Controlling the price of each limited order.
C_λ	$[1.00, 50.000]$	Controlling the price of each limited order.
Δ_S	$[0.00, 0.005]$	Controlling the probability of the side of a market order (ask or bid).
α	$[0.05, 0.500]$	The probability of each liquidity provider submitting a limited order.
μ	$[0.00, 0.050]$	The probability of each liquidity taker submitting a market order.

TABLE II
PARAMETER SETTINGS OF GENERATING 10 SYNTHETIC DATA WITH THE
PGPS MODEL

Parameters	δ	λ_0	C_λ	Δ_S	α	μ
data 1	0.025	100	10	0.001	0.15	0.025
data 2	0.002	200	10	0.002	0.1	0.03
data 3	0.05	152	45	0.003	0.13	0.05
data 4	0.02	288	27	0.003	0.11	0.04
data 5	0.04	62	35	0.003	0.12	0.04
data 6	0.01	171	15	0.0015	0.15	0.03
data 7	0.033	114	14	0.0033	0.1	0.047
data 8	0.01	129	30	0.0017	0.05	0.03
data 9	0.02	62	2	0.003	0.14	0.02
data 10	0.01	87	23	0.001	0.09	0.05

Fig. 5. The illustration of the exponentially decreased non-identifiable set (grey dots). The red dot is the optima.

[6], [20], [30], [31], to optimize ω with respect to (8). For simplicity, the standard version of PSO [32] is considered. The pseudocode for calibration with PSO is shown in Algorithm 1.

Algorithm 1 Calibration with PSO

Input: Number of Iterations to run κ
Parameter: The hyper-parameters of PSO
Output: The optimal candidate ω^*

1: Let $t = 0$.
2: Initialize a population of N particles $\{\omega_i\}_{i=1}^N$.
3: Simulate each of N particles with $M(\omega)$ to obtain N simulated data.
4: Calculate the performance of each of N particles as the Wasserstein distance.
5: Set ω^* to the best performed candidate.
6: **while** $t < \kappa$ **do**
7:     Generate a new population of N particles $\{\omega_i\}_{i=1}^N$ using PSO operators.
8:     Simulate each of N particles with $M(\omega)$ to obtain N simulated data.
9:     Calculate the performance of each of N particles as the Wasserstein distance.
10:     Update ω^* to the best performed candidate.
11: **end while**
12: **return** ω^*

Without further fine-tuning, the hyper-parameters of PSO follow the suggested configurations, where the population size is set to $N = 40$, the inertia weight is set to 0.8, the cognitive and social crossover parameters $c_1 = 0.5, c_2 = 0.5$.

The selected features. Six different yet commonly seen features are considered [23], [33], which mainly concern either the price or volume information of the LOB data and are listed as follows:
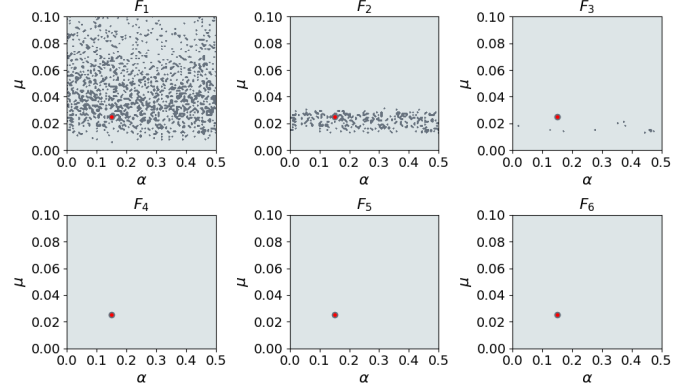
- f_1, the mid-price $m_t = \frac{p_a(t)+p_b(t)}{2}$ at each t-th step;
- f_2, the total traded volume within each t-th step;
- f_3, the price return at each t-th step $\ln m_{t+1} - \ln m_t$;
- f_4, the spread at each t-th step $p_a(t) - p_b(t)$;
- f_5, the volume of the best bid price at each t-th step;
- f_6, the volume of the best ask price at each t-th step.

By incorporating the 6 selected features, 6 objective functions are constructed based on (8), denoted as $F_i, i = 1, 2, ..., 6$. Each F_i indicates that the first i features are used for calibration, i.e., $F_i = \max_{k \in \{1, \cdots, i\}} D(\hat{\mathbf{X}}_T^k, M(\omega))$. In the following empirical studies, it is shown how the increasing number of features influences the simulation fidelity.

The test protocol. The test data consists of 11 targeted time series to be calibrated, including 10 synthetic data sets and 1 real data set. To ensure sufficient diversity, the 10 synthetic time series are generated using the PGPS model by randomly sampling parameters within the ranges specified in table I, as suggested by [12]. The 10 sampled parameters are listed in table II. Each synthetic time series comprises 3600 time steps at a frequency of 1 second. The real data is chosen as 000001.sz from Shenzhen Stock Exchange of China, consisting of 1200 time steps at a frequency of 3 seconds from 9:30 a.m. to 10:30 a.m. of a day in 2019. The time budget of 10^4 simulation evaluations is allowed for the calibration of each targeted data. That is, the PSO runs for 250 iterations with a population size of 40, and the best parameter with the lowest objective function value is applied to the PGPS model to simulate data as $\mathbf{X}_T$.

Calibration performance is evaluated using the Wasserstein (W) distance and Mean Square Error (MSE), two common metrics in time series analysis. Both metrics measure the av-
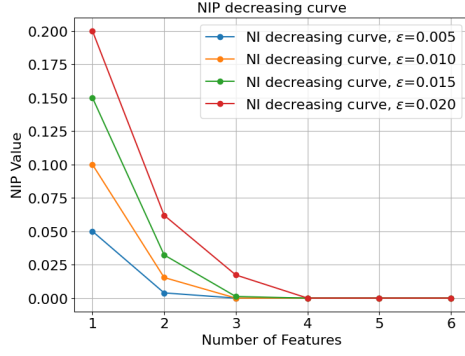
Fig. 6. The non-identifiability (NI) decreases exponentially with the increasing features.
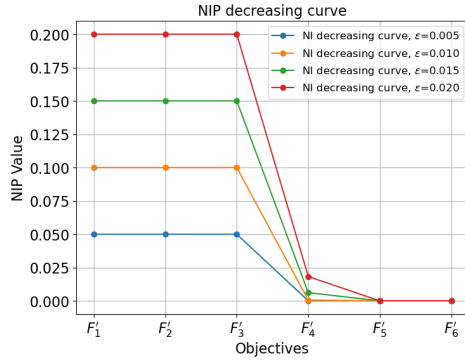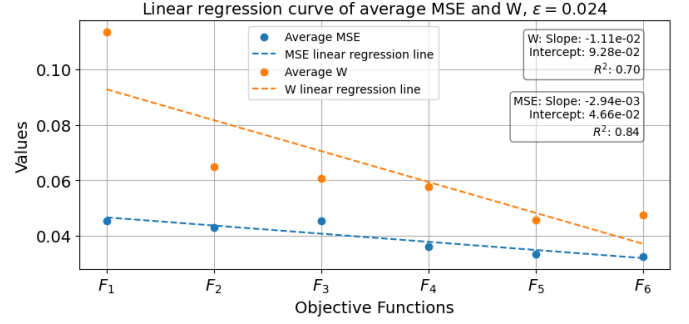

Fig. 8. The distances of the simulated data to the target data decrease almost linearly with the increasing number of features incorporated in F.

Fig. 7. The non-identifiability (NI) decreasing behaviors with new adding features.

erage discrepancy between synthetic and simulated data across 6 features, i.e., $\frac{\sum_{k=1}^{6} D(\hat{\mathbf{X}}_T^k, M(\omega) = \mathbf{X}_T^k)}{6}$, where D denotes W or MSE. Each feature is normalized by the min-max method.

The computational platform. This work builds the simulation environment based on the Multi-Agent Exchange Environment (MAXE) simulator developed by the University of Oxford [34]. The experiments run on a server with 500GB memory, the Intel(R) Xeon(R) Platinum 8358 CPU @ 2.60GHz with 32 physical cores. For a single objective, PSO can be run in parallel across 40 simulators. The calibration of 6 objective functions for 10 synthetic data takes 48 to 60 hours in a data-parallel manner using the CPU.

B. Results on Alleviation of Non-identifiability

Fig. 5 illustrates the changes of the non-identifiable sets in the 2-dimensional parameter space of $[\alpha, \mu]$. This is done by three steps: 1) a target data of 3600 time steps is generated using the PGPS model with the recommended default parameter setting $\omega^* = [0.025, 100, 10, 0.001, 0.15, 0.025]$ [7]; 2) On this default setting, 10000 combinations are created by sampling in $[\alpha, \mu]$ in grid while keeping the values of other 4 parameters unchanged; 3) By using F to measure the discrepancy between the target data and the simulated data of each of the 10000 combinations, it obtains the objective values of the 10000 combinations on each F_i, which are depicted as Fig. 5. The 6 figures in Fig. 5 are depicted according to the objective values of $F_1, F_2, ..., F_6$, respectively. Only the combinations

whose objective values are smaller than $\epsilon = 0.1$ are depicted in grey; otherwise, they will be depicted in blue. For example, in the third figure marked with F_3, the combinations depicted in grey are those whose discrepancy values are smaller than $\epsilon = 0.1$ on all the first 3 features $\{f_i\}_{i=1}^3$. By this means, the grey points indicate the non-identifiable set as they are indistinguishable from ω^* (marked as the red dot). As can be seen, with the increase of the features, the grey non-identifiable set decreases significantly. When 4 are incorporated in F, only the optimum exists, indicating that the non-identifiability issue is indeed alleviated very effectively.

Fig. 6 further depicts how the non-identifiability decreases with the increasing numbers of features with different settings of ϵ. This is done by calculating (7b) using the above 10000 combinations. Clearly, all the curves follow an exponentially decreasing manner. Besides, a larger value of ϵ is, the slower the corresponding curve decreases.

Building on this, an additional experiment investigates the impact of feature dependencies on the alleviation. The theoretical proof in Appendix C suggests that linearly dependent features do not contribute to the alleviation. To validate this, six new test features are constructed based on the original six, where $f_1' = f_1, f_2' = 2f_1, f_3' = -2f_1, f_4' = f_5, f_5' = f_6, f_6' = f_2$. The corresponding new objectives, F_i', are constructed in the same manner as the original F_i. These features include both linearly dependent ones (f_1', f_2', f_3') and non-linearly related ones (f_4', f_5', f_6'), enabling a direct evaluation of their respective effects on NIP reduction. As illustrated in Fig. 7, adding linearly dependent features does
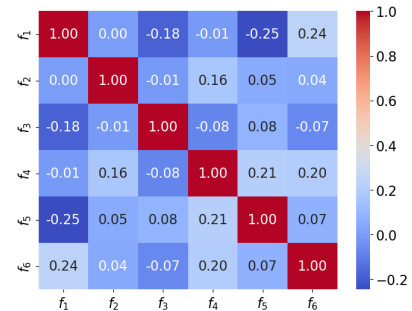

Fig. 9. The average Pearson correlation coefficient of the selected features among 10 simulated data.
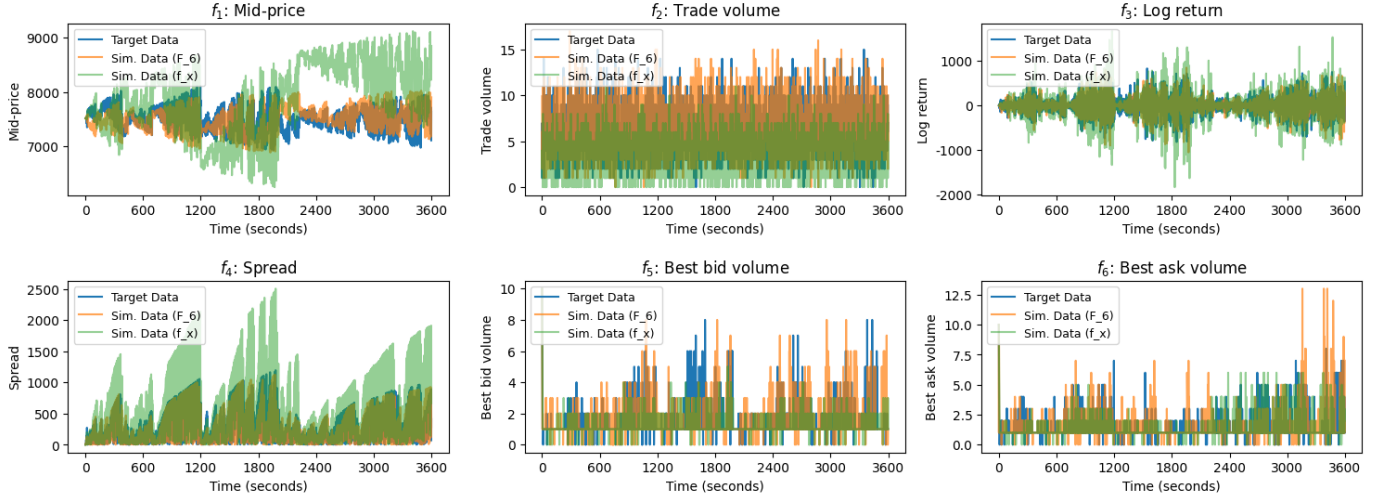
Fig. 10. The simulated data using F_6 resembles the synthetic data on all 6 features.

not reduce non-identifiability, aligning with the theoretical findings in Appendix C. In contrast, introducing non-linearly related features leads to a significant reduction, confirming that independent features impose additional constraints on the non-identifiability set.

While adding more independent features continues for alleviation, the marginal benefit diminishes as more features are included. This effect arises because the initial features already contribute substantially to the reduction, leaving less room for further alleviation. As a result, even if newly added features are entirely independent or possess highly distinctive and informative properties, their impact on further shrinking the non-identifiability set is inherently limited. Mathematically, this corresponds to their associated β values approaching 1, reflecting the decreasing marginal effect of additional features, rather than an indication of feature redundancy.

To further assess the effectiveness of the originally selected six features, the average Pearson correlation coefficient among ten target datasets is computed, as shown in Fig. 9. The results indicate low pairwise correlations among the selected features, supporting their effectiveness in reducing NIP and confirming their direct applicability in calibration.

C. Results on Calibration to the Synthetic Data

Note that, although the non-identifiable set of any F_i is verified to be smaller than that of any F_j, $i > j$, it does not deterministically lead to better calibration performance of F_i than F_j. Because F mainly decides the parameter space, while the calibration performance is also jointly influenced by the optimization algorithms. Given the calibration problem is black-box with multiple local optima, it is non-trivial for the optimization algorithms to find the global optimum even if the parameter space is modeled fully identifiable.

For each of the 10 synthetic data, by optimizing the 6 objective functions $F_i, i = 1, ..., 6$ with PSO, it results in 6 calibrated models and 6 simulated data. Note that, even though a model is calibrated using $F_i, i < 6$, it can still functionally generate the simulated data $\mathbf{X}_T$ with respect to all 6 features.

TABLE III
COMPARISONS OF CALIBRATION PERFORMANCE BETWEEN 6 UNIVARIATE DATA AND F_6.*

	f_1	f_2	f_3	f_4	f_5	f_6	F_6
Data 1	3.99	5.07	6.22	4.62	4.92	3.95	**3.07**
Data 2	3.65	3.95	4.67	3.73	6.18	4.77	**2.90**
Data 3	**2.86**	4.43	3.73	3.87	5.75	4.62	3.08
Data 4	2.83	3.50	3.47	3.87	4.33	2.99	**2.47**
Data 5	2.72	**2.70**	4.93	5.98	3.17	5.83	2.84
Data 6	5.23	5.58	4.02	6.67	4.48	6.77	**3.67**
Data 7	3.08	3.57	**3.06**	3.65	3.60	3.86	3.51
Data 8	4.88	4.50	2.56	3.82	3.91	3.62	**2.26**
Data 9	4.31	3.59	5.07	4.19	5.73	3.82	**3.51**
Data 10	2.53	2.43	2.95	6.93	3.22	2.24	**2.23**
#rank	3.20^+	4.10^+	4.00^+	5.10^+	5.60^+	4.50^+	**1.50**

* Note that the performance is at the $1E10^{-2}$ level. Here + denotes Bonferroni-corrected significance at $\alpha = 0.05$ ($m = 5$, per-test $\alpha = 0.01$) based on pairwise comparisons between F_6 and each baseline $f_1, \cdots, f_5$.

The performance between each simulated data and its corresponding synthetic data is measured by the indicators based on both MSE and W, respectively. In Fig. 8, each (blue or yellow) point displays the average performance on these 10 synthetic data. It shows that by using more features in F, the calibration performance generally improves as the distance values decrease.

Furthermore, it is interesting to see how the simulation model calibrated by all features (F_6) compares to the ones calibrated to every single feature $f_i, i = 1, ..., 6$, instead of in an incremental manner of $F_i, i = 1, ..., 5$. As listed in Table III, by calibrating with F_6, the obtained parameter enjoys the best simulation performance, i.e., the smallest averaged distances on 7 out of the 10 instances and the top 3 averaged distances on the other 3 instances. Besides, the Friedman test ($p = 6.7 \times 10^{-4}$) indicates significant ranking differences

TABLE IV
LINEAR REGRESSION COEFFICIENT TABLE.[*]

ϵ	evaluation D	slope	intercept	R^2
0.016	W	-1.05	9.44	0.61
	MSE	-0.185	4.18	0.59
0.024	W	-1.11	9.28	0.70
	MSE	-0.294	4.66	0.84
0.028	W	-0.746	8.64	0.54
	MSE	-0.154	4.18	0.39
0.030	W	-0.570	8.70	0.61
	MSE	-0.167	4.48	0.24

[*] Note that the unit of slope and intercept is at the $1E10^{-2}$ level.

among methods. Pairwise Wilcoxon tests further confirm that F_6 consistently outperforms f_1 to f_5 with statistical significance. This implies that the proposed theory and new objective function are applicable to various features and not restricted by the feature with more information about the social system, e.g., the mid-price.

We further expand the experiments in Fig. 8 to more different values of ϵ. Note that, ϵ determines how the random noise in the financial market data impacts on the measure of the discrepancy between two time series. The higher the value of ϵ is, the more likely two time series are indistinguishable given the Definition 1. The results are listed in Table IV. It can be seen that the under different ϵ, the calibration performance always decreases linearly with the increasing number of features. This shows that the proposed objective function is robust with respect to different observation noise.

Then, we illustrate the time series of 6 features to show how the simulated data resembles the synthetic data in Fig. 10. Specifically, we calibrate the model by F_6 and f_x on the data generated based on the recommended parameter setting $\omega^* = [0.025, 100, 10, 0.001, 0.15, 0.025]$ [7], where f_x indicates any single feature which helps to calibrate the model and generate the best-simulated data. The simulated data calibrated by F_6 closely resembles the target data on all 6 features. To our best knowledge, this is the first ABM-agnostic method of successfully calibrating the high-frequency market time series at 1 second level and simulating 6 features with high-fidelity. In comparison, the simulated data of f_x performs much poorer. This immediately suggests that calibrating more features can lead to low non-identifiability and high simulation fidelity.

D. Ablation on Aggregation and Discrepancy

The proof in Appendix C establishes that maximization is the unique aggregation function that preserves the intersection property of non-identifiable sets. To empirically validate it, Table V presents an ablation study comparing different aggregation methods (max, min, mean) and calibration measures, i.e., KL divergence (KL), Kolmogorov–Smirnov test (KS), MSE, and W, across 10 synthetic datasets. The results confirm the theoretical properties that treating the aggregation methods under the max aggregation, significantly lower ranks are exhibited by KL, KS, MSE, and W, compared to those under min and mean, as confirmed by the Friedman test

TABLE V
STATISTICAL ABLATION TESTS OF AGGREGATION METHODS AND
CALIBRATION MEASURES[*]

aggregate	max				min	mean
measure	KL	KS	MSE	W	W	W
data 1	3.19	4.47	3.47	**3.07**	6.63	4.77
data 2	6.03	3.08	3.03	**2.90**	3.36	3.89
data 3	6.86	4.30	3.19	**3.08**	4.63	3.47
data 4	3.01	3.10	3.90	**2.47**	4.32	4.33
data 5	**2.55**	6.58	3.75	2.84	3.01	3.79
data 6	3.71	6.69	4.76	**3.67**	3.74	5.73
data 7	3.93	3.10	**1.42**	3.51	3.23	3.52
data 8	3.82	3.18	4.65	**2.26**	3.88	2.54
data 9	8.35	5.48	3.85	**3.51**	5.07	5.40
data 10	3.69	3.63	**2.10**	2.23	5.32	4.04
#rank	3.90+	3.90+	2.90−	**1.50**	4.30+	4.50+

[*] Note that the performance is at the $1E10^{-2}$ level. Here + denotes Bonferroni-corrected significance at $\alpha = 0.05$ ($m = 5$, per-test $\alpha = 0.01$) for pairwise comparisons with max-W, and − indicates no statistical significance.

$(p = 1.74 \times 10^{-4})$. This reinforces the proof that max aggregation maintains strict intersection properties, thus can enhance identifiability. Further pairwise Wilcoxon tests indicate that max-W achieves the lowest rank (1.50), significantly outperforming KL ($p = 4.88 \times 10^{-3}$), KS ($p = 6.84 \times 10^{-3}$), and MSE ($p = 0.0527$), making W the preferable choice for FMS.

E. Results on Calibration to the Real Data

At last, it is verified that the proposed method can also calibrate the real market data at a 3-second level. A 1-hour time series of 000001.sz is calibrated using F_1 and F_3. The simulated data generated by the corresponding best calibrated models is depicted in Fig. 11, together with the real observed data. Note that most of the FMS simulators (including the adopted PGPS model) model the order volume in a coarse-grained manner, i.e., simply setting all the orders with the identical volume of 100. Hence, only the mid-price time series is shown in a time step of 1 minute. It can be seen that the simulated mid-price of F_3 is more similar to the real observed data than F_1 by an improvement of 27.5% in MSE. This confirms that the proposed method indeed can calibrate the
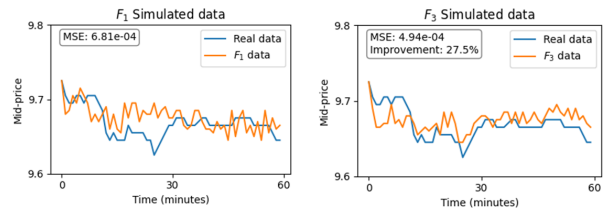

Fig. 11. By calibrating with more features, the simulated data resembles the real market data better.

real market data with multiple time series, thus suggesting a new effective calibration objective function for FMS.

V. CONCLUSIONS AND FUTURE WORK

This paper studies the non-identifiability issue of social simulation. To our best knowledge, this is the first work in which the non-identifiability is formally defined and theoretically alleviated. It is defined as the probability of randomly sampling in the non-identifiable set of the parameter space. It is rigorously analyzed that this probability can be reduced with more time series features in an exponential manner. Based on the analysis, a new objective function is proposed to effectively aggregate multiple features with a maximization function over existing discrepancy metrics. We theoretically and empirically show that the proposed aggregation is the optimal one. Extensive empirical studies have been conducted on 10 synthetic data and 1 real data of the FMS tasks. It has been successfully verified through simulations that, by using multiple features, not only can the non-identifiable set be minimized, but the simulation fidelity is significantly improved. At last, we suggest how to select features via theoretical and empirical studies.

Future work can be multi-directional. First, while the study relies on handcrafted LOB features, learning-based methods, such as advanced representation learning techniques [20] and interpretable alpha factor discovery [35], could help extract more meaningful market signals and refine feature selection. Second, developing adaptive recalibration methods that dynamically adjust parameters based on incoming market data could help maintain high-fidelity simulations over longer time horizons and under changing conditions. Third, beyond financial markets, the approach of reducing non-identifiability through multivariate feature calibration could extend to other agent-based simulations that generate multivariate time series data, like weather forecasting and traffic flow modeling.

APPENDIX

A. Proof of inequality (7c)

Proof 1: Recall that the inequality is

$$\prod_{i=1}^{K} \beta_i \leq \left(\frac{\sum_{i=1}^{K} \beta_i^2}{K} \right)^{\frac{K}{2}}.$$

$\forall i \in \{1, \cdots, K\}$, we have $\beta_i > 0$; by definition, $\beta_i \leq 1$.

Consider $f(x) = \log_e^x \triangleq \ln x$, which is concave on $(0, +\infty)$. Suppose $p_i \geq 0$ and $\sum_{i=1}^{K} p_i = 1$, by Jensen's inequality, we have

$$p_i \sum_{i=1}^{K} \ln x_i \leq \ln \left(\sum_{i=1}^{K} p_i x_i \right). \tag{11}$$

Replace p_i and x_i with $\frac{1}{K}$ and β_i^2, respectively, then

$$\frac{1}{K} \sum_{i=1}^{K} \ln \beta_i^2 \leq \ln \left(\frac{1}{K} \sum_{i=1}^{K} \beta_i^2 \right). \tag{12}$$

Also

$$\frac{1}{K} \sum_{i=1}^{K} \ln \beta_i^2 = \sum_{i=1}^{K} \ln \beta_i^{\frac{2}{K}} = \ln \left(\left(\prod_{i=1}^{K} \beta_i \right)^{\frac{2}{K}} \right). \tag{13}$$

By combining (12) with (13), and noticing the fact that $\ln x$ is strictly increasing on $(0, +\infty)$, we get the desired result.∎

B. Proof of Theorem 1.

Proof 2: To prove the exponential alleviation of the upper bound of non-identifiability under the constraint

$$\beta_K^2 < \frac{\sum_{i=1}^{K-1} \beta_i^2}{K - 1}, \tag{14}$$

We consider two cases with $K - 1$ and K features, respectively. The non-identifiability bounds in these cases are given by

$$P\left(\omega \in \{\mathcal{S}^{D,k,\epsilon}\}_{k=1}^{K-1} \Big| \hat{\mathbf{X}}_T \right) \leq \left(\frac{1}{K-1} \sum_{i=1}^{K-1} \beta_i^2 \right)^{\frac{K-1}{2}}$$

and

$$P\left(\omega \in \{\mathcal{S}^{D,k,\epsilon}\}_{k=1}^{K} \Big| \hat{\mathbf{X}}_T \right) \leq \left(\frac{1}{K} \sum_{i=1}^{K} \beta_i^2 \right)^{\frac{K}{2}}.$$

We then examine the ratio of these two bounds:

$$\frac{\left(\frac{1}{K} \sum_{i=1}^{K} \beta_i^2 \right)^{\frac{K}{2}}}{\left(\frac{1}{K-1} \sum_{i=1}^{K-1} \beta_i^2 \right)^{\frac{K-1}{2}}} = \tag{15}$$

$$\left(\frac{K-1}{K} \cdot \left(1 + \frac{\beta_K^2}{\sum_{i=1}^{K-1} \beta_i^2} \right) \right)^{\frac{K-1}{2}} \cdot \left(\frac{\sum_{i=1}^{K} \beta_i^2}{K} \right)^{\frac{1}{2}}. \tag{16}$$

$$\tag{17}$$

Given the constraint in (14) and the fact that $\forall i \in \{1, \cdots, K-1\}$, $0 < \beta_i \leq 1$, we have

$$\left(\frac{\sum_{i=1}^{K} \beta_i^2}{K} \right)^{\frac{1}{2}} < 1. \tag{18}$$

Furthermore, from (14), it follows that

$$\frac{K-1}{K} \cdot \frac{\beta_K^2}{\sum_{i=1}^{K} \beta_i^2} < \frac{1}{K} \tag{19}$$

and therefore

$$0 < \frac{K-1}{K} \cdot \left(1 + \frac{\beta_K^2}{\sum_{i=1}^{K-1} \beta_i^2} \right) < 1. \tag{20}$$

We define $\Delta_K = 1 - \frac{K-1}{K} \cdot \left(1 + \frac{\beta_K^2}{\sum_{i=1}^{K-1} \beta_i^2} \right)$

Substituting this into the right-hand side of (15), denoted its left-hand side as LHS, we obtain $LHS < (1 - \Delta_K)^{\frac{K}{2}}$.

Therefore,

$$\left(\frac{1}{K} \sum_{i=1}^{K} \beta_i^2 \right)^{\frac{K}{2}} < (1 - \Delta_K)^{\frac{K}{2}} = e^{-\lambda_K \cdot K}, \tag{21}$$

where $\lambda_K = -\frac{\ln(1-\Delta_K)}{2} > 0$. Furthermore, it follows that $\inf_K \Delta_K > 0$, which implies that the corresponding $\inf_K \lambda_K > 0$. Therefore, we define a *uniform constant* $\lambda = \inf_K \lambda_K > 0$.

This ensures that $\forall K, \lambda_K \geq \lambda$. As a result, we have $e^{-\lambda_K \cdot K} \leq e^{-\lambda \cdot K}$, which holds for *all* K. Therefore,

$$\left(\frac{1}{K}\sum_{i=1}^{K}\beta_i^2\right)^{\frac{K}{2}} < (1-\Delta_K)^{\frac{K}{2}} \leq e^{-\lambda \cdot K} \quad (22)$$

∎

C. Proof of Uniqueness of Max Aggregation

Proof 3: We consider a common domain Ω^D for all $\Omega^{D,k}$, as originally stipulated in the definition 3.

Define $F_{\max}\{D_k\} = \max_{1 \leq k \leq K} D_k$, where

$$D_k(\omega) \triangleq D(\hat{\mathbf{X}}_T^k, M(\omega)) \quad (23)$$

for clarity of notation. We first establish the utility of F, followed by its uniqueness.

If $\omega \in \mathcal{S}^{F_{\max},\epsilon}$, then:

$$\max_k D_k(\omega) \leq \epsilon \implies D_k(\omega) \leq \epsilon, \ \forall k \implies$$
$$\omega \in \bigcap_{k=1}^{K}\mathcal{S}^{D,k,\epsilon} \implies \mathcal{S}^{F_{\max}} \subseteq \bigcap_{k=1}^{K}\mathcal{S}^{D,k,\epsilon} \quad (24)$$

If $\omega \in \bigcap_{k=1}^{K}\mathcal{S}^{D,k,\epsilon}$, then:

$$D_k(\omega) \leq \epsilon, \ \forall k \implies \max_k D_k(\omega) \leq \epsilon \implies$$
$$\omega \in \mathcal{S}^{F_{\max},\epsilon} \implies \bigcap_{k=1}^{K}\mathcal{S}^{D,k,\epsilon} \subseteq \mathcal{S}^{F_{\max},\epsilon} \quad (25)$$

Combine the result, we have

$$\mathcal{S}^{F_{\max},\epsilon} = \bigcap_{k=1}^{K}\mathcal{S}^{D,k,\epsilon}, \quad \forall \epsilon > 0, \epsilon \in \mathbb{R} \quad (26)$$

which completes the proof of the utility of F. Next, we establish the uniqueness of F via contradiction. Suppose these exists $F' \neq F_{\max}$ that satisfies (26). Consider two cases:
$\exists\{D_k^*\}$ s.t. $F'(\{D_k^*\}) < \max_k D_k^*$. Take $\epsilon_0 = F'(\{D_k^*\})$:

$$\max_k D_k^* > \epsilon_0 \text{ but } \omega \in \mathcal{S}^{F',\epsilon_0}$$
$$\omega \in \mathcal{S}^{F',\epsilon_0} \implies \omega \in \bigcap_{k=1}^{K}\mathcal{S}^{D,k,\epsilon_0} \implies D_k^* \leq \epsilon_0, \ \forall k. \quad (27)$$

Contradicts $\max_k D_k^* > \epsilon_0$.
$\exists\{D_k^*\}$ s.t. $F'(\{D_k^*\}) > \max_k D_k^*$. Take $\epsilon_0 = \max_k D_k^*$:

$$\omega \in \bigcap_{k=1}^{K}\mathcal{S}^{D,k,\epsilon_0} \implies D_k^* \leq \epsilon_0, \ \forall k \implies F'(\{D_k^*\}) \leq \epsilon_0. \quad (28)$$

Contradicts $F'(\{D_k^*\}) > \epsilon_0$, thus completing the proof and establishing that maximization is the unique aggregation method that preserves the intersection of non-identifiable sets.

Moreover, we analyze two common alternative aggregation methods. The minimum aggregation function is defined as $F_{\min}\{D_k\} = \min_{1 \leq k \leq K} D_k$, This leads to

$$\mathcal{S}^{F_{\min},\epsilon} = \bigcup_{k=1}^{K}\mathcal{S}^{D,k,\epsilon}. \quad (29)$$

The weighted sum aggregation function is defined as $F_{sum}\{D_k\} = \sum_{k=1}^{K}\eta_k D_k$, where $\eta_k > 0$ and $\sum_{k=1}^{K}\eta_k = 1$. This results in

$$\mathcal{S}^{F_{sum},\epsilon} \subseteq \bigcap_{k=1}^{K}\mathcal{S}^{D,k,\epsilon/\eta_k}. \quad (30)$$

∎

D. Proof of Selecting Linearly Independent Features

Proof 4: We demonstrate that if two features are correlated by an affine transformation, there exists a distance metric for which their non-identifiable sets are identical. Therefore, features should be chosen to be linearly independent.

Let $\mathbf{X}_T^k$ and $\hat{\mathbf{X}}_T^k$ (for $k = 1, 2$) represent generated and target time series features related by:

$$\begin{cases} \hat{\mathbf{X}}_T^1 = a\hat{\mathbf{X}}_T^2 + b \\ \mathbf{X}_T^1 = a\mathbf{X}_T^2 + b \end{cases}, \quad a \neq 0, \ b \in \mathbb{R}. \quad (31)$$

The non-identifiable sets are:

$$S^{D,k,\epsilon} = \left\{\omega \Big| D\left(\hat{\mathbf{X}}_T^k(\omega), \mathbf{X}_T^k\right) \leq \epsilon\right\}. \quad (32)$$

Choose D as the standardized MSE distance, denoted as $D_{\text{std}}(X,Y)$,

$$D_{\text{std}}(\mathbf{X}_T^k, \mathbf{Y}_T^k) \triangleq \frac{1}{T}\sum_{t=t_s}^{t_e}\left(\frac{X_t - \mu_{\mathbf{X}_T^k}}{\sigma_{\mathbf{X}_T^k}} - \frac{Y_t - \mu_{\mathbf{Y}_T^k}}{\sigma_{\mathbf{Y}_T^k}}\right)^2 \quad (33)$$

and the mean and standard deviation can be computed as $\mu_{\mathbf{X}_T^k} \triangleq \frac{1}{T}\sum_{t=t_s}^{t_e}X_t$, and $\sigma_{\mathbf{X}_T^k} \triangleq \sqrt{\frac{1}{T}\sum_{t=t_s}^{t_e}(X_t - \mu_{\mathbf{X}})^2}$.

For affine-related features $\frac{\mathbf{X}_T^1 - \mu_{\mathbf{X}_T^1}}{\sigma_{\mathbf{X}_T^1}} = \text{sign}(a)\frac{\mathbf{X}_T^2 - \mu_{\mathbf{X}_T^2}}{\sigma_{\mathbf{X}_T^2}}$, the relationship for $\mathbf{X}_T^k$ and $\hat{\mathbf{X}}_T^k$ is analogous, which yields $D_{\text{std}}\left(\hat{\mathbf{X}}_T^1, \mathbf{X}_T^1\right) = D_{\text{std}}\left(\hat{\mathbf{X}}_T^2, \mathbf{X}_T^2\right)$.

Consequently, the non-identifiable sets are identical, i.e., $S^{D_{\text{std}},1,\epsilon} = S^{D_{\text{std}},2,\epsilon}$.

Under the max-aggregation F, we obtain $S^{F,\epsilon} = S^{D_{\text{std}},1,\epsilon} \cap S^{D_{\text{std}},2,\epsilon} = S^{D_{\text{std}},k,\epsilon}, \quad \forall k$.

Thus, the measure remains unchanged:

$$\mu(S^{F,\epsilon}) = \mu(S^{D_{\text{std}},k,\epsilon}) \quad \forall k. \quad (34)$$

∎

REFERENCES

[1] F. Squazzoni, W. Jager, and B. Edmonds, "Social simulation in the social sciences: A brief overview," *Social Science Computer Review*, vol. 32, no. 3, pp. 279–294, 2014.
[2] J. D. Farmer and D. Foley, "The economy needs agent-based modelling," *Nature*, vol. 460, no. 7256, pp. 685–686, 2009.
[3] A. Fabretti, "On the problem of calibrating an agent based model for financial markets," *Journal of Economic Interaction and Coordination*, vol. 8, pp. 277–293, 2013.

[4] E. Weinstein, "Meltdown modelling," *Nature*, vol. 460, p. 6, 2009.

[5] D. Kim, T.-S. Yun, I.-C. Moon, and J. W. Bae, "Automatic calibration of dynamic and heterogeneous parameters in agent-based models," *Autonomous Agents and Multi-Agent Systems*, vol. 35, no. 2, p. 46, 2021.

[6] D. Platt, "A comparison of economic agent-based model calibration methods," *Journal of Economic Dynamics and Control*, vol. 113, p. 103859, 2020.

[7] K. Goosen, "Calibrating high frequency trading data to agent based models using approximate bayesian computation," Master's thesis, Faculty of Science, 2021.

[8] Y. Bai, H. Lam, T. Balch, and S. Vyetrenko, "Efficient calibration of multi-agent simulation models from output series with bayesian optimization," in *Proceedings of the Third ACM International Conference on AI in Finance*, 2022, pp. 437–445.

[9] R. G. Palmer, W. B. Arthur, J. H. Holland, B. LeBaron, and P. Tayler, "Artificial economic life: a simple model of a stockmarket," *Physica D: Nonlinear Phenomena*, vol. 75, no. 1-3, pp. 264–274, 1994.

[10] A. Ntakaris, M. Magris, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods," *Journal of Forecasting*, vol. 37, no. 8, pp. 852–866, 2018.

[11] G. Fagiolo, A. Moneta, and P. Windrum, "A critical guide to empirical validation of agent-based models in economics: Methodologies, procedures, and open problems," *Computational Economics*, vol. 30, pp. 195–226, 2007.

[12] D. Platt and T. Gebbie, "Can agent-based models probe market microstructure?" *Physica A: Statistical Mechanics and its Applications*, vol. 503, pp. 1092–1106, 2018.

[13] J. Grazzini and M. Richiardi, "Estimation of ergodic agent-based models by simulated minimum distance," *Journal of Economic Dynamics and Control*, vol. 51, pp. 148–165, 2015.

[14] F. Lamperti, "An information theoretic criterion for empirical validation of simulation models," *Econometrics and Statistics*, vol. 5, pp. 83–106, 2018.

[15] S. Barde, "A practical, accurate, information criterion for nth order markov processes," *Computational Economics*, vol. 50, no. 2, pp. 281–324, 2017.

[16] J. Kukacka and J. Barunik, "Estimation of financial agent-based models with simulated maximum likelihood," *Journal of Economic Dynamics and Control*, vol. 85, pp. 21–45, 2017.

[17] J. Grazzini, M. G. Richiardi, and M. Tsionas, "Bayesian estimation of agent-based models," *Journal of Economic Dynamics and Control*, vol. 77, pp. 26–47, 2017.

[18] F. Lamperti, A. Roventini, and A. Sani, "Agent-based model calibration using machine learning surrogates," *Journal of Economic Dynamics and Control*, vol. 90, pp. 366–389, 2018.

[19] K. Jain, N. Firoozye, J. Kochems, and P. Treleaven, "Limit order book simulations: A review," 2024. [Online]. Available: https://arxiv.org/abs/2402.17359

[20] Y. Li, Y. Wu, and P. Yang, "Simlob: Learning representations of limited order book for financial market simulation," 2024. [Online]. Available: https://arxiv.org/abs/2406.19396

[21] J. Paulin, A. Calinescu, and M. Wooldridge, "Agent-based modeling for complex financial systems," *IEEE Intelligent Systems*, vol. 33, no. 2, pp. 74–82, 2018.

[22] S.-H. Chen, C.-L. Chang, and Y.-R. Du, "Agent-based economic models and econometrics," *The Knowledge Engineering Review*, vol. 27, no. 2, pp. 187–219, 2012.

[23] S. Vyetrenko, D. Byrd, N. Petosa, M. Mahfouz, D. Dervovic, M. Veloso, and T. Balch, "Get real: Realism metrics for robust limit order book market simulations," in *Proceedings of the First ACM International Conference on AI in Finance*, 2020, pp. 1–8.

[24] K. Xu, L. Chen, J.-M. Patenaude, and S. Wang, "Rhine: A regime-switching model with nonlinear representation for discovering and forecasting regimes in financial markets," in *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. SIAM, 2024, pp. 526–534.

[25] K. Xu, L. Chen, and S. Wang, "Kan4drift: Are kan effective for identifying and tracking concept drift in time series?" in *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024.

[26] ——, "Kolmogorov-arnold networks for time series: Bridging predictive power and interpretability," *arXiv preprint arXiv:2406.02496*, 2024.

[27] A. Coletta, M. Prata, M. Conti, E. Mercanti, N. Bartolini, A. Moulin, S. Vyetrenko, and T. Balch, "Towards realistic market simulations: a generative adversarial networks approach," in *Proceedings of the Second ACM International Conference on AI in Finance*, 2021, pp. 1–9.

[28] A. Coletta, A. Moulin, S. Vyetrenko, and T. Balch, "Learning to simulate realistic limit order book markets from data as a world agent," in *Proceedings of the Third ACM International Conference on AI in Finance*, 2022, pp. 428–436.

[29] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4. IEEE, 1995, pp. 1942–1948.

[30] M. J. Bazrkar and S. Hosseini, "Predict stock prices using supervised learning algorithms and particle swarm optimization algorithm," *Computational Economics*, vol. 62, no. 1, pp. 165–186, 2023.

[31] M. A. Amal, H. Napitupulu *et al.*, "Particle swarm optimization algorithm for determining global optima of investment portfolio weight using mean-value-at-risk model in banking sector stocks," *Mathematics*, vol. 12, no. 24, p. 3920, 2024.

[32] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *Proceedings of 1998 IEEE international conference on evolutionary computation*. IEEE, 1998, pp. 69–73.

[33] R. Cont, "Empirical properties of asset returns: stylized facts and statistical issues," *Quantitative finance*, vol. 1, no. 2, p. 223, 2001.

[34] P. Belcak, J.-P. Calliess, and S. Zohren, "Fast agent-based simulation framework with applications to reinforcement learning and the study of trading latency effects," in *International Workshop on Multi-Agent Systems and Agent-Based Simulation*. Springer, 2021, pp. 42–56.

[35] J. Zhao, C. Zhang, M. Qin, and P. Yang, "Quantfactor reinforce: Mining steady formulaic alpha factors with variance-bounded reinforce," *arXiv preprint arXiv:2409.05144*, 2024.