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

- Problem: Large number of observations or large dimension
- Consequence: Modern regression approaches reach limits of scalability

Goal: Develop highly efficient regression approaches for massive data

Bayesian Regression

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

- ℓ_p regression
- Hierarchical prior models

Generalised Linear Models (GLMs)

$$h(\mathbb{E}(Y)) = X\beta.$$

Examples:

- Poisson regression for count data
- Logistic regression for binary data

Structural Constraints

$$\min_{\xi \in \mathcal{N}} \|\xi - Y\|, \text{ for some } \mathcal{N} \subseteq \mathbb{R}^n$$

- E.g. monotonicity, unimodality

$$\mathcal{N} = \{\xi \mid \exists i \in [n] \colon \xi_1 \leq \ldots \leq \xi_i \geq \ldots \geq \xi_n\}$$

- Application: peak detection

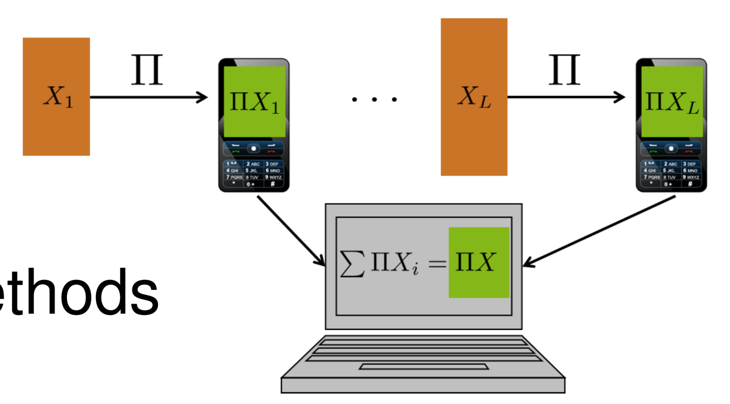
Dimensionality reduction

- Feature selection
- Variable interactions
- SNP and genome-wide data

Methodology

Data Reduction with Guaranteed Little Distortion

- Streaming
- Distributed
- Succinct representation



Sampling and projection methods

- ε-subspace embeddings
- ε-coresets
- $\Pi \in \mathbb{R}^{k \times n}, k \ll n$:

$$\forall \beta \in \mathbb{R}^d \colon (1-\varepsilon)\|X\beta\| \leq \|\Pi X\beta\| \leq (1+\varepsilon)\|X\beta\|$$

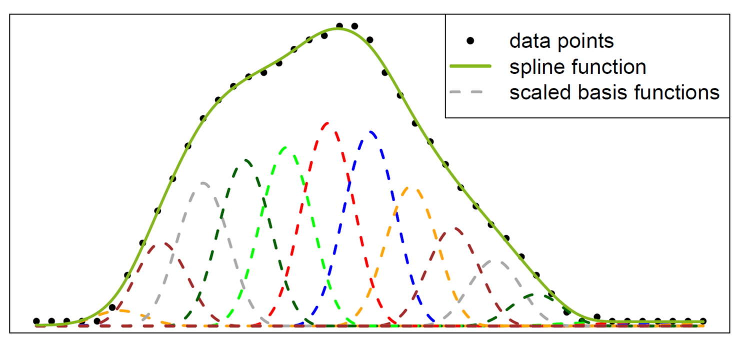
Algorithmic Principle

1. Data reduction
2. Statistical analysis

$$\begin{array}{ccc} [X, Y] & \xrightarrow{\Pi} & [\Pi X, \Pi Y] \\ \downarrow & & \downarrow \\ \pi(\beta \mid X, Y) & \approx_\varepsilon & \pi(\beta \mid \Pi X, \Pi Y) \end{array}$$

Spline Regression Model Structural Constraints

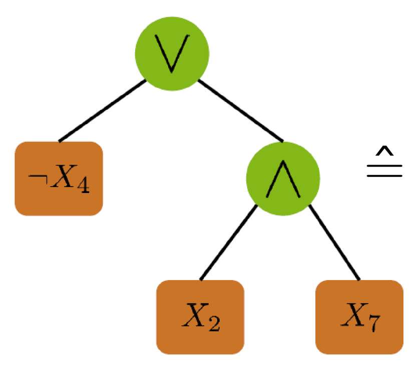
$$s(x) = \sum_{i=1}^{B} \beta_i b_i(x)$$



Logic regression

- Interactions of (binary) variables
- Variable importance measures

$$h(\mathbb{E}(Y)) = \beta_0 + \sum_{t=1}^{T} \beta_t L_t$$



$$\hat{=} \quad \neg X_4 \vee (X_2 \wedge X_7)$$

Results

Scalable Bayesian Regression

- ε-subspace embeddings preserve
  - ℓ_p-generalised linear regression, $p \in [1, \infty)$
  - Hierarchical prior models

GLMs

- Lower bounds $k \in \Omega(n / \log n)$

Poisson regression

- Statistical model relaxation yields data summary
- Close approximation of maximum likelihood estimator

B4

Logistic Regression

- Geometric parameter captures difficulty of reduction

$$\mu(X) = \sup_{\beta \in \mathbb{R}^d \setminus \{0\}} \frac{\|(X\beta)^-\|_1}{\|(X\beta)^+\|_1}$$

A2

- μ-complex data admits coreset of size $O((\mu d \log n/\varepsilon)^{O(1)})$

[Geppert, Ickstadt, Munteanu, ..., Sohler, Stat Comput 2017]
[Müller, MT, 2016]
[Munteanu, Diss, 2018]
[Molina, Munteanu, Kersting, AAAI, 2018]
[Munteanu, Schwiegelshohn, Sohler, ..., NIPS, 2018]

Spline Regression

- Able to model unimodality
- Limited number B of splines (succinct representation)
- Peak detection in IMS spectra

B1

Scalable Logic Regression

- Useful importance measures *cross-leverage-scores*:

$$cl_i = U_i^T U_{d+1}, \text{ for } i \in [d],$$

where U is an orthonormal basis for $[X, Y]$.

- Subsampling retains interesting variables
- Main effects and interactions
- Logic regression operates efficiently on reduced variable set

[Köllmann, Diss, 2016]
[Wollenberg, MT, 2016]

Lehrstuhl Informatik 2
Effiziente Algorithmen und Komplexitätstheorie

Lehrstuhl Mathematische Statistik und Biometrische Anwendungen