Making Short-term High-dimensional Data Predictable

Making accurate forecast or prediction is a challenging task in the big data era, in particular for those datasets involving high-dimensional variables but short-term time series points, which are generally available from real-world systems.

To address this issue, Prof. CHEN Luonan from the Institute of Biochemistry and Cell Biology (SIBCB), Chinese Academy of Sciences (CAS) together with Profs. MA Huanfei (Soochow University), AIHARA Kazuyuki (University of Tokyo) and LIN Wei (Fudan University) proposed a new model-free theoretical framework, namely "Randomly Distributed Embedding" (RDE), for achieving accurate future state prediction based on short-term high-dimensional data.

Specifically, from the observed data of high-dimensional variables, the RDE framework randomly generates a sufficient number of low-dimensional "non-delay embeddings" and maps each of them to a "delay embedding," which is constructed from the data of a target variable to be predicted. Any of these mappings can perform as a low-dimensional weak predictor for future state prediction, and all of such mappings generate a distribution of predicted future states. This distribution actually patches all pieces of association information from various embeddings unbiasedly or biasedly into the whole dynamics of the target variable, which after operated by appropriate estimation strategies, creates a stronger predictor for achieving prediction in a more reliable and robust form.

Through applying the RDE framework to data from both representative models and real-world systems, including the expression level of different genes in the liver, wind speeds across Tokyo and the relation between pollution levels and hospital admissions, the team reveal that a high-dimension feature is no longer an obstacle but a source of

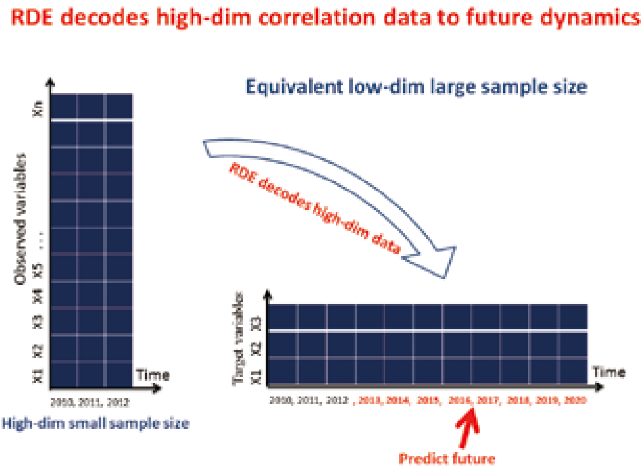

Figure 1: RDE decodes high-dim correlation data to future dynamics. (Image by courtesy of CHEN Luonan's group)

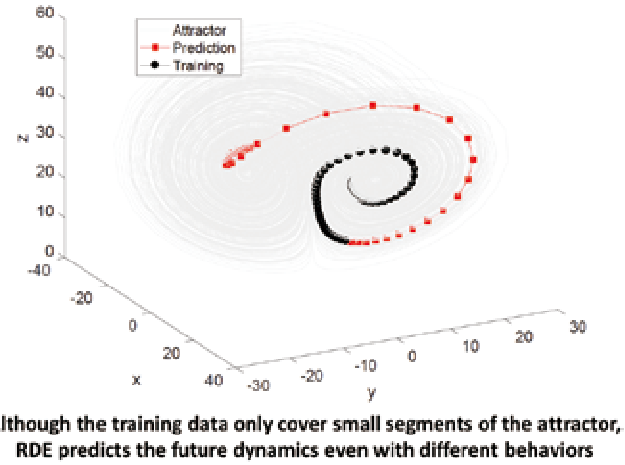

Figure 2: Although the training data only covers small segments of the attractor, RDE predicts the future dynamics even with different behaviors. (Image by courtesy of CHEN Luonan's group)

information crucial to accurate prediction for short-term data, even under noise deterioration. RDE can be expected to be applied to many areas including AI and brain science. In particular, RDE decodes high-dim correlation data to future dynamics of target variables (low-dim), or can be viewed to transform high-dim small sample size into low-dim large size, shown in figure 1.

Although the training data only covers small segments of the attractor, RDE predicts the future dynamics even with different behaviors, as shown in figure 2. Considering the RDE problem as a learning process, we can view it as a "wide-learning" scheme (with small sample size but high-dimension inputs) against the traditional "deep-learning" scheme (with large sample size but usually low-dimension inputs), thus opening a new way to the study of machine learning, AI and brain intelligence.

This work entitled "Randomly Distributed Embedding Making Short-term High-dimensional Data Predictable" was published in *Proceedings of the National Academy of Sciences of the United States of America* on October 8, 2018. This work was supported by the grants from CAS, the National Key R&D Program of China, and the National Natural Science Foundation of China. The publication is available at: http://www.pnas.org/content/early/2018/10/04/1802987115.

Contact:

CHEN Luonan
Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy Sciences,
Shanghai 200031, China
E-mail: lnchen@sibs.ac.cn