Physica A 630 (2023) 129227

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Multiscale major factor selections for complex system data with structural dependency and heterogeneity

Hsieh Fushing [a], Elizabeth P. Chou [b], Ting-Li Chen [c,*]

[a] Department of Statistics, University of California, Davis, 95616, CA, USA
[b] Department of Statistics, National Chengchi University, Taipei, 11605, Taiwan
[c] Institute of Statistical Science, Academia Sinica, Taipei, 11529, Taiwan

ARTICLE INFO

Keywords:
Broken symmetry
Conditional entropy
Contingency table
Major factor selection
Multiclass Classification
Pitching dynamics

ABSTRACT

The unknown multiscale structure hidden in large complex systems is explored bottom-up through discovered heterogeneity under structural dependency embedded within structured data sets. Via two real complex systems, we demonstrate computed hierarchical structures with broken symmetry constituting data's information content. Through graphic displays, such information content indirectly, but efficiently resolves system-related scientific issues that are difficult to resolve directly. All bottom-up explorations and computations are based on conditional entropy and mutual information evaluated upon contingency table platforms after categorizing all quantitative features. Categorical Exploratory Data Analysis (CEDA) first extracts global major factors that share significant mutual information with the targeted response (Re) variable against many covariate (Co) features under the presence of structural dependency. Then each global major factor is taken as one perspective of heterogeneity to subdivide the entire data set according to its categories into sub-collections. This simple "de-associating" protocol significantly reduces structural dependency among the rest of the features such that another run of major factor selection performed on the sub-collection scale can precisely identify which feature sets could provide extra information beyond the global major factor. Finally, informative patterns collected from multiple perspectives of heterogeneity are displayed to explicitly resolve issues of prediction, classification, and detecting minute dynamic changes.

1. Introduction

1.1. Overview and a global goal: a data-driven computing paradigm

In 1859, Charles Darwin presented his groundbreaking theory of evolution, explaining the origin of species through the process of natural selection within ecological systems [1]. Over 140 years later, the concept of self-organization and selection contributing to the emergence of order has provided profound insights into complex systems in various fields, including biology and evolution [2]. Nowadays, complex systems are pervasive in all scientific disciplines [3], engineering [4], and medicine [5]. The recognition of complexity as a multiscale phenomenon has proven to be highly relevant in understanding human societies and global environments [6].

* Corresponding author.
E-mail address: tlchen@stat.sinica.edu.tw (T.-L. Chen).

https://doi.org/10.1016/j.physa.2023.129227
Received 4 December 2022; Received in revised form 10 June 2023
Available online 27 September 2023

What is a common characteristic that contributes to the complexity of a system? It is hierarchical heterogeneity [7]. A complex system is characterized by the presence of diverse and heterogeneous relationships among its non-uniform constituent parts at different hierarchical levels. This characteristic can be observed in social systems, where individuals possess distinct characteristics, in ecological systems with a wide spectrum of species, and even in brain systems composed of inherently different neurons. As a result, experts in complex systems and domain scientists often study these systems with the goal of understanding how the interconnected constituent parts give rise to collective behaviors that define sub-systems at finer scales, and how these sub-systems interact to form large-scale relationships that manifest the entire complex system as a unified whole [8].

Nowadays, in the era of Big Data, experts in complex systems and domain scientists also serve as curators of numerous large structured databases, leveraging their expertise in various fields. While there are countless examples, we will highlight three prominent ones from distinct domains. The first notable example is the extensive collection of databases developed by the National Oceanic and Atmospheric Administration (NOAA). The experts and scientists at NOAA have created a multitude of publicly accessible weather-related structured databases. These databases encompass information on currents, tidal water levels, and other relevant data from hundreds of stations along the Atlantic and Pacific coasts. Detailed minute-by-minute data spanning multiple years can be downloaded for analysis. The second example is the comprehensive health-related database known as the Behavioral Risk Factor Surveillance System (BRFSS), which is maintained by the Center for Disease Control and Prevention (CDC). BRFSS conducts annual phone interviews with approximately 400,000 participants, collecting their responses to over 30 questions related to chronic diseases. This database provides valuable insights into public health trends. The third example is the renowned sport-specific database developed by experts commissioned by Major League Baseball (MLB) called PITCHf/x and STATCAST. MLB collects detailed data on pitching dynamics, capturing information on every pitch thrown by professional pitchers in all games across their 30 stadiums. This includes data on the pitch's location, its outcome, and the involved players. Apart from these diverse examples, the internet is teeming with domain-specific or business-specific websites that offer databases curated by domain scientists. For instance, Physionet.org provides databases focused on heartbeat, respiration, and gait dynamics, catering to researchers in the field. These databases offer a wealth of information and opportunities for analysis in their respective domains.

With the advent of the Internet, structured databases encompassing various complex systems of diverse interest have become readily available to anyone who is curious. In the age of the Internet, the population of individuals with curiosity about different complex systems far exceeds the population of domain scientists and experts in related fields. Therefore, there is a significant demand for genuine information and knowledge about complex systems. Curators of structured databases play a crucial role in making these databases valuable and worth exploring. They leverage their domain knowledge and expertise to select informative features that are included and measured to form the databases, while excluding non-informative and redundant ones. The selection of these features is aimed at establishing structural dependency to capture the intended content of domain knowledge and expertise. In essence, the act of feature selection becomes a means of encoding their domain knowledge and expertise into the structured databases they create. Due to the inherent hierarchical heterogeneity present in complex systems, the collection of selected features has the potential to self-generate the idiosyncratic characteristics of these systems, going beyond the encoded domain knowledge and expertise. From this perspective, such structured databases become highly appealing and worth exploring for a broad spectrum of individuals often referred to as non-experts. It is also stemming from this perspective that the need to build widely applicable data-analyzing paradigms become critical and essential. To fulfill this need, physicists and data scientists alike are the most natural driving forces in resolving the most natural and relevant task: To computationally decode authentic domain knowledge and expertise encoded in a structured database, and at the same time to extract further and discover something more profound and intricate self-generated idiosyncratic characteristics.

In this paper, we specifically focus on databases in which the selected features do not implicitly contain encoded ordinal, spatial, or temporal coordinates. Examples of such databases include BRFSS and MLB, but not all databases from NOAA and Physionet. This is because incorporating spatial and temporal axes adds additional dimensions and complexities to computations and explorations. However, if a structured database involves only a limited number of spatial locations or temporal time points, it can still be included in the framework considered in this paper by introducing additional categorical features to encode spatial locations or temporal time points.

This computational task is equipped with one defining character: a structured database represented as a matrix format. In this format, all selected features are typically arranged along one axis, while subjects are arranged along the other axis. Thus, each entry of this matrix connects a feature on one axis with a subject on the other axis. Moreover, an entry in this matrix can accommodate any data type, including continuous, discrete, or categorical. In other words, some features are quantitative while others are categorical. Despite the variations in data types, the matrix format remains a universal characteristic of all structured databases. Consequently, a computational decoding protocol becomes a unified approach to address the aforementioned computing task when it is designed to be universally applicable to data matrices of any type. Why is a unified decoding protocol necessary?

Complex systems exhibit a wide range of dynamics and encompass diverse knowledge, making it impossible for a universal theory to govern all their dynamics. Nonetheless, a unified decoding protocol, if constructed, must be valid and effective in extracting diverse patterns that carry various forms of knowledge within the data. This ability to extract informative patterns in a universally applicable manner implies that the decoding paradigm must be data-driven and bottom-up in nature. By "data-driven", we mean that a single extracted pattern can inspire multiple directions of exploration. On the other hand, "bottom-up" refers to the derivation of lower-order patterns before higher-order ones. A data-driven and bottom-up computing protocol stands in stark contrast to top-down modeling, which relies on predetermined assumptions and structures. This paper aims to demonstrate the construction of such a bottom-up thinking-based, data-driven computing paradigm to achieve the goal of extracting hidden idiosyncratic knowledge and intelligence from complex systems, as encapsulated within structured databases, in a unified manner.

The potential merits of this unified computing paradigm can be viewed from multiple perspectives. Firstly, it serves as a comprehensive computing paradigm for analyzing all types of structured data. This development holds significant importance in the fields of Statistics and Data Science, especially for scientists across various scientific disciplines who analyze real-world structured data. The resulting information from this paradigm is authentic, devoid of any man-made assumptions, and is both interpretable and explainable. This aspect has the potential to profoundly impact Machine Learning, particularly in terms of result interpretability.

The most crucial aspect is that it provides a new and accessible avenue for scientists and the general public to explore complex systems. Through this paradigm, they may uncover fresh insights into complex systems, whether they are well-studied or not.

1.2. True intelligence in data

The wide availability of databases, coupled with the feasibility of a data-driven and bottom-up computing paradigm, has the potential to expand the spectrum of intelligence in various sciences related to complex systems. Researchers, even those with limited or no domain knowledge or expertise, can effectively act as data analysts by applying a unified computing paradigm to analyze databases of their interest. As the population of such interest-driven researchers grows, the range of extracted patterns will broaden in scope. While some patterns may be informative and others may not, all patterns convey coherent information derived from the data. Through this "more-is-different" phenomenon, new and surprising forms of complex system-specific intelligence may emerge, which could be unimaginable to a few domain scientists and even the curators of such databases. This approach represents a way of enhancing human intelligence across all complex systems. The formation of this human-filtered intelligence stands in stark contrast to the ad hoc or even unnatural criterion-filtered formation of artificial intelligence (AI).

The recent successes of machine learning (ML) and artificial intelligence (AI) in numerous scientific and medical domains undoubtedly signify the advent of computer-algorithm-generated "intelligence". Among these successes, one particularly noteworthy achievement in terms of its magnitude and impact is AlphaFold, developed by DeepMind, a subsidiary of Alphabet. This AI program exhibits remarkable predictive capability and accuracy in predicting the 3D structure of proteins [9]. However, these predictive successes are often accompanied by a significant challenge: the ability to elucidate the rationale behind the AI program's decisions. In other words, there remains a conspicuous gap between predictive success and the understanding of underlying truth. This substantial gap appears to be closely tied to the "Blackbox" nature of ML and AI algorithms.

In general, a vast amount of data is fed into the Blackbox of ML and AI to enhance their predictive capabilities. However, this requirement for a large training dataset may not be feasible for a single database focused on a complex system. By forcibly dividing the database into training and testing subsets, a fundamental issue arises: Are the information contents of the training and testing subsets equivalent? This issue has yet to be resolved from a technical standpoint in the ML and AI literature. Various parsimonious proposals have been put forward to address this problem. For instance, the entire database can be subdivided into multiple (more than two) subsets using simple random sampling, and one subset can take on the role of testing at a time. The crucial point is that the equality between the training and testing subsets is highly questionable when the database is not large in size and involves hierarchical heterogeneity induced by high dimensionality. Considering the aforementioned limitations in explaining and interpreting results, ML and AI are not yet coherent choices for sciences dealing with complex systems until this fundamental issue is resolved.

When scientists, researchers, and data analysts choose to embrace the data-driven bottom-up computing paradigm, numerous challenges stemming from large databases are bound to arise. Among these multifaceted challenges, two stand out as the most critical and essential. The first challenge involves extracting the authentic information content from the data as comprehensively as possible. The second challenge pertains to transforming the data's full information content into data-driven intelligence.

In this paper, we address the first challenge by analyzing databases derived from two complex systems. Specifically, we tackle the task of computing and visualizing the hierarchical heterogeneity of data through the structured dependency among features. Additionally, we briefly touch upon certain aspects of resolving the second challenge, wherein solutions to various detection and prediction questions can be observed in the graphic displays constructed during the process of addressing the first challenge. However, a more comprehensive treatment of resolving the second challenge will be presented in a separate study. In our ongoing research, we develop computational techniques to uncover the feature-based patterns sustained by specific groups of subjects, thereby demonstrating that these linkages collectively reveal knowledge-like conclusions. These conclusions possess a hierarchical and heterogeneous nature. Ultimately, we argue that holistically addressing and merging these two challenges constitutes the fundamental objective of data analysis, particularly concerning databases associated with real-world complex systems.

1.3. MLB-Statcast and CDC-BRFSS databases and related issues

Before delving into the concepts and methodologies essential to our development of a data-driven bottom-up computing paradigm in this paper, we provide a brief introduction to the two example databases analyzed: the CDC's BRFSS and MLB's Statcast.

According to the CDC's web site (https://www.cdc.gov/brfss/index.html), the Behavioral Risk Factor Surveillance System (BRFSS) is the nation's leading system of health-related telephone surveys. Since 1984, the BRFSS has been collecting state data on various health-related risk behaviors, chronic health conditions, and the utilization of preventive services among U.S. residents. With over 400,000 adult interviews conducted annually, the BRFSS stands as the world's largest continuously conducted health survey system.

The reliability and validity of the BRFSS have been thoroughly reviewed and examined in numerous studies. To highlight a few examples, a systematic review focused on the prevalence of various chronic diseases and other factors, comparing surveys that used

physical measures in addition to self-reported data with those that did not [10]. Another study compared national estimates from the National Health Interview Survey (NHIS) and the BRFSS, specifically analyzing 14 measures such as smoking, height, weight, BMI, diabetes, hypertension, and overall health status [11]. The authors concluded that the BRFSS provides national estimates comparable to those of the NHIS, thus offering rapidly available information to guide national policy and program decisions. However, it is important to note that the BRFSS also faces challenges in collecting reliable and valid data due to the rapid changes in personal communication technologies [12].

In this paper, we analyze a cleaned version of the BRFSS database from the year 2015, which consists of more than 250,000 subjects. This structured dataset is available on Kaggle (https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset) and is commonly used in the machine learning literature for predicting heart disease status. The dataset exhibits an imbalance in the non-disease versus disease odds, with a ratio of almost 10-to-1. Due to this imbalanced nature and the underlying hidden heterogeneity, predicting patients' disease status using popular machine learning methodologies has proven to be a challenging task. These methodologies commonly suffer from significantly high error rates, performing worse than simply predicting everyone as non-disease with an accuracy of over 90%. We demonstrate that achieving higher accuracy becomes extremely challenging when ignoring the presence of hierarchical heterogeneity in the data.

The second dataset is extracted from MLB's PITCHf/x and Statcast database. From the years 2006 to 2017, every baseball pitch thrown by a Major League Baseball (MLB) pitcher in any of MLB's 30 stadiums is recorded using two high-speed cameras. These pitches are algorithmically and automatically annotated, resulting in measurements of 22 features along with other relevant information such as batter ID and batting result. For a brief description and a list of all the selected variables, please refer to [13].

On August 15, 2007, an article titled "Pitchf/x, the new technology that will change baseball analysis forever" was published in Slate Magazine, written by Nate DiMeo. The message conveyed in that article still holds true today, as evidenced by several studies. For example, knuckleball trajectories, which are notoriously difficult to hit, have been analyzed [14]. Algorithms for topics in machine learning, such as multiclass classification and Response Manifold Analytics, have been developed for analyzing baseball pitching dynamics [13], and pitching data mimicking has been carried out in [15].

In the year 2015, Statcast was introduced, which combined camera and radar systems installed in MLB stadiums. Following the switch from Pitchf/x to Statcast in 2017 and the implementation of the Hawk-Eye system, the trajectory of a baseball after being hit is tracked by radars and categorized based on its location on the baseball field. Each year, over half a million structured data points of these pitches are recorded and stored in the PITCHf/x and Statcast databases. The radar data, along with information on the arrival of Statcast, is described in an article titled "Statcast Arrives, Offering Way to Quantify Nearly Every Move in Game" written by Richard Sandomir and published in The New York Times on April 21, 2015. Both of these databases can be accessed on the MLB website (https://www.mlb.com).

It is known that each MLB pitcher utilizes several subtypes within their main pitch types, such as fastball, curveball, and slider, and these subtypes may vary across different seasons. This presents a significant challenge when attempting to address important issues such as: (1) precisely linking each pitch to its respective pitcher, known as the multiclass classification (MCC) problem in ML literature; and (2) detecting and revealing minute changes in a pitcher's pitching dynamics across consecutive seasons.

As mentioned earlier, databases like MLB's PITCHf/x and STATCAST, as well as CDC's BRFSS, are examples of the numerous databases currently available on the internet. These databases are created by their respective experts with the goal of gaining a better understanding of the complex dynamics in MLB games and chronic disease dynamics in American societies. The domain knowledge and expertise regarding various complex system dynamics are embedded within these databases. In this paper, the data-driven bottom-up computing paradigm developed aims to explicitly extract the information content of these dynamics, which is sustained by domain knowledge and expertise. The objective is to achieve a better and more accurate understanding of the complex systems under study. As a result, various inferential problems, such as prediction and classification decision-making, as well as the detection of minute changes over different time intervals, can be resolved as byproducts of the extracted information content from the data.

1.4. Why data-driven bottom-up computing paradigm?

It is not unreasonable to claim that, in general, humans lack systemic multiscale knowledge regarding the dynamics of complex systems within our societies and the real world we inhabit. This lack of knowledge hinders our ability to formulate simple and generic analytical or functional expressions for the targeted complex dynamics in a top-down manner. Therefore, when we have access to a structured database, the immediate question arises: Where should we begin? The answer lies in identifying the factors that constitute defining relational associations that characterize the dynamics of the complex system of interest to a significant extent.

One crucial and natural defining relational association is the relationship between response and covariate features. Understanding the constituent factors related to this relationship can greatly enhance our comprehension of the complex system under study. From the perspective of response-covariate relations, we briefly argue why a data-driven bottom-up computing paradigm is necessary.

Consider a structural dataset consisting of data points in the form of an $L + K$ dimensional vector, where each component represents either quantitative measurements or categorical observations, or a mixture of both. The first L components correspond to the response (Re) features, denoted by $\mathcal{Y} = (Y_1, \ldots, Y_L)'$, while the remaining K components are derived from K one-dimensional covariate (Co) features denoted by $V_1, \ldots, V_K$. It should be noted that some or all of the response and covariate features can be categorical.

Conceptually, the response-covariate (Re-Co) dynamics can be understood as the collective set of associative relationships between subsets of the response features and various key subsets of the covariate features. This collective nature of Re-Co dynamics

is not governed by a single functional structure, as different subsets of L response features interacting with different subsets of K covariate features yield distinct perspectives on the Re-Co dynamics. This assertion is motivated by the inherent hierarchical heterogeneity [7] and multiscale deterministic and stochastic structures [16] that naturally exist in complex systems.

Furthermore, it is important to recognize that even with a fixed perspective on Re-Co dynamics, it is unlikely that a broad and unspecified functional construct can fully capture it. Suppose that a collection of M constituent mechanisms, denoted as $\{\mathcal{F}_m\{A_m^*\}|m = 1,\ldots,M\}$, govern a specific Re-Co dynamics, and these mechanisms are incorporated into a global function $\mathcal{G}(\cdot)$ as follows:

$$\mathcal{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_L \end{bmatrix} \cong \mathcal{G}(\mathcal{F}_1\{A_1^*\}, \mathcal{F}_2\{A_2^*\}, ..\mathcal{F}_M\{A_M^*\}, \oplus\varepsilon). \tag{1}$$

If we have no prior knowledge about the number of mechanisms M, the functional forms of $\mathcal{F}_m\cdot$ and $\mathcal{G}(\cdot)$, and how $\oplus\varepsilon$ operates within $\mathcal{G}(\cdot)$, any assumptions made by humans are prone to errors. This fact was well-recognized by Nobel physicist P. W. Anderson in his 1972 Science paper titled "More is Different" [7]. Therefore, top-down approaches, in general, are not valid for analyzing databases of complex systems.

In contrast, bottom-up thinking remains the only feasible and realistic approach for analyzing databases of complex systems. Our previously developed Categorical Exploratory Data Analysis (CEDA) has demonstrated its ability to explicitly and reliably select a collection of major factors, denoted as $\{A_m^*|m = 1,\ldots,M\}$, by employing Theoretical Information Measurements [17]. This protocol for selecting major factors has been successfully applied to several real examples, even without knowledge of the specific mechanisms $\{\mathcal{F}_m\{\cdot\}|m = 1,\ldots,M\}$ [18]. This capability of CEDA makes bottom-up thinking practical and effective in analyzing large databases.

However, all the identified major factors $\{A_m^*|m = 1,\ldots,M\}$ are variables defined across the entire dataset in [17]. In contrast, the heterogeneity considered in [18] is apparent and visible due to its direct link to the IDs of systems constituting Many-System Problems. Therefore, the literature still lacks an insightful concept of heterogeneity and a systematic approach for discovering hidden heterogeneity in any large dataset.

Ideally, according to the specifications of hierarchical heterogeneity in [7], any relevant heterogeneity must exhibit locality-specific characteristics, just like major factors. These locality-centric concepts of major factors and heterogeneity are truly unique. The primary computational achievement of this paper is to establish an explicit link between the task of selecting major factors and the task of discovering hidden hierarchical heterogeneity, thereby operationalizing a data-driven bottom-up computing paradigm. The ensemble of localities that collectively facilitate this data-driven bottom-up computing paradigm is referred to as the heterogeneity contained in the data.

In summary, gathering all relevant locality-specific collections of major factors into an ensemble is a way to address the first front of the challenge in analyzing big data. Equally important is the attempt to synthesize all members of this ensemble of major factors with their data-subjects' linkages, which can potentially provide insights into the operation of $\mathcal{G}(\cdot)$ across all localities. In other words, the construction of this computing paradigm for addressing the first front of the challenge aligns with the resolution of the second front of the challenge in analyzing databases of complex systems.

1.5. Motivations and outline of this paper

Although we expect heterogeneity to be present in the majority of large structured databases, the concept and understanding of "what heterogeneity is" are still lacking in many scientific disciplines that deal with diverse complex systems. The practical issue of how to discover heterogeneity is currently at a primitive stage of human intuition. For example, if we want to investigate gender differences in heart disease dynamics, it may seem appropriate to split the dataset into two subsets: one for males and one for females. But does this data-splitting operation effectively reveal the differences in heart disease dynamics between males and females? Similarly, if we want to uncover subtle changes in the pitching dynamics of an MLB pitcher over several consecutive seasons, it may seem logical to analyze the pitcher's data by categorizing it based on the season ID. But is this approach strategically sound? Unfortunately, the answers to both of these questions are "no". As demonstrated in this paper, the primitive data-splitting operation is neither effective nor strategic in terms of extracting the true information content of the data, which holds the key to understanding the original scientific issues related to dynamics.

However, the concept of heterogeneity is intended to be linked with collectives of localities, which often exhibit a hierarchical structure. Therefore, we need a data-driven computing protocol to derive such hierarchical collections of localities. Moreover, we need a unified fundamental concept of heterogeneity to guide this data-driven protocol operationally. Since this fundamental concept is still not well-conceived or developed, there is currently no operational data-driven computing protocol that can effectively uncover the true heterogeneity present in the data. In this paper, we propose a fundamental concept of heterogeneity from a unique information perspective and subsequently build an easily operable data-driven computing protocol. We motivate this fundamental concept based on the "major factors" of the targeted Re-Co dynamics at both global and fine scales.

As discussed in [17], a covariate feature subset A_m^* is considered a global or locality-specific major factor if it exhibits predictive power for $\mathcal{Y}$ on the global or locality scale, respectively. The specific nature of this predictive relationship will be further elucidated in terms of Information Theoretical Measurements in the subsequent section. Our logical approach begins by addressing the computational task of discovering $\{A_m^*|m = 1,\ldots,M\}$ on the global scale. However, this task is inherently complex due to the

presence of structural dependencies among the covariate features $\{V_1, \ldots, V_K\}$. The associations between the response features $\mathcal{Y}$ and the members of $\{V_1, \ldots, V_K\}$ are typically highly intertwined. A common scenario involves the following example: the association between $\mathcal{Y}$ and one covariate feature, let us say V_1, may primarily stem from its dependence on V_2. As a result, V_2 should be more crucial to $\mathcal{Y}$ than V_1, although not entirely. Untangling such a partially false and partially true factor relationship becomes a central focus in this context.

Here, we propose to disentangle this factorial relationship by accurately assessing the additional information that V_1 can offer beyond what V_2 can provide. If this additional amount of information is minimal, then V_1 cannot be considered an independent major factor. Even in cases where higher-order associations exist between subsets of $\{V_1, \ldots, V_K\}$ and $\mathcal{Y}$, this concept of disentanglement remains applicable. Furthermore, while structural dependencies may also exist on the locality scale, this operational concept is equally effective in capturing fine-scale localities.

It should be noted that this operational concept was absent during the initial identification of major factors in the Re-Co dynamics conducted in [17]. In that study, the decision was based on determining whether the bivariate variable (V_1, V_2) could achieve a significantly larger reduction in uncertainty when predicting $\mathcal{Y}$ compared to the singleton variable V_2. In contrast, this operational concept precisely identifies the specific localities defined by V_2 where V_1 can or cannot provide additional information in terms of uncertainty reduction. This precise operation at the locality level is referred to as "de-associating".

We will provide a brief outline of how to utilize the "de-associating" operation to achieve the desired information content in an indirect manner. To illustrate this plan more concretely, we will use a real example involving a group of 12 MLB pitchers. First, we recognize that the MLB data we are working with primarily encodes information related to pitching dynamics. Therefore, we need to choose a multidimensional response variable that effectively captures the Re-Co dynamics, encompassing the physical rules and mechanical principles of musculoskeletal systems that govern pitching dynamics in general. Second, we identify a small collection of low-order global major factors that collectively capture this Re-Co dynamics to a satisfactory extent. These major factors serve as the foundational elements for understanding the overall system behavior. Third, the collection of low-order major factors gives rise to a spectrum of localities, each representing a specific aspect of the Re-Co dynamics. These localities signify the heterogeneity on a global scale, where different aspects of the system dynamics are highlighted. Fourth, within each identified locality on the global scale, we search for local major factors that can induce heterogeneity on a local scale. These local major factors, either original or altered, provide a means to explore beyond the physical rules and mechanical principles, allowing us to observe manifestations of personal differences and even minute changes across seasons. This series of steps, collectively referred to as the "de-associating" operation, enables us to extract valuable insights from the broken pieces of locality induced by the identified major factors. It provides an exploratory approach to delve into the complexities of the system dynamics and uncover various aspects of heterogeneity that may not be readily apparent based solely on physical principles and rules.

This plan can be potentially applied to a wide range of complex systems that possess structured databases. The knowledge of the multiscale collections of major factors, denoted as $\{A_m^* | m = 1, \ldots, M\}$, plays a crucial role in weaving together the multiscale information content embedded in the data. By employing the Categorical Exploratory Data Analysis (CEDA) paradigm, the hidden heterogeneity-based information content of the data is revealed, offering natural and efficient solutions to predictive and testing inferential issues within complex systems. These resolutions seem to emerge effortlessly from the data's information content as natural byproducts.

However, when attempting to address predictive and testing inferential issues directly under complex systems, several challenges arise. For instance, consider the case of Re-Co dynamics defined by a categorical response variable with 12 categories representing pitcher IDs, commonly known as the Many-System Problem examined in [18]. In this scenario, all major factors are identified on a global scale, and they are likely to be associated with one another due to their shared adherence to the same physical rules and mechanical principles. Assessing the individual contributions of these factors to specific aspects of the information content becomes exceedingly difficult in practice. Achieving precision in such evaluations also proves to be highly challenging. Another example is the multiclass classification (MCC) problem found in the field of statistics and machine learning literature. In this context, classification accuracy serves as the sole criterion for evaluating the performance of classification algorithms, often disregarding the corresponding Re-Co dynamics and its associated information content.

Our computational developments begin by adopting the Categorical Exploratory Data Analysis (CEDA) paradigm, which operates on all categorical or categorized feature variables. The foundation of CEDA relies on the recognition that all data types inherently possess a categorical nature. Consequently, a properly constructed high-dimensional histogram of $\{\mathcal{Y}, V_1, \ldots, V_K\}$, or a hyper-contingency table representation, serves as an almost sufficient statistic that contains nearly complete relational pattern information. This assertion is established through multiple simulation studies based on homogeneity, conducted under varying degrees of structural dependency. The results of these studies are subsequently compared favorably with popular LASSO regression results in [19].

This paper is structured as follows. Section 2 provides a review of the concepts and functions of Information Theoretical measurements necessary for CEDA. Additionally, the "shadowing" and "de-associating" operations are introduced within the context of the contingency table platform. Section 3 utilizes several simulation examples with varying degrees of structural dependency to illustrate the effects of these two operations, with particular emphasis on showcasing the critical role of the "de-associating" operation. In Section 4, the Kaggle version of the BRFSS database is analyzed using the "de-associating" operation to investigate how hierarchical heterogeneity influences a broad range of Re-Co dynamics related to heart disease. In Section 5, fastball data from 12 MLB pitchers' pitching dynamics over three seasons is extracted from the Statcast database. The hidden hierarchical heterogeneity and structured dependency of the data are examined and presented alongside resolutions to the aforementioned inferential questions discussed in the previous subsection. Finally, the Conclusion section reiterates the merits of our data-driven bottom-up computing paradigm and suggests several avenues for future research in analyzing big databases within complex systems.

2. Technical developments

2.1. Information theoretical measurements: concepts and functions

To uncover the multiscale structural dependency and induced hierarchical heterogeneity present in the data, we employ CEDA and a modified version of the major factor selection protocol originally proposed in [17]. The original protocol was designed to effectively operate in globally homogeneous settings where covariate features are relatively independent. However, such settings are rare or even unrealistic in real-world scenarios.

In contrast, most big databases derived from real complex systems exhibit uneven associative relational patterns among covariate features. This implies the existence of groups or communities within the feature set, where members within the same group are highly associated due to structural dependency, while members from different groups have weaker associations. The presence of these feature groups poses computational challenges, even in homogeneous settings, and these difficulties are further magnified when heterogeneity is present. Therefore, it is crucial to develop remedial measures to mitigate the potential impacts of structural dependency.

Among these challenges, the primary one is how to detect and confirm high-order interacting effects involving multiple features. This difficulty arises because conditional dependence relations between members of the same group and the response variable become entangled. In this paper, we explicitly demonstrate this complexity through extensively convoluted Information Theoretical Measurements. Subsequently, we computationally enhance the realism and practicality of the CEDA paradigm and its major factor selection protocol to achieve two interconnected technical objectives simultaneously: (1) identifying a suitable perspective of heterogeneity, and (2) untangling the intricate structural dependency. In this context, heterogeneity is considered to be potentially induced by the structural dependency among low-order major factors. Thus, the range of perspectives on global-scale heterogeneity serves as a manifestation of the key aspects of the dynamics of the complex system under study.

To accomplish the aforementioned technical goals, we utilize Information Theoretical Measurements, specifically conditional entropy and mutual information. However, we focus on the locality scale rather than just the global or marginal scale, as in the original version of the CEDA paradigm. In the following, we provide a brief review of conditional entropy and mutual information, while also illustrating the fact that if highly associated covariate features are assumed to be independent at the global (or marginal) scale, their true effects in predicting the response variable are likely overlooked or misunderstood.

Let us denote the categorized or categorical response variable as $\mathcal{Y}$, which may initially involve multiple features of various data types: continuous, discrete, categorical, or a mixture thereof. It is worth noting that a continuous feature can be coherently categorized based on its own histograms [20], and multiple categorical or categorized features can always be combined into a single variable through a collection of occupied multidimensional hypercubes. This advantage highlights the usefulness of working with the categorical nature of data. Similarly, any covariate feature sets, regardless of size or data type, can be fused into a categorical variable. Therefore, we will employ capital letters, such as A or B, to represent different subsets of categorized or categorical covariate features, while simultaneously denoting different categorical variables obtained by combining various subsets of categorical or categorized covariate features accordingly.

Furthermore, we explicitly construct a contingency table for each pair of categorical variables, such as $(\mathcal{Y}, A)$, $(\mathcal{Y}, B)$, (A, B), and $(\mathcal{Y}, (A, B))$. For example, we denote the contingency table for $(\mathcal{Y}, A)$ as $C[A - vs - \mathcal{Y}]$, where the categories of A and $\mathcal{Y}$ are arranged along the row and column axes, respectively. The same convention is followed for all other pairs. In this paper, we typically arrange the categories of $\mathcal{Y}$ on the column axis. Along the row axis, each row of $C[\mathcal{Y} - vs - A]$, denoted as $A = a$, approximates or defines a conditional multinomial random variable with a conditional (Shannon) entropy (CE) represented by $H[\mathcal{Y}|A = a]$. By calculating the averaged CE $H[\mathcal{Y}|A]$, which is a properly weighted sum of the collection of row-wise CEs $\{H[\mathcal{Y}|A = a]\}$, we obtain the overall conditional entropy. The marginal column-wise and row-wise entropies are denoted as $H[\mathcal{Y}]$ and $H[A]$, respectively.

It is known that $H[\mathcal{Y}|A]$ conveys the expected amount of remaining uncertainty in $\mathcal{Y}$ after knowing A. In reverse, by knowing $\mathcal{Y}$, $H[A|\mathcal{Y}]$ conveys the expected amount of remaining uncertainty in A after seeing $\mathcal{Y}$. The two conditional entropy drops, i.e. differences $H[Y] - H[\mathcal{Y}|A]$ and $H[A] - H[A|\mathcal{Y}]$, indicate the shared amount information between A and $\mathcal{Y}$:

$$\begin{aligned} H[\mathcal{Y}] - H[\mathcal{Y}|A] &= H[A] - H[A|\mathcal{Y}] \\ &= H[A] + H[Y] - H[A, \mathcal{Y}] \\ &= I[\mathcal{Y}; A]. \end{aligned}$$

where $I[\mathcal{Y}; A]$ denotes the mutual information between $\mathcal{Y}$ and A.

Next, we consider the mutual information between the bivariate (A, B) and $\mathcal{Y}$ starting from their conditional mutual information as:

$$I[A; B|\mathcal{Y}] = H[A|\mathcal{Y}] + H[B|\mathcal{Y}] - H[(A, B)|\mathcal{Y}].$$

Further, we decompose this mutual information into the following two key components: (1) the sum of individual CE-drops of A and B and (2) the difference between the conditional and marginal mutual information of A and B:

$$\begin{aligned} H[\mathcal{Y}] - H[\mathcal{Y}|(A, B)] &= H[(A, B)] - H[(A, B)|\mathcal{Y}]; \\ &= H[A] + H[B] - I[A; B] - \{H[A|\mathcal{Y}] + H[B|\mathcal{Y}] - I[A; B|\mathcal{Y}]\}; \\ &= \{H[\mathcal{Y}] - H[\mathcal{Y}|A] + H[\mathcal{Y}] - H[\mathcal{Y}|B]\} + \{I[A; B|\mathcal{Y}] - I[A; B]\}. \end{aligned}$$

The above decomposition precisely conveys the essence of the interpretable meaning of conditional mutual information when the two involving feature sets A and B are indeed marginally independent because $I[A;B]=0$. And if $I[A;B|\mathcal{Y}]$ is relatively large, then we are certain that A and B have a significant interacting effect in reducing the uncertainty of $\mathcal{Y}$. However, if A and B are indeed highly associated, then $I[A;B]>0$, then the last term of the above equation: $\{I[A;B|\mathcal{Y}]-I[A;B]\}$, can be negative. We then face two difficulties: (1) it becomes computationally costly, if not hard, to determine whether the smaller CE-drop by including either A or B is significant or not; (2) it is equally costly to assess whether the A and B have a significant interacting effect or not even when $\{I[A;B|\mathcal{Y}]-I[A;B]\}$ is positive. The first difficulty is about whether A or B are simultaneously present as two stand-alone factors within the dynamics of $\mathcal{Y}$, while the second one is about whether A or B together give rise to an interacting effect. We make simulated examples to explicitly demonstrate such difficulties in the next section below. Also, a feature-pair A and B is termed to achieve the ecological effect if $\{I[A;B|\mathcal{Y}]-I[A;B]\}$ is positive. This notion of "the whole is larger than the sum of its parts" is used heavily in [17,18], so is here.

The above derivations are performed on the global scale, assuming the absence of heterogeneity. However, when heterogeneity is present, as is the case in most real-world scenarios, these difficulties are further amplified. Global evaluations of features' effects are not useful because they are simply weighted averages that pertain to a collection of involved localities. These global measurements do not directly provide meaningful information at the locality scale. Therefore, it is necessary to identify localities where the impact of structural dependency is significantly reduced. However, on the locality scale, new challenges may arise due to smaller sample sizes. This introduces reliability issues that need to be addressed.

In order to address heterogeneity in the data, it is crucial to not only identify relevant perspectives of heterogeneity at the global scale but also develop techniques to ensure that the degree of structural dependency is greatly reduced at the locality scale. These challenges related to structural dependency and heterogeneity are commonly addressed through a process known as "de-associating", which will be discussed in the next subsection.

2.2. Unique operations on contingency table platform: shadowing and de-associating

Within CEDA, the contingency table serves as the central computing platform, enabling the evaluation of conditional entropies and, more importantly, revealing potential and visible associations between the two categorical variables on the row and column axes. In this subsection, we discuss two fundamental and crucial operations on any contingency table: "shadowing" and "de-associating". The "shadowing" operation involves combining a targeted variable's associative random component with the original variable's remaining independent component to create a new variable that retains the same distribution. On the other hand, the purpose of the "de-associating" operation is to decompose and remove the original variable's own random components that are associated with a targeted variable. These two operations, which are unique to contingency tables, have not received much research attention in the literature, and as a result, their merits may be relatively unfamiliar to scientists.

Let us denote two categorical features as A and B, with their contingency table represented as $C[A-vs-B]$, where categories of A and B are arranged on the row and column axes, respectively. The "shadowing" operation performed on $C[A-vs-B]$ is directional, specifically with B being shadowed by A. This operation aims to create a new categorical variable, denoted as $B^*[A]$, which satisfies the following three properties: (1) $B^*[A]$ and B have the same marginal distribution; (2) A and $B^*[A]$ retains the same association as A and B; (3) but $B^*[A]$ is a composition of the projected part of A onto B, along with an independent replicated part that has the same distribution as the part of B that is independent of A. The indirect utility of $B^*[A]$ is that it allows us to examine the information of A contained within B.

The categorical variable $B^*[A]$ is constructed according the following steps:

[De :] Let $\{a_1,..,a_{Ro}\}$ be the entire collection of categories of A. Divide the entire data set into R sub-collections: $D[A=a_r]=\{(a_r,b_{rj})|j=1,\ldots,n_r\}$ with n_r the rth row sum and b_{rj} being an observed category of B.

[MN :] Let $\{b_1,..,b_{Co}\}$ be the entire collection of categories of B. Build a Co-dim multinomial random variable $MN(n_r,\hat{P}[r])$ with $\hat{P}[r]$ the proportion vector from the rth row of $C[A-vs-B]$. Simulate n_r data points based on $MN(n_r,\hat{P}[r])$ and denoted them as $\{(b^*_{rj})|j=1,\ldots,n_r\}$.

[SD :] Create a new rth sub-collection $D^*[A=a_r]=\{(a_r,b^*_{rj})|j=1,\ldots,n_r\}$. Repeat the [De] and [MN] steps for all Ro row. Then the new variable built into the data set $\{D^*[A=a_r]|r=1,\ldots,Ro\}$ is called $B^*[A]$.

The [SD] step of this operation on the contingency table ensures the appropriate simulation of the conditional random variable of B given $A=a_r$. By incorporating the row-sum proportion, the randomness of B given A can be effectively visualized in a randomized manner. Outside the contingency table framework, visualizing the randomness of a conditional random variable is generally challenging. While we do not explicitly employ this operation in this paper, it is implicitly utilized through the [MN] step.

Contrasting with the shadowing operation on the same contingency table $C[A-vs-B]$ platform, the de-associating operation is to explicitly extract the part of B being independent of A. On the $C[A-vs-B]$, as in stated in the above [MN] step, the row-wise Multinomial random variable $MN(n_r,\hat{P}[r])$ is independent of A within the sub-collection $D[A=a_r]$. Thus, the collective of these row-wise Multinomial random variables $\{MN(n_r,\hat{P}[r])|r=1,\ldots,Ro\}$ is marginally independent of A. This is surprising simple de-associating operation is specifically denoted $B^{\perp}[A]$. It is worth emphasizing that this computational simplicity is purely attributed to the categorical nature of the association of (A,B) exhibited in $C[A-vs-B]$. Through the idea of the de-associating operation is

just the conditioning in Probability Theory, its operational simplicity together importance makes it deserve a name within CEDA. It is known that the concept of conditioning in general are neither explicit nor easily operable.

The essential merit of de-associating is seen as follows. When all features, including the possibly multiple dimensional response $\mathcal{Y}$, are commonly made to be de-associating with respect to A, then the entire data set is divided with respect to each category of A into sub-collections, respectively. Upon each sub-collection data, in which A is fixed at a constant category, we can evaluate and reveal information provided by all other features beyond their dependence with A. That is, the new response variable $\mathcal{Y}^{\perp}[A]$ retains only information or uncertainty beyond A, while all new covariate features $\{V_k^{\perp}[A]|k = 1, \ldots, K\}$ become less or much less associative with each other within localities defined by categories of A. This is the setting where all covariate features can demonstrate their pieces of information beyond (independent of) what A can provide. This recognition of its functionality of $B^{\perp}[A]$ upon the contingency table platform is especially important and useful for us to access individual features' effects as well as their joint interacting effects among highly associated features. In summary, the de-associating operation is the tool we use to the two aforementioned technical issues simultaneously: (1) finding a proper perspective of heterogeneity; (2) untangling the convoluted structural dependency. We explicitly demonstrate such functionality of $B^{\perp}[A]$ in both real complex systems of pitching dynamics and heart disease.

Our chief contributions from a computational perspective are achieved by making effective use of de-associating operations to discover where are vital perspectives of hierarchical heterogeneity and what kinds of structural dependency are behind these perspectives on the global and locality scales. When entering the locality scale induced by de-associating operations, the major factor selection protocol proposed in [17,18] becomes effective and precise because all covariate features become much less associative. Via this computing paradigm, CEDA is expected to efficiently extract authentic information content pertaining to a very wide range of complex systems' databases. Consequently, our scientific contribution is summarized as that such computed and recognized perspectives of hierarchical heterogeneity indeed constitute the broken symmetry embedded within complex system dynamics under study [7].

3. Modified major factor selection with motivating examples

In this section, we illustrate our major factor selection protocol and "shadowing" and "de-associating" operations upon contingency table platform under various cases of designed structural dependency, but without heterogeneity. Two linearity based model-examples are used here. Through the first example, we demonstrate that "Shadowing" can provide analytic views of information contributed by any major factors or feature sets. The second example is a slightly modified version of the first one with a specific aim of demonstrating how "de-associating" can computationally confirm whether any major factor candidates could or could not provide information beyond an already confirmed major factors.

The organization of this computational development section is given as follows. In the first subsection, we lay out the two model examples, upon which we also report some computational instability pertaining to LASSO. These settings and results here are meant to be for contrasting purposes only, but not intended to contest the consistency results of LASSO in statistics and machine learning literature [21–23]. In the second subsection, under the first example, we report CE and CE-drop (conditional mutual information) and identify potential major factors, and then "shadowing" and "de-associating" operations are carried out in order for confirming a collection of major factors. In the third subsection, under the second example, we identify a candidate collection of major factors, and then the "de-associating" operation is carried out to confirm that there exist no other major factors being able to offer information beyond this collection of major factors. In the last subsection, we explicitly lay out our modified version of major factor selection protocol with some operational remarks.

3.1. Two illustrating examples with highly associated covariate features

Our first illustrative example, termed as Example-1, is specified by the following linear Re-Co dynamics constituted by 7 highly associated 1D features (or variables) given as follows:

$$\begin{aligned} Y &= 0.8X_1 + X_2 + 1.2X_3 + X_{11} + \epsilon, \\ X_7 &= (X_1 + X_2 + X_3 + X_4 + X_{11})/3.66, \\ X_i &\sim N(0,1),\ \rho(X_i, X_j) = 0.7,\ i, j = 1, \ldots, 6; \\ X_k &\sim N(0,1), i.i.d,\ k = 8, \ldots, 11, \\ X_{11} &: \text{ is the only unobserved hidden variable.} \end{aligned}$$

Due to the presence of unobserved hidden variable X_{11}, it is reasonable to say at least intuitively that $\{X_7, X_4\}$ is the primary order-2 major factor of the Re-Co dynamics. Since $Y = X_7 - X_4 + 0.2(X_3 - X_1) + \epsilon$. It is also intuitive that extra and delicate evidence is certainly needed in order to expand $\{X_7, X_4\}$ into a collection of major factors $\{\{X_7, X_4\}, X_3, X_1\}$. In this subsection, we perform the classic LASSO approach under this linear Re-Co dynamics to prepare for comparisons with what our major factor selection can do in the next two subsections.

Based on a simulated data set with a sample size of 100K, LASSO approach is conducted and results of parameter estimations with respect to L-1 penalty λ are reported in Fig. 1. It is evident that the least square estimation (with $\lambda = 0$) is able to provide the exact and precise model structure. But, as λ moving away from zero, though the presence of X_7 is persistent, its estimated values

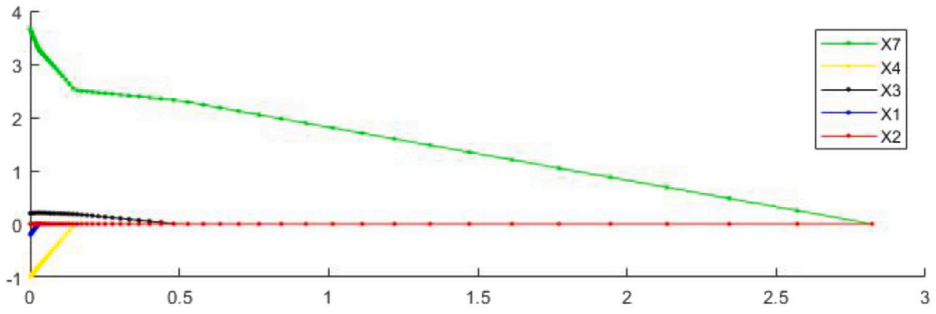

Fig. 1. Results of LASSO estimations with respect to L-1 penalty λ on X-axis.

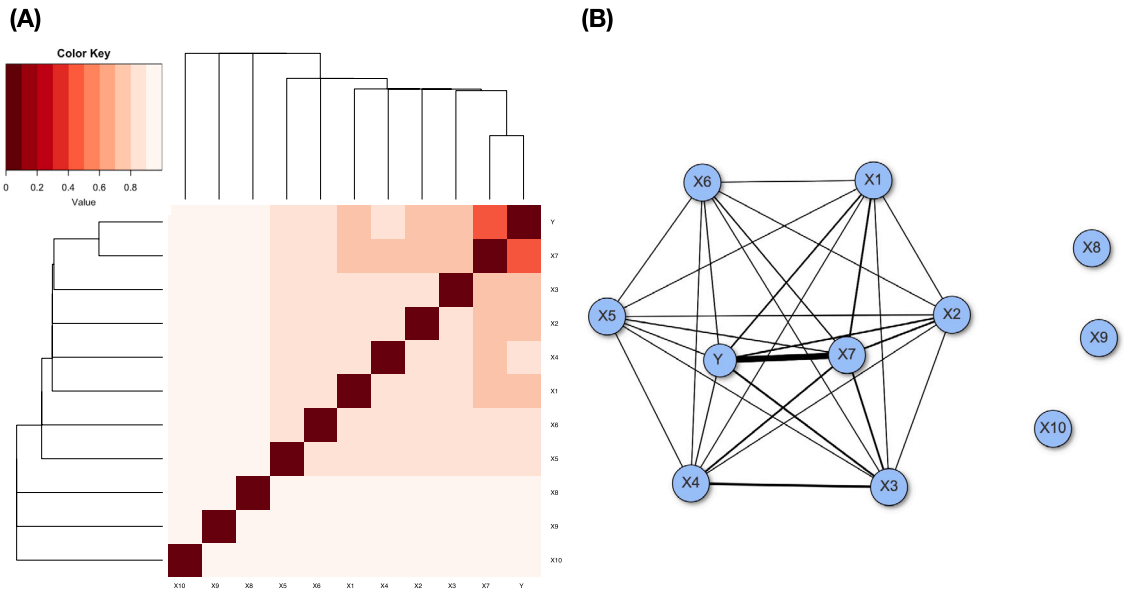

Fig. 2. MCE heatmap and network of 12 features involved in the illustrative example.

are shrinking. Overall LASSO results show diminishing importance of X_4, X_3, and X_1. In summary, even when λ is only slightly positive, all resultant structures are rather off the true structure.

It is important to highlight that the aforementioned results heavily depend on prior knowledge of linear structure. However, even with this prior knowledge, severely biased results are still obtained. Hence, it becomes crucial to explore what can be discovered when such knowledge is not available. In the upcoming subsections, we will develop and present comprehensive solutions to address this fundamental question and subsequently address the following raised questions.

It is typical that the first batch of pieces of information in a structured data set is revealed through its associative heatmap, graph or network among these 12 variables as seen in Fig. 2. This heatmap and its corresponding network apparently contain a complete clique among variables in $\{Y, X_1, \ldots, X_7\}$ due to their strong pairwise correlations, and a triplet of isolated nodes of X_8, X_9 and X_{10}. Such pieces of information, such as communities [24], are beneficial to know. The complete clique indicates that the Re-Co dynamics likely involves with hard to untangled relations. Though X_7 evidently plays a dominant role in the Re-Co dynamics, while X_4 plays a "negative" role within the structure of X_7 and is not directly involved in Y. Can we detect this fact?

Further, while X_1, X_2, and X_3 have slightly increasing coefficients in the linear structure, their roles are not increasingly important. Since X_2 is "covered" by X_7 in the sense that X_2 is not important in the presence of X_7. Further, though X_1 and X_3 are "equal" in the linear structure, their roles of reducing the uncertainty of Y might not be equal. Furthermore, these three features are highly associated. Can we differentiate their intricate differences? On the other hand, X_5 and X_6 play no roles in the Re-Co dynamics, but they are highly correlated (under normality) with $\{X_1, X_2, X_3, X_4, X_7\}$. Can we tease out their roles from the roles played by the 5 directly and 5 indirectly involving covariate features?

Finally, it is clear that the isolated feature or variable nodes $\{X_8, X_9, X_{10}\}$ likely do not play any relational roles in the Re-Co Dynamics. But in real-world data, we need to take into account the potentials that such features could join other features to form essential high-order interacting roles. Therefore, we need to make sure whether such interacting relations exist in data or not. Though they play no roles in the Re-Co dynamics, these three features indeed computationally serve as baselines for our Shannon

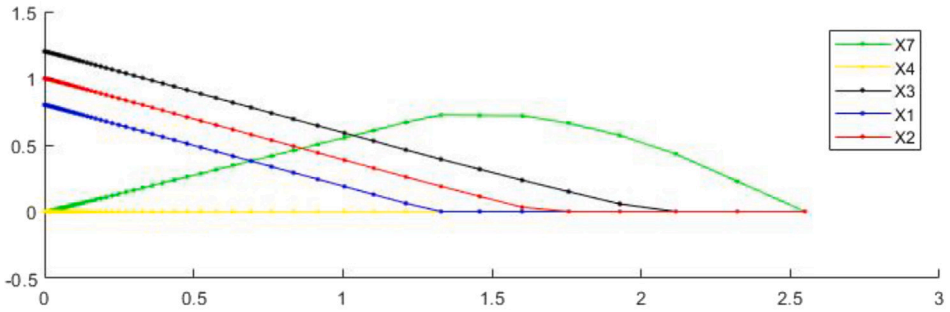

Fig. 3. Results of LASSO estimations of modified example with respect to L-1 penalty λ on X-axis.

entropy evaluations when attempting to keep effects of the finite sample phenomenon or so-called curse of dimensionality at bay, see [19] for practical guidelines on CE and mutual information evaluations.

Next, our second illustrative example is a slightly modified version of the first example. Though this simpler version would reiterate the cause of instability of LASSO within settings that contain two or more "wells" of locally optimal solutions in the optimizing landscape, it is primarily designed to showcase that no information regarding Y could be found beyond a chief collection of major factors.

In this example, termed as Example-2, we take off the hidden factor X_{11} in Y as follow:

$$\begin{aligned} Y &= 0.8X_1 + X_2 + 1.2X_3 + \epsilon, \\ X_7 &= (X_1 + X_2 + X_3 + X_4 + X_{11})/3.66, \\ X_i &\sim N(0,1),\ \rho(X_i, X_j) = 0.7,\ i,j = 1,\ldots,6; \\ X_k &\sim N(0,1), i.i.d,\ k = 8,\ldots,11, \\ X_{11} &: \text{ is an unobserved hidden variable.} \end{aligned}$$

In this second example, the feature-pair $\{X_4, X_7\}$ apparently becomes a less effective order-2 major factor candidate than the collection of three order-1 major factors: $\{X_1, X_2, X_3\}$. This essential fact would be reported based on CEs computations in the subsections below. As a by-product, again the LASSO results with respect to a range of penalty λ are unstable and biased as presented in Fig. 3.

Fig. 3 delivers an even clearer message than that delivered by Fig. 1: "the penalty is not intrinsic in these linear regression settings". The collection of major factors: $\{X_1, X_2, X_3\}$, is clearly needed to be held constant at least for a small range of λ for all practical and realistic reasons. But this pattern does not happen. The implication of such results is simply: LASSO is hardly a realistic tool for extracting information even under man-made linearity when its optimizing landscape contains multiple local minima, which is realistically likely to be the case in the majority of real-world complex systems.

3.2. Major factor selection protocol for Example-1

In our previous works [17,18], we introduced a major factor selection protocol that aimed to discover and identify major factors of varying orders. This protocol relied on two criteria: [C1:confirmable] and [C2:irreplabeable], which would be explicitly listed in the last subsection of this section. These criteria operated under the assumption that the evaluations or estimations of conditional entropies (CE) remained consistently stable across various feature settings.

However, in reality, the CE values may only experience slight variations as the dimensions of contingency tables expand. When the dimensions of contingency tables become too large relative to the sample size, the effects of the finite sample phenomenon or the curse of dimensionality come into play. In such cases, all feature sets tend to behave like random noise variables. To mitigate the impact of the finite sample phenomenon, our major factor selection protocol includes a key characteristic: according to [C1:confirmable], a candidate major factor must exhibit a significantly larger CE-drop (the difference between entropy of the response variable $\mathcal{Y}$ and the conditional entropy of the major factor candidate given $\mathcal{Y}$) than that of a random noise variable under the same dimension of the contingency table.

The above requirement is distinct from the one that requires consistent CE estimates. A detailed and rigorous discussion on the related issues of how to obtain reliable CE-drop evaluations is presented in [19]. The results presented in this paper sharply contrast with the results reported in [25].

As such, a potential major factor candidate's CE-drop must be significantly larger than a random noise variable's under the same dimension of a contingency table with respect to the criterion [C1:confirmable]. We show such implementations of criterion [C1:confirmable] in Table 1 that report calculated Conditional Entropy(CE) of feature-sets ($CE[Y|\{X_i,\ldots,X_j\}]$) across five settings: 1-feature to 5-feature settings. Across these five settings, the dimensions of contingency tables grow with the number of members in $\{X_i,\ldots,X_j\}$. Even though, theoretically $CE[Y] = CE[Y|\{X_i,\ldots,X_j\}]$ when Y is independent of $\{X_i,\ldots,X_j\}$ disregarding the number

Table 1
Example-1 with $N = 10^5$. Each categorized 1-features has 12 bins, so a k-feature setting has $(12)^k$ kD hypercubes.

1Feature	CE	2Feature	CE	3Feature	CE	4Feature	CE
X7	1.0498	X4_X7	0.7648	X3_X4_X7	0.7048	X3_X4_X7_X10	0.6403
X3	1.7927	X3_X7	1.0152	X1_X4_X7	0.7487	X4_X7_X8_X10	0.6524
X2	1.8509	X2_X7	1.0419	X2_X3_X7	0.9948	X2_X3_X4_X7	0.6532
X1	1.8988	X1_X7	1.0459	X7_X8_X9	1.0187	X4_X5_X6_X7	0.6937
X6	2.0473	X2_X3	1.5503	X2_X7_X9	1.0212	X7_X8_X9_X10	0.7934
X4	2.0478	X1_X3	1.5984	X1_X2_X3	1.4122	X5_X7_X8_X9	0.8492
X5	2.0497	X1_X2	1.6505	X1_X3_X4	1.5459	X1_X2_X3_X7	0.9219
X9	2.4089	X3_X9	1.7864	X3_X4_X5	1.6514	X3_X8_X9_X10	1.2531
X8	2.4090	X4_X5	1.9326	X4_X5_X6	1.8291	X3_X4_X5_X6	1.4262
X10	2.4091	X8_X9	2.4014	X8_X9_X10	2.3072	X4_X5_X6_X8	1.5000

of members in $\{X_i, \ldots, X_j\}$, the empirical fact, as shown in Table 1, is that the numerical values of $CE[Y|\{X_i, \ldots, X_j\}]$ would slightly decrease as the number of members of $\{X_i, \ldots, X_j\}$ increasing. Nonetheless, the decreasing amounts is relatively small if the average cell count in the corresponding contingency table remains 10 or more [19]. Thus, we compare CEs within the same feature setting, not cross over different feature settings.

With the above notes in mind, we summarize computed patterns in Table 1 as follows. Starting from the 1-feature setting, one estimated entropy of Y is accordingly calculated as $CE[Y|X_{10}] = 2.4091$ to reflect the dimensionality of the contingency table and stochastic independence of Y and X_{10}. The feature X_7 achieves the lowest CE: 1.0498, so its CE-drop is calculated as $1.3593 = 2.4091 - 1.0498$, which is an estimate of mutual information $\hat{I}^{(1)}[Y, X_7]$. The superscript indicates that the estimated mutual conditional information is calculated within the 1-feature setting. Variables X_3, X_2 and X_1 achieve the top 2nd, 3rd and 4th ranked CE-drops that are significantly larger than the CE-drop of X_4, which is calculated as $\hat{I}^{(1)}[Y, X_4] = 2.4091 - 2.0478 = 0.3613$.

On the 2-feature setting, one estimated entropy of Y is calculated as $CE[Y|X_8, X_9] = 2.4014$. It is known that $X_7 - X_4$ is very close to the functional structure of Y. Since both embrace almost the same linear structure and the common hidden factor X_{11}. Thus, it is as expected that the feature-pair $\{X_4, X_7\}$ achieves the lowest CE with estimated mutual information (CE-drop):

$$\hat{I}^{(2)}[Y, \{X_4, X_7\}] = 2.4014 - 0.7648 = 1.6366.$$

Nonetheless, this CE-drop of $\{X_4, X_7\}$ is smaller than the sum of individual CE-drops of X_7 and X_4, which is equal to 1.7206. This fact indicates that the conditional mutual information of $\{X_4, X_7\}$ given Y is less than the marginal mutual information of $\{X_4, X_7\}$, that is,

$$\hat{I}^{(2)}[\{X_4, X_7\}|Y] - \hat{I}^{(2)}[\{X_4, X_7\}] = -0.0840 < 0.$$

In contrast, in this particular case, the presence of stochastic high dependence between X_7 and X_4 prevents the observation of the ecological effect, which is a key factor in the second criterion [C2:irreplabeable] as discussed in [17,18] under the assumption of independence. However, despite the lack of ecological effect, the pair $\{(X_4, X_7)\}$ remains an evident order-2 major factor in the Re-Co dynamics of Y in this illustrative example. Similarly, the feature-pairs $\{X_1, X_2\}$, $\{X_1, X_3\}$, and $\{X_2, X_3\}$ do not exhibit ecological effects, even though they are part of the linear structure of Y.

Furthermore, the issues surrounding the feature-pairs $\{X_1, X_7\}$ and $\{X_3, X_7\}$ are more intricate. Despite their expected contribution to the linear structure of Y, they do not show significant improvements when compared to X_7 alone. Similarly, in the case of the three-feature setting, the triplets $\{X_1, X_4, X_7\}$ and $\{X_3, X_4, X_7\}$ exhibit limited improvements compared to $\{X_4, X_7\}$. Thus, the subtle but observable effects of including X_1 or X_3 to expand the collection of major factor $\{(X_4, X_7)\}$ are not apparent in this scenario.

In conclusion, the issues discussed above are primarily rooted in the structural dependency among covariate features. Therefore, it is crucial to gain a deeper understanding to enhance the effectiveness of categorical exploratory data analysis (CEDA) in achieving its goal of extracting relevant information from data through major factor selection, thereby providing valuable insights into complex systems under study. In the upcoming sub-subsections, we will present two distinct operational perspectives related to the contingency table, demonstrating the potential of our modified major factor selection protocol to be highly adaptable to diverse complex systems while addressing the aforementioned issues. Through the concepts of "shadowing" and "de-associating" operations, we will illustrate how we can visualize and evaluate subtle effects that may otherwise be overshadowed by strong dependencies among covariate features. These insights will be further elaborated in the subsequent subsections.

3.2.1. The operation of "Shadowing"

In this sub-subsection, we implement the "shadowing" operation on variables used in the Example-1. Consider $A = X_7$ and $B = Y$ under the categorized setting. Likewise, we generate the data of $Y^*[X_7]$. We then use $Y^*[X_7]$ as the response variable with respect to the same collection of covariate features: $\{X_1, \ldots, X_10\}$. The CEs (and CE-drops) are calculated across 1-feature to 3-feature settings in Table 2.

In the 1-feature setting in Table 2, the CE of X_7 almost retains its CE when the response is Y. It is strikingly evident that, with $Y^*[X_7]$ as the response, the CEs of 10 covariate features X_k with $k = 1, .., 10$ indeed reflect exactly the linear structure of X_7. This phenomenon is indeed for the namesake. Further, we see that no other features or feature sets can be coupled with X_7 to achieve

Table 2
Example-1 with $N = 10^5$ and response $Y^*[X_7]$. Each categorized 1-features has 12 bins, so a k-feature setting has $(12)^k$ kD hypercubes.

1Feature	CE	2Feature	CE	3Feature	CE
X7	1.0523	X1_X7	1.0500	X7_X8_X9	1.0207
X3	1.8838	X2_X7	1.0502	X1_X4_X7	1.0356
X2	1.8846	X4_X7	1.0503	X2_X3_X7	1.0356
X1	1.8871	X3_X7	1.0504	X1_X2_X3	1.5115
X4	1.8859	X2_X3	1.6692	X2_X3_X4	1.5119
X6	2.0339	X1_X2	1.6708	X1_X2_X5	1.6063
X5	2.0359	X1_X3	1.6726	X4_X5_X6	1.7052
X8	2.4090	X4_X5	1.7921	X2_X6_X8	1.7371
X10	2.4090	X3_X9	1.8775	X3_X8_X9	1.8113
X9	2.4091	X8_X9	2.4014	X8_X9_X10	2.3066

Table 3
Example-1 with $N = 10^5$ and CEs when $X_7 = 8$ and weighted CEs when X_7 ranging the 12 categories. The response variable is $Y^{\perp}[X_7]$ and covariate features $\{X^{\perp}[X_7]_k | k \neq 7\}$.

1Feature	CE$[X_7 = 8]$	CE$[X_7]$	2Feature	CE$[X_7 = 8]$	CE$[X_7]$
X4	0.7972	0.7648	X3_X4	0.7350	0.7048
X3	1.0827	1.0152	X1_X4	0.7807	0.7487
X5	1.1107	1.0448	X2_X4	0.7834	0.7486
X2	1.1117	1.0419	X4_X6	0.7850	0.7521
X6	1.1125	1.0445	X4_X8	0.7856	0.7527
X1	1.1136	1.0459	X3_X6	1.0596	0.9924
X8	1.1146	1.0473	X3_X8	1.0631	0.9951
X10	1.1147	1.0470	X2_X3	1.0653	0.9948
X9	1.1151	1.0472	X1_X8	1.0938	1.0253
X7	1.1170	1.0498	X7_X9	1.1151	1.0472

significantly improved CEs in the 2-feature and 3-feature settings. In particular, the feature pair $\{X_4, X_7\}$ achieves a much higher CE (=1.0503) than its CE (=0.7648) when the response variable is Y. This fact is strikingly different from the case of having the response variable Y reported in Table 1.

It is also noted that, while CEs of random noise variables X_8, X_9 and X_{10} remain the same, the feature-triplet $\{X_1, X_2, X_3\}$ achieves a higher CE (=1.5115) than its original CE (=1.4122) with Y as the response variable. This CE difference indeed confirms that the triplet $\{X_1, X_2, X_3\}$ can offer extra information beyond X_7 on Y because something is lost after the shadowing via X_7. By putting these facts together, we conclude that some information about Y has gone missing in $Y^*[X_7]$. In other words, we project that the variable $Y^{\perp}[X_7]$ still shares a significant amount of information with all covariate features X_k with $k \neq 7$ as would become more explicit in the next subsection of "de-associating".

3.2.2. The operation of de-associating

Here we illustrate this de-associating concept again on the Example-1. We construct $Y^{\perp}[X_7]$ and all members of $\{X^{\perp}[X_7]_k | k \neq 7\}$ individually and simultaneously lay out their maintained associative relations by conditioning on X_7. To perform such constructions, we only need to divide the entire data set into X_7-specific subsets, denoted $\mathcal{M}[X_7 = j]$ with $J = 1, .., 12$. Very importantly, the $\mathcal{M}[X_7 = j]$ retains the associative relations between $Y^{\perp}[X_7]$ and all members of $\{X^{\perp}[X_7]_k | k \neq 7\}$. It is essential to note also that these variables become much less associated with each other. Thus, we can evaluate the effects of all members of $\{X^{\perp}[X_7]_k | k \neq 7\}$ on reducing the uncertainty of $Y^{\perp}[X_7]$ beyond the effect of X_7. Since such effects are relatively similar in pattern with respect to different values of X_7, we just report such effects in one table with respect to one categorical value of X_7.

On the Table 3, we can evidently see in the 1-feature setting that $X^{\perp}[X_7]_4$ achieves the lowest CE conditioning on $X_7 = 8$. Since this evident pattern occurs across all different values of X_7. We can conclude that X_4 indeed brings extra information beyond X_7. That is, $\{X_4, X_7\}$ is confirmed as an order-2 major factor of Re-Co dynamics of Y.

As for the feature X_3, we observed that the CE of $X^{\perp}[X_7]_3$ is also visible comparing with CEs among $\{X^{\perp}[X_7]_5, X^{\perp}[X_7]_6, X^{\perp}[X_7]_8, X^{\perp}[X_7]_9, X^{\perp}[X_7]_{10}\}$, which are basically noise features after conditioning on $X_7 = 8$. This observed pattern is also persistent across all different values of X_7 because of having no presence of heterogeneity. Further, since the two feature $\{X^{\perp}[X_7]_3, X^{\perp}[X_7]_4\}$ are no longer highly associated, and this feature-pair $\{X^{\perp}[X_7]_3, X^{\perp}[X_7]_4\}$ achieves a CE in 2-feature setting in Table 3 that satisfies the ecological effect. Thus, X_3 is confirmed as an order-1 major factor.

As for features X_1 and X_2, their CEs in the 1-feature setting are rather close to those perceived noise features. But the CEs of feature-pairs $\{X_1^{\perp}[X_7], X_4^{\perp}[X_7]\}$ and $\{X_2^{\perp}[X_7], X_4^{\perp}[X_7]\}$ seem to indicate slight effects of these two features, but not evident enough to achieve the ecological effect. That is, it is not confident to conclude whether X_1 or X_2 will bring extra effects on top of that of $\{X_4, X_7\}$ like X_3 does.

To further confirm whether X_1, X_2 or X_3 indeed have extra effects beyond $\{X_4, X_7\}$, we perform likewise de-associating operation (conditioning) based on bivariate values of $\{X_4, X_7\}$. That is, we look into associative relations between $Y^{\perp}[X_4, X_7]$ and all members of $\{X_k^{\perp}[X_4, X_7] | k \neq 4, 7\}$. It is evident from Table 4, we see X_3 indeed providing extra information beyond what $\{X_4, X_7\}$ can offer

Table 4
Experiment-20220304 with $N = 10^5$ and $(X_4, X_7) = (8, 8)$ with response $Y^{\perp}[X_4, X_7]$ and covariate features $\{X^{\perp}[X_4, X_7]_k | k \neq 4, 7\}$ at 2nd and 5th column. At 3rd column, weighted CEs of features when conditioning on $\{X_4, X_7\}$ across $(12)^2$ 2D categories.

1Feature	$CE^{\perp}[X_4, X_7]$	Weighted $CE^{\perp}[X_4, X_7]$	2Feature	$CE^{\perp}[X_4, X_7]$
X3	0.7441	0.7048	X3_X10	0.7013
X2	0.7978	0.7486	X1_X3	0.7131
X1	0.7931	0.7487	X2_X3	0.7170
X5	0.7978	0.7524	X3_X6	0.7204
X8	0.7982	0.7527	X4_X8	0.7856
X6	0.7992	0.7521	X8_X9	0.7474
X9	0.7993	0.7536	X1_X8	0.7586
X10	0.8002	0.7528	X2_X8	0.7605
X4	0.8032	0.7648	X1_X2	0.7668
X7	0.8032	0.7648	X4_X7	0.8032

Table 5
Experiment-20220408 with $N = 10^5$. Each categorized 1-features has 12 bins, so a k-feature setting has $(12)^k$ kD hypercubes.

1Feature	CE	2Feature	CE	3Feature	CE
X7	1.4205	X2_X3	1.0300	X1_X2_X3	0.5034
X3	1.5841	X1_X3	1.1789	X2_X3_X7	0.8849
X2	1.6807	X3_X7	1.1842	X2_X3_X6	0.9757
X1	1.7576	X2_X7	1.2725	X2_X3_X4	0.9797
X4	1.9700	X1_X2	1.3083	X2_X3_X9	1.0048
X6	1.9705	X1_X7	1.3320	X1_X3_X7	1.0125
X5	1.9729	X6_X7	1.3773	X1_X3_X4	1.1177
X10	2.4103	X4_X7	1.4154	X1_X3_X6	1.1188
X8	2.4104	X4_X5	1.8132	X1_X2_X7	1.1283
X9	2.4104	X8_X9	2.4030	X8_X9_X10	2.3080

on Y. In contrast, it seems that we can confirm to some degrees that X_1 seemingly has a slight amount of effect going beyond that of $\{X_4, X_7\}$ on Y as well. In conclusion, we have the authentic collection of major factors: $\{\{X_4, X_7\}, X_3, X_1\}$, that is, one order-2 and two order-1 major factors for the Re-Co dynamics of Y.

3.3. Major factor selection on the Example-2

First, we recall the structural linearity in the second illustrating example as follows:

$$\begin{aligned} Y &= 0.8X_1 + X_2 + 1.2X_3 + \epsilon, \\ X_7 &= (X_1 + X_2 + X_3 + X_4 + X_{11})/3.66, \\ X_i &\sim N(0,1),\ \rho(X_i, X_j) = 0.7,\ i, j = 1, \ldots, 6; \\ X_k &\sim N(0,1), i.i.d,\ k = 8, \ldots, 11, \\ X_{11} &: \text{ is an unobserved hidden variable.} \end{aligned}$$

Its 100K simulated data set is created by retaining all data of covariate features in Example-1, but only correspondingly changes the Y values according to its formula. Thus, the associative heatmap, graph, or network among these 12 variables is expected to be very much alike Fig. 2. We compute and report CEs across three feature settings in Table 5 and summarize our first phase of major factor selection by pointing out major factor candidates across the three feature settings, respectively.

1-feature setting: The feature X_7 achieves the lowest CE followed by CEs of X_3, X_2 and X_1, in the order of increasing values. They are evident candidates of order-1 major factors. Since their CEs are significantly smaller than CEs of individual features of pure noise $\{X_8, X_9, X_{10}\}$. However, the nearly constant CEs of $\{X_4, X_5, X_6\}$ are smaller than CEs of $\{X_8, X_9, X_{10}\}$, but significantly larger than the 4th ranked CE of X_1. This observed pattern could be entirely due to their high degrees of dependence with $\{X_1, X_2, X_3\}$. Thus, their candidacy for order-1 major factor is not clear and needs a de-associating operation to confirm or dismiss.

2-feature setting: The feature-pair $\{X_2, X_3\}$ achieves the lowest CE, not the feature-pair $\{X_3, X_7\}$, which has a CE even larger than CE of feature-pair $\{X_1, X_3\}$. The feature-pair $\{X_2, X_3\}$ seems to be at the position of the feature-pair $\{X_4, X_7\}$ in the previously discussed Example-1. No pairs achieve the ecological effect. These results seem to indicate that the role X_7 in reducing the uncertainty of Y is out-performed by X_3. For this reason, we need to perform de-associating with respect to X_3 and X_7, respectively, to further confirm the candidacy of order-1 and order-2 major factors. On the other hand, all three feature pairs of members of $\{X_4, X_5, X_6\}$ do not show any significant CE-drops from their individual members. Thus, they are not likely order-2 major factors.

Table 6
Experiment-20220408 with $N = 10^5$ and CEs when $X_7 = 8$ and weighted CEs when X_7 ranging the 12 categories. The response variable is $Y^{\perp}[X_7]$ and covariate features $\{X^{\perp}[X_7]_k | k \neq 7\}$.

1Feature	CE[$X_7 = 8$]	CE[X_7]	2Feature	CE[$X_7 = 8$]	CE[X_7]
X3	1.2598	1.1842	X2_X3	1.0787	0.8849
X2	1.3611	1.2725	X1_X3	1.2111	1.0125
X1	1.4255	1.3320	X1_X2	1.2162	1.1283
X6	1.4726	1.3773	X3_X6	1.2170	1.1390
X5	1.4734	1.3782	X3_X8	1.2404	1.1596
X4	1.5085	1.4154	X3_X7	1.3137	1.1842
X10	1.5090	1.4161	X2_X8	1.3350	1.2457
X9	1.5099	1.4165	X3_X4	1.2598	1.3481
X8	1.5090	1.4161	X4_X7	1.5090	1.4154
X7	1.5130	1.4205	X7_X9	1.5099	1.4165

Table 7
Experiment-20220408 with $N = 10^5$ and CEs when $X_3 = 8$ and weighted CEs when X_3 ranging the 12 categories. The response variable is $Y^{\perp}[X_3]$ and covariate features $\{X^{\perp}[X_3]_k | k \neq 3\}$.

1Feature	CE[$X_3 = 8$]	CE[X_3]	2Feature	CE[$X_3 = 8$]	CE[X_3]
X2	1.0970	1.0300	X1_X2	0.5280	0.5034
X1	1.2505	1.1789	X2_X7	0.9460	0.8849
X7	1.2587	1.1842	X2_X6	1.0436	0.9757
X4	1.5468	1.4617	X2_X4	1.0469	0.9797
X6	1.5500	1.4600	X2_X8	1.0724	1.0046
X5	1.5532	1.4616	X1_X7	1.0735	1.0125
X10	1.6717	1.5793	X1_X6	1.1876	1.1188
X8	1.6721	1.5791	X1_X8	1.2243	1.1488
X9	1.6724	1.5793	X4_X7	1.2403	1.1651
X3	1.6766	1.5841	X3_X9	1.6724	1.5791

Table 8
Experiment-20220408 with $N = 10^5$ and weighted CEs conditioning on (boted by)$\{X_2, X_3\}$ and weighted CEs when conditioning on (boted by) $\{X_1, X_2, X_3\}$ across all 2D and 3D categories. The response variable are $Y^{\perp}[X_2, X_3]$ and $Y^{\perp}[X1, X_2, X_3]$, and covariate features $\{X^{\perp}[X_2, X_3]_k | k \neq 3\}$ and $\{X^{\perp}[X_1, X_2, X_3]_k | k \neq 3\}$, respectively.

1Feature	$CE^{\perp}[X_2, X_3]$	$CE^{\perp}[X_1, X_2, X_3]$
X1	0.5034	0.5034
X2	1.0300	0.5034
X3	1.0300	0.5034
X4	0.9797	0.4518
X5	0.9772	0.4500
X6	0.9758	0.4508
X7	0.8849	0.4686
X8	1.0046	0.4359
X9	1.0048	0.4356
X10	1.0043	0.4344

3-feature setting: The feature-triplet $\{X_1, X_2, X_3\}$ achieves the lowest CE with a significant CE-drop from CE of the feature-pair $\{X_2, X_3\}$. This CE of feature-triplet $\{X_1, X_2, X_3\}$ is significantly smaller than CE of $\{X_2, X_3, X_7\}$. This fact clearly indicates that the role of X_1 with feature-pair $\{X_2, X_3\}$ is much more critical than that of X_7 with feature-pair $\{X_2, X_3\}$. This less important role of X_7 can be further reflected by the observation of the two feature-triplets, $\{X_2, X_3, X_6\}$ and $\{X_2, X_3, X_7\}$, are relatively close. We need further evidences from de-associating with respect to $\{X_2, X_3\}$ and $\{X_1, X_2, X_3\}$.

Next we implement de-associating computations with respect to X_7 first and X_3 secondly, and report CEs of all $X^{\perp}[X_7]_i$ and $X^{\perp}[X_3]_i$ correspondingly in Tables 6 and 7. It is worth noting that members of either $\{X^{\perp}[X_7]_i\}$ or $\{X^{\perp}[X_3]_i\}$ are much less associated with each other. Results pertaining to de-associating with respect to $\{X_2, X_3\}$ and $\{X_1, X_2, X_3\}$ are reported in Table 8. We summarize the results from these three tables to advance our major factor selection protocol for Example 2.

[Table 6 :] From its 3rd and 6th columns, we see the three pairs: $\{X^{\perp}[X_7]_2, X^{\perp}[X_7]_3\}$, $\{X^{\perp}[X_7]_1, X^{\perp}[X_7]_3\}$ and $\{X^{\perp}[X_7]_1, X^{\perp}[X_7]_2\}$, show their ecological effects, simultaneously. So they can be concurrently order-1 major factors, but they are not potential order-2 major factors candidates. The pair $\{X^{\perp}[X_7]_3, X^{\perp}[X_7]_6\}$ does not achieve the ecological effect.

[Table 7 :] From its 3rd and 6th columns, we see pair $\{X^{\perp}[X_3]_1, X^{\perp}[X_3]_2\}$ achieves the ecological effect. Neither $\{X^{\perp}[X_3]_2, X^{\perp}[X_3]_7\}$, nor $\{X^{\perp}[X_3]_2, X^{\perp}[X_3]_6\}$ show ecological effects.

[Table 8 :] When performing de-associating with respect to $\{X_2, X_3\}$, we see that X_1 achieves the lowest weighted CE across all $(12)^2$ categories, which is significantly lower than that of X_7. That is, X_1 still provides an extra amount of uncertainty reduction on Y beyond $\{X_2, X_3\}$. Though, X_7 also provides a small amount of uncertainty reduction on Y beyond $\{X_2, X_3\}$ as expected. Further, when performing de-associating with respect to $\{X_1, X_2, X_3\}$, we see that X_7 achieves a CE even larger than the pure noise features X_8, X_9 or X_{10}. That is, X_7 indeed does not provide any extra information beyond what $\{X_1, X_2, X_3\}$ can provide. Likewise, individual feature like X_4, X_5 or X_6 do not offer any extra information beyond what $\{X_1, X_2, X_3\}$ can offer.

[Conclusion:] It is evident that $\{X_1, X_2, X_3\}$ is a collection of three order-1 major factors. In contrast, feature-pair $\{X_4, X_7\}$ is not alternative collection of order-2 major factors. And X_5 and X_6 do not have any roles as an order-1 major factor.

3.4. Modified major factor selection (MFS) protocol

Through Example-2, which exhibits structural dependency, we provide a comprehensive illustration of the necessity for the "de-associating" operation prior to using our previously developed major factor selection protocol to extract precise major factors of various orders. It is important to note that in a real-world structured database, the primary source of structural dependency is likely to be among the "principle features" that collectively characterize the dynamics of the complex system under study. These principle features demonstrate the system's defining dynamics through structural dependency. However, since subjects in a large complex system cannot participate in its defining dynamics globally, they only participate locally in different parts of the overall dynamics. This perspective offers a simple yet realistic explanation for the generation of heterogeneity.

Furthermore, hierarchical heterogeneity arises as a result of the multiscale nature of complex system dynamics. This reasoning leads to a fundamental assumption that underlies any major factor selection protocol: the Re-Co dynamics must accurately represent the true dynamics of the complex system to a significant extent. We observe that this assumption is often overlooked in many scientific literatures, particularly in Statistics, where the choice of response feature-set is often arbitrary. We term this assumption "Re-Co-coherence". Therefore, satisfying this assumption implies that the hierarchical heterogeneity within the dynamics of a large complex system is indeed coherent with the embraced structural dependency in a proper choice of Re-Co dynamics.

In this subsection, based on the [Re-Co-coherence] assumption, we present a step-by-step operational process for our modified protocol. The objective of this description is to enhance the adaptability of our major factor selection protocol for studying appropriately chosen Re-Co dynamics. As a result, the computational results can provide genuine information content about the complex systems of interest.

MFS-1: Based on a matrix representation of an observed structured data set, we first explore potentially structural associations among all features by using a mutual conditional entropy (MCE) heatmap superimposed with a hierarchical clustering tree [17]. Various compositions of block patterns in MCE heatmap reveal various maps of a community-based structural dependency across all involving features on both response and covariate sides. The response feature(-set) is denoted as $\mathcal{Y}$.

MFS-2: To a great extent, such a manifestation of structural dependency, which maps out which features are highly associated with which features, but not so much with other features, will also help explain and make senses of all calculated CEs of all feature sets across multiple feature-settings on the global scale. Based on explained CEs and corresponding CE-drops (conditional mutual information), we identify highly potential candidates of major factors of low orders: either order-1 or order-2, that are able to achieve significant CE-drops. Denote the collection of low-order major factor candidates as $\{A_1, A_2, .., A_{K^*}\}$ with K^* being a suitable number.

MFS-3: We then perform the de-associating operations with respect to individual A_k with $k = 1, \ldots, K^*$, respectively, to confirm its candidacy and simultaneously further discover which members of $\{A_1, A_2, .., A_{K^*}\}$ or any feature-sets outside of this collection that indeed offer significant CE-drops with respect to dynamics of $\mathcal{Y}^{\perp}[A_k]$ across all categorical values of A_k. This discovery is carried out with help based on the two criteria proposed in [17,18]. These two independence-based criteria become much more relevant and suitable because covariate features are significantly less associative after performing the de-associating operation.

[C1: Confirmable :] A feature-set B is confirmable if a feature-set $\tilde{B}$ is obtained by substituting anyone of feature members of B with a feature that is completely independent of $\mathcal{Y}^{\perp}[A_k]$ and B, we have $I[\mathcal{Y}^{\perp}[A_k];]$ is significantly larger than $I[\mathcal{Y}^{\perp}[A_k]; \tilde{B}]$.

[C2: unreplaceable :] A feature-subset B is replaceable if $I[\mathcal{Y}^{\perp}[A_k]; B] \leq I[\mathcal{Y}^{\perp}[A_k]; B_1] + I[\mathcal{Y}^{\perp}[A_k]; B_2]$ for any compositions of B, i.e. $B = B_1 \bigcup B_2$ and $B_1 \bigcap B_2 = \emptyset$. For B to be declared as unreplaceable, we require that B is not replaceable and simultaneously satisfies the following two extra conditions: (a) its CE-drop is larger than the sum of the top-ranked CE-drop and at least $|B|$-times of its complementary feature-subset' CE-drop; (b) the candidate B joins with any already identified major factor B^*_m must achieve $I[\mathcal{Y}^{\perp}[A_k]; B \bigcup B^*_m] \geq I[\mathcal{Y}^{\perp}[A_k]; B] + I[\mathcal{Y}; B^*_m]$.

MFS-4: The major factor selection results in MFS-3 could vary among all categorical values of A_k. This phenomenon of having varying multiscale major factors reveals the hierarchial heterogeneity in the dynamics of $\mathcal{Y}$ as well as in the data's information content. If some $k' \neq k$ and $A_{k'} \in \{A_1, A_2, .., A_K\}$ never get selected, then apparently $A_{k'}$ does not offer information of $\mathcal{Y}$ beyond A_k. As such the major factor candidacy of $A_{k'}$ is revoked. On the other hand, if $A_{k'}$ uniformly induced significant CE-drops across all categories of A_k, then not only major factor candidacy of $A_{k'}$ is reconfirmed, but also bring out the necessity of performing de-associating with respect to $(A_k, A_{k'})$, that is, we need to look into dynamics of $\mathcal{Y}^{\perp}[A_k, A_{k'}]$ and repeat the MFS-3 step.

The MFS-1 step is designed to assist in selecting appropriate response feature-sets that fulfill the [Re-Co-coherence] assumption, while the MFS-2 step aims to identify a collection of potential principle features underlying the complex system dynamics. It should be noted that the criterion [C1: Confirmable] in MFS-3 serves mainly as a reliability check, while the criterion [C2: Irreplaceable] focuses on determining whether feature-sets provide information beyond a designated major factor candidate, referred to as the ecological effect of these two feature-sets. When a significant ecological effect is found between two feature sets, this criterion serves as an effective tool for confirming the discovery of interacting effects. This approach is crucial in identifying high-order major factors, where a combination of marginally less associated covariate features becomes highly dependent given $\mathcal{Y}$. The various potential scenarios presented in MFS-4 have already been observed in Example-1 and Example-2.

Following our computational developments and illustrations in Example-1 and Example-2, the aforementioned modified major factor selection protocol, which accounts for structural dependency among all features, is applied to extract heterogeneity-based multiscale information content from structured data sets derived from large complex systems. The applicability and practicality of this protocol will be clearly demonstrated through two real-world applications discussed in the following sections.

4. Heart disease's complex dynamics

The objective of BRFSS, as stated on its website, is to collect uniform state-specific data on health risk behaviors, chronic diseases and conditions, access to healthcare, and the use of preventive health services related to the leading causes of death and disability in the United States. Heart disease (HD) has been one of the leading causes of death in the US for several decades, claiming over 600 thousand lives annually. Given the significance of heart disease as a primary concern, it serves as an obvious choice for the response variable in the context of BRFSS. In this section, our focus is on understanding the Re-Co dynamics of heart disease.

We conducted our analysis on a cleaned and consolidated data set obtained from the BRFSS 2015 dataset available on Kaggle (https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset). This data set comprises 253,680 survey responses, with 229,787 respondents indicating that they do not have or have not had heart disease, and 23,893 respondents reporting a history of heart disease. Given the class imbalance with an almost 10-to-1 ratio, the primary objective of utilizing this data set has been to perform binary classification for predicting heart disease risk in Statistics and Machine Learning literatures. This analysis aims to address two key questions posed by Kaggle: (Q1:) To what extent can survey responses from the BRFSS be utilized for predicting heart disease risk? (Q2:) Can a subset of questions from the BRFSS be used for preventative health screening, particularly for diseases like heart disease? Unfortunately, the answers to the aforementioned two questions have not been firmly established thus far. However, we believe that our CEDA-based resolutions, which will be presented below, can have a profound impact on human societies in relation to heart disease and beyond. In order to effectively address these two questions through the Re-Co dynamics, it is crucial to acknowledge that the dynamics of heart disease within a population of over 250,000 subjects encompass a wide range of diverse subtypes. This phenomenon is referred to as hierarchical heterogeneity in this paper. To gain a better understanding of the societal health in the United States as a complex system of interest, it is imperative that we dedicate substantial efforts to identify and gain insights into the complexities of this heterogeneity.

As a contextual background, it is important to emphasize that without investing in such investigative efforts and neglecting the presence of heterogeneity, no inferential approaches can be effective. For instance, when applying various off-the-shelf machine learning approaches such as logistic regression, Random Forest, and different Boosting techniques, the accuracies obtained are not superior to 90%. In the fields of Statistics and Machine Learning, the primary cause of this phenomenon has often been incorrectly attributed to the imbalance between non-diseased and diseased individuals, with ratios of approximately 0.904 to 0.096. Interestingly, even a constant predictive rule of categorizing all subjects as non-diseased would achieve an accuracy greater than 90%. However, it is crucial to recognize that such extreme imbalances are natural and prevalent across all chronic diseases. Therefore, it is imperative to address this data analysis challenge as a matter of urgency.

According to the BRFSS Data Codebook, the variable MICHD represents respondents who have reported a history of coronary heart disease (CHD) or myocardial infarction (MI), commonly known as a heart attack. In this paper, we adopt the variable names used in the Kaggle version of the dataset, where MICHD is replaced by HDAtt and serves as the binary response feature for the targeted Re-Co dynamics. All other variables are considered covariate features in this analysis. The abbreviated names of each covariate feature are defined when they are first mentioned in the text. As HDAtt is a binary variable, the results of our selected major factors are presented and displayed in terms of odds, which represent the ratio of diseased individuals to non-diseased individuals within each category of a covariate feature. Each odds corresponds to one CE.

Upon studying the targeted Re-Co dynamics of HDAtt, a glimpse of the heterogeneity becomes apparent from the MCE-based heatmap of all the involved features shown in Fig. 4. Through the MFS-1 step, we observe complex and structured associative patterns among all the features. Specifically, the response feature HDAtt exhibits strong associations with Age and GenHl (General Health), and moderate associations with many other features.

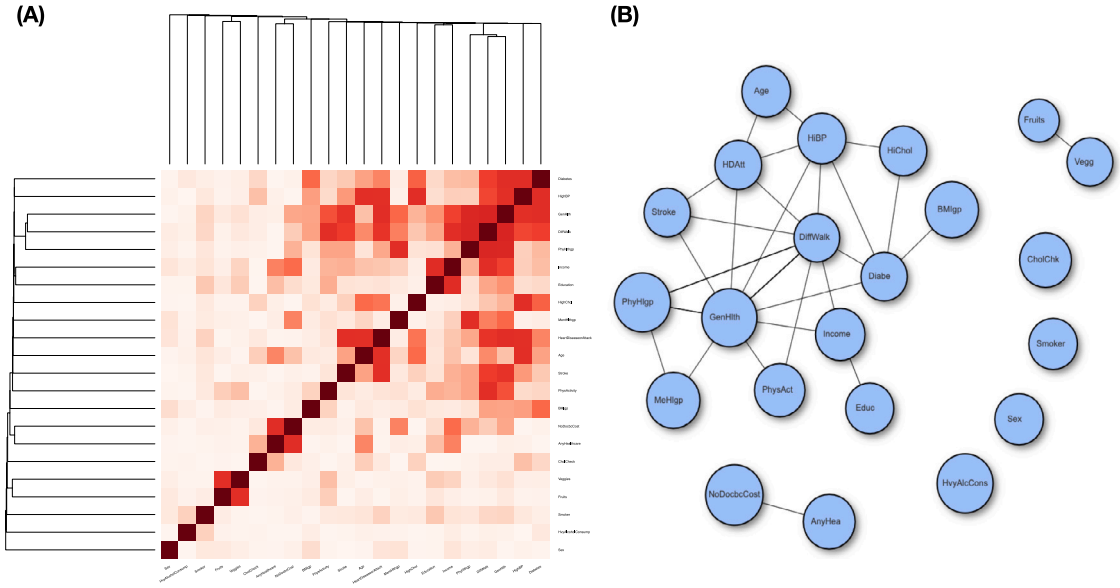

Fig. 4. MCE Heatmap of all features in Kaggle's HD data.

Table 9
Contingency table of (HiBP, HiChol)-vs-HDAtt.

HiBP-HiChol/HD	Non-diseased	Diseased	Prob-vector
0–0	99 044	2876	(0.972, 0.028)
0–1	39 842	3089	(0.929, 0.071)
1–0	39 905	4264	(0.904, 0.096)
1–1	50 996	13 664	(0.733, 0.267)
C-sums	229 787	23 893	(0.906, 0.094)

Another informative aspect of the data can be observed through the contingency table in Table 9, along with the row-wise odds. When considering the single binary feature, high Cholesterol (HiChol), it yields an odds ratio of $3.6 = \frac{16753}{90838} / \frac{7140}{138949}$. On the other hand, the binary feature high Blood Pressure (HiBP) yields an even higher odds ratio of $\frac{17928}{90901} / \frac{5965}{138886} = \frac{0.197}{0.043} = 4.5$. While these two odds ratios effectively convey the relative risks associated with being in the category of $HiChol = 1$ or $HiBP = 1$, it is important to note that the individuals with the disease are still in the minority within both categories. This phenomenon of the diseased minority is also evident in the four bivariate categories of (HiBP, HiChol) presented in Table 9. However, the odds ratio measure is not particularly suitable for situations involving four bivariate categories. Nevertheless, the persistent presence of the diseased minority hiding behind the non-diseased majority remains apparent.

The presence of a persistent and widespread minority hiding behind the majority population sheds light on the complexity of heart disease dynamics. In this section, we aim to address this class imbalance phenomenon by leveraging our major factor selection (MFS) protocol to reveal the hierarchical heterogeneity inherent in the data. Through the application of the MFS protocol, we hope to gain insights into the intricate dynamics of heart disease by exploring the diverse perspectives of heterogeneity encapsulated within the BRFSS dataset.

During the MFS-2 step of our protocol, we identify Age and General Health (GenHl) as the two dominant order-1 major factors. We illustrate these findings by examining the odds expansions across the 13 age categories and 5 GenHl categories in Fig. 5. The odds (represented by blue dots) for 12 of the age categories, except for age-1, exhibit a clear increasing trend as age advances. When centered around GenHl = 3, we observe that the odds for GenHl = 4 and GenHl = 5 expand upwards, while the odds for GenHl = 1 and GenHl = 2 decrease, particularly from age-7 to age-13. These expansions become more pronounced as age increases. Thus, these 65 localities collectively form a spectrum of heterogeneity perspectives that are embedded within the Re-Co dynamics of HDAtt.

We then proceed to the MFS-3 and MFS-4 steps of our MFS protocol to analyze each of the 65 localities resulting from MFS-2. In this study, the feature "Diabetes" (Diabe) is transformed into a binary feature by merging the relatively small category 1 with category 2. Within each locality having sample sizes over 500, we select a triplet of binary features that yield the largest decrease in CE-drop and present our chosen triplet of order-1 major factors in Table 10. These triplets offer the most significant information regarding HDAtt status, surpassing the influence of Age and GenHl. For localities with sample sizes below 500, a NA value is reported.

Upon each selected triplet within a locality, we present the 8 odds and demonstrate their explicit odds-expansions. Fig. 6 illustrates 48 localities categorized by GenHl's 5 categories. The 5 GenHl-specific panels reveal a distinct global pattern of increasing

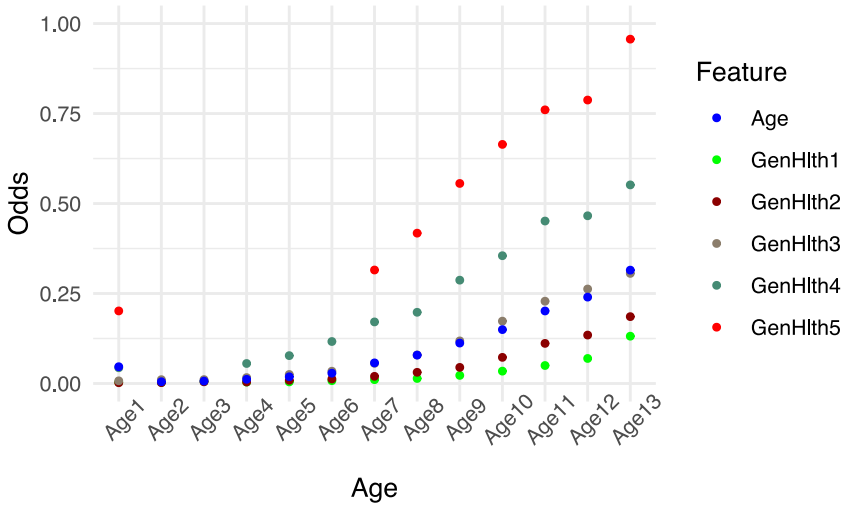

Fig. 5. Odds evolutions w.r.t features.

Table 10
The selected triplets with the response variable is $Y = HD^{\perp}[Age, GenHl]$ and covariate features $\{X^{\perp}[Age, GenHl]_k\}$. B = HBP; C = highChol; S = Stroke; D = Diabetes; X = Sex; O = Smoker; W = DiffWalk; F = Fruits; V = Veggies.

GenHL/Age	1	2	3	4	5	6	7	8	9	10	11	12	13
1	FSV	FSV	DSV	BSV	FOW	CDF	CDS	BDS	BWX	CDS	BDS	BCX	BFX
2	BFO	BFS	BCF	CSX	BDO	BSX	BCD	BCD	BCX	CSX	BCX	CSX	CSX
3	OSX	BOS	BCS	BOS	CSW	BCS	BCS	BCS	BCS	CSX	CSX	CSX	CSX
4	NA	NA	BOS	BOS	BDW	BCS	BCS	BCS	CSX	CSX	CSX	CSX	CSX
5	NA	NA	NA	NA	NA	CDO	BOS	DSX	BDS	DOS	DSX	BDS	BDS

odds-expansions. Within each panel, we observe another pattern of growing odds-expansions along the age axis. Notably, in the 5th panel (GenHl = 5), many odds are greater than 1, indicating that within each of these localities, the prevalence of the disease has become the majority. This figure effectively portrays all localities with heightened odds, emphasizing the alarming probability of developing HD. It represents a landscape of heart disease risk.

From the perspective of heart disease risk assessment, Fig. 6 offers valuable and detailed information contained within the dataset. Its merits and implications are significant. Public health policy makers can gain insights into the distribution of risks on both global and fine scales. Likewise, medical doctors can assess their patients' likelihood of having heart disease more accurately by considering the patients' locality. Furthermore, doctors can effectively guide patients away from the "high-risk" areas indicated by the expanded odds. This type of assessment proves to be more powerful than relying solely on a single odds ratio. It is important to note that an evaluated risk is inherently different from a derived prediction based on an individual's perception. Personal risk evaluation serves as the raw material for generating predictive inferences. However, a simple majority rule based on the risk landscape would outperform blindly applied rules and other machine learning approaches mentioned earlier.

In this manner, we have successfully addressed the two initial questions, Q1 and Q2, posed at the beginning of this section. We believe that the insights provided by Fig. 6, which captures the information content derived from the heterogeneity of the data, will enhance our understanding of behavioral risks associated with heart disease. Moreover, this approach highlights the advantages of a hierarchical heterogeneity-oriented CEDA paradigm. We anticipate that this methodology can be applied effectively to gain a deeper understanding of other chronic diseases within the BRFSS dataset and beyond.

5. MLB fastball pitching dynamics

Beyond being a sport, baseball pitching is a well-known example of the Magnus effect from a physics perspective [26]. This effect elucidates how a spinning object experiences a spin-induced force. This force acts perpendicular to the object's direction of travel and, in essence, causes it to curve based on the spin's direction and rate. For instance, a fastball pitch predominantly exhibits backspin, leading to the Magnus effect that causes the fastball to curve upward, seemingly defying gravity. Fastball pitches are the predominant type in Major League Baseball (MLB) in the US. For more comprehensive descriptions and discussions on the Magnus effect's influence on other pitch types, refer to [13,15,18].

Each professional baseball pitcher in MLB possesses a distinct approach to generating biomechanical forces focused on fastball delivery, often combined with varying backspin. Their pitching motion is a unique representation of how they utilize their musculoskeletal system to create their biomechanics while precisely determining the release point coordinates of the baseball.

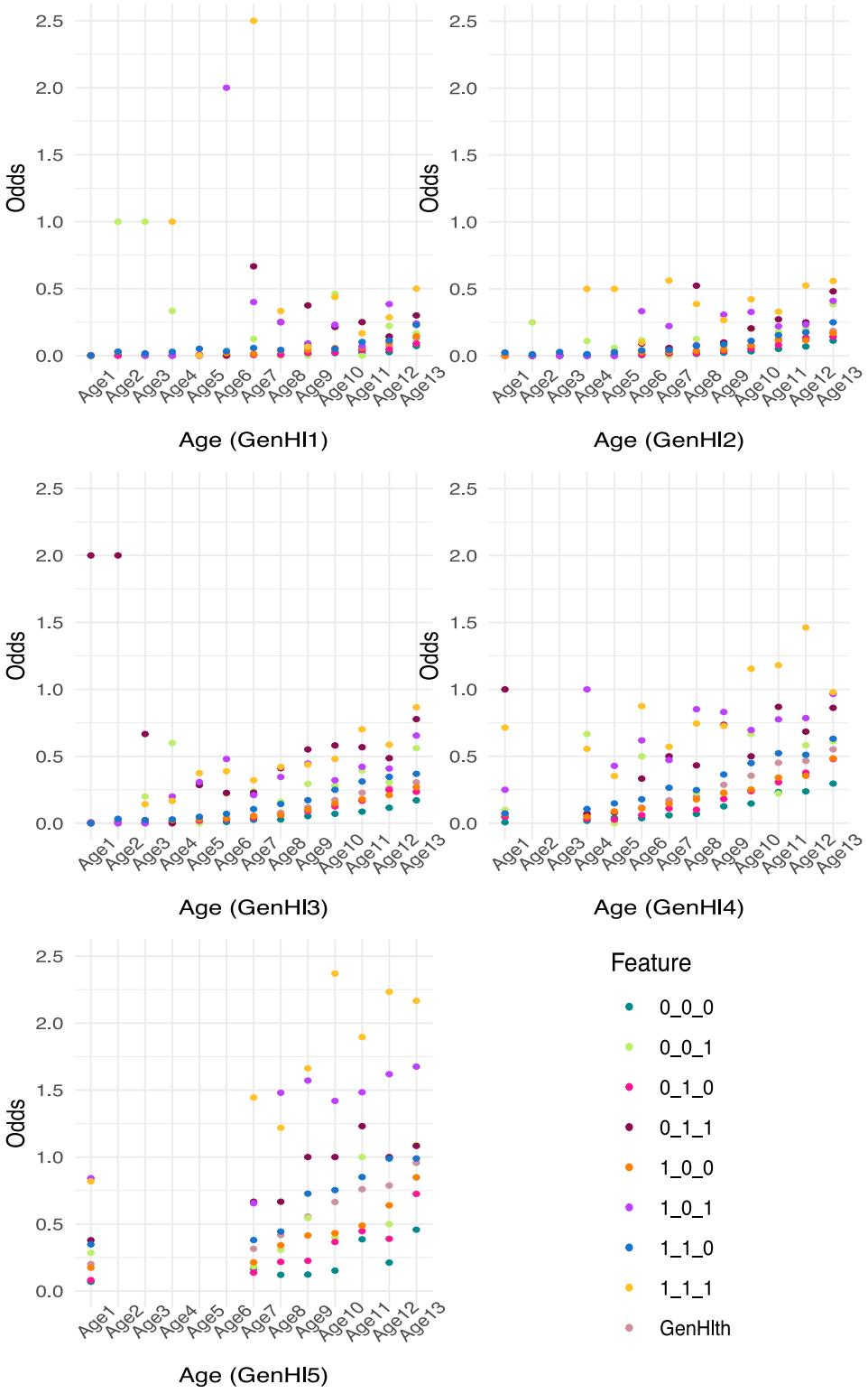

Fig. 6. Odds-expansions w.r.t selected triplet features upon 48 localities defined based on (Age, GenHl).

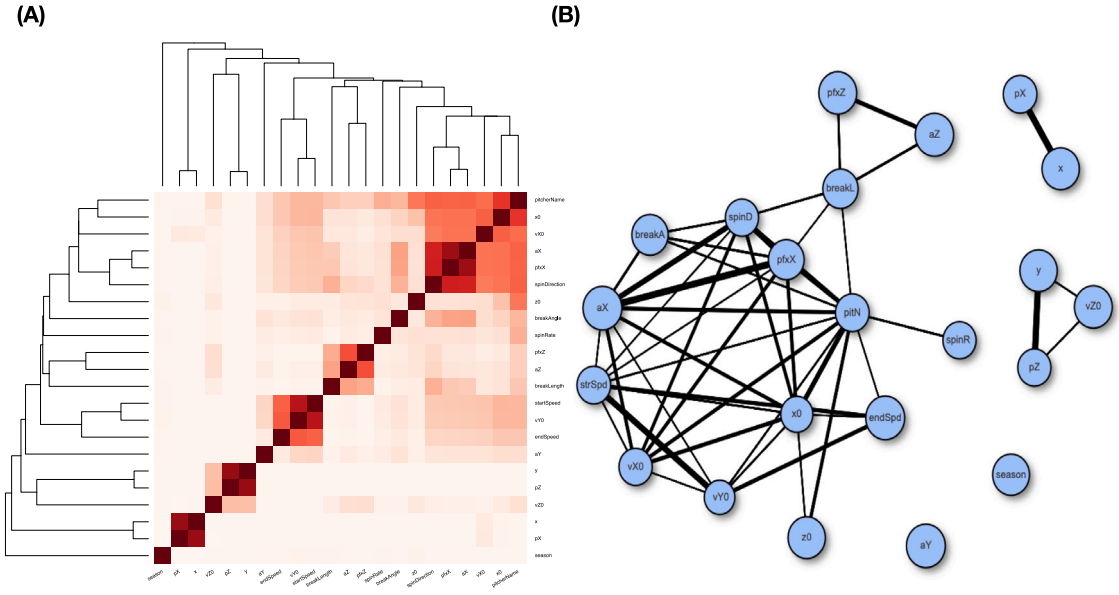

Fig. 7. MCE Heatmap and network of all features in MLB's fastball data.

Table 11
12 fastball pitchers and numbers of pitches across 2017–2019 seasons. (R) for right-handed and (L) for left-handed.

Pitcher-name (handed)/season-pitches	2017	2018	2019
Justin Verlander(R)	2432	2276	1876
Max Scherzer(R)	1589	1813	1596
Chris Archer(R)	1607	951	833
Gerrit Cole(R)	1543	1742	2041
Charlie Morton(R)	367	849	1017
Jacob deGrom(R)	1251	1469	1592
Clayton Kershaw(L)	1385	1194	1170
Chris Sale(L)	1350	1102	908
Jon Lester(L)	1336	1484	910
Mattew Boyd(L)	619	1008	1586
Patric Corbin(L)	840	629	766
Jose Quintana(L)	1293	1480	993

Additionally, the pitcher's grip on the baseball plays a role in determining the spin rate and direction of the pitch, typically concealed within the baseball glove to prevent the batter from anticipating its movement. Furthermore, weather conditions introduce an unpredictable factor, potentially altering the trajectory of the baseball. For instance, precipitation can affect the pitcher's grip on the leather surface of the baseball. Each MLB pitcher exhibits idiosyncratic pitching dynamics that underlie their fastball delivery mechanics.

Furthermore, MLB's regular season typically spans from the beginning of April to the end of October, with playoff games potentially extending into November. During this time, even for a healthy pitcher, their pitching dynamics are expected to evolve in complex and often unpredictable ways over a period of more than six months. Additionally, when considering multiple seasons, it becomes apparent that a professional MLB pitcher's pitching dynamics undergo a distinct and challenging-to-articulate evolution. Hence, it is crucial to acknowledge that the pitching dynamics of an MLB pitcher over multiple seasons is neither well-known nor well-controlled complex system. Moreover, the presence of heterogeneity, resulting from the inclusion of numerous pitchers with different styles and techniques within the same pitch type, further complicates the collective system.

We collected data from 6 right-handed and 6 left-handed pitchers who pitched the top-6 largest numbers of fastballs across the 2017 to 2019 seasons from MLB's Statcast. The names of these 12 pitchers are listed in Table 11 together with their numbers of fastball pitches across the three seasons. As mentioned in the Introduction section, there are 21 features directly linked to pitchers' pitching dynamics. The names of these features can be seen in Fig. 7, while their detailed descriptions can be found in [13,15]. Together with pitcher-name (pitN) and Season as two extra features, we carry out the MFS-1 step of our major factor selection protocol by reporting the MCE-based heatmap and network in two panels of Fig. 7, respectively. Upon the heatmap and network, we see block patterns and community-revealing linkages collectively indicating highly structural dependency among all features. In particular, we see that the two added features: pitN and season, associating with many features located in different blocks and communities. This empirical fact brings out multiple potential challenges when implementing our major factor selection protocol further.

Table 12
Top five feature-sets across three feature settings based on all fastball data of all 12 pitchers.

1Feature	CE	2Feature	CE	3Feature	CE
aX	1.5498	aZ_aX	0.7176	aZ_aX_endSP	0.5728
spinD	1.6596	aZ_spinD	0.9242	aZ_aX_vY0	0.5848
pitN	1.6598	aZ_pitN	0.9828	aZ_aX_startSP	0.5919
aZ	1.8758	aX_BL	1.1969	aZ_aX_pitN	0.6270
x0	2.0851	aZ_x0	1.2286	aZ_aX_z0	0.6684

With each fastball pitch being recorded and described by a 21-dimensional vector labeled by pitcher-name and season, our scientific goal is to explicitly manifest data's information content regarding this complex system of 12 MLB pitchers' collective pitching dynamics. This goal is oriented at least to embrace two chief questions: (Q3:) Which part of the data's information content will allow us to visualize and discover global and fine scales similarity and differences among these 12 individual pitching dynamics? (Q4:) Which part of the data's information content will shed light on fine scale potential changes within a single pitcher's pitching dynamics across the three seasons? Specifically speaking, Q3 obviously contains the major machine learning topic called MultiClass Classification (MCC), while Q4 clearly attempts to discover unspecified fine scale dynamic changes along a targeted complex systems' evolutions. By cohesively addressing Q3 and Q4 here, we want to illustrate an intuitive scientific fact that questions pertaining to a large complex dynamic system are likely connected, so are their resolutions. Here, we stipulate that such connections must chiefly reside in the data's authentic information content.

Again we reiterate that any real-world large complex system likely contains multi-scale structures characterized by multiscale heterogeneous pattern-information, as emphasized in by Nobel physicist P. W. Anderson [7]. As such data derived from such a complex system will likely embrace multiscale information content with hierarchical heterogeneity. Therefore, data analysts and scientists need to search for data' information content from both multiscale and heterogeneity perspectives, and then seek for resolutions to questions. This idea of "data's information first and question's resolution second" might be what the term "data-driven" truly means.

From the multiscale and heterogeneity perspectives, the Q3 indeed embraces that, when each pitcher is commonly defined by a principle physical dynamics and tuned with idiosyncratic characteristics, the MCC resolution can only be efficiently found by pertinently investigating "all chief common factors" underlying all pitchers' dynamics, and then intuitively "subtracting" these chief common factors from each individual dynamics in order to discover individual-specific characteristics. The pieces of information of chief common factors and all of the pitchers' individual dynamics belong to the whole of the data's information content. As for the Q4, it embraces that, even when a question of interest is specifically defined by a fine scale issue within a large complex system, the data's information content involving with many other scales is again needed in order to arrive at a specific fine scale resolution. Indeed, Q1 and Q2 also echo this line of the data-driven message.

5.1. Categorizing Re-Co dynamics and its global information

We make use of the Re-Co dynamics underlying the 2D response variable $\mathcal{Y} = (pfx_x, pfx_z)$ to represent the pitching dynamics commonly shared by 12 pitchers. Since $\mathcal{Y}$, a pitch's horizontal and vertical movements, simultaneously has the two ultimate features that any pitcher wants to deliver when facing a batter at home plate. This Re-Co dynamics surely provides the global scale information content of these 12 fastball pitchers, see the scatter plot of $\mathcal{Y} = (pfx_x, pfx_z)$ in Fig. 8.

We then categorize $\mathcal{Y}$ by taking each occupied cell of 5×5 contingency table $C[pfx_x - vs - pfx_z]$ as one category. That is, pfx_x and pfx_z are categorized via their individual histograms with 5 bins, respectively. So a category of $\mathcal{Y}$ is a rectangle-cell, which is intuitively more proper than irregular-cell resulting from other categorization schemes, such as using K-mean or hierarchical clustering algorithms. Further, the rectangle-cell would be natural for predictive purposes as well.

With the categorized $\mathcal{Y}$ and all covariate features, we carry out the MFS-2 by performing the conditional entropy calculations to look into the major factors underlying the Re-Co dynamics. We report the top five feature sets that achieve the smallest CE across 3 feature settings in Table 12. On the 1-feature setting, we see five potential candidates of order-1 major factors: (1) aX (horizontal acceleration); (2) $spinD$ (spin direction); (3) $pitN$ (pitcher IDs); (4) aZ (vertical acceleration); (5) $x0$ (x-coordinate of leasing point). On the 2-feature setting, we see that the feature-pair (aZ, aX) achieves the smallest CE. But due to dependence between aX and aZ, this feature-pair's CE-drop is slightly less than the sum of aX and aZ's individual CE-drops, that is, the ecological effect is not observed. However, when we perform the de-associating calculations with respect to aX, we see that aZ achieves the smallest CE across all 5 categories of aX. That is, aZ certainly contributes extra information on $\mathcal{Y}$ beyond aX. On the 3-feature setting, we do not see significant CE-drops.

Therefore, we take the collection $\{aX, aZ\}$ as two separate order-1 major factors. That is, $\{aX, aZ\}$ provides the most significant information content of $\mathcal{Y}$ on the global scale. It is also intuitive that categories of $\{aX, aZ\}$ would also provide the spectrum of vital perspectives of heterogeneity because each pitcher's $\{aX, aZ\}$ could only cover just a few, not all localities.

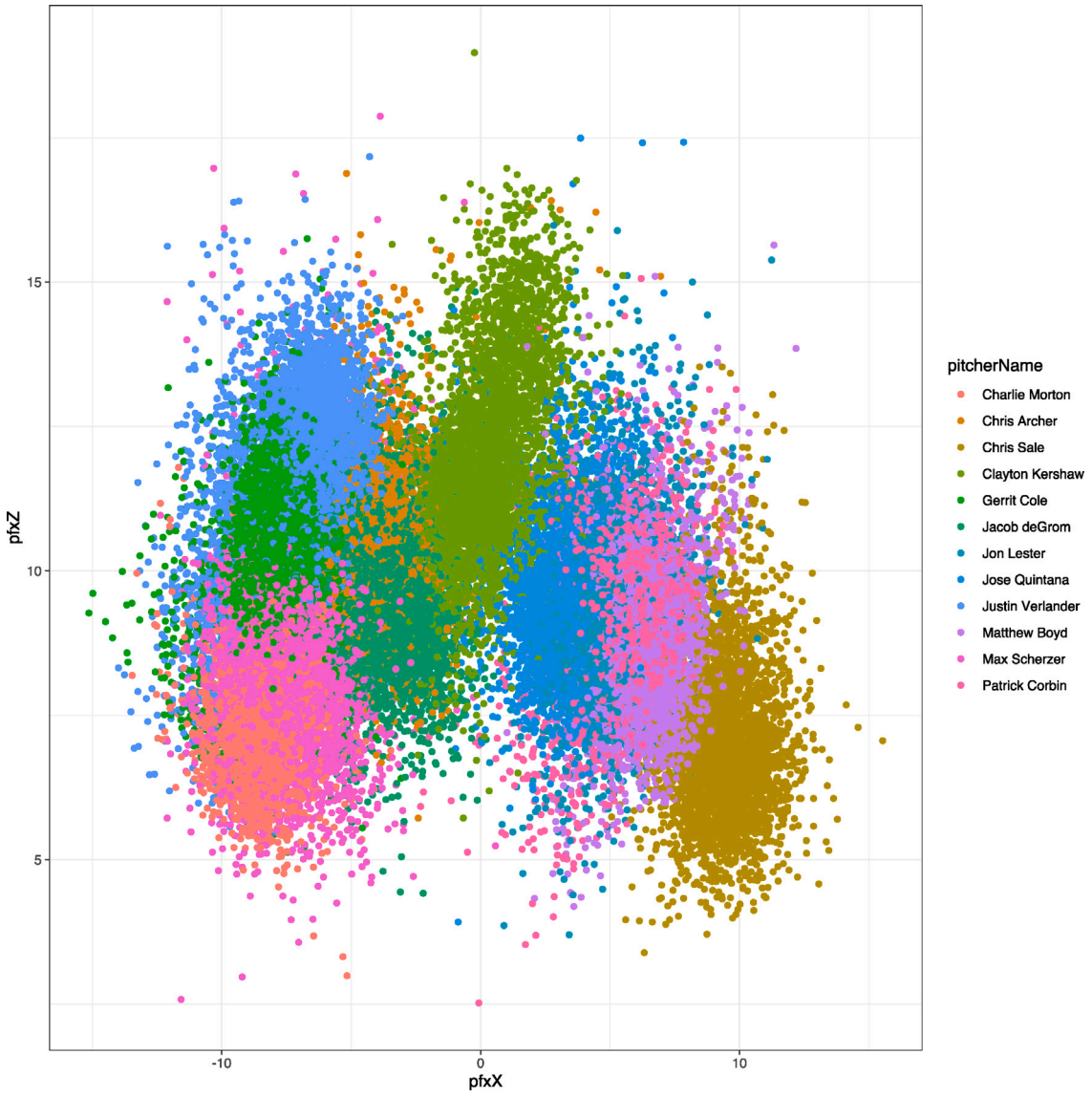

Fig. 8. Scatter plot of $\mathcal{Y} = (pfx_x, pfx_z)$ color-coded with pitcher-name.

5.2. Fine scale information content of Re-Co dynamics of $\mathcal{Y} = (pfx_x, pfx_z)$

To further investigate the fine-scale information content of $\mathcal{Y}$, we perform the MFS-3 and MFS-4 by conducting de-associating calculations with respect to $\{aX, aZ\}$. This involves subdividing the entire dataset based on the contingency table $C[aX - vs - aZ]$, as shown in Table 13. We compute the conditional entropy (CE) for 17 cell-based localities that have more than 800 data points. Among these localities, the spin direction ($spinD$) consistently emerges as the order-1 major factor. The reason behind this observation can be understood by examining the homomorphism between the two 3D scatter plots: $(spinD, aX, aZ)$ and $(spinD, pfx_x, pfx_z)$. The snapshots of these scatter plots in Fig. 9 provide visual evidence of their nearly 2D manifold-like structure, and their homomorphism becomes apparent when they are rotated in any 3D direction. Additional details and visuals can be found in the Appendix, accessible at the link https://rpubs.com/CEDA/factorselect.

Based on the homomorphic relationship between the 3D scatter plots of $(spinD, aX, aZ)$ and $(spinD, pfx_x, pfx_z)$, we can visualize all 25 localities using Fig. 9, or more precisely, through their 3D scatter plots provided in the Appendix. The significance of $spinD$ as an order-1 major factor in all 17 localities can be observed through two sampled snapshots shown in Fig. 10.

Upon examining the 17 localities, each containing more than 800 data points, we also provide a report, as shown in Table 14, on the feature triplets that yield the smallest CE. It is not surprising to find that all of these triplets share a common feature pair, namely $(spinD, endSP)$, where $endSP$ represents the "endspeed" feature. The most frequently occurring third feature in these triplets

Table 13
Contingency table of $C[ax - vs - az]$ based on 12 fastball pitchers' across 2017–2019 seasons.

ax/az	1	2	3	4	5
1	675	1612	1233	859	301
2	1039	3387	5336	6326	2909
3	105	681	2890	2275	835
4	943	5451	3906	882	566
5	1925	1726	767	215	53

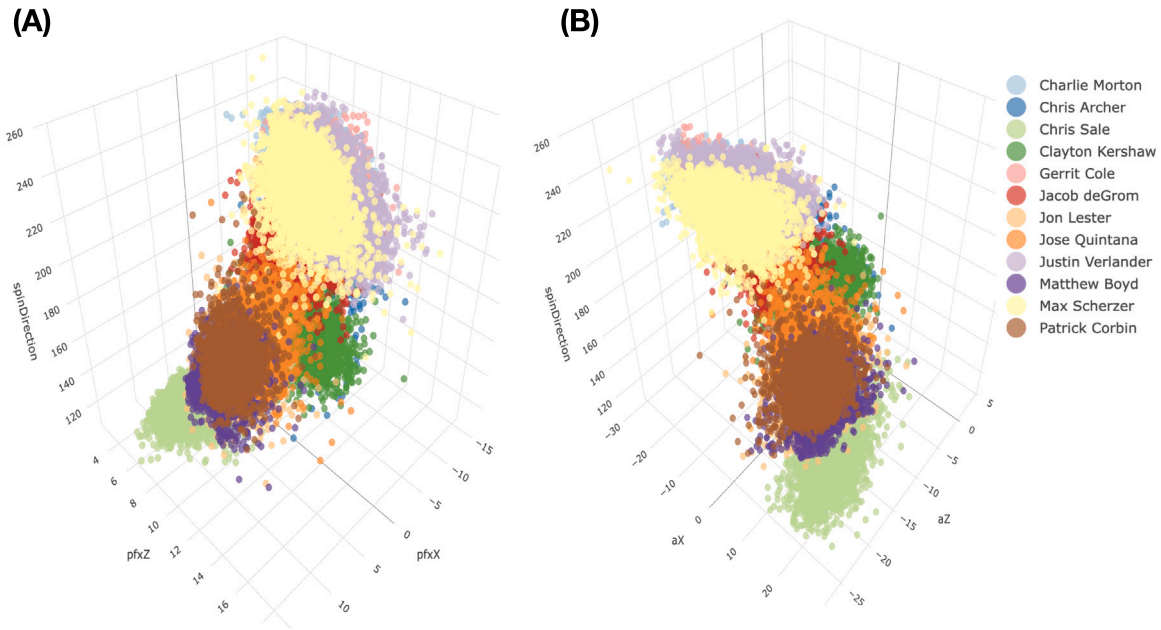

Fig. 9. Homomorphic 2D snapshot of 3D manifolds of $(spinD, pfx_x, pfx_z)$ and $(spinD, aX, aZ)$. Also see rotatable 3D manifolds in Appendix (https://rpubs.com/CEDA/factorselect).

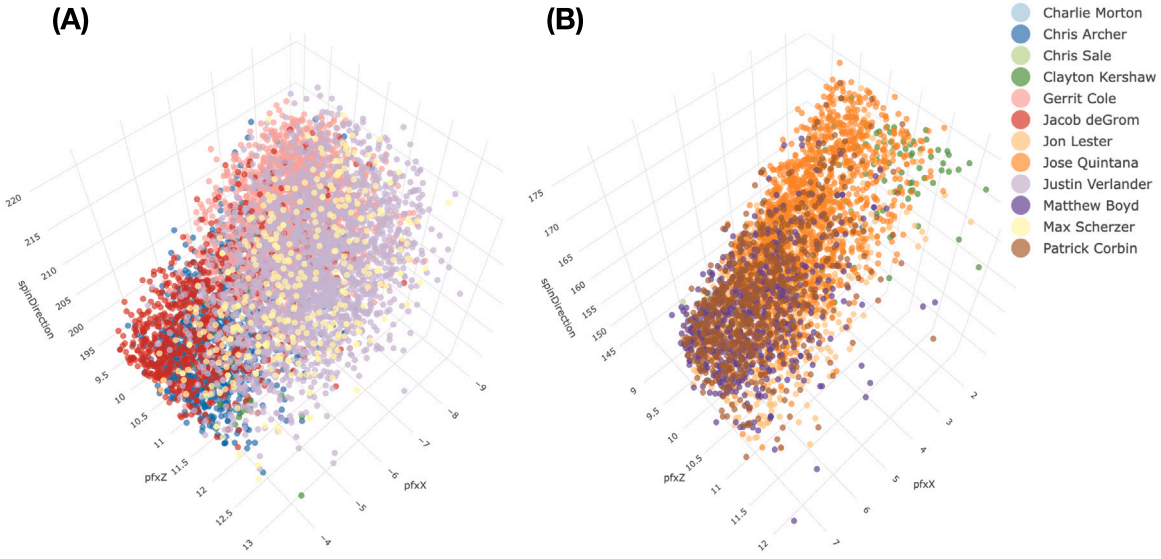

Fig. 10. 2D snapshot of 3D manifolds of $(spinD, pfx_x, pfx_z)$ at the localities: (A) $(aX, aZ) = (2, 4)$ for right-handed pitchers; (B) $(aX, aZ) = (4, 3)$ for left-handed pitchers; Also see rotatable 3D manifolds in Appendix (https://rpubs.com/CEDA/factorselect).

Table 14
Table of best triplets within the framework of $C[ax - vs - az]$ based on 12 fastball pitchers' across the 2017–2019 seasons.

ax/az	1	2	3	4	5
1	NA	BA_spinD_endSP	BA_spinD_endSP	BA_spinD_endSP	NA
2	BA_spinD_endSP	BA_spinD_endSP	BA_spinD_endSP	BA_spinD_endSP	BA_spinD_endSP
3	NA	NA	BA_spinD_pitN	vX0_spinD_endSP	BA_spinD_endSP
4	BA_spinD_endSP	BA_spinD_endSP	BA_spinD_endSP	vX0_spinD_endSP	NA
5	BA_spinD_endSP	BA_spinD_endSP	NA	NA	NA

Table 15
Table of best triplets within localities of (aX, aZ) based on 12 fastball pitchers' across the 2017–2019 seasons. (*) denoting the (x0, z0, spinR) is ranked next to the best. (o) (*) denoting the (x0, z0, season) is ranked next to the best.

ax/az	1	2	3	4	5
1	NA	x0_z0_spinR	x0_z0_spinR	z0_spinR_season(*)	NA
2	x0_z0_season(*)	x0_z0_spinR	x0_z0_spinR	x0_z0_season(*)	x0_z0_spinR
3	NA	NA	x0_z0_spinR(o)	x0_z0_spinR(o)	x0_z0_spinR
4	x0_z0_spinR	x0_z0_spinR	x0_z0_spinR	x0_z0_spinR	NA
5	aY_vX0_x0	x0_z0_vX0	NA	NA	NA

is $break-angle(BA)$. These triplets provide insights into the fine-scale information content of the Re-Co dynamics of $\mathcal{Y} = (pfx_x, pfx_z)$ within these localities.

In conclusion, based on Fig. 10 and the 3D plots in the Appendix, it is evident that combining the global-scale information content of (aX, aZ) with the fine-scale information content of the feature triplet within each locality leads to highly accurate predictive results for $\mathcal{Y} = (pfx_x, pfx_z)$. Within a given locality defined by (aX, aZ), the information provided by $spinD$ specifies a narrow strip within the local region of $\mathcal{Y} = (pfx_x, pfx_z)$ on the manifold of $(pfx_x, pfx_z, spinD)$, as observed in the two panels of Fig. 10 and their corresponding 3D plots in the Appendix. Additional information from two selected features, forming a triplet with $spinD$, further narrows down the predicted region, resulting in a precise inference for $\mathcal{Y} = (pfx_x, pfx_z)$. This approach demonstrates how predictive inferences can be made in a complex dynamic system, utilizing the data's information content. Overall, this data-driven inferential approach, based on the inherent information contained in the data, is natural and highly effective.

5.3. Fine scale information content of local Re-Co dynamics of $Y = pitN$

In this subsection, we address the multiclass classification (MCC) problem, which aims to classify pitches based on 21 features into 12 pitcher labels. A common and straightforward approach is to construct a decision-making framework, such as a label-embedding tree, using a training dataset. Previous research, including the developments and literature review in [13], has explored various approaches such as Random Forest and Boosting methods for MCC.

Another direct approach to MCC involves treating the pitcher ID $Y = pitN$ as the response variable and using 22 features as covariates (including "season"). In our analysis, we found that the best feature-triplet achieving the lowest conditional entropy (CE) is $(x0, z0, spinR)$. However, when we examine the effectiveness of $(x0, z0, spinR)$ on a global scale, as shown in the 2D projection of the 3D scatter plot of all 12 pitchers in Fig. 11, we observe that it is not particularly effective. This lack of effectiveness is also evident in the original 3D plot in the Appendix.

While left-handed and right-handed pitchers are partially separated, the six pitcher-specific (color-coded) point clouds overlap heavily in some regions and separate to some extent in other regions. To achieve a more effective resolution for this MCC problem, a multiscale perspective of heterogeneity is needed when examining $(x0, z0, spinR)$. In fact, the lowest CE of $(x0, z0, spinR)$ is quite close to the CEs of the other top five feature-triplets: $(x0, z0, spinD)$, $(z0, spinD, spinR)$, $(x0, z0, aX)$, and $(x0, z0, pfx_x)$. These three features, namely pfx_x, $spinD$, and aX, are strongly associated with the pitching dynamics across all 12 pitchers. By taking the direct approach to the MCC problem, we encounter the fact that each pitch to be classified is a result of the specific pitcher's pitching dynamics.

Would not it be intuitive and reasonable to take a reverse approach by first examining the pitching dynamics and then performing the computation for $Y = pitN$ within each locality specified by (aX, aZ)? Surprisingly, this reverse approach yields excellent results.

In this subsection, we present a natural and effective data-driven resolution for multiclass classification (MCC) developed from the perspective of a complex dynamic system. Based on the global Re-Co dynamics of $\mathcal{Y} = (pfx_x, pfx_z)$, each locality is defined by (aX, aZ). Within each locality, we then analyze the local Re-Co dynamics of $Y = pitN$ for the purpose of MCC and compute the collection of major factors specific to each locality. The results are summarized in Table 15.

For each of the 17 localities, a 3D plot is constructed based on the computed feature-triplet. Remarkably, each feature-triplet yields a nearly perfect separation of all the pitchers within its respective locality. We present four panels showing 2D projections of four out of the 17 (3D) plots in Fig. 12. It should be noted that the classification results are even more evident when viewing the 3D plots included in the Appendix.

This remarkable success in MCC resolution strongly supports the idea that the authentic multiscale information content based on data's hierarchical heterogeneity should be the foundation for classification-oriented inferences. The precision of classification inference is accompanied by visible evidence, allowing for interpretability. All the evidence consistently points to an intuitive

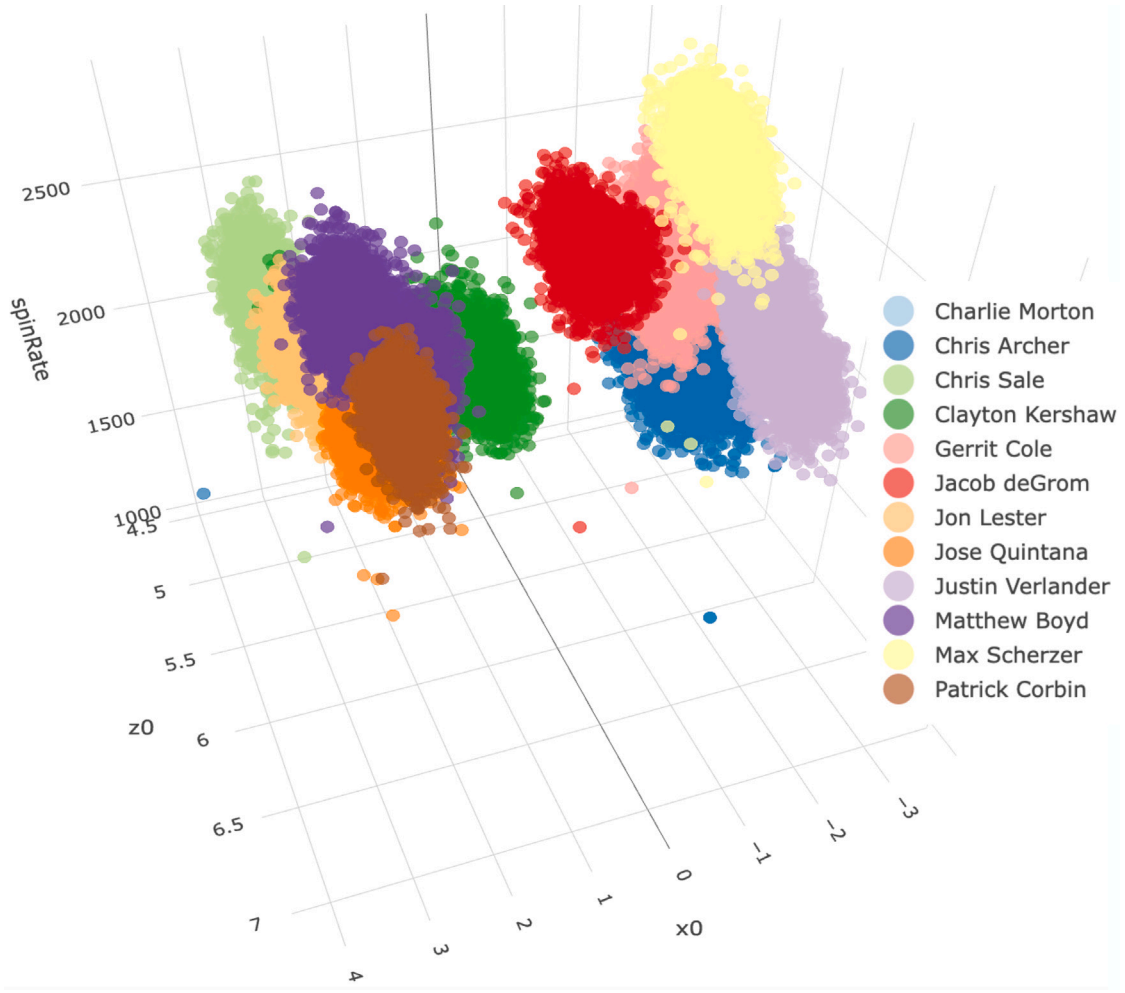

Fig. 11. 2D snapshot of 3D manifolds of $(x0, z0, spinR)$ based on all the 12 pitchers' pitches data. Also see rotatable 3D manifolds in Appendix (https://rpubs.com/CEDA/factorselect).

conjecture: "First build data's hierarchical heterogeneity-based information content, and then address any inferential issues related to complex systems". This conjecture is likely to hold universally because most inferential issues are inherently local in nature and require information about the multiscale heterogeneity of the data. Therefore, our CEDA-based approach appears to be the most effective method for MCC. The strength of our MCC results is further supported by the information provided by the three categories of $season$, as we will see in the next subsection.

5.4. Fine scale information content of Re-Co dynamics of $Y = season$

It is conceivable that the pitching dynamics of MLB pitchers may undergo subtle and unknown changes from season to season. These changes are often fine-scaled and may be challenging, if not impossible, to directly identify within each pitcher's data. This task becomes even more arduous when only a limited number of pitcher-specific data points are available. To illustrate these challenges, we focus on two out of the twelve pitchers: Jacob deGrom and Chris Archer.

We will first consider the case of Jacob deGrom. In order to gather information about potential changes in Jacob deGrom's fastball pitching dynamics across three seasons, a direct approach is employed based on the dynamics of $Y = season$. The top three triplets in terms of conditional entropy (CE) are identified as $(pfx_z, z0, spinD)$, $(pfx_x, vY0, z0)$, and $(vY0, z0, aX)$. These triplets collectively indicate that the dynamics of $\mathcal{Y} = (pfx_x, pfx_z)$ play a significant role. The implication is clear: we need to delve into the localities defined by (aX, aZ) to investigate the specific changes within the heterogeneity-based localities pertaining to the dynamics $\mathcal{Y} = (pfx_x, pfx_z)$. By considering a range of heterogeneity-based perspectives, the feature $z0$ remains important, as does $vY0$ in the dynamics of $Y = season$.

As we examine the localities defined by (aX, aZ) for all pitchers, not just the focal one, we can clearly separate the data points according to the seasons: 2017, 2018, and 2019. We can visually observe distinct geometries of the point clouds that involve pitchers,

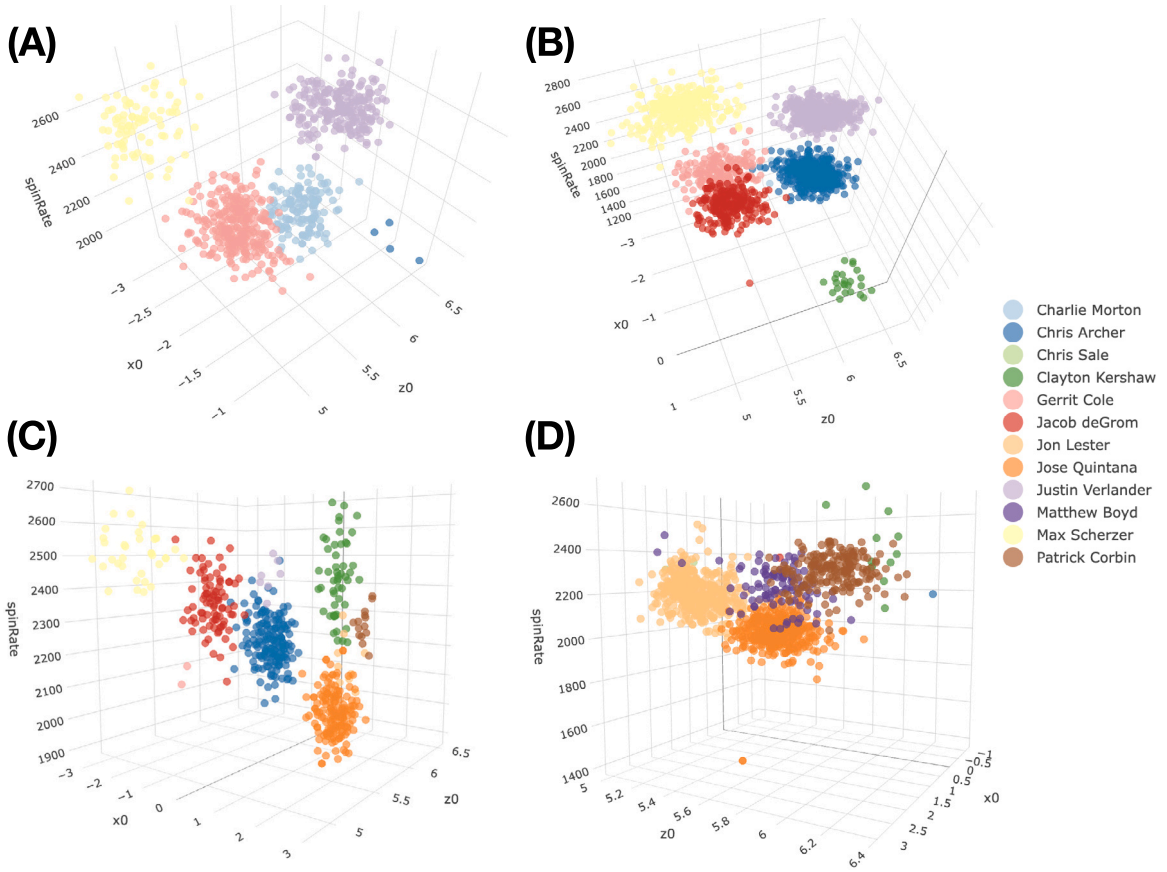

Fig. 12. 2D project of 3D plot of (x0, z0, spinR) at the localities: (A) $(aX, aZ) = (1, 2)$; (B) $(aX, aZ) = (2, 4)$; (C) $(aX, aZ) = (3, 3)$; (D) $(aX, aZ) = (4, 3)$. Also, see rotatable 3D manifolds in Appendix (https://rpubs.com/CEDA/factorselect).

particularly their relative positions in Euclidean space. It is remarkable that the information regarding changes is already present by examining their relative geometric positions across the three seasons. For instance, in the case of the locality $(aX, aZ) = (2, 4)$, as demonstrated in the three panels of Fig. 13, we observe that Max Scherzer's 2017 point cloud (depicted in yellow) has nearly disappeared in 2018 and 2019. Jacob deGrom's 2019 point cloud exhibits significant expansions with respect to the horizontal coordinate of the releasing point $x0$. The separations between the point clouds of deGrom and Gerrit Cole in 2017 and 2018 are no longer present. The three point clouds of Chris Archer appear to contract in both the $(x0, z0)$ dimensions from 2017 to 2019. On the other hand, Justin Verlander's point clouds seem to shift with progressively larger $z0$ values. These informative patterns are most prominently observed through rotatable 3D plots, which can be found in the Appendix as the counterpart to Fig. 13.

Next, let us continue with Jacob deGrom as the focal pitcher in this subsection. We now focus on the direct search based on the dynamics of $Y = season$ within the locality $(aX, aZ) = (2, 4)$. We perform conditional entropy (CE) calculations for three settings: 1-feature to 3-feature. Interestingly, the feature triplet that achieves the lowest CE is $(vY0, vZ0, z0)$. We present the 3D plot of Jacob deGrom's point clouds with respect to $(vY0, vZ0, z0)$ across the three seasons. To depict the patterns contained in this 3D plot, we display four snapshots in the four panels of Fig. 14 (the complete 3D plot can be found in the Appendix). In panel (A), by adjusting the perspective, we observe that the three season-specific (color-coded) point clouds of Jacob deGrom are clearly located and centered in different positions. In panels (B), (C), and (D), we distinctly observe the pairwise differences across the three pairs of seasons. In summary, we can clearly discern Jacob deGrom's changes across the three seasons within the locality $(aX, aZ) = (2, 4)$. Similarly, we can explore other localities for similar or distinct changes. Such pieces of information can be extremely valuable to both the pitcher himself and his pitching coach.

In contrast, let us examine the locality of $(aX, aZ) = (2, 3)$ where Chris Archer is one of the involved pitchers. Once again, we perform our conditional entropy (CE) calculations with $Y = season$ for three settings: 1-feature to 3-feature. Interestingly, the feature triplet that achieves the lowest CE is $(aY, vX0, x0)$, which differs significantly from the feature-triplet found in Jacob deGrom's data. To visualize the patterns specific to Chris Archer contained in his 3D plot, which can be found in the Appendix, we present four snapshots in the four panels of Fig. 15.

In panel (A), by adjusting the perspective, we observe that the three season-specific (color-coded) point clouds of Chris Archer are distinctly located and centered in different positions. In panels (B), (C), and (D), we clearly observe the pairwise differences across

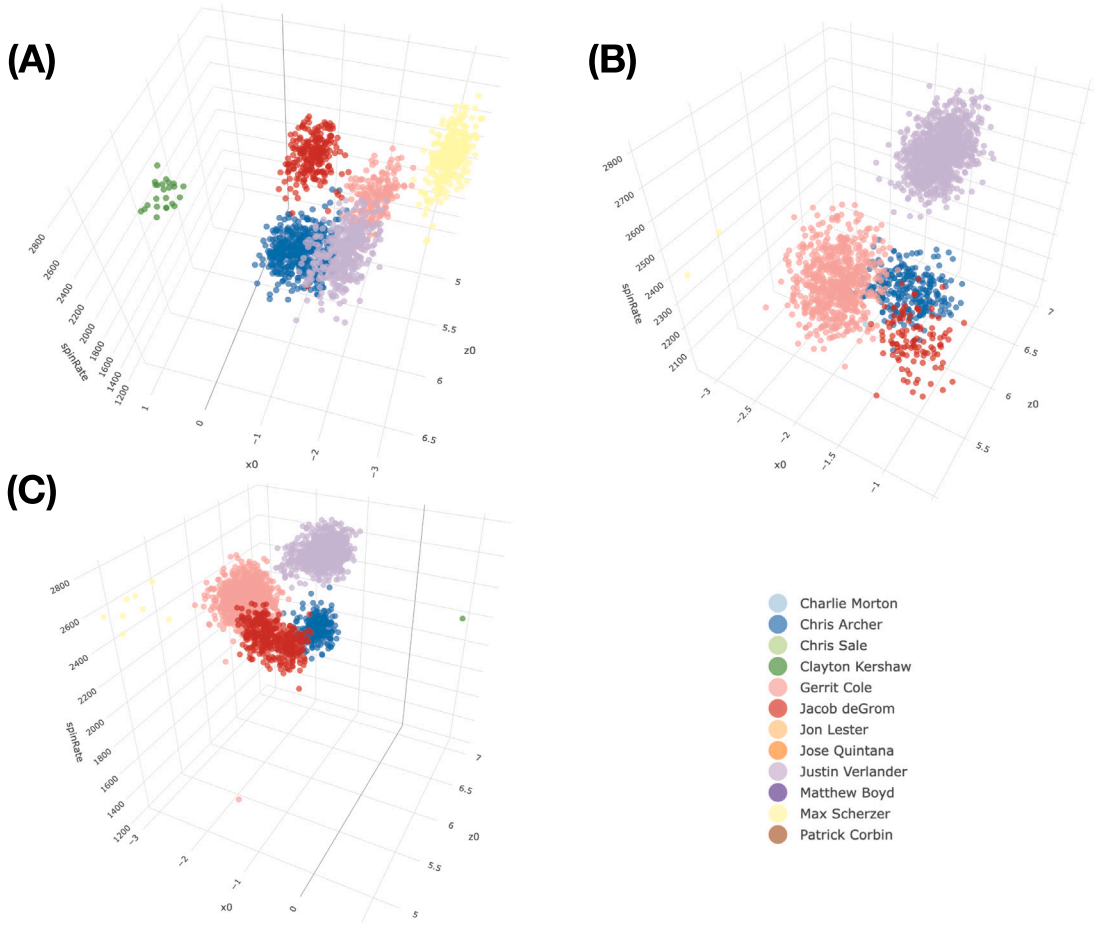

Fig. 13. 2D project of 3D plot of (x0, z0, spinR) at the locality of $(aX, aZ) = (2, 4)$: (A) 2017; (B) 2018; (C) 2019. Also see rotatable 3D manifolds in Appendix (https://rpubs.com/CEDA/factorselect).

the three pairs of seasons. This allows us to observe and analyze Chris Archer's changes throughout the three seasons within the locality $(aX, aZ) = (2, 3)$. Moreover, by employing this approach, we can uncover the season-specific changes in pitching dynamics for different pitchers across various localities.

As a final remark in this subsection, it is evident that the point clouds belonging to all pitchers involved in the localities $(aX, aZ) = (2, 3)$ and $(aX, aZ) = (2, 4)$, as depicted in Figs. 15 and 14 and their counterparts in the Appendix, are clearly distinct from the three perspectives of the three seasons. This observation further confirms that addressing the MCC topic as a complex system dynamics problem and achieving an effective and efficient resolution should naturally emerge from the multiscale information content of the data guided by the discovered hierarchical heterogeneity associated with the system dynamics.

6. Conclusions

We illustrate the CEDA-based data-driven bottom-up computing paradigm through vivid examples from the BRFSS and MLB datasets. The paradigm consists of four MFS-steps, showcasing the process: 1. Manifesting the presence of structural dependency among all involved features (MFS-1). 2. Recognizing and setting up a coherent Re-Co dynamics that represents a vital part of the complex system's dynamics, inducing pertinent perspectives of hierarchical heterogeneity (MFS-2). 3. Applying de-associating operations to significantly reduce structural dependency on the local level (MFS-3). 4. Precisely and effectively carrying out major factor selection protocols on the original Re-Co dynamics or other inferential tasks-oriented Re-Co dynamics, facilitating data's information content and resolving inferential issues (MFS-4). We successfully demonstrate the multiscale information contents of these two complex systems and address four essential questions in a precise and efficient manner. We believe that the CEDA paradigm is applicable to a wide range of large complex systems with structured databases. Scientists in various fields, including Physics, Statistics, and Computer Science, can leverage this paradigm to extract authentic information from their data-driven explorations into complex systems.

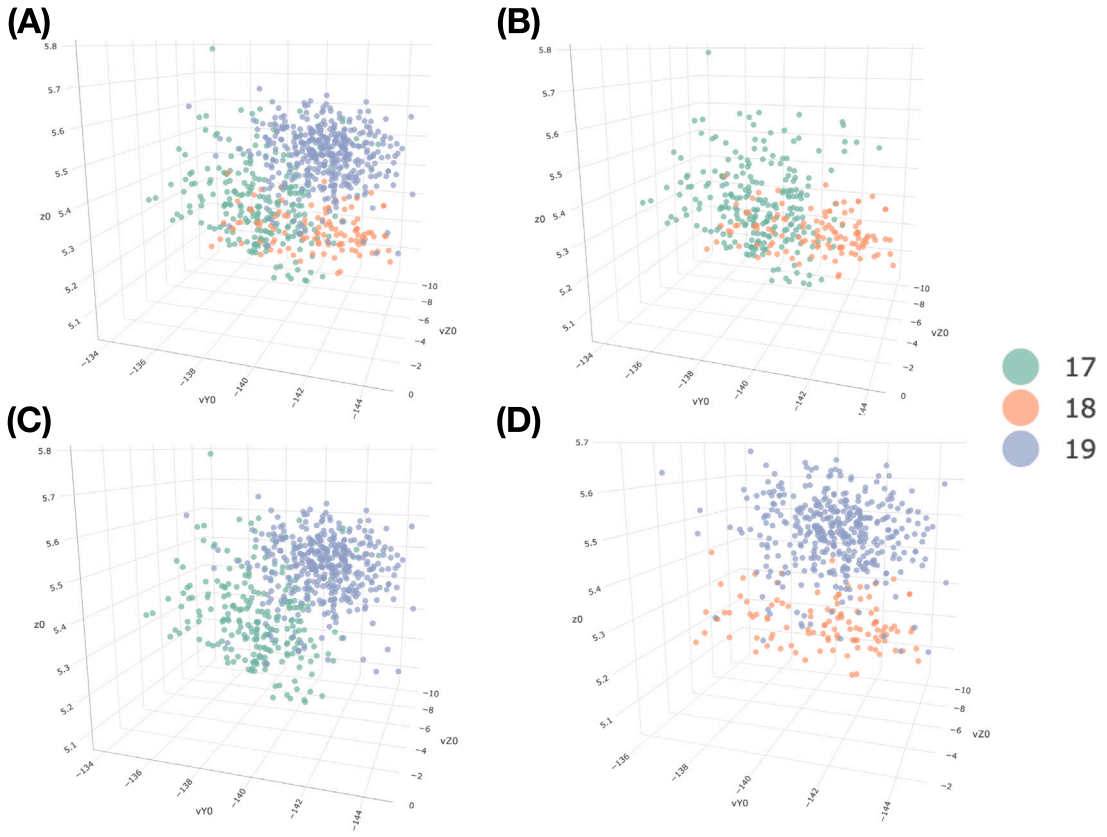

Fig. 14. Jacob deGrom's 2D project of 3D plot of $(vY0, vZ0, z0)$ at the locality of $(aX, aZ) = (2, 4)$: (A) 2017–2019; (B) 2017–2018; (C) 2017–2019; (D) 2018–2019. Also, see rotatable 3D manifolds in Appendix (https://rpubs.com/CEDA/factorselect).

It is important to acknowledge that when analyzing big datasets from complex systems, computing challenges arise due to the complex and interconnected relationships among all features on a global scale. To mitigate these challenges, we utilize the CEDA approach along with the previously developed major factor selection (MFS) protocol. The goal is to extract only a few lower-order major factors that are already known to domain scientists or are evidently representative of a crucial part of the complex system dynamics we are interested in.

By focusing on these lower-order major factors, we gain valuable insights into the essential perspectives of heterogeneity. These perspectives span across different localities, where the remaining features exhibit reduced associations. This enables us to precisely and effectively identify and confirm higher-order interacting effects. Our findings suggest that adopting a global-to-local approach is an effective strategy for addressing the complexities and interdependencies present in the data. Overall, this approach allows us to navigate the challenges posed by the intricate structural dependency within complex systems.

The concept of going from global-to-local is not commonly employed in model-selection or feature-selection problems within the fields of machine learning and statistics. Most scientists and researchers tend to follow a top-down approach, where their models involve numerous quantitative features. These features are often combined under a single modeling framework, such as linearity, which typically disregards potential interacting effects.

To select the "optimal" model, researchers typically employ various optimization algorithms and ad hoc criteria within a pre-specified space of models. However, this approach is limited in its exploratory nature, as the optimality is determined within a confined ad hoc framework. As a result, there is a gap between this common ad hoc computing practice and exploratory analysis.

Furthermore, when dealing with high dimensionality and multiple response and covariate features, traditional modeling frameworks often struggle to handle a response variable defined by multiple response features simultaneously. They also face challenges in capturing both low and high-order interacting effects. These limitations arise from the functional constraints of the modeling framework and the abundance of potential candidates for interacting effects.

In contrast, the Re-Co dynamics approach is particularly relevant when a proper response variable is used. Although the specific forms of interacting effects are typically unknown, their presence holds valuable information within the data. Our CEDA-based data-driven bottom-up computing paradigm, in stark contrast, is not constrained by these issues and provides a solution that is free from these limitations.

Moreover, the traditional modeling practice with a fixed functional framework faces significant validity issues when dealing with categorical features. These features cannot be easily accommodated within the modeling framework, leading to the common practice

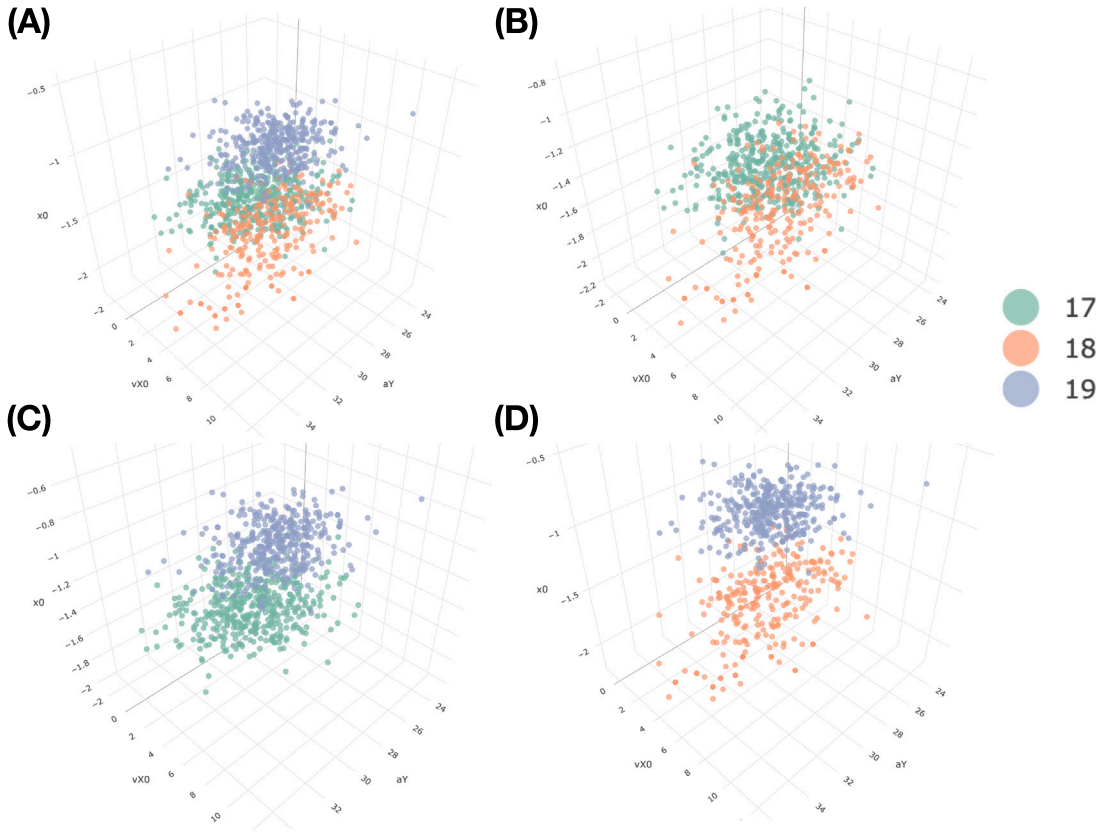

Fig. 15. Chris-Archer's 2D project of 3D plot of $(aY, vX0, x0)$ at the locality of $(aX, aZ) = (2, 3)$: (A) 2017–2019; (B) 2017–2018; (C) 2017–2019; (D) 2018–2019. Also see rotatable 3D manifolds in Appendix (https://rpubs.com/CEDA/factorselect).

of numerically encoding categorical features using arbitrary coding schemes. For example, using "0" for female and "1" for male. Treating these numerically coded categorical features as quantitative variables is a flawed practice that has the potential to generate misleading results. It is crucial to emphasize that hierarchical heterogeneity is likely to exist in all complex system dynamics. This fact should not be ignored or assumed away. Properly addressing hierarchical heterogeneity is essential for scientists to uncover knowledge and insights about the complex systems they study.

When we delve into each locality and perspective of heterogeneity, where the structural dependency among covariate features is significantly reduced, we can uncover the effects of individual features as well as the interacting effects of feature sets. It is important to note that these effects go beyond what the low-order major factors, which define heterogeneity, can provide in terms of reducing uncertainty in response variables. Furthermore, these resulting effects are observable and interpretable through contingency tables and scatter plots in Euclidean spaces. They form essential components of the information content within the data. This CEDA-based data-driven bottom-up computing paradigm demonstrates how data analysis can be conducted on structured databases of complex systems, presenting a clear and significant advantage.

In this paper, we also highlight the fact that precise and efficient solutions to inferential tasks, such as predictions, multiclass classifications (MCC), and detection of minute structural changes, are natural outcomes based on the information content within the data. This indirect approach to achieving inferential resolutions through the CEDA paradigm critically sheds light on the limitations and inefficiencies of direct approaches advocated in the fields of machine learning and statistics. As we enter the era of Big Data, this contrasting fact will become increasingly evident as databases grow in size and become more accessible. We are confident that further exploratory efforts will be necessary to fully exploit the information content within the data. While our CEDA paradigm is well-equipped to play a crucial role in future exploratory endeavors, the task of attaining complete information content in structured databases, such as CDC's BRFSS and MLB's Statcast, still faces numerous challenges. In our future research, we outline two of these challenges. The first challenge is how to compute, extract, and gather the majority of relevant information from the data. The second challenge is how to synthesize all the computed information into actionable knowledge. The first challenge pertains to computing, while the second challenge involves effective graphic representation. We provide our thoughts on these two challenges below.

Regarding the first challenge, when the primary objective of data analysis is to interpret computed results and gain a comprehensive understanding of a specific complex system, it is essential to acknowledge that the full information content of the data comprises a vast array of interacting effects of different orders. However, due to the limitations of finite-sized datasets, exploring

high-order interacting effects becomes increasingly challenging. Yet, these high-order effects are undeniably the most captivating and significant aspects of the information content within the data. Thus, we find ourselves in a dilemma. On one hand, we strive to identify and validate as many reliable high-order interacting effects as possible. On the other hand, we must effectively store and organize these computed pieces of information. Balancing these two conflicting goals is crucial.

Regarding the second challenge, before obtaining a systemic knowledge and understanding of the targeted complex system, the computed pieces of relevant information within the data often appear fragmented and dispersed. This is particularly true for heterogeneity-based data information content computed through CEDA and its MFS protocol. Therefore, significant efforts are required to organize and synthesize these scattered pieces of information to facilitate a comprehensive understanding of the complex system. The organization task necessitates the use of creative graphic displays, such as bipartite and other network representations, to connect computed patterns with subjects exhibiting specific categorical traits. Patterns derived from such networks are likely to stimulate potential and intriguing discoveries. The synthesis task, on the other hand, may rely on the cognitive capabilities of humans complemented by computer assistance in interpreting and synthesizing the information presented through graphic displays. These scientific endeavors are challenging. Nevertheless, they have the potential to unveil multifaceted insights into the dynamics of any targeted complex system beyond the knowledge possessed by data curators alone.

Lastly, in terms of computational considerations, we emphasize the significance of the contingency table as the computing platform for CEDA. Through comprehensive illustrations in Example-1 and Example-2, we demonstrate the major factor selection protocol by utilizing two operations, namely "shadowing" and "de-associating", conducted on the contingency table platform. While these operations explicitly employ the concept of statistical conditioning, it is important to note that beyond categorical variables, statistical conditioning often leads to implicit resultant random variables that are challenging to characterize, particularly in cases involving continuous non-Normal distributions. In contrast, the contingency table platform enables us to explicitly and operationally express the conditioning results for any pair of feature sets, which may consist of multiple 1D variables of any data type. This capability allows us to explicitly address issues arising from the presence of structural dependency among all features, which remained unresolved in our previous works on major factor selection [17,18].

Another noteworthy advantage of the "shadowing" and "de-associating" operations on the contingency table platform is the ability to assess and uncover which feature sets, at specific localities, provide information that surpasses a targeted major factor or feature set. This capability facilitates the identification of major factors that contribute unique information not found in others. Hence, we conclude that our CEDA paradigm demonstrates significant potential and utility for extending Granger causality to complex dynamic systems with structured data of various types. This presents an intriguing future research direction for us.

CRediT authorship contribution statement

Hsieh Fushing: Conceptualization, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Elizabeth P. Chou:** Formal analysis, Investigation, Methodology, Software, Visualization, Writing – review & editing. **Ting-Li Chen:** Conceptualization, Investigation, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

All authors have read and agreed to the published version of the manuscript.

References

[1] C. Darwin, On the Origin of Species By Means of Natural Selection, Murray, London, 1859.
[2] S.A. Kauffman, The Origins of Order: Self Organization and Selection in Evolution, Oxford University Press, New York, 1993.
[3] M. Gell-Mann, What is complexity? Complexity 1 (1995) 16–19.
[4] K. Tumer, D. Wolpert, Collectives and the Design of Complex Systems, Springer, 2004.
[5] C. Adami, What is complexity? BioEssays 24 (2002) 1085–1094.
[6] M. Gell-Mann, The Quark and the Jaguar, W. H. Freeman & Co, 1994.
[7] P.W. Anderson, More is different, Science 177 (1972) 393–396.
[8] Y. Bar-Yam, Dynamics of Complex Systems, Perseus Press, Cambridge, MA, 1997.
[9] J. Jumper, R. Evans, A. Pritzel, et al., Highly accurate protein structure prediction with AlphaFold, Nature 596 (2021) 583–589.
[10] C. Pierannunzi, S.S. Hu, L.A. Balluz, Systematic review of publications assessing reliability and validity of the behavioral risk factor surveillance system (BRFSS), in: 2004-2011. BMC Med Res Methodol, Vol. 13, 2013, p. 49.
[11] D.E. Nelson, E. Powell-Griner, M. Town, M.G. Kovar, A comparison of national estimates from the national health interview survey and the behavioral risk factor surveillance system, Am J Public Health 93 (2003) 1335–1341.

[12] A.H. Mokdad, D.F. Stroup, W.H. Giles, Public health surveillance for behavioral risk factors in a changing environment: recommendations from the behavioral risk factor surveillance team, MMWR Recomm Rep. 52 (RR-9) (2003) 1–12.
[13] H. Fushing, E.P. Chou, Categorical exploratory data analysis: From multiclass classification and response manifold analytics perspectives of baseball pitching dynamics, Entropy 23 (7) (2021) 792.
[14] A.M. Nathan, Analysis of knuckleball trajectories, Procedia Eng. 34 (2012) 116–121.
[15] H. Fushing, E.P. Chou, T.-L. Chen, Mimicking structured data matrix for categorical exploratory data analysis, Entropy 23 (5) (2021) 594.
[16] P.J. Crutchfield, Between order and chaos, Nat. Phys. 8 (2012) 17–24.
[17] T.-L. Chen, E.P. Chou, Fushing Hsieh, Categorical nature of major factor selection via information theoretic measurements, Entropy 23 (12) (2022) 1684.
[18] E.P. Chou, T.-L. Chen, Fushing Hsieh, Unraveling hidden major factors by breaking heterogeneity into homogeneous parts within many-system problems, Entropy 24 (2) (2022) 170.
[19] T.-L. Chen, H. Fushing, E.P. Chou, Practical guidelines on evaluating information theoretical measurements for discovering major factors and making inferences in categorical exploratory data analysis, Entropy 24 (10) (2022) 1382.
[20] H. Fushing, T. Roy, Complexity of possibly-gapped histogram and analysis of histogram (ANOHT), Royal Soc.-Open Sci. (2018).
[21] L. Meier, S. van de Geer, P. Bühlmann, The group lasso for logistic regression, J. Royal Stat. Soc. Ser. B, Methodol. 70 (2007) 53–71.
[22] N. Meinshausen, B. Yu, Lasso-type recovery of sparse representations for high-dimensional data, Ann. Statist. 37 (2009) 246–270.
[23] P. Zhao, B. Yu, On model selection consistency of lasso, J. Mach. Learn. Res. 7 (2007) 2541–2567.
[24] C. Chen, H. Fushing, Multi-scale community geometry in network and its application, Phys. Rev. E 86 (2012) 041120.
[25] L. Paninski, Estimation of entropy and mutual information, Neural Comput. 15 (2003) 1191–1253.
[26] L. Briggs, Effect of spin and speed on the lateral deflection (curve) of a baseball; and the magnus effect for smooth spheres, Am. J. Phys. 27 (1959) 589–596.