Statistical Inference for Differentially Private Stochastic Gradient Descent

Zhanrui Cai[1]

Faculty of Business and Economics, University of Hong Kong

ABSTRACT

Privacy preservation in machine learning, particularly through Differentially Private Stochastic Gradient Descent (DP-SGD), is critical for sensitive data analysis. However, existing statistical inference methods for SGD predominantly focus on cyclic subsampling, while DP-SGD requires randomized subsampling. This paper first bridges this gap by establishing the asymptotic properties of SGD under the randomized rule and extending these results to DP-SGD. For the output of DP-SGD, we show that the asymptotic variance decomposes into statistical, sampling, and privacy-induced components. Two methods are proposed for constructing valid confidence intervals: the plug-in method and the random scaling method. We also perform extensive numerical analysis, which shows that the proposed confidence intervals achieve nominal coverage rates while maintaining privacy.

Keywords: Statistical inference; Stochastic gradient descent; Differential privacy

Greedy Model Selection under Sparsity and Covariate Shift

Ching-Kang Ing

Institute of Statistics and Data Science, National Tsing Hua University

ABSTRACT

Sparsity assumptions are central to high-dimensional model selection, yet the true sparsity level is typically unknown in practice. We investigate the convergence of the Chebyshev Greedy Algorithm (CGA) under weak sparsity conditions and propose a data-driven stopping rule, based on a high-dimensional information criterion (HDIC), to adaptively determine the number of iterations. After briefly noting that the CGA+HDIC framework attains the optimal convergence rate without prior knowledge of sparsity, we turn to the more challenging setting where covariates are subject to distributional change. To address this, we extend CGA to an importance-weighted version (IWCGA) by incorporating importance weights into the algorithm. In parallel, we develop an importance-weighted variant of HDIC, termed HDIWIC, to improve model selection under distribution shift. We establish convergence guarantees for the joint IWCGA-HDIWIC procedure and demonstrate its effectiveness through simulations and real-data applications.

Keywords: Chebyshev Greedy Algorithm; Covariate shift; High-dimensional model selection; Importance weighting

Tying Maximum Likelihood Estimation for Dependent Data

Masamune Iwasawa[1], **Qingfeng Liu[2]**, Ziyan Zhao[3]

[1]*Doshisha University, Faculty of Economics*

[2]*Department of Industrial and Systems Engineering, Hosei University*

[3]*Management School, Lancaster University*

ABSTRACT

This study proposes a tying maximum likelihood estimation (TMLE) method to improve the estimation performance of parametric models for dependent data where some time series have long sample periods, while the others are significantly shorter. The TMLE achieves this by flexibly tying some parameters of the long time series to those of the short ones, facilitating the transfer of valuable information to improve the parameter estimation accuracy for the short series. We derive the asymptotic properties of the TMLE and its finite-sample risk bound under a tuning parameter that determines the strength of the tying. The theoretical analysis shows that the TMLE not only substantially outperforms the standard MLE under the correct tying but also can maintain this strength under a local misspecification setting. We propose a feasible bootstrapping procedure for selecting the tuning parameter to reduce the finite-sample risk, with a supporting finite-sample theory that can guide effective implementation of the procedure. Extensive simulations and empirical applications demonstrate that the TMLE exhibits superior performance compared to alternative methods.

Keywords: Tying; MLE; Finite-sample theory; Local misspecification

LLM-Powered Prediction Inference with Online Text Time Series

Yingying Fan[1], **Jinchi Lv[1]**, Ao Sun[1] and Yurou Wang[2]

[1]*University of Southern California*

[2]*Xiamen University*

ABSTRACT

Time series prediction inference is an important yet challenging task in economics and business, where existing approaches often rely on low-frequency, survey-based data. With the recent advances of large language models (LLMs), there is growing potential to leverage high-frequency online text data for improved time series prediction, an area still largely unexplored. This paper proposes LLM-TS, an LLM-based approach for time series prediction inference incorporating online text data. The LLM-TS is based on a joint time series framework that combines survey-based low-frequency data with LLM-generated high-frequency surrogates. The framework relies only on an error correlation assumption, combining a text-embedding-augmented ARX model for the observed gold-standard measurements with a VARX model for the LLM-generated surrogates. LLM-TS employs LLMs such as ChatGPT and the trained BERT models to construct LLM surrogates. Online text embeddings are extracted via LDA and BERT. We establish the asymptotic properties of the method and provide two forms of constructed prediction intervals. To demonstrate the practical power of LLM-TS, we apply it to a critical real-world example: inflation forecast. We collect a large set of high-frequency online texts from a widely used Chinese social media platform and employ LLMs to construct inflation labels for posts that are related to inflation. The finite-sample performance and practical advantages of LLM-TS are illustrated through simulations and this noisy real data example, highlighting its potential to improve time series prediction in economic applications. This is a joint work with Yingying Fan, Ao Sun and Yurou Wang.

Keywords: large language models, CPI prediction, Online texts, Asymptotic distributions, Time series