Diffusion Models for High-Dimensional Digital Twins

Georg Gottwald, Shuigen Liu, Yang Lyu, Youssef Marzouk, Tan Nguyen, Yuchun Qian, Sebastian Reich, **Xin T. Tong**

Department of Mathematics, National University of Singapore

ABSTRACT

Diffusion model is a popular tool to generate new data samples with possible applications in digital twin training. However, rigorous understanding of the diffusion model is still lacking. One issue is how to train these models for high dimensional problems as score function estimation is subject to the curse of dimension. Another issue is how to avoid the memorization effect, where the diffusion model is bound to generate an exact copy from the training data. We will provide solutions to the first issue by focusing on high dimensional distributions with sparse dependence. We will leverage the sparse dependence to provide a local estimation of the score functions. As for the second issue, we will modify the diffusion model in the final stage and generate new samples close to the same manifold where the training data is originated.

Keywords: Diffusion models; Curse of dimension; manifold hypothesis, Sparse dependence

Two-Sample Tests for Equal Distributions in Separable Metric Spaces: A Unified Semimetric-Based Approach

Jin-Ting Zhang[1], Meichen Qian[1], and Tianming Zhu[2]

[1]*Department of Statistics and Data Science, National University of Singapore, Singapore*

[2]*National Institute of Education, Nanyang Technological University, Singapore*

ABSTRACT

With the advancement of data collection techniques, researchers frequently encounter complex data objects within separable metric spaces across various domains. One common interest lies in determining whether two groups of complex data objects originate from the same population. This paper introduces and examines a fast and accurate unified semimetric-based approach designed to tackle this challenge. The approach exhibits broad applicability across a wide range of research areas, such as bioinformatics, audiology, environmentology, finance, and more. It effectively identifies differences between the distributions of two complex datasets, including both high-dimensional data and functional data. The asymptotic null and alternative distributions of the proposed test statistic are established. Unlike the permutation approach, a unified, rapid and precise method to approximate the null distribution is described. Furthermore, the proposed test is shown to be root-n consistent. Numerical results are presented for illustrating the excellent performance of the proposed test in terms of size control, power, and computational cost. Additionally, the applications of the proposed test are showcased through examples involving both high-dimensional data and functional data.

Keywords: Two-sample test; Equal distribution; Unified semimetric-based approach; Three-cumulant matched chi-square-approximation

A Fast and Accurate Kernel-Based Independence Test with Applications to High-Dimensional and Functional Data

Jin-Ting Zhang[1], **Tianming Zhu[2,*]**

[1]*Department of Statistics and Data Science, National University of Singapore*

[2]*National Institute of Education, Nanyang Technological University*

ABSTRACT

Testing the dependency between two random variables is an important inference problem in statistics since many statistical procedures rely on the assumption that the two samples are independent. To test whether two samples are independent, a so-called HSIC (Hilbert--Schmidt Independence Criterion)-based test has been proposed. Its null distribution is approximated either by permutation or a Gamma approximation. In this paper, a new HSIC-based test is proposed. Its asymptotic null and alternative distributions are established. It is shown that the proposed test is root-n consistent. A three-cumulant matched chi-squared-approximation is adopted to approximate the null distribution of the test statistic. By choosing a proper reproducing kernel, the proposed test can be applied to many different types of data including multivariate, high-dimensional, and functional data. Three simulation studies and two real data applications show that in terms of level accuracy, power, and computational cost, the proposed test outperforms several existing tests for multivariate, high-dimensional, and functional data.

Keywords: complicated data objects; Hilbert-Schmidt independence criterion; three-cumulant matched chi-squared-approximation; two-sample independence test