technische universität dortmund

SFB 876 Providing Information by Resource-Constrained Data Analysis

Project C4

Regression approaches for large-scale high-dimensional data

Prof. Dr. Katja Ickstadt, Prof. Dr. Christian Sohler

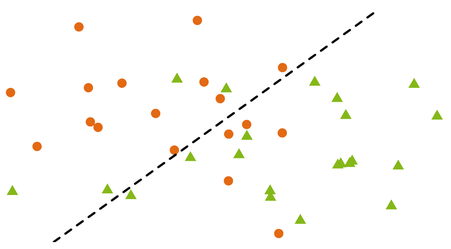
Problem

- Goals: develop highly efficient regression approaches for massive data
  obtain a common and unified understanding of sketching and sampling methods

Three cornerstones to develop unified solutions with guarantees for regression models

Geometric Relaxations

- Overlap of data
- Small perturbations
- Avoid worst-case
- Complexity parameter
- Captures geometric structure



Algorithmic Techniques

- Streaming and distributed computing
- Sketching and sampling
- Random projections

$$\begin{bmatrix} -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -x_1 + x_3 \\ x_2 - x_4 \end{bmatrix}$$

- Sensitivity framework
  quantifies importance of each point by its contribution in the worst case

$$\varsigma_i = \sup_{\beta \in \mathbb{R}^d \setminus \{0\}} \frac{f(x_i\beta)}{\sum_{i=1}^{n} f(x_i\beta)}$$

Statistical Modelling

- Exponential family

$$f(x \mid \theta) = h(x)g(\theta)\exp(\eta(\theta) \cdot t(x))$$

- Common form of distributions
- Prior as relaxation technique
- Prior as penalisation

$$\mathcal{L}(Y \mid X, \beta) \cdot \pi(\beta)$$

- Distributions that allow efficient inference

Planned Research

WP1: Generalised Linear Models

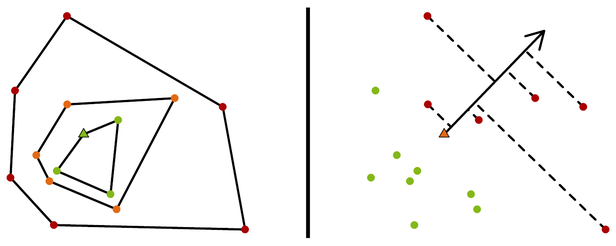
- Common loss function over the exponential family

$$\ell(\beta|X, Y) = \sum_{i=1}^{n} g(x_i\beta) - Y^T X\beta + C$$

- Sensitivity sampling

$$\varsigma_i = \sup_{\beta \in \mathbb{R}^d \setminus \{0\}} \frac{\ell(\beta|x_i, y_i)}{\ell(\beta|X, Y)}$$

- Bound sensitivities using convex layers and Tukey depth



- Smoothed analysis

A2

WP2: Bayesian Generalised Linear Models

Bayesian Sensitivities

$$\varsigma_i = \sup_{\beta \in \mathbb{R}^d \setminus \{0\}} \frac{\ell(\beta|x_i, y_i) + \log(\pi(\beta))/n}{\ell(\beta|X, Y) + \log(\pi(\beta))}$$

- Prior as relaxation admits sublinear sampling scheme
- Bound error in expectation, $(1+\varepsilon)$-guarantee in expectation

Logistic regression and Cox regression

- Genome data and survival analysis
- Prior models ℓ_1 and ℓ_2 penalization

C1

Graph models and generalised linear models

- Evaluate GLM coresets in a Bayesian setting
- Develop coresets for sum-product-networks

B4

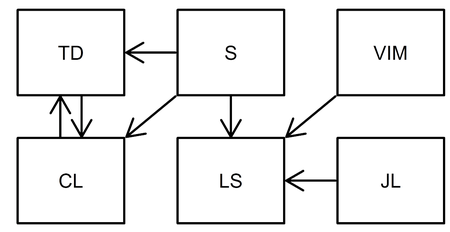
WP3: Importance Measures

Aims:

- Impact on sampling
- Variable selection
- Study properties and relationships

Techniques:

- Leverage scores, cross-leverage scores (LS)
- Jacobian leverage (JL)
- Tukey depth, Tukey median (TD)
- Convex-layer depth (CL)
- Sensitivity (S)
- Variable importance measures (VIM)



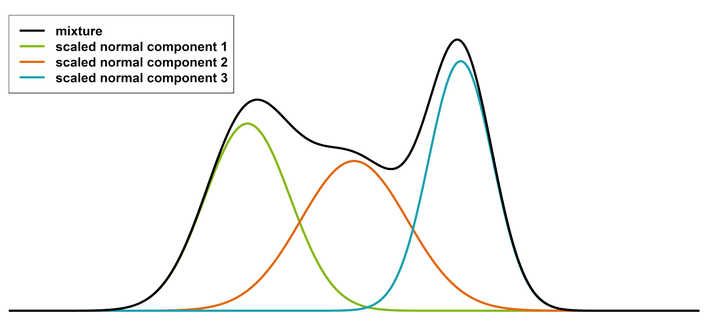
WP4: Mixtures of Normal Distributions

Motivation:

- Regression on normal mixture models
- Approximate arbitrary continuous distributions
- Handle outliers

Challenges:

- Allocation to mixture components and regression problem
- Trade-off: number of components $\leftrightarrow$ model accuracy
- Relaxing the allocation problem via oracle techniques



Lehrstuhl Informatik 2
Effiziente Algorithmen und Komplexitätstheorie

Lehrstuhl Mathematische Statistik und Biometrische Anwendungen