Optimal Subsampling Design for Big Data Regression

Torsten Reuter

Institute for Mathematical Stochastics, Otto von Guericke University Magdeburg, 39016 Magdeburg, Germany

Summary

While modern information technology allows for collecting large amounts of data (X_i, Y_i), $i = 1\dots, N$, limitations in terms of statistical methods and costs for acquiring such data may make it desirable to perform statistical analysis on a subsample only. We construct D-optimal subsample designs, based on the density of the independent variables and on the regression model.

Optimal Bounded Design Measures

Consider the general regression model

$$Y_i = \mu(X_i, \beta) + \varepsilon_i, \quad i = 1, \dots, N.$$

- N is very large.
- The X_i follow a given distribution with density f_X.
- The response function μ is known except for β.

We want to find a subsampling design ξ^* that

- has density g^* with total mass $\alpha < 1$, which is the percentage of the full data we want to sample.
- g^* is bounded from above by f_X.
- minimizes the D-criterion, which aims at minimizing the volume of the asymptotic confidence ellipsoid for the parameter β.

Figure 1 shows the density function g^* (red) of such a D-optimal design in multilinear regression, $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$, given standard normal two-dimensional data X_i with density f_X (blue).

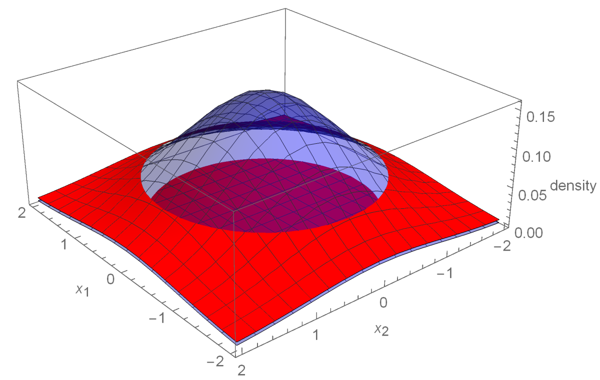

Figure 1

Easy Implementation

- Regions where $g^*(x) = f_X(x) \rightarrow$ accept all x
- Regions where $g^*(x) = 0 \rightarrow$ reject all x

Subsample Design in Quadratic Regression

Consider quadratic regression, i.e. $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$ and assume $\mathbb{E}[X_i^4] < \infty$. We determine an optimal subsampling design ξ^* with density g^* by doing the following.

- Determine the directional derivative $F(\xi^*, x)$ of the D-criterion from ξ^* in the direction of a single-point measure at point x.
- Determine the threshold c such that $P(F(\xi^*, X_i) \leq c) = \alpha$. Then $g^*(x) = f_X(x)$ when $F(\xi^*, x) \leq c$ and $g^*(x) = 0$ otherwise.

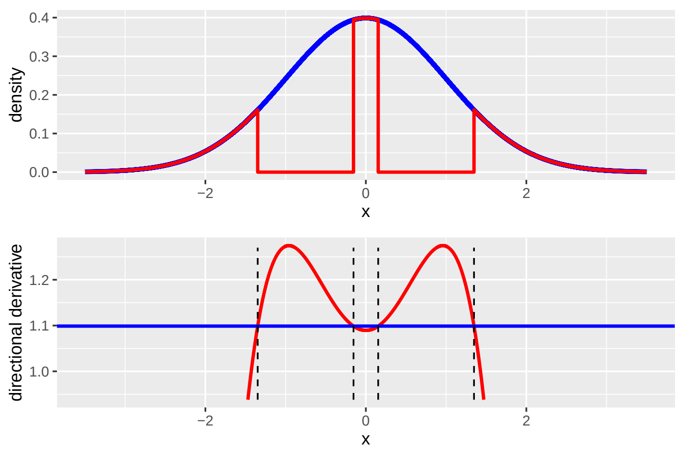

Figure 2: upper panel: D-optimal subsample density g^* (red) and full sample density f_X (blue, standard normal), lower panel: directional derivative (red) and threshold c (blue).

Discussion

So far: Subsample designs in various linear regression models.
Next step: Extend to generalized linear models, e.g.

- Poisson regression.
- Logistic regression.

Challenge: Optimal design is dependent on the unknown β. Locally optimal designs have to be obtained first before sequential or multi-stage algorithms can be developed.

Acknowledgement Joint work with Rainer Schwabe. This work is supported by the Deutsche Forschungsgemeinschaft DFG under grant 314838170, GRK 2297 MathCoRe.