Optimal Subdata Selection for Large-Scale Linear Regression Under Model Misspecification

Yundi Kong, **Min Yang**, and Ping-Shou Zhong

University of Illinois at Chicago

ABSTRACT

Large-scale regression problems frequently arise in real world applications, and its substantial computational cost motivates the need for subdata selection methods which balance estimation efficiency and accuracy. Many existing work and approaches are designed to optimize parameter estimation efficiency but can be sensitive to model misspecification or rely on assumptions that are difficult to justify in practice.

Motivated by the need for methods that remain effective under both well-specified and misspecified models, we develop a new subset selection framework and derive its corresponding objective function. We show that minimizing the expected prediction error in both well-specified and misspecified linear models leads to an A-optimality-type criterion that unifies model-based and prediction-oriented perspectives. To illustrate the practical implications of this result, we instantiate the framework using the existing IBOSS+ algorithm, which selects informative subsets according to our derived criterion. Our empirical studies confirm that the proposed method achieves superior performance over existing methods under model misspecification setting.

Keywords: A-optimality, Coreset, IBOSS

Influence-Guided Active Subsampling for High-Dimensional Ridge Regression with Application in GWAS

Lin Wang

Department of Statistics, Purdue University

ABSTRACT

Despite the availability of extensive data sets, it is often impractical to collect labels for all data points in many applications due to various measurement constraints. Subsampling approaches can be employed to select a subset of design points from a large pool, resulting in substantial savings in experimental costs. However, existing subsampling methods are primarily designed for low-dimensional data or rely on the assumption of sparse significant predictors. In this study, we propose a computationally tractable sampling method that enables the selection of a small subset from a large data set without assuming sparsity. Our method acknowledges the possibility that the number of significant predictors can be as large as or even larger than the sample size of the full data set. Specifically, our focus lies on ridge regression, for which we develop sampling probabilities that minimize the mean squared predictive risk on the full data set. The efficacy of our proposed approach is substantiated through theoretical analysis and extensive simulations. The results demonstrate its superiority over existing subsampling methods when dealing with high-dimensional data containing numerous significant predictors. Additionally, we illustrate the advantages of our new approach through its application to genome-wide association studies, highlighting its potential to yield valuable insights in this domain.

Keywords: Experimental design; Active learning

Robust Data Fusion via Subsampling

HaiYing Wang[1], Jing Wang, Kun Chen

Department of Statistics, University of Connecticut

ABSTRACT

Data fusion and transfer learning are rapidly growing fields that enhance model performance for a target population by leveraging other related data sources or tasks. The challenges lie in the various potential heterogeneities between the target and external data, as well as various practical concerns that prevent a naïve data integration. We consider a realistic scenario where the target data is limited in size while the external data is large but contaminated with outliers; such data contamination, along with other computational and operational constraints, necessitates proper selection or subsampling of the external data for transfer learning. To our knowledge, transfer learning and subsampling under data contamination have not been thoroughly investigated. We address this gap by studying various transfer learning methods with subsamples of the external data, accounting for outliers deviating from the underlying true model due to arbitrary mean shifts. Two subsampling strategies are investigated: one aimed at reducing biases and the other at minimizing variances. Approaches to combine these strategies are also introduced to enhance the performance of the estimators. We provide non-asymptotic error bounds for the transfer learning estimators, clarifying the roles of sample sizes, signal strength, sampling rates, magnitude of outliers, and tail behaviors of model error distributions, among other factors. Extensive simulations show the superior performance of the proposed methods. Additionally, we apply our methods to analyze the risk of hard landings in A380 airplanes by utilizing data from other airplane types, demonstrating that robust transfer learning can improve estimation efficiency for relatively rare airplane types with the help of data from other types of airplanes.

Keywords: Mean shifts; Non-asymptotic error bounds; Outliers; Transfer learning

Efficient Subdata Selection for Parameter Estimation

Min Yang[1], **John Stufken[2]**, Ming-Chung Chang[3]

[1]*Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago*

[2]*Department of Statistics, George Mason University*

[3]*Institute of Statistical Science Academia Sinica*

ABSTRACT

When, in terms of the number of data points, the size of a dataset exceeds available computing resources, or when labeling is expensive, an attractive solution consists of selecting only some of the data points (subdata) for further consideration. A central question for selecting subdata of size n from N available data points is which n points to select. While an answer to this question depends on the objective, one approach for a parametric model and a focus on parameter estimation is to select subdata that retains maximal information. Identifying such subdata is a classical NP-hard problem due to its inherent discreteness. Based on optimal approximate design theory, we propose a new methodology for information-based subdata selection, resulting in subdata that approaches the optimal solution. The proposed method is supported by a novel algorithm that is proven to converge and that applies to a general model, accommodates arbitrary choices of N and n, and supports multiple optimality criteria. This is joint work with Min Yang (University of Illinois Chicago) and Ming-Chung Chang (Academia Sinica).

Keywords: Approximate design, Equivalence theorem, Exact design, IBOSS, Optimal design