Optimal Subsampling for Linear Models with Heteroscedasticity

Strategies for Large-Scale Data Analysis

Based on the paper by Jiayi Zheng, GMU et al.

February 4, 2026

Outline

The Big Data Challenge

Context

- Modern datasets are massive (e.g., millions of rows, $N \gg p$).
- Standard OLS matrix operations ($\mathbf{X}^T\mathbf{X}$) become computationally expensive or impossible due to memory constraints (RAM).

The Solution: Subsampling

- Select a small, informative subset of size $n \ll N$.
- Fit the model to the subsample to approximate the full-data estimator efficiently.

The Specific Problem: Heteroscedasticity

The Issue:

- Most existing optimal subsampling methods (e.g., standard IBOSS, Information-Based Optimal Subdata Selection) assume **homoscedasticity** (constant error variance: $\text{Var}(\epsilon) = \sigma^2$).

Heteroscedasticity in Real Data:

- Variance often depends on covariates: $\text{Var}(\epsilon_i|\mathbf{x}_i) = \sigma^2(\mathbf{x}_i)$.

Consequence

Ignoring non-constant variance leads to inefficient subsamples and compromised statistical efficiency. Points with high variance provide ***less*** information but might be selected by standard methods.

Optimality Criteria

We aim to optimize the subsample ξ based on two statistical criteria:

1. D-Optimality

- **Focus:** Parameter Estimation.
- **Goal:** Maximize the determinant of the Information Matrix,

$$\max_{\xi} \det(\mathbf{X}_{\xi}^{T}\mathbf{W}_{\xi}\mathbf{X}_{\xi})$$

- Minimizes the generalized variance of regression coefficients ($\hat{\boldsymbol{\beta}}$).

2. I-Optimality

- **Focus:** Prediction Accuracy.
- **Goal:** Minimize the integrated prediction variance,

$$\min_{\xi} \mathrm{Tr}\left((\mathbf{X}_{\xi}^{T}\mathbf{W}_{\xi}\mathbf{X}_{\xi})^{-1}\boldsymbol{\Omega}\right)$$

- Where $\boldsymbol{\Omega} = \int \mathbf{x}\mathbf{x}^{T}\,\mathrm{d}\mathbf{x}$ is the moment matrix of covariates.

The Challenge: Unknown Variance

To select an optimal subsample under heteroscedasticity, observations should be weighted by their **inverse variance** $(1/\sigma_i^2)$.

Problem: The true variance function $g(\mathbf{x})$ is usually unknown.

Solution: Estimate variances $\hat{\sigma}^2$ using a pilot sample before performing the main selection.

Algorithm 1: Variance Estimation via LHD

Step 1: Pilot Sample via Latin Hypercube Design (LHD)

- Select a small pilot subsample n_1.
- **Why LHD?** It is a space-filling design, ensuring the pilot sample covers the range of the data well, capturing the structure of the variance.

Step 2: Nearest Neighbors via k-d Trees

- For points in the pilot sample, find nearest neighbors in the full dataset.
- **Efficiency:** k-d trees (binary space-partitioning) make neighbor search computationally feasible ($O(\log N)$).

Step 3: Kernel Smoothing

- Fit a local linear model to the pilot data to estimate $E(y|\mathbf{x})$.
- Derive variance estimates $\hat{\sigma}^2$ for the *entire* dataset.

Standard IBOSS Refresher

Information-Based Optimal Subdata Selection (IBOSS)

- A deterministic method for homoscedastic linear models.
- **Strategy:** Selects data points with extreme covariate values (min and max).
- **Intuition:** Points far from the center provide the most leverage for defining the regression slope.

The Innovation: Weighted IBOSS (Algorithm 2)

Logic: Under heteroscedasticity, extreme points might be "noisy." Observations with *smaller* variances provide more precise information and should be prioritized.

The Weighted IBOSS Method

1. **Standardize Covariates:** Divide the covariates by the estimated standard deviation:
$$\mathbf{x}^* = \frac{\mathbf{x}}{\hat{\sigma}}$$
2. **Apply IBOSS Strategy:** Perform the standard IBOSS selection (picking extremes) on these new *weighted* covariates $\mathbf{x}^*$.

Outcome: A deterministic subsample that maximizes **D-efficiency** by explicitly accounting for the estimated variance structure.

The Limitation of IBOSS

D-Opt vs. I-Opt

IBOSS is excellent for **parameter estimation** (D-optimality) because it picks corner points.

However, it is often suboptimal for **prediction** (I-optimality), which requires a more balanced coverage of the interior of the data space to minimize prediction error variance.

Algorithm 3: ANNSA Mechanism

Objective: Minimize I-optimality loss $L(\xi)$ via iterative swapping.

The "Swapping" Process

1. **Selection:** Randomly select a point $\mathbf{x}_{out}$ from the current subsample to remove.
2. **Candidate Generation (The "ANN" Boost):**
   - Instead of scanning the entire dataset (N) for a replacement, use k-**d trees** to find a set of Approximate Nearest Neighbors (ANN) to $\mathbf{x}_{out}$ in the full data.
   - Select a candidate $\mathbf{x}_{in}$ from this local neighborhood.
3. **Evaluation:** Calculate the change in the criterion ΔL.
4. **Simulated Annealing Acceptance:**
   - If $\Delta L < 0$ (Improvement): **Always Accept**.
   - If $\Delta L > 0$ (Worse): Accept with probability $P = \exp(-\Delta L/T)$, where T is a "temperature" parameter that decays over time.

Why ANN? It reduces the search complexity from $O(N)$ to $O(\log N)$, making iterative optimization feasible for millions of rows.

Simulation Study Results

Setup:

- $N = 1,000,000$ observations.
- Tested against 4 variance functions (1 homoscedastic, 3 heteroscedastic).

Key Findings:

Method	Performance
Weighted IBOSS	Consistently achieved higher **D-optimality** than standard IBOSS when heteroscedasticity existed.
ANNSA	Consistently achieved the lowest (best) **I-optimality**, making it superior for prediction tasks.

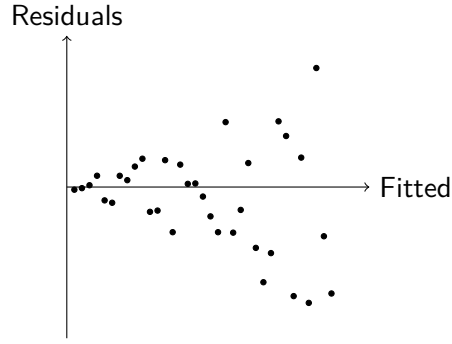
Real Data Application: Airline Dataset

Data Overview:

- 7 million U.S. flight records (2007).
- **Goal:** Predict *ArrDelay* based on *CRSElapsedTime*, *DepDelay*, and *Distance*.

Heteroscedasticity Check:

- Residual plots revealed a clear "fan shape."
- Variance decreased as flight time/distance increased (long flights average out delays).



(Stylized "Fan Shape")

Real Data Performance

Experimental Setup:

- Models fit to subsamples of size $n = 400$ to 500.
- Validated against a large test set.

Result: Superior Prediction
ANNSA provided the lowest Mean Squared Prediction Error (MSPE) on the test set compared to IBOSS, Uniform Subsampling, and Weighted IBOSS.

Conclusion: The proposed methods effectively handle massive, real-world data with complex variance structures.