D-optimal Subsampling Design for Massive Data

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

von M.Sc. Torsten Reuter
geb. am 17.04.1994 in Frankfurt am Main

genehmigt durch die Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg

Gutachter: Prof. Dr. Rainer Schwabe
Prof. Dr. Alexandra Carpentier
Prof. Dr. John Stufken

eingereicht am: 24.05.2024
Verteidigung am: 05.12.2024

Abstract

Subsampling is a central problem in big data analysis when classical statistical methods are not applicable due to computational limitations. The goal of subsampling is to select an informative subset of the full data that allows the regression parameter to be estimated as precisely as possible.

In this thesis, we study subsampling from the perspective of optimal design of experiments. The focus is on massive data with an extraordinarily large number of observations but few covariates. First, we introduce the statistical models we investigate throughout the thesis and give an overview of the relevant optimal design theory.

After the introductory chapter, we construct D-optimal subsampling designs in the setting of polynomial regression and Poisson regression in one covariate, as well as multiple linear regression in several covariates. Specific to the given setup, we present equivalence theorems based on convex optimization that establish a representation of the support of the D-optimal subsampling design. We make use of fundamental concepts from optimal design theory and an equivalence theorem from constrained convex optimization. We study theoretical properties of the constructed D-optimal subsampling design. For the given models, location-scale transformations of the covariate and the simultaneous transformation of the D-optimal subsampling design are investigated in order to extend the results for standardized covariates to general covariates.

The obtained D-optimal subsampling designs provide simple rules for whether to accept or reject a data point. Throughout the thesis we propose methods of implementation. We study these methods theoretically through efficiency considerations. For multiple linear regression, we present a simulation study comparing our method to others.

Zusammenfassung

Subsampling ist ein zentrales Problem der Big-Data-Analyse, wenn klassische statistische Methoden aufgrund technischer Einschränkungen nicht anwendbar sind. Das Ziel von Subsampling ist es, eine informative Teilmenge der Gesamtdaten auszuwählen, die eine möglichst präzise Schätzung des Regressionsparameters erlaubt.

In dieser Dissertation wird Subsampling aus der Perspektive der optimalen Versuchsplanung betrachtet. Der Schwerpunkt liegt dabei auf „massive data“ mit außergewöhnlich vielen Beobachtungen, aber nur wenigen Kovariablen. Zunächst werden die statistischen Modelle, die im Verlauf der Dissertation untersucht werden, vorgestellt sowie ein Überblick über die relevante Theorie der optimalen Versuchsplanung gegeben.

Nach dem einleitenden Kapitel werden D-optimale Subsampling Designs für polynomiale und Poisson Regression in einer Kovariablen sowie für multiple lineare Regression in mehreren Kovariablen konstruiert. Für das jeweilige Modell werden auf konvexer Optimierung basierende Äquivalenzsätze präsentiert, die eine Darstellung des Trägers des D-optimalen Subsampling Designs liefern. Dabei werden grundlegenden Konzepten aus der optimalen Versuchsplanung und ein Äquivalenztheorem aus der konvexen Optimierung unter Nebenbedingungen verwendet. Theoretische Eigenschaften der D-optimalen Subsampling Designs wwerden untersucht. Für die gegebenen Modelle werden Lokations-Skalen-Transformationen der Kovariable und die gleichzeitige Transformation des D-optimalen Subsampling Designs betrachtet, um die Ergebnisse für standardisierte Kovariablen auf allgemeine Kovariablen zu übertragen.

D-optimalen Subsampling Designs bieten einfache Regeln zur Annahme oder Ablehnung eines Datenpunktes. Methoden zur Implementierung von D-optimalen Subsampling Designs werden vorgeschlagen und theoretisch durch Effizienzbetrachtungen untersucht. Für die multiple lineare Regression wird eine, zu anderen Methoden vergleichende, Simulationsstudie präsentiert.

Declaration of Honor

I hereby declare that I produced this thesis without prohibited assistance and that all sources of information that were used in producing this thesis, including my own publications, have been clearly marked and referenced. In particular I have not willfully:

- Fabricated data or ignored or removed undesired results.
- Misused statistical methods with the aim of drawing other conclusions than those warranted by the available data.
- Plagiarized data or publications or presented them in a distorted way.

I know that violations of copyright may lead to injunction and damage claims from the author or prosecution by the law enforcement authorities. This work has not previously been submitted as a doctoral thesis in the same or a similar form in Germany or in any other country. It hast not previously been published as a whole. The contents of Chapter 2 have been published in Reuter and Schwabe (2023a). The contents of Chapters 3 and 4 have been published as electronic preprints, see Reuter and Schwabe (2023b) and Reuter and Schwabe (2024) respectively.

Torsten Reuter	11.02.2025
	Date

Contents

Chapter 1

Introduction and General Considerations

In this introduction, we first establish the relevant concepts of optimal design of experiments, also known as 'design of experiments' or 'experimental design' in the literature. We then provide an overview of the field of subsampling. The core idea of optimal design is that a practitioner must establish a method of data collection when conducting an experiment. Specifically, an experimenter chooses which experimental settings to use and how many observations to take at each of these experimental settings. This allocation of observations to experimental settings is called a design. Typical applications range from the choice of dosages in pharmacology to the selection of test items in educational testing or the choice of experimental conditions in engineering (see e. g. Dean et al. (2015) or Berger and Wong (2005) for an overview on applications). The natural goal when deciding between designs is to gain as much information as possible about the unknown parameters. Ronald Fisher first formalized this information through the so called Fisher information matrix. The information matrix innately quantifies the quality of the design as its inverse corresponds to the covariance matrix of the estimator of the unknown parameter. Finding a design that maximizes suitable functions of the information matrix, called optimality criteria, is the main focus of the optimal design theory (Silvey, 1980, Section 2.2) and the following sections in this introduction. The field of optimal design has by many accounts seen its first work explicitly on an experimental design in a paper by Smith (1918). The book by Fisher (1935) and the article by Kiefer (1959) are predominantly named as the foundational works of the field.

Subsampling as a branch of optimal design arose from finite population sampling (see e. g. Wynn, 1977). The field has gained popularity in recent years with the rapid growth of very large data sets. Performing statistical analysis on the full data may no longer be viable because of computational limitations. Consequently, data reduction is a key factor in big data analysis. The size of a data set is usually determined by the number of observations, covariates, and response variables. In the present thesis we consider only the case of a single response variable. We speak of massive data when the number of observations is extraordinarily large, while the number of covariates is relatively small. This is the focus of this thesis. The goal of subsampling is to extract a subset of observations from the full data that contains as much information as possible for estimating the unknown parameter. The key concept of optimal design, maximizing some suitable function of the information matrix, may be adopted to identify the most informative subset of the full data. Earlier studies focus mostly on probabilistic subsampling schemes, where observations are sampled according to some sampling distribution (see e. g. Ma et al., 2014). In the work of Wang et al. (2019) the information-based optimal subdata selection (IBOSS) method is introduced. There, the subset of the full data is formed based on deterministic rules to insure only the most informative observations are selected. Since then many works have used similar deterministic subsampling strategies (e. g. Cheng et al. (2020), Deldossi and Tommasi (2021), Wang et al. (2021)). This is also the main focus of the present thesis. Application of subsampling to flight data can be found in Wang et al. (2019). He et al. (2024) apply subsampling to financial data. Liu et al. (2026, forthcoming) use structural protein data to study their method. Recent reviews on subsampling are Yao and Wang (2021) and Yu et al. (2024). We provide more literature in the introductory Sections 2.1, 3.1, and 4.1 of the corresponding chapters.

In Section 1.1 we present the relevant theory on optimal design in the classical setup when there are no constraints on the choice of the experimental settings. There, we consider the linear model with normality assumption. These concepts are extended to the generalized linear model in Section 1.2. Then, we point out the distinct assumptions we make for subsampling design in Section 1.3. More specific introductions to the following chapters are also given in Section 1.3. In Section 1.4 we make general and concluding remarks about the whole thesis. The Chapters 2, 3 and 4 constitute the main part of this thesis. They have been published as Reuter and Schwabe (2023a), Reuter and Schwabe (2023b) and Reuter and Schwabe (2024), respectively. While the layout and formatting have been revised for coherence within this thesis, the content remains unchanged otherwise, except for minor editorial changes. However, the respective abstracts have been removed and these chapters

now share a common bibliography with the remainder of the thesis.

1.1 Theory of Optimal Design under a Linear Model

This section gives a brief overview of the concepts from optimal design theory that we make use of in Chapters 2 and 3. We refer to the textbooks by Silvey (1980) and, primarily, Pukelsheim (1993).

We consider the situation of data $(\mathbf{x}_i, y_i)$, where $i = 1, \ldots, n$. The component y_i is the value of the response variable Y_i. The covariate value $\mathbf{x}_i$, or experimental setting, may be chosen by the experimenter from the design region $\mathcal{X}$, i. e. $\mathbf{x}_i \in \mathcal{X}$. Suppose that the dependence of the response variable Y_i on the covariate $\mathbf{x}_i$ is given by the linear model

$$Y_i = \mathbf{f}(\mathbf{x}_i)^\top \boldsymbol{\beta} + \varepsilon_i, \tag{1.1}$$

where $\mathbf{f}$ is a p-dimensional vector of regression functions and $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter to be estimated. We assume the random errors ε_i to be independent and normally distributed with mean 0 and equal variance $\sigma_\varepsilon^2 > 0$. In particular, we focus on the two models that correspond to Chapters 2 and 3. We present them in the following two examples.

Example 1.1 (Polynomial regression in one covariate)**.** We consider the linear model with normality assumption (1.1) for a covariate $x \in \mathcal{X} \subset \mathbb{R}$, when $\mathbf{f}(x) = (1, x, x^2, \ldots, x^q)^\top$. We speak of polynomial regression of degree q if q is the largest power of the monomials in x contained in the regression function $\mathbf{f}$. Hence, we assume

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_q x_i^q + \varepsilon_i,$$

where the error terms $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ are independent and homoscedastic. The $p = (q+1)$-dimensional parameter vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_q)^\top$ is to be estimated.

Example 1.2 (Multiple linear regression)**.** We assume the linear model (1.1) for $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{X} \subset \mathbb{R}^d$, when $\mathbf{f}(\mathbf{x}) = (1, x_1, \ldots, x_d)^\top$. More specifically, we assume that the response variable Y_i depends on the covariate $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^\top$ through

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id} + \varepsilon_i,$$

where, again, $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ are independent, homoscedastic random errors. The $p = (d+1)$-dimensional parameter vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)^\top$ is to be estimated.

Usually the experimenter is able to control the frequency $\xi(\mathbf{x})$ how often specific values of $\mathbf{x}$ are applied in the experiment. This distribution ξ over the possible experimental settings is called a design. The set $\mathcal{X}$ of possible experimental settings is called the design region.

We relax the practically essential requirement that $n\xi(\mathbf{x})$ is a nonnegative integer to allow for continuous analytical tools. This was suggested by Kiefer (1959) and is usually referred to as an approximate design in the literature. Since we only use approximate designs in this thesis they are simply referred to as designs.

Definition 1.3 (Design)**.** A probability distribution ξ which assigns all of its mass on a finite number of points on the design region $\mathcal{X}$ is called a design.

For data $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$ consider the least squares estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{F}^\top \boldsymbol{F})^{-1} \boldsymbol{F}^\top \mathbf{y}$, where $\boldsymbol{F} = (\mathbf{f}(\mathbf{x}_1), \ldots, \mathbf{f}(\mathbf{x}_n))^\top$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top$. The existence of $\hat{\boldsymbol{\beta}}$ is guaranteed if $\boldsymbol{F}$ has full column rank. Then, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is normally distributed with mean zero and covariance matrix proportional to $(\boldsymbol{F}^\top \boldsymbol{F})^{-1}$. Naturally, the goal of an experimenter is to maximize the information matrix $\boldsymbol{F}^\top \boldsymbol{F} = \sum_{i=1}^n \mathbf{f}(\mathbf{x}_i)\mathbf{f}(\mathbf{x}_i)^\top$ related to the data. For a design ξ this is formalized by the information matrix.

Definition 1.4 (Information matrix)**.** The information matrix of a design ξ is the $p \times p$ matrix defined by

$$\mathbf{M}(\xi) = \int \mathbf{f}(\mathbf{x})\mathbf{f}(\mathbf{x})^\top \xi(\mathrm{d}\mathbf{x}),$$

where we assume all elements of $\mathbf{M}(\xi)$ to be finite.

Note that $\mathbf{M}(\xi) \in \mathrm{NND}(p)$ for any design ξ, where $\mathrm{NND}(p)$ denotes the closed cone of nonnegative definite $p \times p$ matrices. The goal of experimental design is to maximize the information matrix. However, it is not possible, outside of degenerate cases, to find a design ξ^* which is best in the strong sense that $\mathbf{M}(\xi^*) - \mathbf{M}(\xi)$ is nonnegative for all ξ (see Silvey, 1980, Section 1.3). Instead, we look at functions that measure the "largeness" of the information matrix, called optimality criteria.

Definition 1.5 (Optimality criterion and optimal design)**.** Let Φ be a function from the closed cone of nonnegative definite $p \times p$ matrices into the union of the real line and $\{-\infty\}$, $\Phi : \mathrm{NND}(p) \to \mathbb{R} \cup \{-\infty\}$, such that Φ is isotonic with respect to the Loewner ordering, i. e. $\Phi(\mathbf{M}_2) \geq \Phi(\mathbf{M}_1)$ if $\mathbf{M}_2 - \mathbf{M}_1 \in \mathrm{NND}(p)$. Then, Φ is an optimality criterion.

A design ξ^* that maximizes $\Phi(\mathbf{M}(\xi))$ is said to be Φ-optimal.

The support of a Φ-optimal design ξ^* is denoted by $\mathcal{X}^*$. The optimality criterion most used in practice is the D-criterion (determinant criterion). We pay special attention

to D-optimality, as it is the main focus of all following chapters. We present it here in its homogeneous form. An optimality criterion Φ is said to be homogeneous when $\Phi(\lambda \mathbf{M}) = \lambda \Phi(\mathbf{M})$ for every $\mathbf{M} \in \mathrm{NND}(p)$ and $\lambda > 0$.

Definition 1.6 (D-optimality)**.** The design ξ^* that maximizes the D-optimality criterion

$$\Phi_D(\mathbf{M}(\xi)) = \det(\mathbf{M}(\xi))^{1/p}$$

is called D-optimal.

The D-optimality criterion may appear in equivalent forms such as $\log(\det(\mathbf{M}(\xi)))$ or $\det(\mathbf{M}(\xi))$ when convenient. The confidence ellipsoid of $\boldsymbol{\beta}$, given the least squares estimator $\hat{\boldsymbol{\beta}}$, has volume inversely proportional to $\det(\mathbf{M}(\xi))^{1/2}$. Hence a large value of $\det(\mathbf{M}(\xi))$ secures a small volume of the confidence ellipsoid (see Pukelsheim, 1993, Chapter 6.2).

We assume that β_0 is the intercept parameter related to the constant term of the model and denote $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_{p-1})^\top$. In practice, the main interest often lies on the parameters $\boldsymbol{\beta}_1$ only, rather than the full parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$. We assume $\mathbf{M}(\xi)$ to be nonsingular. Then, the covariance matrix $\mathbf{M}(\xi)^{-1}$ of $\hat{\boldsymbol{\beta}}$ can be written as a block matrix as

$$\mathbf{M}(\xi)^{-1} = \begin{pmatrix} c_0(\xi) & \boldsymbol{c}_1(\xi)^\top \\ \boldsymbol{c}_1(\xi) & \mathcal{C}_1(\xi) \end{pmatrix}.$$

Then, an experimenter may only be interested in the $(p-1) \times (p-1)$ covariance matrix $\mathcal{C}_1(\xi)$ of $\hat{\boldsymbol{\beta}}_1$.

Definition 1.7 (D_1-optimality)**.** The design ξ^* that maximizes the D_1-optimality criterion

$$\Phi_{D_1}(\mathbf{M}(\xi)) = \begin{cases} \det(\mathcal{C}_1(\xi))^{-1/(p-1)} & \text{if } \mathbf{M}(\xi) \text{ is nonsingular,} \\ 0 & \text{if } \mathbf{M}(\xi) \text{ is singular} \end{cases}$$

is called D_1-optimal.

Remark 1.8. A well-known result establishes the equivalence of the two criteria. In a model containing a constant term, e. g. Example 1.2, a design ξ^* is D_1-optimal if and only if ξ^* is D-optimal. For a proof see the lecture notes by Schwabe (1996, Theorem 3.3).

Another criterion that is referred to in this thesis is the A-criterion (average-variance criterion). It aims to minimize the average variance $p^{-1} \sum_{j=0}^{p-1} \mathrm{Var}(\hat{\beta}_j)$ of $\hat{\boldsymbol{\beta}}$.

Definition 1.9 (A-optimality)**.** The design ξ^* that maximizes the A-optimality criterion

$$\Phi_A(\mathbf{M}(\xi)) = \begin{cases} \text{trace}(\mathbf{M}(\xi)^{-1})^{-1} & \text{if } \mathbf{M}(\xi) \text{ is nonsingular,} \\ 0 & \text{if } \mathbf{M}(\xi) \text{ is singular} \end{cases}$$

is called A-optimal.

Definition 1.10 (A_1-optimality)**.** The design ξ^* that maximizes the A_1-optimality criterion

$$\Phi_{A_1}(\mathbf{M}(\xi)) = \begin{cases} \text{trace}(\mathcal{C}_1(\xi))^{-1} & \text{if } \mathbf{M}(\xi) \text{ is nonsingular,} \\ 0 & \text{if } \mathbf{M}(\xi) \text{ is singular} \end{cases}$$

is called A_1-optimal.

The D-, D_1-, A- and A_1-criteria share some desirable properties, including being isotonic with respect to the Loewner ordering as required by Definition 1.5.

Lemma 1.11. *Let Φ be the D-, D_1-, A- or A_1-criterion. Then Φ satisfies the following.*

(i) Φ is strictly concave and strictly isotonic relative to the Loewner ordering on $\text{NND}(p)$.

(ii) If there exists a design with nonsingular information matrix, then the matrix $\mathbf{M}(\xi^*)$ is nonsingular and unique.

For a proof of (i) see Pukelsheim (1993, Chapter 6.13). Statement (ii) follows immediately from the fact that Φ is strictly concave at $\mathbf{M}(\xi^*)$ (see Pronzato, 2013). Note that while the optimal matrix $\mathbf{M}(\xi^*)$ is unique, the Φ-optimal design might not be. To evaluate the quality of design ξ, it is common to study its efficiency. The efficiency is a number between zero and one that measures the performance of ξ in the sense of some optimality criterion Φ, relative to the Φ-optimal design ξ^*.

Definition 1.12 (Φ-efficiency)**.** Let ξ^* be the Φ-optimal design for a homogeneous optimality criterion Φ. Then, the Φ-efficiency of a design ξ is defined as

$$\text{eff}_\Phi(\xi) = \frac{\Phi(\mathbf{M}(\xi))}{\Phi(\mathbf{M}(\xi^*))} \,.$$

Next, we introduce the main tool we use for convex optimization of the design criterion Φ. The directional derivative of Φ at a design ξ with non-singular information matrix $\mathbf{M}(\xi)$ in the direction of a design η is defined by

$$F_\Phi(\xi, \eta) = \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \left(\Phi(\mathbf{M}((1-\epsilon)\xi + \epsilon\eta)) - \Phi(\mathbf{M}(\xi)) \right).$$

Here, we use the equivalent concave form $\log(\det(\mathbf{M}(\xi)))$ of the D-criterion to ensure a simpler directional derivative. For the D-criterion $\log(\det(\mathbf{M}(\xi)))$ we find $F_D(\xi, \eta) = \text{trace}(\mathbf{M}(\xi)^{-1}\mathbf{M}(\eta)) - p$ (compare Silvey, 1980, Example 3.8). Equivalently, one may consider only the essential part of the directional derivative $\text{trace}(\mathbf{M}(\xi)^{-1}\mathbf{M}(\eta))$. As will become clear in Theorem 1.14, we are particularly interested in the directional derivative in the direction of a one-point design $\xi_{\mathbf{x}}$ that puts all its mass on one point $\mathbf{x}$.

Definition 1.13 (Sensitivity function)**.** The sensitivity function ψ w.r.t. the D-criterion is defined by

$$\psi(\mathbf{x}, \xi) = \text{trace}(\mathbf{M}(\xi)^{-1}\mathbf{M}(\xi_{\mathbf{x}})) = \mathbf{f}(\mathbf{x})^\top \mathbf{M}(\xi)^{-1}\mathbf{f}(\mathbf{x}).$$

One of the most central results in the theory of optimal design of experiments is an equivalence theorem by Kiefer and Wolfowitz (1960). They establish the equivalence between maximization of the D-criterion and minimization of the maximum of the sensitivity function.

Theorem 1.14 (Kiefer-Wolfowitz equivalence theorem)**.** *A design ξ^* is D-optimal if and only if*

$$\psi(\mathbf{x}, \xi^*) \leq p \text{ for all } \mathbf{x} \in \mathcal{X}.$$

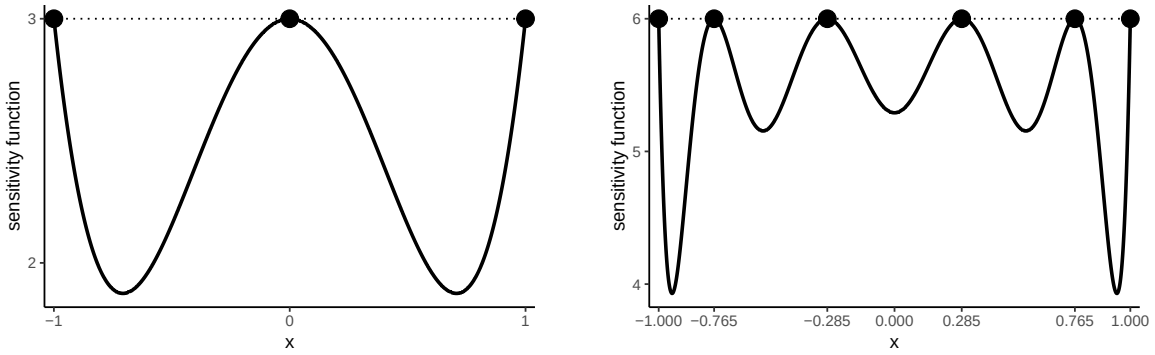
In particular, equality is achieved for the support points of ξ^*.

For a proof see Pukelsheim (1993, chapter 9.4). Similar equivalence theorems exist for other optimality criteria as well, e. g. for A-optimality see Pukelsheim (1993, chapter 9.7). We illustrate the above concepts using the two examples presented earlier. The D-optimal designs presented here directly relate to Chapters 2 and 3, as they typically represent the limit of the D-optimal subsampling designs when the subsampling proportion tends to zero.

Example 1.15 (Polynomial regression, continuation of Example 1.1)**.** We return to the example of polynomial regression of degree q in one covariate. Let the design region be $\mathcal{X} = [-1, 1]$. Recall that the regression function is $\mathbf{f}(x) = (1, x, x^2, \ldots, x^q)^\top$. We assume that the information matrix $\mathbf{M}(\xi)$, and thus $\mathbf{M}(\xi)^{-1}$, is positive definite. Then, the sensitivity function $\psi(x, \xi)$ is a polynomial of even degree $2q$ with positive leading term. In particular, $\psi(x, \xi)$ is symmetric around zero, when the design ξ is invariant w.r.t. the sign change, $\xi(x) = \xi(-x)$. These invariant designs form an essentially complete class (see Pukelsheim, 1993, chapter 13.1) for any concave optimality criteria invariant w.r.t. the sign change, e. g. A- and D-optimality. This allows us to construct D-optimal designs with the help of the optimality condition in Theorem 1.14. The support $\mathcal{X}^*$ of the D-optimal design ξ^* only contains the minimally required p points. This is because $\psi(x, \xi^*)$ has at most $p - 2$ local maxima in the interior of $\mathcal{X}$ and $\psi(x, \xi^*) = p$ for any $x \in \mathcal{X}^*$. The support points are the

$p-2$ points where the local maxima of $\psi(x, \xi^*)$ are attained and the boundary points ± 1. The p support points of ξ^* have equal weight $1/p$.

Here, we give examples for D-optimal designs for polynomial regression of degree 2 and 5, as presented by Pukelsheim (1993, chapter 9.5). When $q = 2$, the D-optimal design ξ^* is equally supported with weight $1/3$ on the points $\pm 1, 0$. When $q = 5$, the D-optimal design ξ^* is equally supported with weight $1/6$ on the points $\pm 1, \pm 0.765, \pm 0.285$. In Figure 1.1 we show the support points of the D-optimal designs alongside the corresponding sensitivity functions. The horizontal dotted line represents the dimension p of the parameter $\boldsymbol{\beta}$, which is simultaneously the upper bound of the sensitivity function in the condition in Theorem 1.14. The large points depict the support points of ξ^* together with their value of the sensitivity function, which is equal to p for all support points. Note that in general other optimality



(a) polynomial regression of degree $q = 2$

(b) polynomial regression of degree $q = 5$

Figure 1.1: Sensitivity function for the D-optimal design for polynomial regression of degree two (left) and degree five (right)

criteria give different optimal designs. The A-optimal design is equal to the D-optimal design for simple linear regression, i. e. $q = 1$. For $q = 2$ the A-optimal design shares the same support points $-1, 0, 1$ with the D-optimal design, but assigns weights $0.25, 0.5, 0.25$ to them. For $q \geq 3$ the support points of the A-optimal design differ from those of the D-optimal design (for $q \leq 10$, see Pukelsheim, 1993, chapters 9.5 and 9.9).

Example 1.16. (Multiple linear regression, continuation of Example 1.2) We examine the case of multiple linear regression, with d covariates in $[-1, 1]$, i. e. the design region is $\mathcal{X} = [-1, 1]^d$. A D-optimal design ξ^* places equal weight 2^{-d} on all the vertices of the hypercube (compare Kiefer, 1960). Thus $\mathcal{X}^* = \{-1, 1\}^d$ is the support of ξ^*. The information matrix of ξ^* is equal to the $p \times p$ identity matrix $\mathbb{I}_p$. The sensitivity function is then given by $\psi(\mathbf{x}, \xi^*) = 1 + \mathbf{x}^\top \mathbf{x}$. We see that the optimality condition in Theorem 1.14

is indeed satisfied. Note that for $d \geq 3$ the D-optimal design is not unique and can be reduced to a fractional factorial design with fewer support points (see Pukelsheim, 1993, chapter 15.11). Here, we present the D-optimal design when the entire parameter $\boldsymbol{\beta}$ is to be estimated. By Remark 1.8, ξ^* is still optimal when only the slope parameters $\boldsymbol{\beta}_1$ are to be estimated. The design ξ^* is also A- and A_1-optimal.

1.2 Theory of Optimal Design under a Generalized Linear Model

We now consider the generalized linear model (GLM) and point out the relevant optimal design theory we use in Chapter 4. In line with before, we consider data $(\mathbf{x}_i, y_i)$, where $i = 1, \ldots, n$. The value of the response variable Y_i is denoted by y_i, and the covariate by $\mathbf{x}_i$. We denote the variance $\mathrm{Var}(Y_i)$ of the response variable by $\sigma^2(\mathbf{x}_i, \boldsymbol{\beta})$. We assume the response variable Y_i to follow a distribution from the natural exponential family and the mean $\mu(\mathbf{x}_i, \boldsymbol{\beta})$ of Y_i to depend on the linear predictor $\mathbf{f}(\mathbf{x}_i)^\top \boldsymbol{\beta}$ through the inverse of a bijective link function g. The p-dimensional vector of regression functions is denoted by $\mathbf{f}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter to be estimated. Specifically, we assume that the mean of Y_i depends on $\mathbf{x}_i$ through

$$\mathrm{E}(Y_i) = \mu(\mathbf{x}_i, \boldsymbol{\beta}) = g^{-1}(\mathbf{f}(\mathbf{x}_i)^\top \boldsymbol{\beta}). \tag{1.2}$$

The information matrix is again defined with the goal to quantify the information of a given design ξ. In the GLM setting, a design ξ^* is only locally Φ-optimal at parameter $\boldsymbol{\beta}$, rather than Φ-optimal over the entire parameter space. This results from the dependence of the information matrix on $\boldsymbol{\beta}$. We assume the inverse link function g^{-1} to be differentiable.

Definition 1.17 (Information matrix (GLM))**.** The intensity function λ is defined as $\lambda(\mathbf{x}, \boldsymbol{\beta}) = (g^{-1})'(\mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta})^2 / \sigma^2(\mathbf{x}, \boldsymbol{\beta})$. Then, the information matrix of a design ξ is the $p \times p$ matrix defined by

$$\mathbf{M}(\xi, \boldsymbol{\beta}) = \int \lambda(\mathbf{x}, \boldsymbol{\beta}) \mathbf{f}(\mathbf{x}) \mathbf{f}(\mathbf{x})^\top \xi(\mathrm{d}\mathbf{x}),$$

where we assume all elements of $\mathbf{M}(\xi, \boldsymbol{\beta})$ to be finite.

Again, the information matrix is nonnegative definite for any design ξ and parameter $\boldsymbol{\beta}$.

Remark 1.18. In the previous section the information matrix for the linear model is defined by $\mathbf{M}(\xi) = \int \mathbf{f}(\mathbf{x}) \mathbf{f}(\mathbf{x})^\top \xi(\mathrm{d}\mathbf{x})$. Definition 1.17 produces the same information matrix for the special case of a linear model, up the variance $\sigma^2(\mathbf{x}, \boldsymbol{\beta})$ of Y_i. For the linear model we have

$\lambda(\mathbf{x}, \boldsymbol{\beta}) = 1/\sigma_\varepsilon^2$, because we are in the special case that the link function $g(\mu) = \mu$ is the identity. However, σ_ε^2 is negligible for optimization here, as it is constant in the linear model.

We consider Poisson regression as an example, as it is the subject of Chapter 4.

Example 1.19 (Poisson regression in one covariate)**.** We assume that the response variable Y_i follows a Poisson distribution with rate $\mathrm{E}(Y_i)$ dependent on the covariate $x_i \in \mathcal{X} \subset \mathbb{R}$ via a log-link and a linear component $\beta_0 + \beta_1 x_i$. For the model (1.2) we have $g^{-1}(\mu) = \exp(\mu)$ and $\mathbf{f}(x) = (1, x)^\top$ such that $\mathrm{E}(Y_i) = \exp(\beta_0 + \beta_1 x)$. The parameter vector $\boldsymbol{\beta}$ to be estimated is of dimension $p = 2$. Unlike in the previous examples, the variance of the response variable Y_i depends on x_i. Specifically, $\sigma^2(x, \boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x)$. We find for the intensity function $\lambda(x, \boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x)$. Subsequently, the information matrix of a design ξ is given by $\mathbf{M}(\xi, \boldsymbol{\beta}) = \int \exp(\beta_0 + \beta_1 x)\mathbf{f}(x)\mathbf{f}(x)^\top \xi(\mathrm{d}x)$.

Let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimator of $\boldsymbol{\beta}$. Under regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically normally distributed with mean zero and covariance matrix $\mathbf{M}(\xi, \boldsymbol{\beta})^{-1}$. This again gives motivation to maximize $\mathbf{M}(\xi, \boldsymbol{\beta})$ in the sense of an optimality criterion. Definition 1.3 of a design as well as the definitions on the optimality criteria and efficiency (Definitions 1.5, 1.6, 1.7, 1.9, 1.10 and 1.12) stay as they are, except that a design ξ^* is now called locally Φ-optimal at $\boldsymbol{\beta}$, when ξ^* maximizes $\Phi(\mathbf{M}(\xi, \boldsymbol{\beta}))$. Lemma 1.11 still holds. The sensitivity function ψ w.r.t. the D-criterion is defined similarly as in Definition 1.13 by $\psi(\mathbf{x}, \xi, \boldsymbol{\beta}) = \mathrm{trace}(\mathbf{M}(\xi, \boldsymbol{\beta})^{-1}\mathbf{M}(\xi_{\mathbf{x}}, \boldsymbol{\beta}))$, only with the additional notation to express the dependence on the parameter $\boldsymbol{\beta}$. Here the sensitivity function reduces to

$$\psi(\mathbf{x}, \xi, \boldsymbol{\beta}) = \lambda(\mathbf{x}, \boldsymbol{\beta})\mathbf{f}(\mathbf{x})^\top \mathbf{M}(\xi, \boldsymbol{\beta})^{-1}\mathbf{f}(\mathbf{x})^\top.$$

The Kiefer-Wolfowitz equivalence theorem can be extended to local D-optimality in GLMs. Even though Theorem 1.14 does not change except for the additional dependence on $\boldsymbol{\beta}$ of the sensitivity function ψ, we present it once more for the sake of completeness. A proof can be found in the textbook by Fedorov (1972, Theorem 2.2.1).

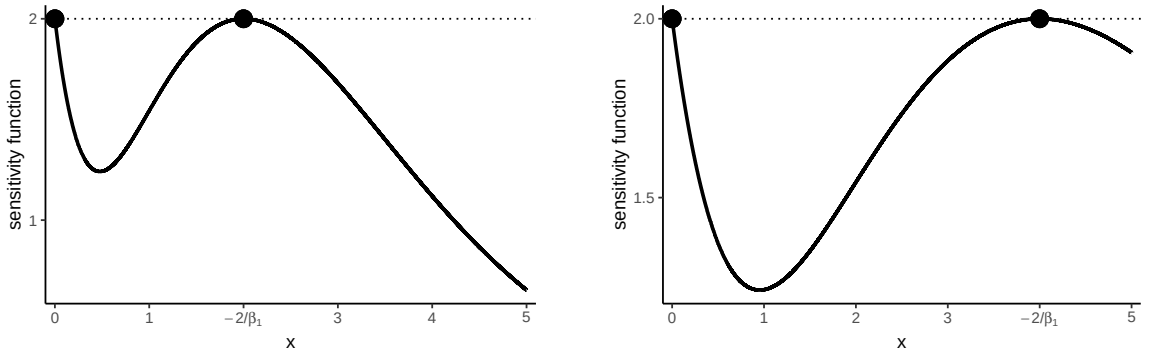
Theorem 1.20 (Equivalence theorem)**.** *A design ξ^* is locally D-optimal if and only if*

$$\psi(\mathbf{x}, \xi^*, \boldsymbol{\beta}) \leq p \text{ for all } \mathbf{x} \in \mathcal{X}.$$

In particular, equality is achieved for the support points of ξ^*.

Example 1.21. (Poisson regression, continuation of Example 1.19) We consider Poisson regression as defined above. In applications the rate $\mathrm{E}[Y_i]$ more commonly decreases for

increasing x_i, thus we assume $\beta_1 < 0$. Let $\mathcal{X} = [x_{\min}, x_{\max}]$, where $x_{\min} < x_{\max}$. Recall that $\mathbf{M}(\xi, \boldsymbol{\beta}) = \int \exp(\beta_0 + \beta_1 x)\mathbf{f}(x)\mathbf{f}(x)^\top \xi(\mathrm{d}x)$. Subsequently, the sensitivity function is given by $\psi(x, \xi, \boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x)\mathbf{f}(x)^\top \mathbf{M}(\xi, \boldsymbol{\beta})^{-1}\mathbf{f}(x)$, the product of an exponential function and a quadratic polynomial in x. Russell et al. (2009) show that the D-optimal design is equally supported on $x_{\min}$ and $\min(x_{\min} - 2/\beta_1, x_{\max})$. In Figure 1.2 we present the support points of the D-optimal design on $\mathcal{X} = [0, \infty)$ for two different values of the slope parameter β_1. Here, as $x_{\max} = \infty$, ξ^* is equally supported on $x_{\min}$ and $x_{\min} - 2/\beta_1$ for any $\beta_1 < 0$. Alongside the support points of ξ^*, depicted by the large points, the corresponding sensitivity function is shown. The horizontal dotted line represents the upper bound $p = 2$ of the sensitivity function in the condition in Theorem 1.20. Note that the graphics on the



(a) $\beta_1 = -1$ (b) $\beta_1 = -0.5$

Figure 1.2: Sensitivity function for the locally D-optimal design for Poisson regression on $\mathcal{X} = [0, \infty)$ for $\beta_1 = -1$ (left) and $\beta_1 = -0.5$ (right)

left and right hand side of Figure 1.2 only differ by scaling of the x-axis. Such equivariance considerations will be studied in detail in the following chapters. For equivariance in the case of Poisson regression specifically, see Theorem 4.6.

1.3 Subsampling Design

We now shift our attention to the main subject of this thesis: subsampling. The first key deviation from the above sections concerns the model. We consider the full data $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, where y_i is the value of the response variable Y_i as before. Additionally, we now assume that the covariate value $\mathbf{x}_i$ is generated by a d-dimensional continuous random variable $\boldsymbol{X}_i$ to model the distribution of the full data. The density of $\boldsymbol{X}_i$ is denoted by $f_{\boldsymbol{X}}$. The dependence of the response variable Y_i on the linear predictor $\mathbf{f}(\boldsymbol{X}_i)^\top \boldsymbol{\beta}$ is

modeled through the conditional mean that is described by a bijective link function g such, that analogous to Section 1.2,

$$\mathrm{E}[Y_i|\boldsymbol{X}_i] = \mu(\boldsymbol{X}_i, \boldsymbol{\beta}) = g^{-1}(\mathbf{f}(\boldsymbol{X}_i)^\top \boldsymbol{\beta}). \tag{1.3}$$

We aim to select a fixed percentage $\alpha \in (0, 1)$, called subsampling proportion, of the full data $(\mathbf{x}_i, y_i), i = 1, \ldots, n$. The goal is to find the subsample that yields the most precise estimation of the parameter $\boldsymbol{\beta} \in \mathbb{R}^p$. In the linear model, the link function g and its inverse g^{-1} are the identity function $g(\mu) = \mu$. Then $\mathrm{E}[Y_i|\boldsymbol{X}_i] = \mathbf{f}(\boldsymbol{X}_i)^\top \boldsymbol{\beta}$.

A key assumption in the following chapters is that the density $f_{\boldsymbol{X}}$ of the covariate is known. This is made to obtain analytical results for the (locally) D-optimal subsampling design. However, subsampling methods that do not require the density of the covariate to be known are studied as well. A continuous design that describes subsampling from the covariate $\boldsymbol{X}_i$ should be bounded from above by the distribution of $\boldsymbol{X}_i$. To ensure the boundedness as well as the subsampling proportion α, we define a subsampling design as follows.

Definition 1.22 (Subsampling design)**.** Given a subsampling proportion $\alpha \in (0, 1)$, a continuous distribution ξ with density f_ξ on the design region $\mathcal{X}$ is called subsampling design with respect to the distribution of $\boldsymbol{X}_i$ if and only if it satisfies

(*i*) $\int_{\mathcal{X}} \xi(\mathrm{d}\mathbf{x}) = \alpha$,

(*ii*) $f_\xi(\mathbf{x}) \leq f_{\boldsymbol{X}}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$.

Other relevant definitions of the information matrix, optimality criteria, and sensitivity function are inherited from Sections 1.1 and 1.2 by replacing the discrete measure ξ by a continuous one with density f_ξ.

Early studies on such constrained designs include Wynn (1977), Fedorov (1989) and Pronzato (2004). Miller (2002, Chapter 4.4) gives examples for D- and c-optimal constrained designs for polynomial regression in one covariate for up to degree four, where the density is bounded from above and below, though the lower bound may be zero, thus including subsampling designs as defined above.

Theorem 1.14 was used to verify the D-optimality of a given design in the classical theory of Section 1.1. Similarly we use an equivalence theorem to help us construct subsampling designs and verify their D-optimality. The following Theorem 1.23 is adapted to our setting from Sahm and Schwabe (2001, Corollary 1).

Theorem 1.23 (General equivalence theorem for constrained designs)**.**
Let $\mathrm{P}(\psi(\boldsymbol{X}_i, \xi, \boldsymbol{\beta}) = s) = 0$ *for any* ξ *and* s *be satisfied. Then, the subsampling design* ξ^* *is locally D-optimal at* $\boldsymbol{\beta}$ *if and only if there exist a set* $\mathcal{X}^*$ *and a threshold* s^* *such that*

(i) ξ^* *has density* $f_{\xi^*}(\mathbf{x}) = f_{\boldsymbol{X}}(\mathbf{x}) \mathbb{1}_{\mathcal{X}^*}(\mathbf{x})$

(ii) $\psi(\mathbf{x}, \xi^*, \boldsymbol{\beta}) \geq s^*$ *for* $\mathbf{x} \in \mathcal{X}^*$*, and*

(iii) $\psi(\mathbf{x}, \xi^*, \boldsymbol{\beta}) < s^*$ *for* $\mathbf{x} \notin \mathcal{X}^*$*.*

In principle a subsample can be generated according to a continuous design ξ by accepting units i into the subsample with probability $f_\xi(\mathbf{x}_i)/f_{\boldsymbol{X}}(\mathbf{x}_i)$. However, by Theorem 1.23 (i) the probability $f_{\xi^*}(\mathbf{x}_i)/f_{\boldsymbol{X}}(\mathbf{x}_i)$ is equal to one for all $\mathbf{x}_i \in \mathcal{X}^*$ and equal to zero otherwise for the locally D-optimal subsampling design ξ^*. Therefore, the D-optimal subsample can be selected deterministically.

Theorems similar to Theorem 1.23 appear extensively throughout the following chapters, specific to the given situation, see e. g. Theorems 2.1 and 3.1, and is restated as Theorem 4.11 in the setting of Chapter 4. The condition $\mathrm{P}(\psi(\boldsymbol{X}_i, \xi, \boldsymbol{\beta}) = s) = 0$ for any ξ and s is discussed thoroughly in the corresponding proofs. Next, we make introductory remarks to the three following chapters.

1.3.1 Introduction to Chapter 2 on Optimal Subsampling Design for Polynomial Regression in one Covariate

In Chapter 2 we present the work published as Reuter and Schwabe (2023a). We construct D-optimal subsampling designs for polynomial regression of degree q in one covariate. Specifically, this corresponds to the model (1.1) with $\mathbf{f}(x) = (1, x, \ldots, x^q)^\top$, except that we now assume that the covariate X_i is a random variable. The model is described in detail in Section 2.2.

In Section 2.3 we adapt Theorem 1.23 to the present setup to establish the equivalence Theorem 2.1. It is shown that the density of the D-optimal subsampling design is concentrated on, at most, $p = q + 1$ intervals. This is consistent with Example 1.15, where the D-optimal design has p support points. Corollary 2.3 additionally assumes that the covariate follows a symmetric distribution. As a result, we find that the support of ξ^* is also symmetric, similar to how the support points in Example 1.15 are symmetrically placed around zero. We also consider the location-scale transformation $Z_i = \sigma_Z X_i + \mu_Z$ for $\sigma_Z \neq 0, \mu_Z \in \mathbb{R}$ of the covariate. Theorem 2.2 shows how the D-optimal subsampling design for covariate X_i can be transformed to be D-optimal for the covariate Z_i.

Then, in Section 2.4 we give examples for linear regression ($q = 1$). This case is also discussed in Section 3.3.1, when there is only one covariate in the setup of multiple linear regression in Chapter 3. For a symmetric distribution of the covariate, the D-optimal subsampling design is the theoretical counterpart of the IBOSS method proposed in Wang et al. (2019). The IBOSS method is a subsampling method that selects $k < n$ data points. For a single covariate, the IBOSS method selects the farthest $k/2$ data points from each tail, i. e. the $k/2$ data points with the smallest value of the covariate and the $k/2$ data points with the largest value of the covariate, respectively.

In Section 2.5 we give examples for several distributions of the covariate for quadratic regression ($q = 2$). Through the examples we observe properties of the D-optimal design such as the following. Assume a uniform distribution on $[-1, 1]$ of the covariate X_i. Then, the weight $\xi^*(\mathcal{I}_r)/\alpha$ on the three intervals $\mathcal{I}_r,\ r = 0, 1, 2$ of the D-optimal subsampling design ξ^* converges to $1/3$ as the subsampling proportion α tends to zero, see Theorem 2.9. Thus when $\alpha \to 0$ the rescaled D-optimal subsampling design ξ^*/α converges in distribution to the D-optimal design from Example 1.15. All D-optimal subsampling designs discussed in Sections 2.4 and 2.5 for a covariate following the normal, exponential, or uniform distribution are supported on exactly p intervals, given $q \leq 2$. However, for $q = 2$ and a heavy-tailed distribution like the t-distributed covariate with ν degrees of freedom, there exists a threshold α^* such that for $\alpha \geq \alpha^*$ the D-optimal design is supported only on $(-\infty, -t_{\nu,1-\alpha/2}] \cup [t_{\nu,1-\alpha/2}, \infty)$. Here, $t_{\nu,1-\alpha/2}$ denotes the $1-\alpha/2$-quantile of the t-distribution with ν degrees of freedom. We state this specifically for $\nu = 5$ in Theorem 2.11.

In Section 2.6 we first study the efficiency of uniform subsampling to illustrate the advantage of using a D-optimal subsampling design. Then, we propose an IBOSS-like design for quadratic regression and study its efficiency. The IBOSS-like method takes proportions $\alpha/3$ from each of the two tails of the data as well as from the center of the data. This procedure can be used without any prior knowledge of the distribution of the covariate. We find that this IBOSS-like design is highly efficient over the whole range of the subsampling proportion α. He et al. (2024) developed their method for extending the IBOSS method to quadratic regression and applied it to financial data. Their method extends the IBOSS method to quadratic regression in the case of a d-dimensional covariate. As in the IBOSS method, the d covariates are considered successively. For each covariate the subsample is selected similarly to the method we propose here.

1.3.2 Introduction to Chapter 3 on D-optimal Subsampling Design for Massive Data Linear Regression

Chapter 3 contains the work published as Reuter and Schwabe (2023b). Here, we consider a multiple linear regression, i. e. model (1.1) with $\mathbf{f}(\mathbf{x}) = (1, x_1, \ldots, x_d)^\top$ and $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$. However, we now assume that the covariate $\boldsymbol{X}_i$ is a d-dimensional random variable. In Section 3.2 we present the model.

While the focus of this chapter is on a d-dimensional covariate, we briefly discuss the construction of D-optimal subsampling designs for linear regression in one covariate in Section 3.3.1. This connects Chapter 3 to the preceding Chapter 2.

In Section 3.3.2 we extend the results to d-dimensional covariates. Initially, we assume that the covariate $\boldsymbol{X}_i$ follows a centered, spherical distribution, i. e. a distribution that is invariant with respect to the special orthogonal group $SO(d)$. This assumption is later relaxed to non-centered elliptical distributions of the covariate. One of the key results of Section 3.3.2 is Theorem 3.6. It shows that any subsampling design ξ can be symmetrized such that the symmetrized subsampling design $\overline{\xi}$ is superior to ξ in the sense of D-optimality. This allows us to restrict the search for a D-optimal subsampling design to such symmetrized subsampling designs that are invariant with respect to $SO(d)$. Theorem 3.7 then gives a representation of the density of the D-optimal subsampling design ξ^* for centered spherical distributions. We find that f_{ξ^*} is equal to zero in the interior of a sphere of radius $q_{1-\alpha}$ and equal to the bounding density $f_{\boldsymbol{X}}$ outside that sphere. Here, $q_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the distribution of $\|\boldsymbol{X}_i\|_2^2$. Similarly, the D-optimal design in Example 1.16 has support points in the corners of the design region. To allow the relaxation to elliptical distributions of the covariate we consider the transformation $\boldsymbol{X}_i = \boldsymbol{A}\boldsymbol{Z}_i + \boldsymbol{\mu}$ of the covariate in Lemma 3.9, where $\boldsymbol{A}$ is a nonsingular $d \times d$ matrix and $\boldsymbol{\mu} \in \mathbb{R}^d$. The density of the D-optimal subsampling design for non-centered elliptical distributions of $\boldsymbol{X}_i$ is given in Theorem 3.10. We propose subsampling methods that implements the D-optimal subsampling design in Algorithms 1 and 2. The algorithms differ only slightly. The subsample size K_n of Algorithm 1 is random, while the subsample size k_n of Algorithm 2 is deterministic. Both methods require knowledge of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ and the mean $\boldsymbol{\mu}_{\boldsymbol{X}}$ of the covariate.

In Section 3.4 we treat the scenario of a fixed subsample size k for a growing full sample size n, rather than a fixed subsampling proportion α. We discuss the covariance matrix and mean squared error of the least squares estimator of the d-dimensional slope parameter vector $\hat{\boldsymbol{\beta}}_1$. We also study the efficiency of uniform random subsampling to illustrate the advantage of using the D-optimal subsampling design. Here, we also propose a simplified subsampling method that requires less knowledge of the covariate $\boldsymbol{X}_i$ and has lower computational

complexity $\mathcal{O}(nd)$. Specifically, only the variances of the covariates are needed, i. e. the diagonal entries of $\boldsymbol{\Sigma_X}$.

In Section 3.5 we present a simulation study that compares our proposed subsampling methods with the IBOSS method (Wang et al., 2019). The IBOSS method is a subsampling method for multiple linear regression with d covariates that selects $k < n$ data points. The method considers the d covariates successively. For each covariate the IBOSS method selects the farthest $k/(2d)$ remaining data points with the smallest and largest values in the currently considered covariate. Unsurprisingly, the method that requires full knowledge of the covariance matrix $\boldsymbol{\Sigma_X}$ and mean $\boldsymbol{\mu_X}$ of $\boldsymbol{X}_i$ outperforms the IBOSS method, which requires no such prior knowledge. Outside of cases where there is knowledge of $\boldsymbol{\Sigma_X}$ and $\boldsymbol{\mu_X}$ the method can serve as a benchmark for other methods. We find that the proposed simplified subsampling method can outperform the IBOSS method in some scenarios of low correlation between the covariates.

1.3.3 Introduction to Chapter 4 on Poisson Regression in one Covariate on Massive Data

In Chapter 4 we present the work published as Reuter and Schwabe (2024). We assume that conditional on the covariate X_i, the response Y_i follows a Poisson distribution. The conditional mean of Y_i given by model 1.3, with inverse link function g^{-1} equal to the natural exponential function, and $\mathbf{f}(x) = (1, x)^\top$. The model is discussed in detail in Section 4.2.

In Section 4.3 we first work under the assumption that $\beta_1 < 0$, analogous to Example 1.21. We establish a representation of the support of the D-optimal subsampling design for $\beta_1 < 0$ in Theorem 4.2. There are two possible cases. Firstly, the D-optimal subsampling design ξ^* may be supported on two intervals, similar to the D-optimal design having two support points in Example 1.21. Alternatively, ξ^* may be supported only on the interval $(-\infty, q_\alpha]$, where q_α denotes the α-quantile of X_i. Theorem 4.6 then treats transformations of the covariate and simultaneous transformation of the D-optimal subsampling design, as we have also seen in Chapters 2 and 3. Here, however, the transformation requires a simultaneous transformation of the parameter β_1. We use this transformation of the parameter β_1 to derive Corollary 4.7, which gives a representation of the support of ξ^* for positive $\beta_1 > 0$. The case $\beta_1 = 0$ is equivalent to linear regression in one covariate, which is treated in Sections 2.4 and 3.3.1, as mentioned previously. We give examples of the D-optimal subsampling design ξ^* for exponential and uniform distributions of the covariate X_i throughout Section 4.3. Here, we place special emphasis on the transition between the cases of one and two intervals that make up the support of ξ^*. The transition occurs at a crossover point that depends on both

the subsampling proportion α and the parameter β_1.

As in Chapters 2 and 3 we study the efficiency of uniform random subsampling in Section 4.4. Further, we consider both a one-sided and a two-sided IBOSS-like design. The one-sided design has support $(-\infty, q_\alpha]$. The two-sided design is the theoretical counterpart of the IBOSS method and is supported on $(-\infty, q_{\alpha/2}] \cup [q_{1-\alpha/2}, \infty)$. While we found that such heuristic designs can perform very well in the polynomial regression setting in Chapter 2, the one- and two-sided IBOSS-like designs are not efficient over the entire range of subsampling proportions α. In particular, for small α, these heuristic designs do not perform well.

In Chapter 4 we consider only the linear predictor $\beta_0 + \beta_1 X_i$. The results can be extended to polynomial Poisson regression. There, the linear predictor is given by $\mathbf{f}(X_i)^\top \boldsymbol{\beta}$ with $\mathbf{f}(\mathbf{x}) = (1, x, \ldots, x^q)^\top$ and $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$. Then, the sensitivity function is the product of an exponential function and a polynomial of degree q in x. Thus, the support of ξ^* is the union of $q+1$ intervals, similar to how the number of intervals that make up the support of ξ^* grows with the degree q in Chapter 2.

1.4 Concluding Remarks

Throughout this thesis we construct and analyze D-optimal subsampling designs for various regression models under different distributional assumptions. The main contribution of this thesis is the theoretical derivation of the D-optimal subsampling designs. We find that D-optimal subsampling designs strongly outperform uniform random subsampling, particularly for small subsample proportions with unbounded covariate distributions. Additionally, we propose a generalized IBOSS method for quadratic regression that does not require prior knowledge of the distribution and showed competitive performance. For multiple linear regression, we construct optimal designs for centered spherical distributions and their location-scale transformations, providing two implementation methods with varying assumptions and computational complexity. Comparative simulations show the superiority of our methods over the IBOSS method in specific scenarios. We showed that IBOSS-like subsampling can be reasonably efficient for some settings of Poisson regression, but perform poorly for small subsampling proportion. We investigate misspecification of the regression parameter in the case of Poisson regression.

The emphasis in this work is on D-optimal subsampling designs. We note that many of the results may be extended to other optimality criteria like A-optimality and other criteria from the Kiefer's Φ_q-class of optimality criteria (Kiefer, 1974). Alternatively, the $IMSE$-optimality may be considered when predicting the mean response. The theoretical

results that establish a representation of the support of D-optimal subsampling designs require strong assumptions. In practice, it is not reasonable to assume that the distribution of the covariate is known, or even that it is invariant with respect to rotations about the origin after transformation, as we did in Chapter 3. In this thesis we solely focused on massive data with few covariates, and just a single covariate in Chapters 2 and 4. While we propose subsampling methods that do not require prior knowledge of the distribution of the covariate throughout this thesis, future work should focus on reducing the assumptions on the covariate, e. g. in the style of the work by Pronzato and Wang (2021). Extensions of the subsampling methods presented here to d covariates are relatively straight-forward for some settings such as quadratic regression (see He et al., 2024). Extending the theoretical results to higher dimensionality requires extensive work, as we have seen in Chapter 3. However, as high-dimensional data is a key aspect of big data analysis, more investigation is needed. A recent work by Singh and Stufken (2023) tackles this issue by combining LASSO with subsampling. Further, the D-optimal subsampling design heavily depends on the chosen regression model. The subsampling methods we propose may have a consequential loss in efficiency, when the model is misspecified. Finding subsampling methods that are robust to model misspecification is valuable, as prior knowledge of the regression model is difficult to justify in big data settings. Current work on this includes Joseph and Mak (2021) for regression problems and Singh (2024) for classification.

Chapter 2

Optimal Subsampling Design for Polynomial Regression in one Covariate

In this chapter we present the work titled "Optimal Subsampling Design for Polynomial Regression in one Covariate" (Reuter and Schwabe, 2023a) published in the journal *Statistical Papers*.

2.1 Introduction

Data Reduction is a major challenge as technological advances have led to a massive increase in data collection to a point where traditional statistical methods fail or computing power can not keep up. In this case we speak of big data. We typically differentiate between the case where the number of covariates is much larger than the number of observations and the case where the massive amount of observations is the problem. The first case is well studied, most notably by Tibshirani (1996) introducing LASSO, which utilizes ℓ_1 penalization to find sparse parameter vectors, thus fusing subset selection and ridge regression. The second case, often referred to as massive data, can be tackled in two ways. Firstly in a probabilistic fashion, creating random subsamples in a non-uniform manner. Prominent studies include Drineas et al. (2006), Mahoney (2011) and Ma et al. (2014). They present subsampling methods for linear regression models called algorithmic leveraging that sample according to probabilities based on the normalized statistical leverage scores of the covariate matrix. More recently Dereziński and Warmuth (2018) study volume sampling, where

subdata is chosen proportional to the squared volume of the parallelepiped spanned by its observations. Conversely to these probabilistic methods one can select subdata by applying deterministic rules. Shi and Tang (2021) present such a method, that maximizes the minimal distance between two observations in the subdata. Wang et al. (2021) propose orthogonal subsampling inspired by orthogonal arrays. Most prominently, Wang et al. (2019) introduce the information-based optimal subdata selection (IBOSS) to tackle big data linear regression in a deterministic fashion based on D-optimality.

In this paper we study D-optimal subsampling designs for polynomial regression in one covariate, where the goal is to select a percentage α of the full data that maximizes the determinant of the information matrix. For the conventional study of approximate designs in this setting we refer to Gaffke and Heiligers (1996). Heiligers and Schneider (1992) consider specifically cubic regression on a ball. We consider D-optimal designs with measure α that are bounded from above by the distribution of the known covariate. Such directly bounded designs were first studied by Wynn (1977) and Fedorov (1989). Pronzato (2004) considers this setting using a form of the subsampling design standardized to one and bounded by α^{-1} times the distribution of the covariates. More recently, Pronzato and Wang (2021) studies the same in the context of sequential subsampling. For the characterization of the optimal subsampling designs we make use of an equivalence theorem by Sahm and Schwabe (2001). This equivalence theorem enables us to construct such subsampling designs for various settings of the distributional assumptions on the covariate. Here we will only look at distributions of the covariate that are invariant to a sign change, i.e. symmetric about the vertical axis. We discuss the shape of D-optimal subsampling subsampling designs for polynomial regression of degree q first. We then study quadratic regression under several distributional assumptions more closely, after showing two examples for simple linear regression. In particular we take a look at the percentage of mass of the optimal subsampling design on the outer intervals compared to the inner one, which changes drastically given the distribution of the covariate, particularly for heavy-tailed distributions. In addition we examine the efficiency of uniform random subsampling to illustrate the advantage of the optimal subsampling designs. All numerical results are obtained by the Newton method implemented in the **R** package *nleqslv* by Hasselman (2018). All relevant **R** scripts are available on a GitHub repository `https://github.com/TorstenReuter/polynomial_regression_in_one_covariate`.

The rest of this paper is organized as follows. In Section 2.2 we specify the polynomial model. In Section 2.3 we introduce the concept of continuous subsampling designs and give characterizations for optimization. In Sections 2.4 and 2.5 we present optimal subsampling designs in the case of linear and quadratic regression, respectively, for various classes of

distributions of the covariate. Section 2.6 contains some efficiency considerations showing the strength of improvement of the performance of the optimal subsampling design compared to random subsampling. The paper concludes with a discussion in Section 2.7. Proofs are deferred to an Appendix.

2.2 Model Specification

We consider the situation of pairs (x_i, y_i) of data, where y_i is the value of the response variable Y_i and x_i is the value of a single covariate X_i for unit $i = 1, \ldots, n$, for very large numbers of units n. We assume that the dependence of the response on the covariate is given by a polynomial regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_q X_i^q + \varepsilon_i$$

with independent, homoscedastic random errors ε_i having zero mean ($\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2 > 0$). The largest exponent $q \geq 1$ denotes the degree of the polynomial regression, and $p = q + 1$ is the number of regression parameters $\beta_0, \ldots, \beta_q$ to be estimated, where, for each $k = 1, \ldots, q$, the parameter β_k is the coefficient for the kth monomial x^k, and β_0 denotes the intercept. For example, for $q = 1$, we have ordinary linear regression, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, with $p = 2$ parameters β_0 (intercept) and β_1 (slope) and, for $q = 2$, we have quadratic regression, $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$, with $p = 3$ and an additional curvature parameter β_2. Further, we assume that the units of the covariate X_i are identically distributed and that all X_i and random errors $\varepsilon_{i'}$ are independent.

For notational convenience, we write the polynomial regression as a general linear model

$$Y_i = \mathbf{f}(X_i)^\top \boldsymbol{\beta} + \varepsilon_i \,,$$

where $\mathbf{f}(x) = (1, x, \ldots, x^q)^\top$ is the p-dimensional vector of regression functions and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)^\top$ is the p-dimensional vector of regression parameters.

2.3 Subsampling Design

We are faced with the problem that the responses Y_i are expensive or difficult to observe while the values x_i of all units X_i of the covariate are available. To overcome this problem, we consider the situation that the responses Y_i will be observed only for a certain percentage α of the units ($0 < \alpha < 1$) and that these units will be selected on the basis of the knowledge

of the values x_i of the covariate for all units. As an alternative motivation, we can consider a situation where all pairs (x_i, y_i) are available but parameter estimation is computationally feasible only on a percentage α of the data. In either case we want to find the subsample of pairs (x_i, y_i) that yields the most precise estimation of the parameter vector $\boldsymbol{\beta}$.

To obtain analytical results, the covariate X_i is supposed to have a continuous distribution with density $f_X(x)$, and we assume that the distribution of the covariate is known. The aim is to find a subsample of this distribution that covers a percentage α of the distribution and that contains the most information. For this, we will consider continuous designs ξ as measures of mass α on $\mathbb{R}$ with density $f_\xi(x)$ bounded by the density $f_X(x)$ of the covariate X_i such that $\int f_\xi(x)\,\mathrm{d}x = \alpha$ and $f_\xi(x) \leq f_X(x)$ for all $x \in \mathbb{R}$. A subsample with mean subsample size αn can then be generated according to such a continuous design by accepting units i with probability $f_\xi(x_i)/f_X(x_i)$.

For a continuous design ξ, the information matrix $\mathbf{M}(\xi)$ is defined as $\mathbf{M}(\xi) = \int \mathbf{f}(x)\mathbf{f}(x)^\top f_\xi(x)\,\mathrm{d}x$. In the present polynomial setup, $\mathbf{M}(\xi) = \big(m_{j+j'}(\xi)\big)_{j=0,\ldots,q}^{j'=0,\ldots,q}$, where $m_k = \int x^k f_\xi(x)\,\mathrm{d}x$ is the kth moment associated with the design ξ. Thus, it has to be required that the distribution of X_i has a finite moment $\mathrm{E}(X_i^{2q})$ of order $2q$ in order to guarantee that all entries in the information matrix $\mathbf{M}(\xi)$ exist for all continuous designs ξ for which the density $f_\xi(x)$ is bounded by $f_X(x)$.

The information matrix $\mathbf{M}(\xi)$ measures the performance of the design ξ in the sense that the covariance matrix of the least squares estimator $\hat{\boldsymbol{\beta}}$ based on a subsample according to the design ξ is proportional to the inverse $\mathbf{M}(\xi)^{-1}$ of the information matrix $\mathbf{M}(\xi)$ or, more precisely, $\sqrt{\alpha n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is normally distributed with mean zero and covariance matrix $\sigma_\varepsilon^2 \mathbf{M}(\xi)^{-1}$, at least asymptotically. Note that for continuous designs ξ the information matrix $\mathbf{M}(\xi)$ is always of full rank and, hence, the inverse $\mathbf{M}(\xi)^{-1}$ exists. Based on the relation to the covariance matrix, it is desirable to maximize the information matrix $\mathbf{M}(\xi)$. However, as well-known in design optimization, maximization of the information matrix cannot be achieved uniformly with respect to the Loewner ordering of positive-definiteness. Thus, commonly, a design criterion which is a real valued functional of the information matrix $\mathbf{M}(\xi)$ will be maximized, instead. We will focus here on the most popular design criterion in applications, the D-criterion, in its common form $\log(\det(\mathbf{M}(\xi)))$ to be maximized. Maximization of the D-criterion can be interpreted in terms of the covariance matrix to be the same as minimizing the volume of the confidence ellipsoid for the whole parameter vector $\boldsymbol{\beta}$ based on the least squares estimator or, equivalently, minimizing the volume of the acceptance region for a Wald test on the whole model. The subsampling design ξ^* that maximizes the D-criterion $\log(\det(\mathbf{M}(\xi)))$ will be called D-optimal, and its density is

denoted by $f_{\xi^*}(x)$.

To obtain D-optimal subsampling designs, we will make use of standard techniques coming from constrained convex optimization and symmetrization. For convex optimization we employ the directional derivative

$$F_D(\xi, \eta) = \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \left(\log(\det(\mathbf{M}((1-\epsilon)\xi + \epsilon\eta))) - \log(\det(\mathbf{M}(\xi))) \right)$$

of the D-criterion at a design ξ with non-singular information matrix $\mathbf{M}(\xi)$ in the direction of a design η, where we allow here η to be a general design of mass α that has not necessarily a density bounded by $f_X(x)$. In particular, $\eta = \xi_x$ may be a one-point design which assigns all mass α to a single setting x in $\mathbb{R}$. Evaluation of the directional derivative yields $F_D(\xi, \eta) = \text{trace}(\mathbf{M}(\xi)^{-1}\mathbf{M}(\eta)) - p$ (compare Silvey, 1980, Example 3.8) which reduces to $F_D(\xi, \xi_x) = \alpha \mathbf{f}(x)^\top \mathbf{M}(\xi)^{-1}\mathbf{f}(x) - p$ for a one-point design $\eta = \xi_x$. Equivalently, for one-point designs $\eta = \xi_x$, we may consider the sensitivity function $\psi(x, \xi) = \alpha \mathbf{f}(x)^\top \mathbf{M}(\xi)^{-1}\mathbf{f}(x)$ which incorporates the essential part of the directional derivative ($\psi(x, \xi) = p + F_D(\xi, \xi_x)$). For the characterization of the D-optimal continuous subsampling design, the constrained equivalence theorem under Kuhn-Tucker conditions (see Sahm and Schwabe, 2001, Corollary 1 (c)) can be reformulated in terms of the sensitivity function and applied to our case of polynomial regression.

Theorem 2.1. *In polynomial regression of degree q with density $f_X(x)$ of the covariate X_i, the subsampling design ξ^* with support $\mathcal{X}^*$ is D-optimal if and only if there exist a threshold s^* and settings $a_1 > \cdots > a_{2r}$ for some r $(1 \le r \le q)$ such that*

(i) the D-optimal subsampling design ξ^* is given by

$$f_{\xi^*}(x) = \begin{cases} f_X(x) & \text{if } x \in \mathcal{X}^* \\ 0 & \text{otherwise} \end{cases}$$

(ii) $\psi(x, \xi^*) \ge s^*$ for $x \in \mathcal{X}^*$, and

(iii) $\psi(x, \xi^*) < s^*$ for $x \notin \mathcal{X}^*$,

where $\mathcal{X}^* = \bigcup_{k=0}^{r} \mathcal{I}_k$ and $\mathcal{I}_0 = [a_1, \infty)$, $\mathcal{I}_r = (-\infty, a_{2r}]$, and $\mathcal{I}_k = [a_{2k+1}, a_{2k}]$, for $k = 1, \ldots, r-1$, are mutually disjoint intervals.

The density $f_{\xi^*}(x) = f_X(x)\mathbb{1}_{\mathcal{X}^*}(x) = \sum_{k=0}^{r} f_X(x)\mathbb{1}_{\mathcal{I}_k}(x)$ of the D-optimal subsampling design ξ^* is concentrated on, at most, $q+1$ intervals $\mathcal{I}_k$, where $\mathbb{1}_A(x)$ denotes the indicator function on the set A, i. e. $\mathbb{1}_A(x) = 1$ for $x \in A$, and $\mathbb{1}_A(x) = 0$ otherwise. The density $f_{\xi^*}(x)$

has a 0-1-property such that it is either equal to the density $f_X(x)$ of the covariate (on $\mathcal{X}^*$) or equal to 0 (on the complement of $\mathcal{X}^*$). Thus, the generation of a subsample according to the optimal continuous subsampling design ξ^* can be implemented easily by accepting all units i for which the value x_i of the covariate is in $\mathcal{X}^*$ and rejecting all other units with $x_i \notin \mathcal{X}^*$. The threshold s^* can be interpreted as the $(1-\alpha)$-quantile of the distribution of the sensitivity function $\psi(X_i, \xi^*)$ as a function of the random variable X_i (see Pronzato and Wang, 2021).

A further general concept to be used is equivariance. This can be employed to transform the D-optimal subsampling design simultaneously with a transformation of the distribution of the covariate. More precisely, the location-scale transformation $Z_i = \sigma X_i + \mu$ of the covariate and its distribution is conformable with the regression function $\mathbf{f}(x)$ in polynomial regression, and the D-criterion is equivariant with respect to such transformations.

Theorem 2.2. *Let $f_{\xi^*}(x)$ be the density for a D-optimal subsampling design ξ^* for covariate X_i with density $f_X(x)$. Then $f_{\zeta^*}(z) = \frac{1}{\sigma} f_{\xi^*}(\frac{z-\mu}{\sigma})$ is the density for a D-optimal subsampling design ζ^* for covariate $Z_i = \sigma X_i + \mu$ with density $f_Z(z) = \frac{1}{\sigma} f_X(\frac{z-\mu}{\sigma})$.*

As a consequence, also the optimal subsampling design ζ^* is concentrated on, at most, $p = q+1$ intervals, and its density $f_{\zeta^*}(z)$ is either equal to the density $f_Z(z)$ of the covariate Z_i (on $\mathcal{Z}^* = \sigma\mathcal{X}^* + \mu$) or it is equal to 0 (elsewhere) such that, also here, the optimal subsampling can be implemented quite easily.

A further reduction of the optimization problem can be achieved by utilizing symmetry properties. Therefore, we consider the transformation of sign change, $g(x) = -x$, and assume that the distribution of the covariate is symmetric, $f_X(-x) = f_X(x)$ for all x. For a continuous design ξ, the design ξ^g transformed by sign change has density $f_{\xi^g}(x) = f_\xi(-x)$ and, thus, satisfies the boundedness condition $f_{\xi^g}(x) \leq f_X(x)$, when the distribution of X_i is symmetric, and has the same value for the D-criterion as ξ, $\log(\det(\mathbf{M}(\xi^g))) = \log(\det(\mathbf{M}(\xi)))$. By the concavity of the D-criterion, standard invariance arguments can be used as in Pukelsheim (1993, Chapter 13) and Heiligers and Schneider (1992). In particular, any continuous design ξ is dominated by its symmetrization $\bar{\xi} = (\xi + \xi^g)/2$ with density $f_{\bar{\xi}}(x) = (f_\xi(x) + f_\xi(-x))/2 \leq f_X(x)$ such that $\log(\det(\mathbf{M}(\bar{\xi}))) \geq \log(\det(\mathbf{M}(\xi)))$ (Pukelsheim, 1993, Chapter 13.4). Hence, we can restrict the search for a D-optimal subsampling design to symmetric designs $\bar{\xi}$ with density $f_{\bar{\xi}}(-x) = f_{\bar{\xi}}(x)$ which are invariant with respect to sign change ($\bar{\xi}^g = \bar{\xi}$). For these symmetric subsampling designs $\bar{\xi}$, the moments $m_k(\bar{\xi})$ are zero for odd k and positive when k is even. Hence, the information matrix $\mathbf{M}(\bar{\xi})$ is an even checkerboard matrix (see Jones and Willms, 2018) with positive entries $m_{j+j'}(\bar{\xi})$ for even index sums and entries equal to zero when the index sum is odd. The inverse $\mathbf{M}(\bar{\xi})^{-1}$ of the

information matrix $\mathbf{M}(\bar{\xi})$ shares the structure of an even checkerboard matrix. Thus, the sensitivity function $\psi(x, \bar{\xi})$ is a polynomial with only terms of even order and is, hence, a symmetric function of x. This leads to a simplification of the representation of the optimal subsampling design in Theorem 2.1 because the support $\mathcal{X}^*$ of the optimal subsampling design ξ^* will be symmetric, too.

Corollary 2.3. *In polynomial regression of degree q with a symmetrically distributed covariate X_i with density $f_X(x)$, the D-optimal subsampling design ξ^* with density $f_{\xi^*}(x) = \sum_{k=0}^{r} f_X(x)\mathbb{1}_{\mathcal{I}_k}(x)$ has symmetric boundaries $a_1, \ldots, a_{2r}$ of the intervals $\mathcal{I}_0 = [a_1, \infty)]$, $\mathcal{I}_k = [a_{2k+1}, a_{2k}]$, and $\mathcal{I}_r = (-\infty, a_{2r}]$, i. e. $a_{2r+1-k} = -a_k$ and, accordingly, $\mathcal{I}_{r-k} = -\mathcal{I}_k$.*

This characterization of the optimal subsampling design ξ^* will be illustrated in the next two sections for ordinary linear regression ($q = 1$) and for quadratic regression ($q = 2$).

2.4 Optimal Subsampling for Linear Regression

In the case of ordinary linear regression $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, we have

$$\mathbf{M}(\xi) = \begin{pmatrix} \alpha & m_1(\xi) \\ m_1(\xi) & m_2(\xi) \end{pmatrix},$$

for the information matrix of any subsampling design ξ. The inverse $\mathbf{M}(\xi)^{-1}$ of the information matrix is given by

$$\mathbf{M}(\xi)^{-1} = \frac{1}{\alpha m_2(\xi) - m_1(\xi)^2} \begin{pmatrix} m_2(\xi) & -m_1(\xi) \\ -m_1(\xi) & \alpha \end{pmatrix},$$

and the sensitivity function

$$\psi(x, \xi) = \frac{1}{\alpha m_2(\xi) - m_1(\xi)^2}(m_2(\xi) - 2m_1(\xi)x + \alpha x^2) \tag{2.1}$$

is a polynomial of degree two in x. The D-optimal continuous subsampling design ξ^* has density $f_\xi(x) = f_X(x)$ for $x \leq a_2$ and for $x \geq a_1$ while $f_\xi(x) = 0$ for $a_2 < x < a_1$. The corresponding subsampling design then accepts those units i for which $x_i \leq a_2$ or $x_i \geq a_1$, and rejects all units i for which $a_2 < x_i < a_1$.

To obtain the D-optimal continuous subsampling design ξ^* by Theorem 2.1, the boundary

points a_1 and a_2 have to be determined to solve the two non-linear equations

$$\mathrm{P}(X_i \leq a_2) + \mathrm{P}(X_i \geq a_1) = \alpha \tag{2.2}$$

and

$$\psi(a_2, \xi^*) = \psi(a_1, \xi^*)\,.$$

By equation (2.1), the latter condition can be written as

$$\alpha a_2^2 - 2m_1(\xi^*)a_2 = \alpha a_1^2 - 2m_1(\xi^*)a_1\,,$$

which can be reformulated as

$$\alpha(a_1 + a_2) = 2m_1(\xi^*)\,. \tag{2.3}$$

When the distribution of X_i is symmetric, Corollary 2.3 provides symmetry $a_2 = -a_1$ of the boundary points. This is in agreement with condition (2.3) because $m_1(\xi^*) = 0$ in the case of symmetry. Further, by the symmetry of the distribution, $\mathrm{P}(X_i \leq a_2) = \mathrm{P}(X_i \geq a_1) = \alpha/2$, and a_1 has to be chosen as the $(1-\alpha/2)$-quantile of the distribution of X_i to obtain the D-optimal continuous subsampling design.

Example 2.4 (normal distribution)**.** If the covariate X_i comes from a standard normal distribution, then the optimal boundaries are the $(\alpha/2)$- and the $(1-\alpha/2)$-quantile $\pm z_{1-\alpha/2}$, and unit i is accepted when $|x_i| \geq z_{1-\alpha/2}$.

For X_i having a general normal distribution with mean μ and variance σ^2, the optimal boundaries remain to be the $(\alpha/2)$- and $(1-\alpha/2)$-quantile $a_2 = \mu - \sigma z_{1-\alpha/2}$ and $a_1 = \mu + \sigma z_{1-\alpha/2}$, respectively, by Theorem 2.2.

This approach applies accordingly to all distributions which are obtained by a location or scale transformation of a symmetric distribution: units will be accepted if their values of the covariate lie in the lower or upper $(\alpha/2)$-tail of the distribution. This procedure can be interpreted as a theoretical counterpart in one dimension of the IBOSS method proposed by Wang et al. (2019).

However, for an asymmetric distribution of the covariate X_i, the optimal proportions for sampling from the upper and lower tail may differ. By condition (2.7), there will be a proportion α_1, $0 \leq \alpha_1 \leq \alpha$, for the upper tail and $\alpha_2 = \alpha - \alpha_1$ for the lower tail such that a_1 is the $(1-\alpha_1)$-quantile and a_2 is the α_2-quantile of the distribution of the covariate X_i, respectively. In view of condition (2.3), neither α_1 nor α_2 can be zero. Hence, the optimal subsampling design will have positive, but not necessarily equal mass at both tails. This will be illustrated in the next example.

Example 2.5 (exponential distribution)**.** If the covariate X_i comes from a standard exponential distribution with density $f_X(x) = e^{-x},\ x \geq 0$, we conclude from Theorem 2.1 that $f_{\xi^*}(x) = f_X(x)\mathbb{1}_{[0,b]\cup[a,\infty)}(x)$ with $a = a_1$ and $b = a_2$ when $a_2 \geq 0$. Otherwise, when $a_2 < 0$, the density $f_X(x)$ of the covariate X_i vanishes on the left interval $\mathcal{I}_1 = (-\infty, a_2]$ because the support of the distribution of X_i does not cover the whole range of $\mathbb{R}$. In that case, we may formally let $b = 0$. Then, we can calculate the entries of $\mathbf{M}(\xi^*)$ as functions of a and b as

$$\begin{aligned} m_1(\xi^*) &= 1 + (a+1)e^{-a} - (b+1)e^{-b} \\ m_2(\xi^*) &= 2 + (a^2+2a+2)e^{-a} - (b^2+2b+2)e^{-b}\,. \end{aligned}$$

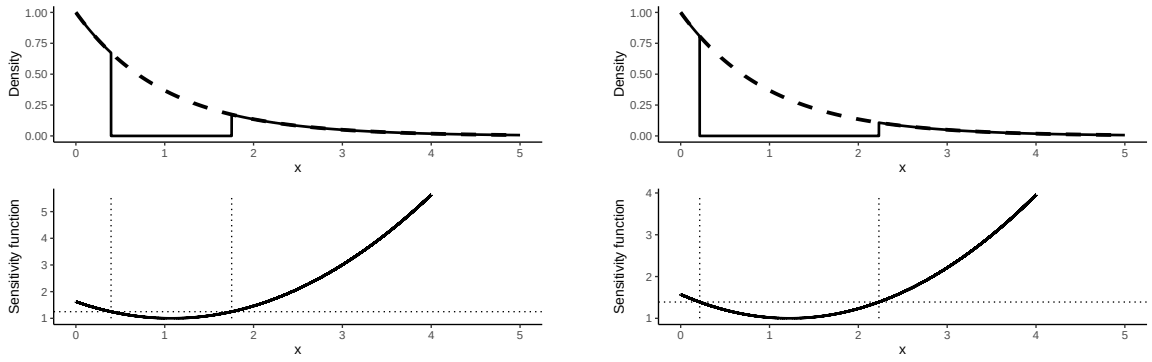
To obtain the optimal solutions for a and b in the case $a_2 \geq 0$, the two non-linear equations (2.2) and (2.3) have to be satisfied which become here $e^{-b} - e^{-a} = 1 - \alpha$ and $\alpha(a+b) = 2m_1(\xi^*)$.

If $a_2 < 0$ would hold, the first condition reveals $a = -\log(\alpha)$ and, hence, $m_1(\xi^*) = \alpha(a+1)$. There, similar to the proof of Theorem 2.7 below, the second condition has to be relaxed to $\psi(a, \xi^*) \geq \psi(0, \xi^*)$ which can be reformulated to $\alpha a \geq 2m_1(\xi^*) = 2\alpha(a+1)$ and yields a contradiction. Thus, this case can be excluded, and a_2 has to be larger than 0 for all α.

For selected values of α, numerical results are presented in Table 2.1. Additionally to the optimal values for a and b, also the proportions $\mathrm{P}(X_i \leq b)$ and $\mathrm{P}(X_i \geq a)$ are presented in Table 2.1 together with the percentage of mass allocated to the left interval $[0, b]$. In Figure 2.1, the density f_{ξ^*} of the optimal subsampling design ξ^* and the corresponding sensitivity function $\psi(x, \xi^*)$ are exhibited for $\alpha = 0.5$ and $\alpha = 0.3$. Vertical lines indicate the positions of the boundary points a and b, and the dotted horizontal line displays the threshold s^*. As could have been expected, less mass is assigned to the right tail of the

Table 2.1: Numeric values for the boundary points a and b for selected values of the subsampling proportion α in the case of standard exponential X_i

α	b	$\mathrm{P}(X_i \leq b)$	a	$\mathrm{P}(X_i \geq a)$	% of mass on $[0, b]$
0.5	0.39572	0.32681	1.75335	0.17319	65.36
0.3	0.21398	0.19264	2.23153	0.10736	64.21
0.1	0.06343	0.06146	3.25596	0.03854	61.46
0.01	0.00579	0.00577	5.46588	0.00423	57.71



(a) $\alpha = 0.5$ (b) $\alpha = 0.3$.

Figure 2.1: Density of the optimal subsampling design (solid line) and the standard exponential distribution (dashed line, upper panels), and sensitivity functions (lower panels) for subsampling proportions $\alpha = 0.5$ (left) and $\alpha = 0.3$ (right)

right-skewed distribution because observations from the right tail are more influential and, thus, more observations seem to be required on the lighter left tail for compensation.

For X_i having an exponential distribution with general intensity $\lambda > 0$ (scale $1/\lambda$), the optimal boundary points remain to be the same quantiles as in the standard exponential case, $a_1 = a/\lambda$ and $a_2 = b/\lambda$ associated with the proportion α, by Theorem 2.2.

2.5 Optimal Subsampling for Quadratic Regression

In the case of quadratic regression $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$ we have

$$\mathbf{M}(\bar{\xi}) = \begin{pmatrix} \alpha & 0 & m_2(\bar{\xi}) \\ 0 & m_2(\bar{\xi}) & 0 \\ m_2(\bar{\xi}) & 0 & m_4(\bar{\xi}) \end{pmatrix}, \tag{2.4}$$

for the information matrix of a symmetric subsampling design $\bar{\xi}$. The inverse $\mathbf{M}(\bar{\xi})^{-1}$ of the information matrix is given by

$$\mathbf{M}(\bar{\xi})^{-1} = \frac{1}{\alpha m_4(\bar{\xi}) - m_2(\bar{\xi})^2} \begin{pmatrix} m_4(\bar{\xi}) & 0 & -m_2(\bar{\xi}) \\ 0 & \alpha \frac{m_4(\bar{\xi})}{m_2(\bar{\xi})} - m_2(\bar{\xi}) & 0 \\ -m_2(\bar{\xi}) & 0 & \alpha \end{pmatrix},$$

and the sensitivity function

$$\psi(x, \bar{\xi}) = \frac{1}{\alpha m_4(\bar{\xi}) - m_2(\bar{\xi})^2}(m_4(\bar{\xi}) - 3m_2(\bar{\xi})x^2 + \alpha\frac{m_4(\bar{\xi})}{m_2(\bar{\xi})}x^2 + \alpha x^4) \tag{2.5}$$

is a polynomial of degree four and is symmetric in x.

According to Corollary 2.3, the density $f_{\xi^*}(x)$ of the D-optimal continuous subsampling design ξ^* has, at most, three intervals that are symmetrically placed around zero, where the density is equal to the bounding density $f_X(x)$, and $f_{\xi^*}(x)$ is equal to zero elsewhere. Thus the density $f_{\xi^*}(x)$ of the D-optimal subsampling design has the shape

$$f_{\xi^*}(x) = f_X(x)\mathbb{1}_{(-\infty,-a]\cup[-b,b]\cup[a,\infty)}(x)\,, \tag{2.6}$$

where $a > b \geq 0$. We formally allow $b = 0$ which means that $\psi(0, \xi^*) \leq s^* = \psi(a, \xi^*)$ and that the density $f_{\xi^*}(x)$ is concentrated on only two intervals, $f_{\xi^*}(x) = f_X(x)\mathbb{1}_{(-\infty,-a]\cup[a,\infty)}(x)$. Although the information matrix will be non-singular even in the case of two intervals ($b = 0$), the optimal subsampling design will include a non-degenerate interior interval $[-b, b]$ in many cases, $b > 0$, as illustrated below in Examples 2.6 and 2.8. However, for a heavy-tailed distribution of the covariate X_i, the interior interval may vanish in the optimal subsampling design as shown in Example 2.10.

To obtain the D-optimal continuous subsampling design ξ^* by Corollary 2.3, the boundary points $a = a_1$ and $b = a_2 \geq 0$ have to be determined to solve the two non-linear equations

$$\mathrm{P}(|X_i| \leq b) + \mathrm{P}(|X_i| \geq a) = \alpha \tag{2.7}$$

and

$$\psi(b, \xi^*) = \psi(a, \xi^*)\,. \tag{2.8}$$

By equation (2.5), the latter condition can be written as

$$\alpha m_2(\xi^*)b^4 + \left(\alpha m_4(\xi^*) - 3m_2(\xi^*)^2\right)b^2 = \alpha m_2(\xi^*)a^4 + \left(\alpha m_4(\xi^*) - 3m_2(\xi^*)^2\right)a^2\,,$$

which can be reformulated as

$$\alpha m_2(\xi^*)(a^2 + b^2) = 3m_2(\xi^*)^2 - \alpha m_4(\xi^*)\,. \tag{2.9}$$

For finding the optimal solution, we use the Newton method implemented in the **R** package *nleqslv* by Hasselman (2018) to calculate numeric values for a and b based on

equations (2.7) and (2.8) for various symmetric distributions.

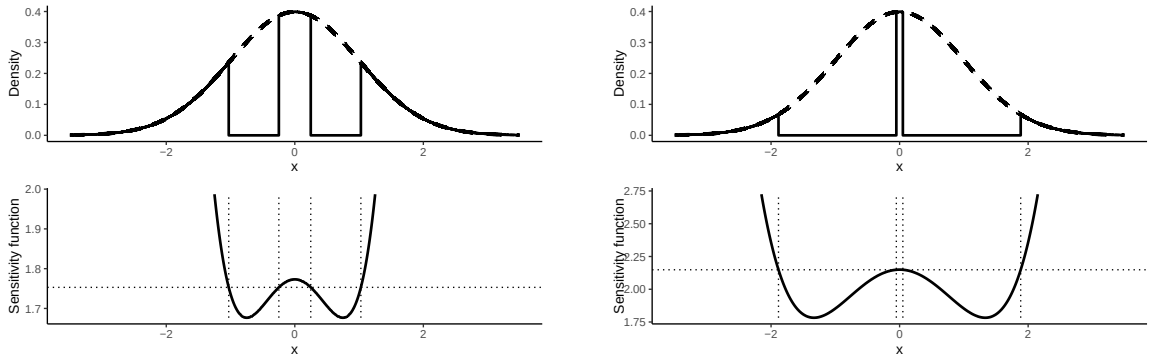
The case $b = 0$ relates to the situation of only two intervals ($r = 1 < q$). There, condition (2.7) simplifies to $a = q_{1-\alpha/2}$, where $q_{1-\alpha/2}$ is the $(1 - \alpha/2)$-quantile of the distribution of the covariate X_i, and equation (2.8) has to be relaxed to $\psi(0, \xi^*) \leq \psi(a, \xi^*)$, similar to the case $b = 0$ in Example 2.5.

Example 2.6 (normal distribution)**.** For the case that the covariate X_i comes from a standard normal distribution, results are given in Table 2.2 for selected values of α. Additionally to

Table 2.2: Numeric values for the boundary points a and b for selected values of the subsampling proportion α in the case of standard normal X_i

α	a	$1 - \Phi(a)$	b	$2\Phi(b) - 1$	% of mass on $[-b, b]$
0.5	1.02800	0.15198	0.24824	0.19605	39.21
0.3	1.34789	0.08885	0.15389	0.12231	40.77
0.1	1.88422	0.02977	0.05073	0.04046	40.46
0.01	2.73996	0.00307	0.00483	0.00386	38.55

the optimal values for a and b, also the proportions $\mathrm{P}(X_i \geq a) = \mathrm{P}(X_i \leq -a) = 1 - \Phi(a)$ and $\mathrm{P}(-b \leq X_i \leq b) = 2\Phi(b) - 1$ are presented in Table 2.2 together with the percentage of mass $(2\Phi(b) - 1)/\alpha$ allocated to the interior interval $[-b, b]$. In Figure 2.2, the density f_{ξ^*} of the optimal subsampling design ξ^* and the corresponding sensitivity function $\psi(x, \xi^*)$ are exhibited for $\alpha = 0.5$ and $\alpha = 0.1$. Vertical lines indicate the positions of the boundary



(a) $\alpha = 0.5$ (b) $\alpha = 0.1$

Figure 2.2: Density of the optimal subsampling design (solid line) and the standard normal distribution (dashed line, upper panels), and sensitivity functions (lower panels) for subsampling proportions $\alpha = 0.5$ (left) and $\alpha = 0.1$ (right)

points $-a$, $-b$, b, and a, respectively. In the subplots of the sensitivity function, the dotted horizontal line displays the threshold s^*. For other values of α, the plots are looking similar.

The numerical results in Table 2.2 suggest that the interior interval $[-b, b]$ does not vanish for any α $(0 < \alpha < 1)$. This will be established in the following theorem.

Theorem 2.7. *In quadratic regression with standard normal covariate X_i, for any subsampling proportion $\alpha \in (0, 1)$, the D-optimal subsampling design ξ^* has density $f_{\xi^*}(x) = f_X(x)\mathbb{1}_{(-\infty,-a]\cup[-b,b]\cup[a,\infty)}(x)$ with $a > b > 0$.*

For X_i having a general normal distribution with mean μ and variance σ^2, the optimal boundary points remain to be the same quantiles as in the standard normal case, $a_1, a_4 = \mu \pm \sigma a$ and $a_2, a_3 = \mu \pm \sigma b$, by Theorem 2.2.

Example 2.8 (uniform distribution)**.** If the covariate X_i is uniformly distributed on $[-1, 1]$ with density $f_X(x) = \frac{1}{2}\mathbb{1}_{[-1,1]}(x)$, we can obtain analytical results for the dependence of the subsampling design on the proportion α to be selected.

The distribution of X_i is symmetric. By Corollary 2.3, the density of the D-optimal continuous subsampling design ξ^* has the shape

$$f_{\xi^*}(x) = \frac{1}{2}\mathbb{1}_{[-1,-a]\cup[-b,b]\cup[a,1]}(x)\,, \tag{2.10}$$

where we formally allow $a = 1$ or $b = 0$ resulting in only one or two intervals of support. The relevant entries in the information matrix $\mathbf{M}(\xi^*)$ are $m_2(\xi^*) = \frac{1}{3}(1 - a^3 + b^3)$ and $m_4(\xi^*) = \frac{1}{5}(1 - a^5 + b^5)$. If, in Corollary 2.3, the boundary points a_1 and a_2 satisfy $a_1 \le 1$ and $a_2 \ge 0$, then $a = a_1$ and $b = a_2$ are the solution of the two equations $a - b = 1 - \alpha$ and $\alpha m_2(\xi^*)(a^2 + b^2) = \alpha m_4(\xi^*) - 3m_2(\xi^*)^2$ arising from conditions (2.7) and (2.9). On the other hand, if there exist solutions a and b of these equations such that $0 < b < a < 1$, then these are the boundary points in the representation (2.10), and the density of the optimal subsampling design is supported by three proper intervals. Solving the two equations results in

$$a(\alpha) = \frac{1}{2}(1 - \alpha) + \left(\frac{1}{180(1-\alpha)} \Big(45 - 15\alpha + 15\alpha^2 - 45\alpha^3 + 20\alpha^4 \right.$$
$$\left. - 4\alpha\sqrt{5}\sqrt{45 - 90\alpha + 90\alpha^2 - 75\alpha^3 + 57\alpha^4 - 27\alpha^5 + 5\alpha^6} \Big) \right)^{1/2} \tag{2.11}$$

and

$$b(\alpha) = a(\alpha) - (1 - \alpha) \tag{2.12}$$

for the dependence of a and b on α. The values of a and b are plotted in Figure 2.3. There

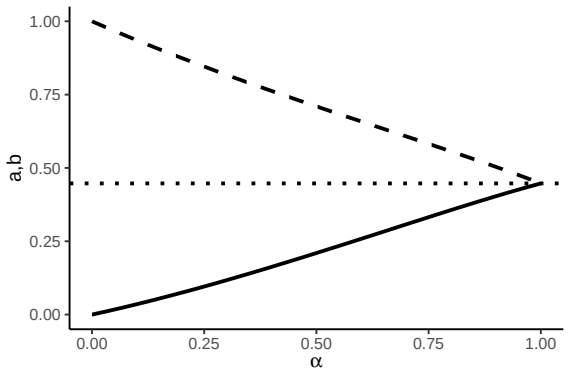


Figure 2.3: Boundary points a (dashed) and b (solid) of the D-optimal subsampling design in the case of uniform X_i on $[-1, 1]$ as functions of α

it can be seen that $0 < a < b < 1$ for all α and that a and b both tend to $1/\sqrt{5}$ as α tends to 1. Similar to the case of the normal distribution, the resulting values and illustrations are given in Table 2.3 and Figure 2.4. Note that the mass of the interior interval $\mathrm{P}(-b \leq X_i \leq b)$ is equal to b itself as X_i is uniformly distributed on $[-1, 1]$. Also here, in Figure 2.4,

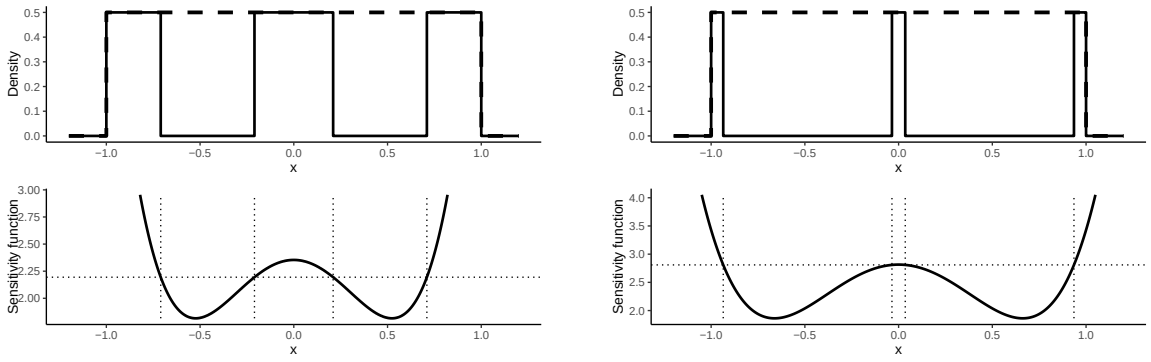
Table 2.3: Values for the boundary points a and b for selected values of the subsampling proportion α in the case of uniform X_i on $[-1, 1]$

α	a	$\mathrm{P}(X_i \geq a)$	$b = \mathrm{P}(-b \leq X_i \leq b)$	% of mass on $[-b, b]$
0.5	0.70983	0.14508	0.20983	41.97
0.3	0.81737	0.09132	0.11737	39.12
0.1	0.93546	0.03227	0.03546	35.46
0.01	0.99336	0.00332	0.00336	33.55

vertical lines indicate the positions of the boundary points $-a$, $-b$, b, and a, and the dotted horizontal line displays the threshold s^*. Moreover, the percentage of mass at the different intervals is displayed in Figure 2.5.

The results in Table 2.3 and Figure 2.5 suggest that the percentage of mass on all three intervals $[-1, -a]$, $[-b, b]$, and $[a, 1]$ tend to $1/3$ as α tends to 0. We establish this in the following theorem.

Theorem 2.9. *In quadratic regression with covariate X_i uniformly distributed on $[-1, 1]$, let ξ_α^* be the optimal subsampling design for subsampling proportion α, $0 < \alpha < 1$, defined in equations* (2.11) *and* (2.12)*. Then* $\lim_{\alpha \to 0} \xi_\alpha^*([-b, b])/\alpha = 1/3$.



(a) $\alpha = 0.5$ (b) $\alpha = 0.1$

Figure 2.4: Density of the optimal subsampling design (solid line) and the uniform distribution on $[-1, 1]$ (dashed line, upper panels), and sensitivity functions (lower panels) for subsampling proportions $\alpha = 0.5$ (left) and $\alpha = 0.1$ (right)

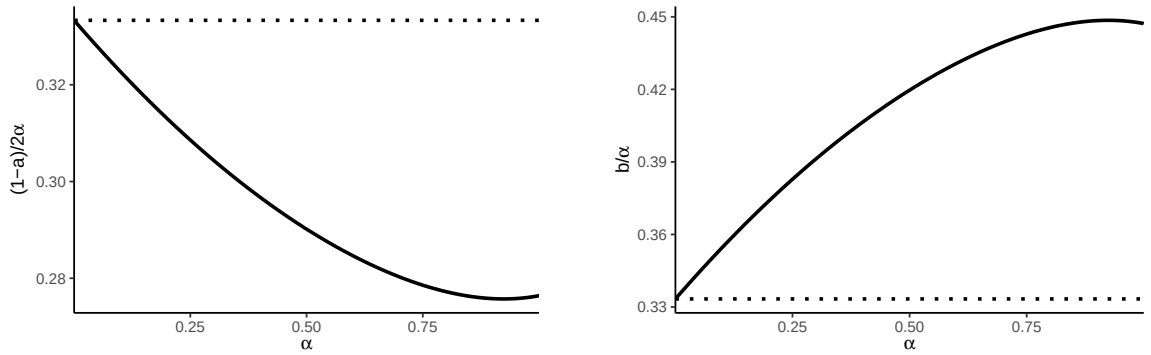


Figure 2.5: Percentage of mass on the support intervals $[a, 1]$ (left) and $[-b, b]$ (right) of the D-optimal subsampling design in the case of uniform X_i on $[-1, 1]$ as a function of α

It is worth-while mentioning that the percentages of mass displayed in Figure 2.5 are not monotonic over the whole range of $\alpha \in (0, 1)$, as, for example the percentage of mass at the interior interval $[-b, b]$ is increasing from 0.419666 at $b = 0.50$ to 0.448549 at $b = 0.92$ and then slightly decreasing back again to 0.447553 at $b = 0.99$.

Finally, it can be checked that, for all α, the solutions satisfy $0 < b < a < 1$ such that the optimal subsampling designs are supported on three proper intervals.

In the two preceding examples it could be noticed that the mass of observations is of comparable size for the three supporting intervals in the case of a normal and of a uniform distribution with light tails. This may be different in the case of a heavy-tailed distribution for the covariate X_i as the t-distribution.

Example 2.10 (t-distribution)**.** For the case that the covariate X_i comes from a t-distribution with ν degrees of freedom, we observe a behavior which differs substantially from the normal case of Example 2.6. The interior interval typically has less mass than the outer intervals and may vanish for some values of α. We show this in the case of the least possible number $\nu = 5$ of degrees of freedom to maintain an existing fourth moment, which appears in the information matrix of the D-optimal continuous subsampling design ξ^* while maximizing the dispersion.

Theorem 2.11. *In quadratic regression with t-distributed covariate $X_i \sim t_5$ with five degrees of freedom, there is a critical value $\alpha^* \approx 0.082065$ of the subsampling proportion α such that the D-optimal subsampling design ξ^* has*

(i) density $f_{\xi^*}(x) = f_X(x)\mathbb{1}_{(-\infty,-a]\cup[-b,b]\cup[a,\infty)}(x)$ with $a > b > 0$ for $\alpha < \alpha^*$.

(ii) density $f_{\xi^*}(x) = f_X(x)\mathbb{1}_{(-\infty,-t_{5,1-\alpha/2}]\cup[t_{5,1-\alpha/2},\infty)}(x)$, where $t_{5,1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the t_5-distribution, for $\alpha \geq \alpha^*$.

For illustration, numerical results are given in Table 2.4. The percentage of mass on the interior interval $[-b, b]$ is equal to zero for all larger values of α as stated in Theorem 2.11. The percentage of mass on $[-b, b]$ decreases with increasing subsampling proportion α before vanishing entirely.

Table 2.4: Values for the boundary points a and b for selected values of the subsampling proportion α in the case of t_5-distributed X_i

α	a	$\mathrm{P}(X_i \geq a)$	b	$\mathrm{P}(-b \leq X_i \leq b)$	% of mass on $[-b, b]$
0.10	2.01505	0.05000	0	0	0
0.07	2.31512	0.03423	0.00202	0.00153	2.03
0.03	3.09141	0.01356	0.00380	0.00288	4.74
0.01	4.18942	0.00429	0.00187	0.00142	14.23

Further calculations provide that the critical value α^*, where the D-optimal subsampling design switches from a three-interval support to a two-interval support, increases with the number of degrees ν of freedom of the t-distribution and converges to one when ν tends to infinity. This is in accordance with the results for the normal distribution in Example 2.6 as the t-distribution converges in distribution to a standard normal distribution for $\nu \to \infty$. We have given numeric values for the crossover points for selected degrees of freedom in Table 2.5, where $\nu = \infty$ relates to the normal distribution. The corresponding value $\alpha^* = 1$

indicates that the D-optimal subsampling design is supported by three intervals for all α in this case.

Table 2.5: Values of the critical value α^* for selected degrees of freedom ν of the t-distribution

ν	5	6	7	8	9	30	∞
α^*	0.08207	0.34670	0.50374	0.60125	0.66670	0.92583	1

2.6 Efficiency

To exhibit the gain in using a D-optimal subsampling design compared to random subsampling, we consider the performance of the uniform random subsampling design ξ_α of size α, which has density $f_{\xi_\alpha}(x) = \alpha f_X(x)$, compared to the D-optimal subsampling design ξ_α^* with mass α.

More precisely, the D-efficiency of any subsampling design ξ with mass α is defined as

$$\text{eff}_{D,\alpha}(\xi) = \left(\frac{\det(\mathbf{M}(\xi))}{\det(\mathbf{M}(\xi_\alpha^*))} \right)^{1/p},$$

where p is the dimension of the parameter vector $\boldsymbol{\beta}$. For this definition the homogeneous version $(\det(\mathbf{M}(\xi)))^{1/p}$ of the D-criterion is used which satisfies the homogeneity condition $(\det(\lambda\mathbf{M}(\xi)))^{1/p} = \lambda(\det(\mathbf{M}(\xi)))^{1/p}$ for all $\lambda > 0$ (see Pukelsheim, 1993, Chapter 6.2).

For uniform random subsampling, the information matrix is given by $\mathbf{M}(\xi_\alpha) = \alpha\mathbf{M}(\xi_1)$, where $\mathbf{M}(\xi_1)$ is the information matrix for the full sample with raw moments $m_k(\xi_1) = \mathrm{E}(\boldsymbol{X}_i^k)$ as entries in the (j, j')th position, $j + j' - 2 = k$. Thus, the D-efficiency $\text{eff}_{D,\alpha}(\xi_\alpha)$ of uniform random subsampling can be nicely interpreted: the sample size (mass) required to obtain the same precision (in terms of the D-criterion), as when the D-optimal subsampling design ξ_α^* of mass α is used, is equal to the inverse of the efficiency $\text{eff}_{D,\alpha}(\xi_\alpha)^{-1}$ times α. For example, if the efficiency $\text{eff}_{D,\alpha}(\xi_\alpha)$ is equal to 0.5, then twice as many observations would be needed under uniform random sampling than for a D-optimal subsampling design of size α. Of course, the full sample has higher information than any proper subsample such that, obviously, for uniform random subsampling, $\text{eff}_{D,\alpha}(\xi_\alpha) \geq \alpha$ holds for all α.

For the examples of Sections 2.4 and 2.5, the efficiency of uniform random subsampling is given in Table 2.6 for selected values of α and exhibited in Figure 2.6 for the full range of α between 0 and 1 (solid lines). Here the determinant of the information matrix is determined

Table 2.6: Efficiency of uniform subsampling w.r.t. D-optimality for selected values of the subsampling proportion α

		α			
		0.5	0.3	0.1	0.01
linear regression	normal	0.73376	0.61886	0.47712	0.34403
	exponential	0.73552	0.61907	0.46559	0.30690
quadratic regression	normal	0.73047	0.59839	0.41991	0.24837
	uniform	0.78803	0.70475	0.62411	0.58871
	t_5	0.66400	0.50656	0.29886	0.10941
	t_9	0.70390	0.56087	0.36344	0.17097

as in the examples of Sections 2.4 and 2.5 for the optimal subsampling designs ξ_α^* either numerically or by explicit formulas where available.

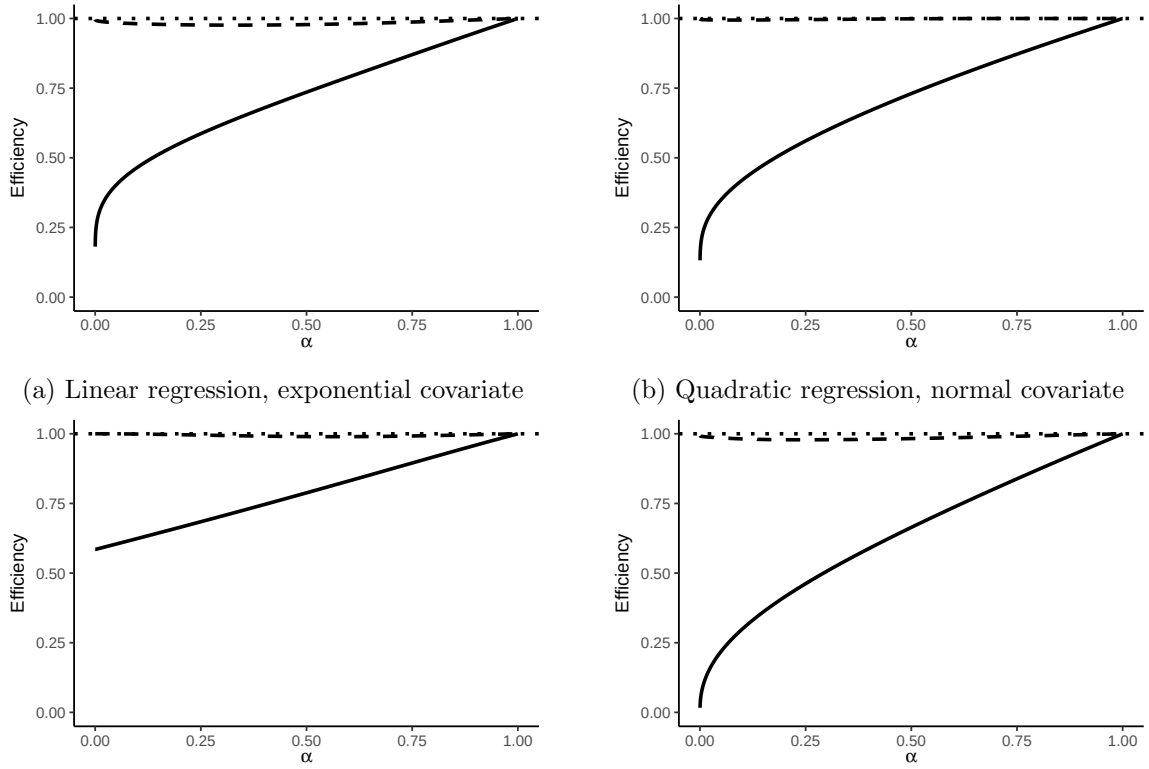
Both Table 2.6 and Figure 2.6 indicate that the efficiency of uniform random subsampling is decreasing in all cases when the proportion α of subsampling gets smaller. In the case of quadratic regression with uniformly distributed covariate, the decrease is more or less linear with a minimum value of approximately 0.58 when α is small. In the other cases, where the distribution of the covariate is unbounded, the efficiency apparently decreases faster, when the proportion α is smaller than 10%, and tends to 0 for $\alpha \to 0$.

The latter property can be easily seen for linear regression and symmetric distributions: there, the efficiency $\mathrm{eff}_{D,\alpha}(\xi_\alpha)$ of uniform random sampling is bounded from above by $c/q_{1-\alpha/2}$, where $c = \mathrm{E}(X_i^2)^{1/2}$ is a constant and $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the distribution of the covariate. When the distribution is unbounded like the normal distribution, then these quantiles tend to infinity for $\alpha \to 0$ and, hence, the efficiency tends to 0. Similar results hold for quadratic regression and asymmetric distributions.

In any case, as can be seen from Table 2.6, the efficiency of uniform random subsampling is quite low for reasonable proportions $\alpha \leq 0.1$ and, hence, the gain in using the D-optimal subsampling design is substantial.

By equivariance arguments as indicated above in the examples of Sections 2.4 and 2.5, the present efficiency considerations carry over directly to a covariate having a general normal, exponential, or uniform distribution, respectively.

In the IBOSS approach by Wang et al. (2019), half of the proportion α is taken from both tails of the data. The corresponding continuous subsampling design ξ'_α would be to have two intervals $(-\infty, b]$ and $[a, \infty)$ and to choose the boundary points a and b to be the



(a) Linear regression, exponential covariate

(b) Quadratic regression, normal covariate

(c) Quadratic regression, uniform covariate

(d) Quadratic regression, t_5 covariate

Figure 2.6: Efficiency of uniform random subsampling (solid line) and of an IBOSS-type subsampling design (dashed line) w.r.t. D-optimality

$(1-\alpha/2)$- and $(\alpha/2)$-quantile of the distribution of the covariate, respectively. For linear regression, it can been seen from Corollary 2.3 that the subsampling design ξ'_α is D-optimal when the distribution of the covariate is symmetric. As the IBOSS procedure does not use prior knowledge of the distribution, it would be tempting to investigate the efficiency of the corresponding continuous subsampling design ξ'_α under asymmetric distributions. For the exponential distribution, this efficiency $\mathrm{eff}_{D,\alpha}(\xi'_\alpha)$ is added to the upper left panel in Figure 2.6 by a dashed line. There the subsampling design ξ'_α shows a remarkably high efficiency over the whole range of α with a minimum value 0.976 at $\alpha = 0.332$.

As an extension of IBOSS for quadratic regression, we may propose a procedure which takes proportions $\alpha/3$ from both tails of the data as well as from the center of the data. This procedure can be performed without any prior knowledge of the distribution of the covariate. The choice of the proportions $\alpha/3$ is motivated by the standard case D-optimal

design on an interval where one third of the weight is allocated to each of the endpoints and to the midpoint of the interval, respectively. For a symmetric distribution, the corresponding continuous subsampling design ξ_α'' can be defined by the boundary points a and b to be the $(1-\alpha/3)$- and $(1/2+\alpha/6)$-quantile of the distribution of the covariate, respectively. In the case of the uniform distribution, the subsampling design ξ_α'' is the limiting D-optimal subsampling design for $\alpha \to 0$ by Theorem 2.9. In Figure 2.6, the efficiency $\text{eff}_{D,\alpha}(\xi_\alpha'')$ is shown by dashed lines for the whole range of α for the uniform distribution as well as for the normal and for the t- distribution in the case of quadratic regression. In all three cases, the subsampling design ξ_α'' is highly efficient over the whole range of α with minimum values 0.994 at $\alpha = 0.079$ for the normal distribution, 0.989 at $\alpha = 0.565$ for the uniform distribution, and 0.978 at $\alpha = 0.245$ for the t_5-distribution, respectively. This is of particular interest for the t_5-distribution, where the interior interval of the D-optimal subsampling design ξ_α^* is considerably smaller than of the IBOSS-like subsampling design ξ_α'' and even vanishes entirely for $\alpha > \alpha^* \approx 0.08$. However, we only tested this extension of IBOSS for quadratic regression for symmetric distributions of the covariate. Further investigations for non-symmetric distributions is necessary.

2.7 Concluding Remarks

In this paper we have considered a theoretical approach to evaluate subsampling designs under distributional assumptions on the covariate in the case of polynomial regression on a single explanatory variable. We first reformulated the constrained equivalence theorem under Kuhn-Tucker conditions in Sahm and Schwabe (2001) to characterize the D-optimal continuous subsampling design for general distributions of the covariate. For symmetric distributions of the covariate we concluded the following. The D-optimal subsampling design is equal to the bounding distribution in its support and the support of the optimal subsampling design will be the union of at most $q+1$ intervals that are symmetrically placed around zero. Further we have found that in the case of quadratic regression the D-optimal subsampling design has three support intervals with positive mass for all $\alpha \in (0,1)$, whereas the interior interval vanishes for some α for a t-distributed covariate. In contrast to that, for linear regression, always two intervals are required at the tails of the distribution.

The main emphasis in this work was on D-optimal subsampling designs. But many of the results may be extended to other optimality criteria like A- and E-optimality from the Kiefer's Φ_q-class of optimality criteria, $IMSE$-optimality for predicting the mean response, or optimality criteria based on subsets or linear functionals of parameters.

The D-optimal subsampling designs show a high performance compared to uniform random subsampling. In particular, for small proportions, the efficiency of uniform random subsampling tends to zero when the distribution of the covariate is unbounded. This property is in accordance with the observation that estimation based on subsampling according to IBOSS is "consistent" in the sense that the mean squared error goes to zero with increasing population size even when the size of the subsample is fixed.

We propose a generalization of the IBOSS method to quadratic regression which does not require prior knowledge of the distribution of the covariate and which performs remarkably well compared to the optimal subsampling design. However, an extension to higher order polynomials does not seem to be obvious.

2.A Proofs

Before proving Theorem 2.1, we establish two preparatory lemmas on properties of the sensitivity function $\psi(x, \xi)$ for a continuous subsampling design ξ with density $f_\xi(x)$ and reformulate an equivalence theorem on constraint design optimality by Sahm and Schwabe (2001) for the present setting. The first lemma deals with the shape of the sensitivity function.

Lemma 2.12. *The sensitivity function $\psi(x, \xi)$ is a polynomial of degree $2q$ with positive leading term.*

Proof of Lemma 2.12. For a continuous subsampling design ξ with density $f_\xi(x)$, the information matrix $\mathbf{M}(\xi)$ and, hence, its inverse $\mathbf{M}(\xi)^{-1}$ is positive definite. Thus the last diagonal element $m^{(pp)}$ of $\mathbf{M}(\xi)^{-1}$ is positive and, as $\mathbf{f}(x) = (1, x, \ldots, x^q)^\top$, the sensitivity function $\psi(x, \xi) = \mathbf{f}(x)^\top \mathbf{M}(\xi)^{-1} \mathbf{f}(x)$ is a polynomial of degree $2q$ with coefficient $m^{(pp)} > 0$ of the leading term. □

The second lemma reveals a distributional property of the sensitivity function considered as a function in the covariate X_i.

Lemma 2.13. *The random variable $\psi(X_i, \xi)$ has a continuous cumulative distribution function.*

Proof of Lemma 2.13. As the sensitivity function $\psi(x, \xi)$ is a non-constant polynomial by Lemma 2.12, the equation $\psi(x, \xi) = s$ has only finitely many roots $x_1, \ldots, x_\ell$, $\ell \leq 2q$, say, by the fundamental theorem of algebra. Hence, $\mathrm{P}(\psi(X_i, \xi) = s) = \sum_{k=1}^{\ell} \mathrm{P}(X_i = x_k) = 0$ by the continuity of the distribution of X_i which proves the continuity of the cumulative distribution function of $\psi(X_i, \xi)$. □ □

With the continuity of the distribution of $\psi(X_i, \xi^*)$ the following equivalence theorem can be obtained from Corollary 1(c) in Sahm and Schwabe (2001) for the present setting by transition from the directional derivative to the sensitivity function and considering $\mathbb{R}$ as the design region.

Theorem 2.14 (Equivalence Theorem)**.** *The subsampling design ξ^* is D-optimal if and only if there exist a threshold s^* and a subset $\mathcal{X}^*$ of $\mathbb{R}$ such that*

(i) the D-optimal subsampling design ξ^* is given by

$$f_{\xi^*}(x) = f_X(x)\mathbb{1}_{\mathcal{X}^*}(x)$$

(ii) $\psi(x, \xi^*) \geq s^*$ for $x \in \mathcal{X}^*$, and

(iii) $\psi(x, \xi^*) < s^*$ for $x \notin \mathcal{X}^*$.

As $\mathrm{P}(\psi(X_i, \xi^*) \geq s^*) = \mathrm{P}(X_i \in \mathcal{X}^*) = \int f_{\xi^*}(x)\,\mathrm{d}x = \alpha$, the threshold s^* is the $(1-\alpha)$-quantile of the distribution of $\psi(X_i, \xi^*)$.

Proof of Theorem 2.1. By Lemma 2.12 the sensitivity function $\psi(x, \xi)$ is a polynomial in x of degree $2q$ with positive leading term. Using the same argument as in the proof of Lemma 2.13 we obtain that there are at most $2q$ roots of the equation $\psi(x, \xi^*) = s^*$ and, hence, there are at most $2q$ sign changes in $\psi(x, \xi^*) - s^*$. As $\psi(x, \xi^*)$ is a polynomial of even degree, also the number of (proper) sign changes has to be even, and they occur at $a_1 > \cdots > a_{2r}$, say, $r \leq q$. Moreover, for $0 < \alpha < 1$, $\mathcal{X}^*$ is a proper subset of $\mathbb{R}$ and, thus, there must be at least one sign change, $r \geq 1$. Finally, as the leading coefficient of $\psi(x, \xi^*)$ is positive, $\psi(x, \xi^*)$ gets larger than s^* for $x \to \pm\infty$ and, hence, the outmost intervals $[a_1, \infty)$ and $(-\infty, a_{2r}]$ are included in the support $\mathcal{X}^*$ of ξ^*. By the interlacing property of intervals with positive and negative sign for $\psi(x, \xi^*) - s^*$, the result follows from the conditions on the D-optimal subsampling design ξ^* in Theorem 2.14. □

Proof of Theorem 2.2. First note that for any μ and $\sigma > 0$, the location-scale transformation $z = \sigma x + \mu$ is conformable with the regression function $\mathbf{f}(x)$, i. e. there exists a non-singular matrix $\mathbf{Q}$ such that $\mathbf{f}(\sigma x + \mu) = \mathbf{Q}\mathbf{f}(x)$ for all x. Then, for any design ξ bounded by $f_X(x)$, the design ζ has density $f_\zeta(z) = \frac{1}{\sigma} f_\xi(\frac{z-\mu}{\sigma})$ bounded by $f_Z(z) = \frac{1}{\sigma} f_X(\frac{z-\mu}{\sigma})$. Hence, by the

transformation theorem for measure integrals, it holds that

$$\begin{aligned}\mathbf{M}(\zeta) &= \int \mathbf{f}(z)\mathbf{f}(z)^\top \zeta(\mathrm{d}z) \\ &= \int \mathbf{f}(\sigma x + \mu)\mathbf{f}(\sigma x + \mu)^\top \xi(\mathrm{d}x) \\ &= \int \mathbf{Q}\mathbf{f}(x)\mathbf{f}(x)^\top \mathbf{Q}^\top \xi(\mathrm{d}x) \\ &= \mathbf{Q}\mathbf{M}(\xi)\mathbf{Q}^\top.\end{aligned}$$

Therefore $\det(\mathbf{M}(\zeta)) = \det(\mathbf{Q})^2 \det(\mathbf{M}(\xi))$. Thus ξ^* maximizes the D-criterion over the set of subsampling designs bounded by $f_X(x)$ if and only if ζ_* maximizes the D-criterion over the set of subsampling designs bounded by $f_Z(z)$. □

Proof of Corollary 2.3. The checkerboard structure of the information matrix $\mathbf{M}(\xi^*)$ carries over to its inverse $\mathbf{M}(\xi^*)^{-1}$. Hence, the sensitivity function $\psi(x, \xi^*)$ is an even polynomial, which has only non-zero coefficients for even powers of x, and is thus symmetric with respect to 0, i. e. $\psi(-x, \xi^*) = \psi(x, \xi^*)$. Accordingly, also the roots of $\psi(x, \xi^*) = s^*$ are symmetric with respect to 0. □

Proof of Theorem 2.7. In view of the shape (2.6) of the density and by Corollary 2.3, the tails are included in the optimal subsampling design such that $a < \infty$.

Next, we consider the symmetric design ξ' which is supported only on the tails and which will be the optimal subsampling design when $b = 0$. This design has density $f_{\xi'}(x) = \mathbb{1}_{(-\infty,-a]\cup[a,\infty)}(x) f_X(x)$ with $a = z_{1-\alpha/2}$ for given α. The information matrix $\mathbf{M}(\xi')$ is of the form (2.4) with relevant entries

$$\begin{aligned}m_2(\xi') &= \alpha + \sqrt{2/\pi} a \exp(-a^2/2), \\ m_4(\xi') &= 3m_2(\xi') + \sqrt{2/\pi} a^3 \exp(-a^2/2)\,.\end{aligned}$$

For the sensitivity function (2.5), we have

$$\psi(0, \xi') = \frac{\alpha m_4(\xi')}{\alpha m_4(\xi') - m_2(\xi')^2}$$

and

$$\psi(a,\xi') = \frac{\alpha m_4(\xi')}{\alpha m_4(\xi') - m_2(\xi')^2} - \frac{\alpha 2 m_2(\xi') a^2}{\alpha m_4(\xi') - m_2(\xi')^2} + \frac{\alpha a^2}{m_2(\xi')} + \frac{\alpha^2 a^4}{\alpha m_4(\xi') - m_2(\xi')^2}\,.$$

Let $c(\alpha) = \psi(0,\xi') - \psi(a,\xi')$ be the difference between the values of the sensitivity function at $x = 0$ and $x = a$, then

$$c(\alpha) = \alpha a^2 \left(\frac{2 m_2(\xi')}{\alpha m_4(\xi') - m_2(\xi')^2} - \frac{a^2 \alpha}{\alpha m_4(\xi') - m_2(\xi')^2} - \frac{1}{m_2(\xi')} \right). \tag{2.13}$$

$c(\alpha)$ is continuous in α and does not have any roots in $(0,1)$. Further, it can be checked that $c(0.1) > 0$, say. Thus $c(\alpha) > 0$ which means that $\psi(0,\xi') > \psi(a,\xi')$ for all α. Hence, by Theorem 2.14, the subsampling design ξ' cannot be optimal and, as a consequence, the optimal subsampling design ξ^* has support on three proper intervals with $b > 0$ for all α. □

Proof of Theorem 2.9. Let

$$u(\alpha) = 45 - 15\alpha + 15\alpha^2 - 45\alpha^3 + 20\alpha^4 - 4\sqrt{5}\sqrt{45\alpha^2 - 90\alpha^3 + 90\alpha^4 - 75\alpha^5 + 57\alpha^6 - 27\alpha^7 + 5\alpha^8}$$

and

$$v(\alpha) = 180(1-\alpha)\,.$$

Then

$$b(\alpha) = \left(\frac{u(\alpha)}{v(\alpha)} \right)^{1/2} - \frac{1}{2}(1-\alpha)\,.$$

We have $u(0) = 45$, $v(0) = 180$, and $b(\alpha)$ can be continuously extended to $b(0) = 0$ at $\alpha = 0$. The derivative of b is given by

$$b'(\alpha) = \frac{1}{2} + \frac{1}{2} \frac{u'(\alpha)v(\alpha) - u(\alpha)v'(\alpha)}{v(\alpha)^2} \sqrt{\frac{v(\alpha)}{u(\alpha)}}\,, \tag{2.14}$$

where

$$u'(\alpha) = -15 + 30\alpha - 135\alpha^2 + 80\alpha^3 - w(\alpha), \tag{2.15}$$

$$v'(\alpha) = -180, \tag{2.16}$$

and

$$w(\alpha) = 2\sqrt{5}\frac{90 - 270\alpha + 360\alpha^2 - 375\alpha^3 + 342\alpha^4 - 189\alpha^5 + 40\alpha^6}{\sqrt{45 - 90\alpha + 90\alpha^2 - 75\alpha^3 + 57\alpha^4 - 27\alpha^5 + 5\alpha^6}}.$$

We have $v'(0) = -180$. To determine $u'(0)$ we note that $w(0) = 60$ and thus $u'(0) = -75$. Hence, also the derivative $b'(\alpha)$ can be continuously extended at $\alpha = 0$ and the value for $b'(0)$ can be obtained by plugging in the values of $u(0)$, $v(0)$, $u'(0)$, and $v'(0)$ into formula (2.14),

$$b'(0) = \frac{1}{2} + \frac{1}{2}\frac{-75 \cdot 180 + 45 \cdot 180}{180^2}\sqrt{\frac{180}{45}} = \frac{1}{3}\,.$$

Finally, we note that $b(\alpha)/\alpha$ is the percentage of mass on the interior interval $[-b(\alpha), b(\alpha)]$ and that $\lim_{\alpha\to 0} b(\alpha)/\alpha$ is the derivative $b'(0)$ of $b(\alpha)$ at $\alpha = 0$. Hence, the percentage of mass on the interior interval tends to $b'(0) = 1/3$ when the subsampling proportion α goes to 0. □

Proof of Theorem 2.11. The proof will follow the idea of the proof of Theorem 2.7. For $\alpha \in (0, 1)$, we consider the symmetric design ξ' which is supported only on the tails and which will be the optimal subsampling design when $b = 0$. This design has density $f_{\xi'}(x) = \mathbb{1}_{(-\infty,-a]\cup[a,\infty)(x)} f_X(x)$ with $a = t_{5,1-\alpha/2}$. The relevant entries of the information matrix $\mathbf{M}(\xi')$ are

$$m_2(\xi') = \frac{5}{3\pi}\left(\pi - \frac{2\sqrt{5}a(a^2-5)}{(a^2+5)^2} - 2\arctan(a/\sqrt{5})\right),$$

$$m_4(\xi') = \frac{25}{3\pi}\left(3\pi + \frac{10\sqrt{5}a(a^2+3)}{(a^2+5)^2} - 6\arctan(a/\sqrt{5})\right).$$

The sensitivity function $\psi(x, \xi')$ and the difference $c(\alpha) = \psi(0, \xi') - \psi(a, \xi')$ between the values of the sensitivity function at $x = 0$ and $x = a$ are defined as for the normal distribution with the above moments $m_2(\xi')$ and $m_4(\xi')$ related to the t-distribution inserted. The function $c(\alpha)$ defined by (2.13) then looks as shown in Figure 2.7. The vertical dotted line indicates the position of the critical value $\alpha^* \approx 0.082065$, where the curve of the function

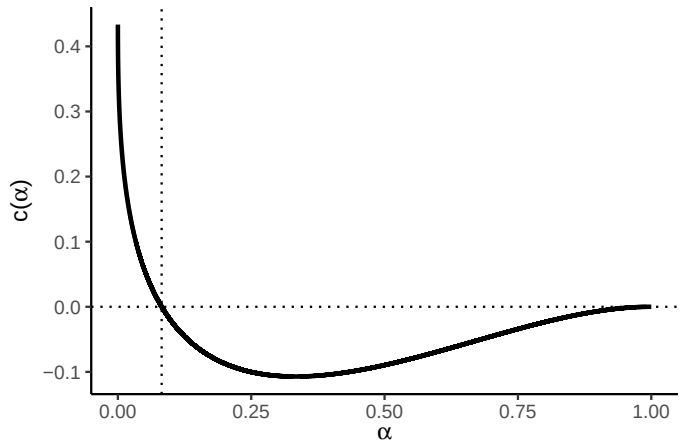


Figure 2.7: Difference $c(\alpha) = \psi(0, \xi') - \psi(a, \xi')$ (solid) for the case of a t-distributed covariate with 5 degrees of freedom

$c(\alpha)$ intersects the horizontal dotted line indicating $c = 0$.

Thus for $\alpha < \alpha^* \approx 0.082065$ we have $\psi(0, \xi') > \psi(a, \xi')$ and the design ξ' cannot be optimal by Theorem 2.14. In this situation, an inner interval has to be included in the optimal subsampling design ξ^* with $b > 0$.

Conversely, for $\alpha \geq \alpha^* \approx 0.082065$ we have that $\psi(0, \xi') \leq \psi(a, \xi')$. Hence, the design ξ' is optimal by Theorem 2.14, and no inner interval has to be added to the optimal subsampling design $\xi^* = \xi'$ $(b = 0)$. □

Chapter 3

D-optimal Subsampling Design for Massive Data Linear Regression

In this chapter we present the work titled "D-optimal Subsampling Design for Massive Data Linear Regression" (Reuter and Schwabe, 2023b) published as an electronic preprint.

3.1 Introduction

Data reduction is a fundamental challenge of modern technology, which allows us to collect huge amounts of data. Often, technological advances in computing power do not keep pace with the amount of data, creating a need for data reduction. We speak of big data whenever the full data size is too large to be handled by traditional statistical methods. We usually distinguish between the case where the number of covariates is large and the case where there are very many observations. The first case is referred to as high-dimensional data and numerous methods have been studied to deal with such data, most notably LASSO by Tibshirani (1996), which utilizes ℓ_1 penalization to find sparse parameter vectors, thus fusing subset selection and ridge regression. We consider the second case, referred to as massive data. To deal with huge amounts of observations typically one of two methods is applied: One strategy is to divide the data into several smaller datasets and compute them separately, known as divide-and-conquer, see Lin and Xi (2011). Alternatively one can find an informative subsample of the full data. This can be done in a probabilistic fashion, creating random subsamples in a nonuniform manner. Among the prominent studies are Drineas et al. (2006), Mahoney (2011) and Ma et al. (2014). They present subsampling methods for linear regression models called algorithmic leveraging, which draw samples

according to probabilities based on the normalized statistical leverage scores of the covariate matrix. More recently, Dereziński and Warmuth (2018) studied volume sampling, where subsamples are chosen proportional to the squared volume of the parallelepiped spanned by its observations. Conversely, subdata can be selected in a deterministic way. Shi and Tang (2021) present such a method, that maximizes the minimal distance between two observations in the subdata. Most prominently, Wang et al. (2019) have introduced the information-based optimal subdata selection (IBOSS) to tackle big data linear regression in a deterministic fashion based on D-optimality. The IBOSS approach selects the outer-most data points of each covariate successively. Other subsampling methods for linear regression include the works by Wang et al. (2021), who have introduced orthogonal subsampling inspired by orthogonal arrays, which selects observations in the corners of the design space and the optimal design based subsampling scheme by Deldossi and Tommasi (2021). Subsampling becomes increasingly popular, leading to more work outside linear models. Cheng et al. (2020) extent the idea of the IBOSS method from the linear model to logistic regression and other work on generalized linear regression include the papers by Zhang et al. (2021) and Ul Hassan and Miller (2019). Su et al. (2022) have recently considered subsampling for missing data, whereas Joseph and Mak (2021) focused on non-parametric models and make use of the information in the dependent variables. Various works consider subsampling when the full data is distributed over several data sources, among them Yu et al. (2022) and Zhang and Wang (2021) For a more thorough recent review on design inspired subsampling methods see the work by Yu et al. (2024).

In this paper we assume that both the model and the shape of the joint distribution of the covariates are known. We search for D-optimal continuous subsampling designs of total measure α that are bounded from above by the distribution of the covariates. Wynn (1977) and Fedorov (1989) were the first to study such directly bounded designs. Pronzato (2004) considered this setting using subsampling designs standardized to one and bounded by $1/\alpha$ times the distribution of the covariates. More recently, the same has been studied by Pronzato and Wang (2021) in the context of sequential subsampling. In Reuter and Schwabe (2023a) we have studied bounded D-optimal subsampling designs for polynomial regression in one covariate, using many similar ideas as we use here. We stay with the unstandardized version emphasizing the subsampling character of the design. For the characterization of the optimal subsampling design, we will make use of an equivalence theorem from Sahm and Schwabe (2001). This equivalence theorem allows us to construct such subsampling designs for different settings of the distributional assumptions on the covariates. Based on this, we propose a simple subsampling scheme for selecting observations. This method

includes all data points in the support of the optimal subsampling design and rejects all other observations. Although this approach is basically probabilistic, as it allows selection probabilities, the resulting optimal subsampling design is purely deterministic, since it depends only on the acceptance region defined by the optimal subsampling design. We make comments on the asymptotic behavior of the ordinary least squares estimator based on the D-optimal subsampling design that selects the data points with the largest Mahalanobis distance from the mean of the data.

Since the proposed algorithm requires computational complexity of the same magnitude as calculating the least squares estimator on the full data, we also propose a simplified version with lower computational complexity, that takes the variances of the covariates into account while disregarding the covariances between them.

The rest of this paper is organized as follows. After introducing the model in Section 3.2 we present the setup and establish necessary concepts and notation in Section 3.3. Section 3.3.1 illustrates our methodology for linear regression in one explanatory variable. We construct optimal subsampling designs for multiple linear regression in Section 3.3.2. In Section 3.4 we consider the case of a fixed subsample size, then examine the performance of our method in simulation studies in Section 3.5. We make concluding remarks in Section 3.6. Technical details and proofs are deferred to an Appendix.

3.2 Model Specification

We treat the situation of data $(\mathbf{x}_i, y_i)$, where y_i is the value of the response variable Y_i and the $\mathbf{x}_i$ are realizations of the d-dimensional i.i.d. random vectors $\boldsymbol{X}_i$ of covariates with probability density function $f_{\boldsymbol{X}}$ for unit $i = 1, \ldots, n$. We assume the covariates $\boldsymbol{X}_i$ have an elliptical distribution. We suppose that the dependence of the response variable on the covariates $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^\top$ is given by the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_d X_{id} + \varepsilon_i$$

with independent, homoscedastic errors ε_i with zero mean and $\mathrm{Var}[\varepsilon_i] = \sigma_\varepsilon^2 < \infty$ which we assume to be independent of all $\boldsymbol{X}_{i'}$.

We assume that the number of observations n is very large. The aim is to estimate the regression parameter $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)^\top$, where β_0 is the intercept and β_j is the slope parameter in the j-th component x_j of $\mathbf{x} = (x_1, \ldots, x_d)^\top$ for $j = 1, \ldots, d$. For notational

convenience we write the multiple linear regression model as a general linear model

$$Y_i = \mathbf{f}(\boldsymbol{X}_i)^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n\,,$$

where $\mathbf{f}(\mathbf{x}) = (1, \mathbf{x}^\top)^\top$.

3.3 Subsampling Design

We consider a scenario where the y_i are expensive to observe and therefore only a percentage α ($0 < \alpha < 1$) of the y_i are observed, given all $\mathbf{x}_i$. Another possible setting is that all y_i and $\mathbf{x}_i$ are available, but parameter estimation is only computationally feasible on a percentage α of the data. Either setup leads to the question which subsample of the data $(\mathbf{x}_i, y_i)$ yields the best estimation of the parameter $\boldsymbol{\beta}$ or essential parts of it.

Throughout this section we assume, that the distribution of $\boldsymbol{X}_i$ and its density $f_{\boldsymbol{X}}$ are known. We consider continuous designs ξ with total measure α on $\mathbb{R}^d$ with density functions f_ξ that are bounded from above by the density of the covariates $f_{\boldsymbol{X}}$ such that $\int f_\xi(\mathbf{x})\,\mathrm{d}\mathbf{x} = \alpha$ and $f_\xi(\mathbf{x}) \leq f_{\boldsymbol{X}}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$. The resulting set of all such designs ξ is denoted by $\Xi^{f_{\boldsymbol{X}}}$. A subsample can then be generated according to such a continuous design by accepting units i with probability $f_\xi(\mathbf{x}_i)/f_{\boldsymbol{X}}(\mathbf{x}_i)$.

Let $\mathbf{M}(\xi) = \int \mathbf{f}(\mathbf{x})\mathbf{f}(\mathbf{x})^\top \xi(\mathrm{d}\mathbf{x})$ be the information matrix of ξ. We require $\mathrm{E}[\|\boldsymbol{X}_i\|_2^2] < \infty$ as some entries of the information matrix can be infinite otherwise. $\mathbf{M}(\xi)$ measures the quality of the least squares estimator $\hat{\boldsymbol{\beta}}$ based on a subsample according to ξ in the sense that $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ asymptotically follows a normal distribution with mean zero and covariance matrix $\sigma_\varepsilon^2 \mathbf{M}(\xi)^{-1}$ when n tends to infinity. To find an appropriate subsampling design $\xi \in \Xi^{f_{\boldsymbol{X}}}$, we aim to minimize the design criterion for D-optimality $\Psi(\xi) = -\ln(\det(\mathbf{M}(\xi)))$. Then, the D-optimal design minimizes the determinant of the asymptotic covariance matrix of the parameter least squares estimator and can be interpreted as minimizing the volume of the respective confidence ellipsoid of $\boldsymbol{\beta}$. The optimal subsampling design that minimizes $\Psi(\xi)$ in $\Xi^{f_{\boldsymbol{X}}}$ is denoted by ξ^* with density f_{ξ^*}. We make use of the sensitivity function $\psi(\mathbf{x}, \xi) = \alpha \mathbf{f}(\mathbf{x})^\top \mathbf{M}(\xi)^{-1} \mathbf{f}(\mathbf{x})$ (see Lemma 3.16). For the characterization of the D-optimal continuous subsampling design, we apply the constrained equivalence theorem under Kuhn-Tucker conditions (see Sahm and Schwabe, 2001, Corollary 1 (c)) to the present case of multiple linear regression in the following theorem.

Theorem 3.1. *In multiple linear regression with $d \geq 2$ covariates with density $f_{\boldsymbol{X}}(\mathbf{x})$ of the covariates $\boldsymbol{X}_i$, the subsampling design ξ^* is D-optimal if and only if there exist a subset*

$\mathcal{X}^* \subset \mathbb{R}^d$ *and a threshold* s^* *such that*

(i) ξ^* *has density* $f_{\xi^*}(\mathbf{x}) = f_{\boldsymbol{X}}(\mathbf{x})\mathbb{1}_{\mathcal{X}^*}(\mathbf{x})$

(ii) $\psi(\mathbf{x}, \xi^*) \geq s^*$ *for* $\mathbf{x} \in \mathcal{X}^*$*, and*

(iii) $\psi(\mathbf{x}, \xi^*) < s^*$ *for* $\mathbf{x} \notin \mathcal{X}^*$*.*

Here, $\mathbb{1}_A(\mathbf{x})$ denotes the indicator function, i. e. $\mathbb{1}_A(\mathbf{x}) = 1$, if $\mathbf{x} \in A$ and $\mathbb{1}_A(\mathbf{x}) = 0$ otherwise. Before treating the general case of subsampling design in multiple linear regression, we briefly present some results from Reuter and Schwabe (2023a) for the case of ordinary linear regression in one covariate for illustrative purposes.

3.3.1 Subsampling Design in a single Covariate

In the case of linear regression in one covariate X_i, we have $d = 1$, $\mathbf{f}(x) = (1, x)^\top$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$. We assume the known distribution of the covariate X_i to be symmetric ($f_X(-x) = f_X(x)$) and to have a finite second moment ($\mathrm{E}[X_i^2] < \infty$). We use the linear equivariance of the regression function, $\mathbf{f}(h(x)) = \mathrm{diag}(1, -1)\mathbf{f}(x)$, where $\mathrm{diag}(\cdot)$ denotes a diagonal matrix, and the invariance of the D-criterion w.r.t. the sign change $h(x) = -x$ to show that any design ξ is dominated by its symmetrization $\bar{\xi} = (\xi + \xi^h)/2$ such that $\Psi(\bar{\xi}) \leq \Psi(\xi)$ (see Pukelsheim, 1993, Chapter 13.11.). Thus we can restrict our search for a D-optimal subsampling design ξ^* to designs in Ξ^{f_X} that are invariant to the sign change. For an invariant ξ^* we find for the off-diagonal entries of the information matrix $\int x f_{\xi^*}(x)\,\mathrm{d}x = 0$. $\mathbf{M}(\xi^*) = \mathrm{diag}(\alpha, m)$ is thus a 2×2 diagonal matrix, where $m = \int x^2 f_{\xi^*}(x)\,\mathrm{d}x$. As a consequence the sensitivity function $\psi(x, \xi^*) = 1 + \alpha x^2/m$ is a polynomial of degree two as a function in x which is symmetric in x, $\psi(-x, \xi^*) = \psi(x, \xi^*)$. Obviously, the coefficient of the leading term of $\psi(x, \xi^*)$ is positive. We use from Theorem 3.1 that there exists a threshold s^* such that $f_{\xi^*}(x) = f_X(x)$ if $\psi(x, \xi^*) \geq s^*$ and $f_{\xi^*}(x) = 0$ elsewhere. Paired with the symmetry of $\psi(x, \xi^*)$, we find $\mathcal{X}^* = (-\infty, -a] \cup [a, \infty)$, where $a \geq 0$ and conclude that the density f_{ξ^*} of the D-optimal subsampling design is of the form

$$f_{\xi^*}(x) = f_X(x)\mathbb{1}_{(-\infty,-a]\cup[a,\infty)}(x).$$

Since we require $\xi^*(\mathbb{R}) = \alpha$, we can easily see that a is equal to the $(1 - \alpha/2)$-quantile of the distribution of X_i. To select a subsample we accept all units where the absolute value of the covariate is equal or greater than a.

This approach is not limited to centered symmetric distributions, but applies accordingly to all symmetric distributions: units will be accepted if their values of the covariate lie in

the lower or upper $(\alpha/2)$-tail of the distribution. This procedure can be interpreted as a theoretical counterpart in one dimension to the IBOSS method proposed by Wang et al. (2019).

Example 3.2 (normal distribution)**.** If the covariate X_i comes from a standard normal distribution, then the optimal boundaries are the $(\alpha/2)$- and the $(1-\alpha/2)$-quantile $\pm z_{1-\alpha/2}$, and unit i is accepted when $|x_i| \geq z_{1-\alpha/2}$. We find $\mathbf{M}(\xi^*) = \operatorname{diag}(\alpha, m)$, where $m = \alpha + \sqrt{2/\pi} z_{1-\alpha/2} \exp(-z_{1-\alpha/2}^2/2)$.

For X_i having a general univariate normal distribution with mean μ_X and variance σ_X^2, the optimal boundaries remain to be the $(\alpha/2)$- and $(1-\alpha/2)$-quantile $\mu_X \pm \sigma_X z_{1-\alpha/2}$ (see Reuter and Schwabe, 2023a, Theorem 3.2).

3.3.2 Multiple Linear Regression Subsampling Design

We now examine the case of multiple linear regression $Y_i = \mathbf{f}(\boldsymbol{X}_i)^\top \boldsymbol{\beta} + \varepsilon_i$, $i = 1, \ldots, n$, where $\boldsymbol{X}_i$ is a d-dimensional random vector with $d \geq 2$ and $\mathbf{f}(\mathbf{x}) = (1, \mathbf{x}^\top)^\top$. In this work, we assume the $\boldsymbol{X}_i$ have an elliptical distribution with density $f_{\boldsymbol{X}}$. In this section we start with the special case that the $\boldsymbol{X}_i$ follow a centered spherical distribution, i. e. a distribution invariant w.r.t. the special orthogonal group $SO(d)$ (rotations about the origin in $\mathbb{R}^d$), but relax this to the case of non-centered and elliptical distributions later. For $\boldsymbol{X}_i$ to be centered and spherical implies, in particular, that $\mathrm{E}[\boldsymbol{X}_i] = \mathbf{0}$, $\boldsymbol{X}_i$ has covariance matrix $\sigma_{\boldsymbol{X}}^2 \mathbb{I}_d$, where $\mathbb{I}_d$ denotes the identity matrix of dimension d, and all d covariates follow the same symmetric distribution. For instance, the multivariate standard normal distribution satisfies this condition with $\sigma_{\boldsymbol{X}}^2 = 1$.

To make use of the rotational invariance, we characterize subsampling designs in their hyperspherical coordinate representation, where a point in $\mathbb{R}^d$ is represented by a radial coordinate or radius r and a $(d-1)$-dimensional vector of angular coordinates $\boldsymbol{\theta}$, indicating the direction in the space. Details are deferred to the appendix. The design in hyperspherical coordinates $\xi_{R,\boldsymbol{\Theta}}$ can be decomposed into the product $\xi_R \otimes \xi_{\boldsymbol{\Theta}|R}$ of the marginal design ξ_R on the radius, and the conditional design $\xi_{\boldsymbol{\Theta}|R}$ on the vector of angles given $R = r$ as a Markov kernel. In particular for $\mathbb{B} = [0, \pi]^{d-2} \times [0, 2\pi)$ we have $\xi_{\boldsymbol{\Theta}|R=r}(\mathbb{B}) = 1$ for any $r \geq 0$. Subsequently, for $\xi_{\boldsymbol{\Theta},R} \in \Xi^{f_{\boldsymbol{X}}}$ it must hold that $\xi_R([0, \infty)) = \alpha$ and the density of ξ_R is bounded from above by the marginal density $f_{R(\boldsymbol{X})}$ of $\boldsymbol{X}_i$ on the radius. In the case $d = 2$, the transformation is a mapping to the standard polar coordinates and we can decompose the subsampling design into a measure on the radius R and a conditional one on the single angle θ.

To start, we want to show that there exists a continuous D-optimal subsampling design that is invariant w.r.t. $SO(d)$. This requires to employ a left Haar measure μ on $SO(d)$. For the representation in hyperspherical coordinates this is, up to a constant, a product of Lebesgue measures λ on the components of the angle vector $\boldsymbol{\theta}$ (see Cohn, 2013, Example 9.2.1.(c) for the case $d = 2$). We set $\mu = \bigotimes_{i=1}^{d-2} \lambda/\pi \otimes \lambda/(2\pi)$ such that $\mu(\mathbb{B}) = 1$, where $\otimes$ denotes the common product of measures.

Now, we prove the equivalence between invariance w.r.t. the special orthogonal group $SO(d)$ of a subsampling design ξ and decomposing ξ in a measure on the radius and a uniform measure on the angle.

Lemma 3.3. *In multiple linear regression with $d \geq 2$ covariates, a design ξ is invariant with respect to $SO(d)$ if and only if ξ can be decomposed into the marginal measure ξ_R on the radius and the Haar measure μ on the angle, i. e. $\xi = \xi_R \otimes \mu$.*

For a subsampling design $\xi \in \Xi^{f\boldsymbol{x}}$ with marginal design ξ_R on the radius, we denote the symmetrized measure $\xi_R \otimes \mu$ of ξ by $\overline{\xi}$.

Lemma 3.4. *In multiple linear regression with $d \geq 2$ covariates, let $\xi = \xi_R \otimes \xi_{\boldsymbol{\Theta}|R} \in \Xi^{f\boldsymbol{x}}$. Then its symmetrization $\overline{\xi} = \xi_R \otimes \mu$ is also in $\Xi^{f\boldsymbol{x}}$.*

Note that $\overline{\xi}$ is invariant w.r.t. $SO(d)$ by Lemma 3.3. Next, we establish an equality between the arithmetic mean of information matrices of rotated subsampling designs and the information matrix of $\overline{\xi}$.

Lemma 3.5. *In multiple linear regression with $d \geq 2$ covariates, let G be the finite group of rotations about the d axes that map the d-dimensional cross-polytope onto itself. Then*

$$\frac{1}{|G|} \sum_{g \in G} \mathbf{M}(\xi^g) = \mathbf{M}(\overline{\xi}).$$

We make use of this to prove that any subsampling design can be improved by its symmetrized subsampling design $\overline{\xi}$, which allows us to restrict the search for an optimal subsampling design from $\Xi^{f\boldsymbol{x}}$ to the essentially complete class of rotation invariant subsampling designs in $\Xi^{f\boldsymbol{x}}$.

Theorem 3.6. *In multiple linear regression with $d \geq 2$ covariates, let Φ be a convex optimality criterion that is invariant w.r.t. $SO(d)$, i. e. $\Phi(\xi^h) = \Phi(\xi)$ for any $\mathbf{h} \in SO(d), \xi \in \Xi^{f\boldsymbol{x}}$. Then for any $\xi = \xi_R \otimes \xi_{\boldsymbol{\Theta}|R} \in \Xi^{f\boldsymbol{x}}$ it holds that $\Phi(\overline{\xi}) \leq \Phi(\xi)$, with $\overline{\xi} = \xi_R \otimes \mu$.*

The regression model is linearly equivariant w.r.t. $SO(d)$ as

$$\mathbf{f}(\mathbf{h}(\mathbf{x})) = \mathbf{Q}_\mathbf{h}\mathbf{f}(\mathbf{x}) = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{H} \end{pmatrix} \mathbf{f}(\mathbf{x}),$$

for $\mathbf{h} \in SO(d)$ and $\mathbf{H}$ its respective orthogonal matrix with determinant one, i. e. $\mathbf{h}(\mathbf{x}) = \mathbf{H}\mathbf{x}$. Further note that $\Psi(\xi) = \Psi(\xi^\mathbf{h})$ for any $\xi \in \Xi^{f_\boldsymbol{X}}, \mathbf{h} \in SO(d)$, since $\det(\mathbf{M}(\xi^\mathbf{h})) = \det(\mathbf{Q}_\mathbf{h})^2 \det(\mathbf{M}(\xi))$ and $\det(\mathbf{Q}_\mathbf{h}) = 1$ for all $\mathbf{h} \in SO(d)$. The D-optimality criterion $\Psi(\xi) = \det(\mathbf{M}(\xi)^{-1})$ is indeed convex and invariant w.r.t. $SO(d)$. Theorem 3.6 applies to other optimality criteria as well such as Kiefer's Φ_q-criteria including the A-criterion or the integrated mean squared error (IMSE) criterion. Before we construct the optimal subsampling design in the subsequent theorem we make some preliminary remarks.

By Theorem 3.6 we can restrict our search for a D-optimal subsampling design to invariant designs. We study the shape of the sensitivity function $\psi(\mathbf{x}, \xi^*)$ of an invariant D-optimal subsampling design $\xi^* \in \Xi^{f_\boldsymbol{X}}$. Since ξ^* is composed of the Haar measure on the vector of angles, one can easily verify that all off-diagonal entries $\int x_j \xi^*(\mathrm{d}\mathbf{x})$ and $\int x_j x_{j'} \xi^*(\mathrm{d}\mathbf{x})$ of the information matrix of ξ^* are equal to zero, $j, j' = 1, \ldots, d$, $j \neq j'$. The $(d+1) \times (d+1)$ information matrix is thus $\mathbf{M}(\xi^*) = \operatorname{diag}(\alpha, m, \ldots, m)$, where $m = \int x_1^2 \xi^*(\mathrm{d}\mathbf{x})$. As a consequence, the sensitivity function simplifies to

$$\psi(\mathbf{x}, \xi^*) = \alpha \left(1, \mathbf{x}^\top\right) \operatorname{diag}(1/\alpha, 1/m, \ldots, 1/m) \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix} = 1 + \frac{\alpha}{m} \|\mathbf{x}\|_2^2 . \tag{3.1}$$

The sensitivity function is thus invariant to $SO(d)$ in the sense that $\psi(\mathbf{h}(\mathbf{x}), \xi^*) = \psi(\mathbf{x}, \xi^*)$ for all $\mathbf{h} \in SO(d)$ because $\|\mathbf{h}(\mathbf{x})\|_2^2 = \|\mathbf{x}\|_2^2$. Theorem 3.1 states that for a subsampling design to be optimal, it must hold that $\inf_{\mathbf{x} \in \mathcal{X}^*} \psi(\mathbf{x}, \xi^*) \geq \sup_{\mathbf{x} \notin \mathcal{X}^*} \psi(\mathbf{x}, \xi^*)$, where $\mathcal{X}^*$ is the support of ξ^*. Given that $\psi(\mathbf{x}, \xi^*)$ is constant on the d-sphere for all radii $r > 0$, this suggests that the optimal subsampling design is equal to zero in the interior of a d-sphere around the origin with radius r^* and equal to the bounding distribution on $\mathcal{X}^*$, outside of this sphere. Since the total measure of ξ^* is α, r^* is the $(1-\alpha)$-quantile of the radius R.

Theorem 3.7. *For multiple linear regression with $d \geq 2$ covariates and any $SO(d)$ invariant distribution of the covariates $\boldsymbol{X}_i$, the density of the continuous D-optimal subsampling design ξ^* is*

$$f_{\xi^*}(\mathbf{x}) = f_{\boldsymbol{X}}(\mathbf{x}) \mathbb{1}_{[q_{1-\alpha}, \infty)}(\|\mathbf{x}\|_2^2),$$

where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of the distribution of $\|\boldsymbol{X}_i\|_2^2$.

Note that this corresponds to the optimal subsampling design derived in Section 3.3.1 for $d = 1$.

Example 3.8 (multivariate standard normal distribution)**.** We apply Theorem 3.7 to the case of $\boldsymbol{X}_i \sim \mathcal{N}_d(\mathbf{0}, \mathbb{I}_d)$. Then $q_{1-\alpha} = \chi^2_{d,1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2-distribution with d degrees of freedom and the $(d+1) \times (d+1)$ information matrix $\mathbf{M}(\xi^*)$ is of the form $\text{diag}(\alpha, m, \ldots, m)$ with

$$m = \alpha + \frac{2\chi^2_{d,1-\alpha}}{d} f_{\chi^2_d}(\chi^2_{d,1-\alpha}), \tag{3.2}$$

where $f_{\chi^2_d}$ is the density of the χ^2-distribution with d degrees of freedom. Note that $m > \alpha$ for all $\alpha \in (0,1)$, because $\chi^2_{d,1-\alpha} > 0$ for $\alpha \in (0,1)$ and $f_{\chi^2_d}(w) > 0$ for all $w > 0$. In view of Example 3.2, we see that equation (3.2) also holds for $d = 1$. We will use this example to examine the performance of the subsampling design in Section 3.4. In Figure 3.1 we consider a 2-dimensional standard normal covariate $\boldsymbol{X}_i$ and $\alpha = 0.1$ and we depict the marginal optimal subsampling design ξ^*_R and the corresponding sensitivity function $\psi_R(r, \xi^*_R)$ as a function of the radius $r = \|\mathbf{x}\|_2$. We find from equation 3.1 that $\psi_R(r, \xi^*_R) = 1 + r^2\alpha/m$. The dotted vertical line describes the $(1-\alpha)$-quantile $\sqrt{\chi^2_{d,1-\alpha}}$ of the marginal distribution on the radius R. The horizontal dotted line indicates the threshold s^*.

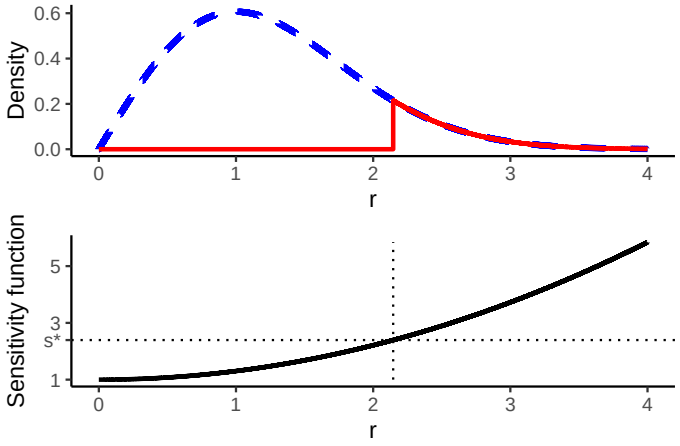


Figure 3.1: Density of the marginal optimal subsampling design ξ^*_R on the radius (red solid) and of the marginal distribution on the radius (blue dashed) of a 2-dimensional standard normal distribution (upper panel), and sensitivity function (lower panel) for subsampling proportion $\alpha = 0.1$

So far we have assumed that the covariates are centered and spherical. Let $\boldsymbol{Z}_i$ be such covariates that are invariant w.r.t. $SO(d)$. Now we consider covariates $\boldsymbol{X}_i = \boldsymbol{A}\boldsymbol{Z}_i + \boldsymbol{\mu}$ which

are location-scale transformations of $\boldsymbol{Z}_i$ with non-singular transformation matrix $\boldsymbol{A}$. The covariates $\boldsymbol{X}_i$ have elliptical distribution with mean $\boldsymbol{\mu}$ and non-singular covariance matrix $\boldsymbol{\Sigma_X} = \boldsymbol{A}\boldsymbol{A}^\top$. Because of equivariance of the D-criterion w.r.t. to such transformations, we find D-optimal subsampling designs in this case by transforming the observations back to the former situation by subtracting the mean $\boldsymbol{\mu}$ and multiplying with $\boldsymbol{A}^{-1}$. We show that this indeed constitutes a D-optimal subsampling design in the following lemma and derive the respective density in the subsequent theorem.

Lemma 3.9. *In multiple linear regression with $d \geq 2$ covariates, let the distribution of covariates $\boldsymbol{Z}_i \in \mathbb{R}^d$ be invariant w.r.t. $SO(d)$ and let $\zeta^* \in \Xi^{f_{\boldsymbol{Z}}}$ be the corresponding D-optimal subsampling design. Let $\boldsymbol{A}$ be a non-singular $d \times d$ matrix and $\boldsymbol{\mu}$ a constant in $\mathbb{R}^d$. Then the D-optimal subsampling design $\xi^* \in \Xi^{f_{\boldsymbol{X}}}$ for the covariates $\boldsymbol{X}_i = \boldsymbol{A}\boldsymbol{Z}_i + \boldsymbol{\mu}$ is given by $\xi^*(B) = \zeta^*(\boldsymbol{A}^{-1}(B - \boldsymbol{\mu}))$ for any measurable set $B \subset \mathbb{R}^d$.*

Note that ξ^* is the measure theoretic image of ζ^* under the transformation $\mathbf{z} \mapsto \mathbf{x} = \boldsymbol{A}\mathbf{z} + \boldsymbol{\mu}$.

Theorem 3.10. *In multiple linear regression with $d \geq 2$ covariates, let the distribution of the covariates $\boldsymbol{Z}_i \in \mathbb{R}^d$ be invariant w.r.t. $SO(d)$. Let $\boldsymbol{A}$ be a non-singular $d \times d$ matrix and $\boldsymbol{\mu}$ a constant in $\mathbb{R}^d$. Then the density of the D-optimal subsampling design ξ^* for covariates $\boldsymbol{X}_i = \boldsymbol{A}\boldsymbol{Z}_i + \boldsymbol{\mu}$ is*

$$f_{\xi^*}(\mathbf{x}) = f_{\boldsymbol{X}}(\mathbf{x}) \mathbb{1}_{[q_{1-\alpha}, \infty)} \left((\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right),$$

where $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^\top$ and $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of $\|\boldsymbol{Z}_i\|_2^2$.

To implement the continuous D-optimal subsampling design ξ^* from Theorem 3.10 we suggest Algorithm 1, a simple acceptance-rejection method, where all data points that lie in the support of ξ^* are accepted into the subdata and all others are rejected.

Algorithm 1: Subsample selection according to D-optimal subsampling design ξ^*

Data: Covariates $\mathbf{x}_i, i = 1, \ldots, n$, mean $\boldsymbol{\mu_X}$, covariance matrix $\Sigma_{\boldsymbol{X}}$.
Step 1: Calculate $c_i = (\mathbf{x}_i - \boldsymbol{\mu_X})^\top \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1} (\mathbf{x}_i - \boldsymbol{\mu_X})$;
Step 2: Select $\mathbf{x}_i$ when $c_i \geq q_{1-\alpha}$;

The resulting subsample is denoted by $(\mathbf{x}'_1, \ldots, \mathbf{x}'_{k_n})$ as realization of $(\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_{K_n})$, where the $\boldsymbol{X}'_i$ are i.i.d. random variables. For Algorithm 1, the subsample size K_n is random as it depends on the $\boldsymbol{X}_i$, but K_n is independent of the $\boldsymbol{X}'_i$. In view of limit theorems on

stopped random walks (see e. g. Gut, 2009, Theorem 1.1.) one can reasonably presume that the least squares estimator $\hat{\boldsymbol{\beta}}$ based on $(\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_{K_n})$ asymptotically follows a normal distribution with covariance matrix $\sigma_\varepsilon^2 \mathbf{M}(\xi^*)^{-1}$.

Example 3.11 (multivariate normal distribution)**.** Here we extend our findings from Example 3.8 to the case of a general multivariate normal distribution of the covariates with mean $\boldsymbol{\mu_X}$ and covariance matrix $\boldsymbol{\Sigma_X}$, i. e. $\boldsymbol{X}_i = \boldsymbol{A}\boldsymbol{Z}_i + \boldsymbol{\mu_X}$, where $\boldsymbol{A}$ is a root of $\boldsymbol{\Sigma_X}$, i. e. $\boldsymbol{\Sigma_X} = \boldsymbol{A}\boldsymbol{A}^\top$, and $\boldsymbol{Z}_i$ follows a multivariate standard normal distribution. By Theorem 3.10 we know that the D-optimal subsampling design is equal to the distribution of the $\boldsymbol{X}_i$ outside of an ellipsoid given by $(\mathbf{x}-\boldsymbol{\mu_X})^\top \boldsymbol{\Sigma_X}^{-1}(\mathbf{x}-\boldsymbol{\mu_X}) < q_{1-\alpha}$, where $q_{1-\alpha} = \chi^2_{d,1-\alpha}$ is equal to the $(1-\alpha)$-quantile of the χ^2-distribution with d degrees of freedom.

To guarantee the subsampling proportion α as well as to avoid reliance on the $(1-\alpha)$-quantile of $\|\boldsymbol{Z}_i\|_2^2$ in Algorithm 1 we suggest Algorithm 2. Here, the subsample size k_n is deterministic and this algorithm only depends on the first two moments of the distribution and the distribution to be elliptical.

Algorithm 2: Subsample selection according to design $(\mathbf{x}_{1:n}, \ldots, \mathbf{x}_{k_n:n})$

Let $k_n = [\alpha n]$ be the integer part of αn;
Data: Covariates $\mathbf{x}_i, i = 1, \ldots, n$, mean $\boldsymbol{\mu_X}$, covariance matrix $\boldsymbol{\Sigma_X}$.
Step 1: Calculate $c_i = (\mathbf{x}_i - \boldsymbol{\mu_X})^\top \boldsymbol{\Sigma_X}^{-1}(\mathbf{x}_i - \boldsymbol{\mu_X})$;
Step 2: Select $\mathbf{x}_{1:n}, \ldots, \mathbf{x}_{k_n:n}$ corresponding to the k_n largest c_i;

The notation $\mathbf{x}_{i:n}$ is chosen to indicate a generalized (reverse) order statistics based on the standardized distance c_i such that $(\mathbf{x}_{1:n}, ..., \mathbf{x}_{n:n})$ is a permutation of $(\mathbf{x}_1, ..., \mathbf{x}_n)$ and $(\mathbf{x}_{i:n} - \boldsymbol{\mu_X})^\top \boldsymbol{\Sigma_X}^{-1}(\mathbf{x}_{i:n} - \boldsymbol{\mu_X}) \geq (\mathbf{x}_{i+1:n} - \boldsymbol{\mu_X})^\top \boldsymbol{\Sigma_X}^{-1}(\mathbf{x}_{i+1:n} - \boldsymbol{\mu_X})$. Because the distribution of the covariates is continuous, these inequalities are strict almost surely. The selection in Step 2 of Algorithm 2 can e. g. be done using partial quicksort (see Martínez, 2004).

Remark 3.12. For multiple linear regression the selection criterion $c_i = (\mathbf{x}_i - \boldsymbol{\mu_X})^\top \boldsymbol{\Sigma_X}^{-1}(\mathbf{x}_i - \boldsymbol{\mu_X})$ for a D-optimal subsample is equivalent to the theoretical leverage scores $h_i = \mathbf{f}(\mathbf{x}_i)^\top \mathbf{M_X}^{-1} \mathbf{f}(\mathbf{x}_i)$, where $\mathbf{M_X} = \mathrm{E}(\mathbf{f}(\boldsymbol{X}_i)\mathbf{f}(\boldsymbol{X}_i)^\top)$, as $h_i = c_i + c_0$, where $c_0 = \boldsymbol{\mu_X}^\top \boldsymbol{\Sigma_X}^{-1} \boldsymbol{\mu_X} \geq 0$ is a constant. Subsampling via algorithmic leveraging as described in e. g. Ma et al. (2014) uses a sampling distribution proportional to the leverage scores h_i, rather than selecting a subsample deterministically as we do here.

3.4 Fixed Sample Size

Unlike in the previous section, where we selected a certain percentage of the full data, we now want to select a fixed, sufficiently large, number of k instances out of the total n data points. This implies that we want to select a decreasing percentage $\alpha_n = k/n$ of the full data when n increases. The subsampling design ξ_n with total measure α_n has non-standardized (per subsample) information matrix $\mathbf{M}_n(\xi_n) = n \int \mathbf{f}(\mathbf{x})\mathbf{f}(\mathbf{x})^\top \xi_n(\mathrm{d}\mathbf{x})$, such that $n \int \xi_n(\mathrm{d}\mathbf{x}) = k$. Here we use the non-standardized information matrix to to allow for comparison of the performance for varying n.

If k is large, the asymptotic properties in the previous section may give rise to consider the inverse information matrix $\mathbf{M}_n(\xi_n)^{-1}$ as an approximation to the covariance matrix of $\hat{\boldsymbol{\beta}}$ based on k out of n observations. Hence, it seems to be reasonable to make use of the optimal continuous subsampling design ξ_n^* for subsampling a proportion $\alpha_n = k/n$ according to Theorem 3.10. Here we adapt Algorithm 2 to select the fixed number $k_n = k$ data points $\mathbf{x}_i$ that correspond to the k largest c_i. This design will be denoted by $(\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})$ with non-standardized information matrix $\mathbf{M}((\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})) = \sum_{i=1}^{k} \mathbf{f}(\mathbf{x}_{i:n})\mathbf{f}(\mathbf{x}_{i:n})^\top$.

The computational complexity for the selection of $\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n}$ is $\mathcal{O}(nd^2)$. Note that finding the inverse root of the covariance matrix is negligible, as computation of the inverse only depends on the number of covariates d and we work under the assumption that $d \ll n$. Computing the least squares estimator based on k observations uses computational complexity $\mathcal{O}(kd^2)$. When $n \gg d$ it is reasonable to assume that $k \leq n/d$. Then the computational complexity for the entire procedure is $\mathcal{O}(nd^2)$, the same magnitude as computing the least squares estimator $\hat{\boldsymbol{\beta}}$ on the full data, making it only viable in a scenario where the focus is on the expense of observing the response variable Y_i.

For scenarios where computational complexity is the main issue, we propose a second simplified method. Here we merely standardize each covariate X_{ij} by its standard deviation σ_j. We use the matrix $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}$, containing only the diagonal entries of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, for transformation of the data. To implement this we adapt Algorithm 2 by replacing the c_i with $\tilde{c}_i = (\mathbf{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}})^\top \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}})$ and select a fixed number k of points. This design will be denoted by $(\tilde{\mathbf{x}}_{1:n}, \ldots, \tilde{\mathbf{x}}_{k:n})$. Here, the entire procedure has computational complexity $\mathcal{O}(nd)$, as the matrix multiplication $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}})$ only requires computational complexity $\mathcal{O}(nd)$, because $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}^{-1}$ is a diagonal matrix. The simplified method has one more advantage. It is easier to implement in practice when there is no prior knowledge of the covariance matrix of the covariates as estimating only the variances of the covariates on a small uniform random subsample (prior to the actual subsampling procedure) is much easier than estimating the entire covariance matrix. We will see in the simulation study in Section 3.5 that this second

method is indeed viable.

For now, however, we first want to study the performance of the initial method, where the full covariance matrix of the covariates is used for the transformation of the data. As a measure of quality of the method with a fixed sample size k we use the covariance matrix $\mathrm{Cov}[\hat{\boldsymbol{\beta}}; (\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})]$ of the least squares estimator $\hat{\boldsymbol{\beta}}$ given a subsample according to $(\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})$. For large k the subsample size k is expected to be close to the random subsample size generated by Algorithm 1 according to ξ_n^*, and the covariance matrix may be approximated by the inverse of the information matrix of the corresponding optimal continuous design ξ_n^*,

$$\mathrm{Cov}[\hat{\boldsymbol{\beta}}; (\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})] \approx \sigma_\varepsilon^2 \mathbf{M}_n(\xi_n^*)^{-1}. \tag{3.3}$$

In the literature, the main interest is often only in the slope parameter $\boldsymbol{\beta}_{\text{slope}} = (\beta_1, \ldots, \beta_d)$ and the covariance matrix of the vector $\hat{\boldsymbol{\beta}}_{\text{slope}} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)^\top$ of slope parameter estimators. Therefore, we will adopt this approach here. Note that the D-optimal subsampling design for $\boldsymbol{\beta}_{\text{slope}}$ is the same as for the full parameter vector $\boldsymbol{\beta}$ because $\det(\mathbf{M}^{-1})$ and the determinant of the lower right $d \times d$ submatrix of $\mathbf{M}^{-1}$ differ only by the constant factor $1/k$. Then, under the D-optimal subsampling design ξ_n^* from Theorem 3.10, we find for $\boldsymbol{X}_i$ with mean $\boldsymbol{\mu}_{\boldsymbol{X}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$

$$\mathrm{Cov}[\hat{\boldsymbol{\beta}}_{\text{slope}}; (\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})] \approx \sigma_\varepsilon^2 (n m_n)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\,, \tag{3.4}$$

where m_n is the design specific term in $\mathbf{M}_n(\zeta_n^*) = n\,\mathrm{diag}(k/n, m_n, \ldots, m_n)$ with $\xi_n^*(B) = \zeta_n^*(\boldsymbol{A}^{-1}(B - \boldsymbol{\mu}_{\boldsymbol{X}}))$.

Example 3.13 (multivariate standard normal distribution)**.** We consider the design $(\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})$. In the case of normally distributed covariates $\boldsymbol{X}_i \sim \mathcal{N}_d(\mathbf{0}, \mathbb{I}_d)$, we find the following approximation of the covariance matrix $\mathrm{Cov}[\hat{\boldsymbol{\beta}}_{\text{slope}}; \xi_{k,n}^*]$ from equations (3.2) and (3.4)

$$\mathrm{Cov}[\hat{\boldsymbol{\beta}}_{\text{slope}}; (\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})] \approx \sigma_\varepsilon^2 \left(k + \frac{2n\chi^2_{d,1-(k/n)}}{d} f_{\chi^2_d}(\chi^2_{d,1-(k/n)}) \right)^{-1} \mathbb{I}_d. \tag{3.5}$$

With this we can approximate the trace of the covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{slope}}$, which is equal to the mean squared error $\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\text{slope}}) = \mathrm{E}[\|\hat{\boldsymbol{\beta}}_{\text{slope}} - \boldsymbol{\beta}_{\text{slope}}\|_2^2]$, since the least squares estimator is unbiased. In order to compare the behavior between different dimensions d we find $\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\text{slope}})$ divided by d is equal to any of the diagonal entries of the covariance matrix, e.g. the variance $\mathrm{Var}[\hat{\beta}_1; (\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})]$ of the slope parameter estimator of the

first covariate. In Figure 3.2 the lines depict the approximation from equation (3.5) of $\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{slope}})/d$, indicated on the left vertical axis of Figure 3.2, for standard normal covariates in dependence of the size of the full data n given a fixed mean subsample size $k = 1000$ of the subsample. The symbols depict the respective simulated values. The simulation procedure is given in section 3.5, with the only difference that the number of simulation runs S for each combination of number of covariates d and full sample size n here is only $S = 1000$, since the computations for $n = 10^7$ take infeasibly long. We see that $\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{slope}})/d$ tends to zero as $n \to \infty$, but substantially slower for higher dimensions d as more parameters need to be estimated. Moreover, the approximation in equations (3.4) and (3.5) turn out to be useful because they are very close to the values obtained by simulation, at least, for small to moderate dimensions d.

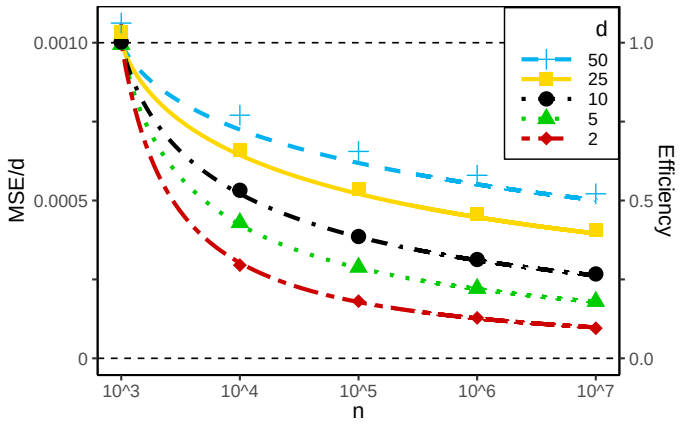


Figure 3.2: Approximated (lines) and simulated (symbols) MSE/d of the slope parameter estimator and approximated efficiencies of uniform random subsampling (lines) in the case of standard normal covariates in dependence of the size of the full data n given a fixed mean subsample size $k = 1000$ and $\sigma_\varepsilon^2 = 1$ for various numbers of covariates $d = 2, 5, 10, 25$ and 50

To demonstrate the advantage of the design $(\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})$, we consider uniform random subsampling as a natural choice to compare with. The uniform random subsampling design ξ_n^u has density $f_{\xi_n^u}(\mathbf{x}) = (k/n) f_{\boldsymbol{X}}(\mathbf{x})$. As a measure of quality, we study the D-efficiency of ξ_n^u w.r.t. the D-optimal subsampling design $(\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})$. For estimating the slopes, the D-efficiency of a subsampling design ξ_n with subsampling proportion $\alpha_n = k/n$ is defined as

$$\mathrm{eff}_{D_{\mathrm{slope}}}(\xi_n) = \left(\frac{\det(\mathrm{Cov}[\hat{\boldsymbol{\beta}}_{\mathrm{slope}}; (\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})])}{\det(\mathrm{Cov}[\hat{\boldsymbol{\beta}}_{\mathrm{slope}}; \xi_n])} \right)^{1/d}$$

and can be approximated by replacing the covariances by the inverse information matrices for the slopes. For this definition the homogeneous version $(\det(\text{Cov}[\hat{\boldsymbol{\beta}}_{\text{slope}}; \xi_n]))^{1/d}$ of the D-criterion is used, satisfying the homogeneity condition $(\det(\lambda \, \text{Cov}[\hat{\boldsymbol{\beta}}_{\text{slope}}; \xi_n]))^{1/d} = \lambda(\det(\text{Cov}[\hat{\boldsymbol{\beta}}_{\text{slope}}; \xi_n]))^{1/d}$ for all $\lambda > 0$ (see Pukelsheim, 1993, Chapter 6.2).

As mentioned in Reuter and Schwabe (2023a), the D-efficiency $\text{eff}_{D_{\text{slope}}}(\xi_n^u)$ of uniform random subsampling can be nicely interpreted: the sample size required to obtain the same precision (in terms of the D_{slope}-criterion), as when the D_{slope}-optimal subsampling design ξ_n^* with subsample size k is used, is equal to the inverse of the efficiency $\text{eff}_{D_{\text{slope}}}(\xi_n^u)$ times k. For example, if the efficiency $\text{eff}_{D_{\text{slope}}}(\xi_n^u)$ is equal to 0.5, then twice as many observations would be needed under uniform random sampling than for a D_{slope}-optimal subsampling design. The information matrix for uniform random subsampling is given by

$$\mathbf{M}_n(\xi_n^u) = k \int \mathbf{f}(\mathbf{x})\mathbf{f}(\mathbf{x})^\top f_{\boldsymbol{X}}(\mathbf{x}) \, \mathrm{d}\mathbf{x} = k \begin{pmatrix} 1 & \boldsymbol{\mu}_{\boldsymbol{X}}^\top \\ \boldsymbol{\mu}_{\boldsymbol{X}} & \boldsymbol{\Sigma}_{\boldsymbol{X}} + \boldsymbol{\mu}_{\boldsymbol{X}}\boldsymbol{\mu}_{\boldsymbol{X}}^\top \end{pmatrix}$$

such that $\text{Cov}[\hat{\boldsymbol{\beta}}_{\text{slope}}; \xi_n^u] = k^{-1}\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$.

Corollary 3.14. *The D_{slope}-efficiency of the design $(\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})$ can be approximated by $\text{eff}_{D_{\text{slope}}}(\xi_n^u) \approx k/(nm_n)$, where m_n are the diagonal entries of the information matrix $\mathbf{M}_n(\zeta_n^*) = n \, \text{diag}(k/n, m_n, \ldots, m_n)$.*

Example 3.15 (normal distribution)**.** Consider $\boldsymbol{X}_i \sim \mathcal{N}_d(\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$. By Corollary 3.14 we have $\text{eff}_{D_{\text{slope}}}(\xi_n^u) \approx k/(nm_n)$, where

$$m_n = \frac{k}{n} + \frac{2\chi^2_{d,1-(k/n)}}{d} f_{\chi^2_d}(\chi^2_{d,1-(k/n)}),$$

from equation (3.2).

The approximated efficiency $\text{eff}_{D_{\text{slope}}}(\xi_n^u)$ is thus equal to the approximated $\text{MSE}(\hat{\boldsymbol{\beta}}_{\text{slope}})/d$ given the design $(\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})$ as depicted in Figure 3.2, multiplied by k/σ_ε^2. The efficiency ranges between zero and one and is indicated by the vertical axis on the right of Figure 3.2.

3.5 Simulation Study

We divide our simulation study into two parts. First, we study the performance of the optimal subsampling designs $(\mathbf{x}_{1:n}, \ldots, \mathbf{x}_{k:n})$ derived from Theorem 3.10 in the case of multivariate normally distributed and multivariate t-distributed covariates with three degrees of freedom, respectively, both with and without correlation between the covariates. For the t-distribution,

we choose three degrees of freedom to maximize dispersion, while maintaining existence of the variance. Second, we use the simplified design $(\tilde{\mathbf{x}}_{1:n}, \ldots, \tilde{\mathbf{x}}_{k:n})$ discussed in Section 3.4 that only takes the variance of the covariates into account while ignoring the correlation. The latter has lower computational complexity, $\mathcal{O}(nd)$. For better comparability, the simulation is structured similar to those in the work by Wang et al. (2019). The data is generated from the linear model $Y_i = \beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}_{\text{slope}} + \varepsilon_i,\ i = 1, \ldots, n$, with $d = 50$. The parameter vector $\boldsymbol{\beta}$ was generated from a multivariate normal distribution in each iteration. Note, however, that the value of $\boldsymbol{\beta}$ does not have any influence on the results. For the errors we choose independent $\varepsilon_i \sim \mathcal{N}(0, 1)$. The subdata is of fixed size $k = 1000$, whereas the size of the full data takes the values $n = 5000, 10000, 100000$, and one million. For each value of n, we apply our subsampling methods and calculate the least squares estimator $\hat{\boldsymbol{\beta}}$ for each method. We repeat this $S = 10000$ times. We select subdata based on our approach (D-OPT) and the IBOSS method (IBOSS) by Wang et al. (2019). Further we select subdata by uniform sampling (UNIF) and give a comparison to estimates based on the full data (FULL) to give context to our approach and the IBOSS method. In each iteration s, we form the subsample in the $k \times d$ matrix $\boldsymbol{X}_{(s)}$ (based on the respective method) and calculate its inverse information matrix $\mathcal{C}_{(s)} = ((\mathbf{1}_k, \boldsymbol{X}_{(s)})^\top (\mathbf{1}_k, \boldsymbol{X}_{(s)}))^{-1}$, where $\mathbf{1}_k$ is a k-dimensional vector with all entries equal to one. We then take the average of these $(d+1) \times (d+1)$ covariance matrices $\mathcal{C} = 1/S \sum_{s=1}^{S} \mathcal{C}_{(s)}$ and partition this matrix the following way.

$$\mathcal{C} = \begin{pmatrix} c_0 & \boldsymbol{c}_1^\top \\ \boldsymbol{c}_1 & \mathcal{C}_1 \end{pmatrix},$$

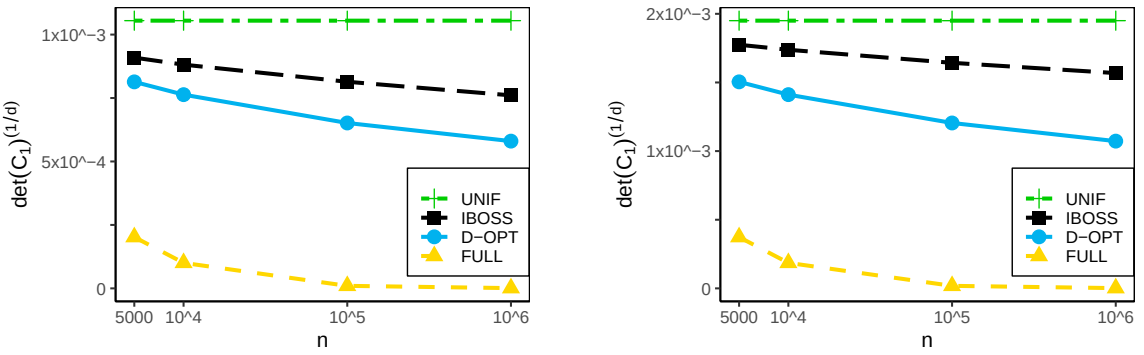
where $c_0 \geq 1/k$ with equality if $\boldsymbol{c}_1 = \mathbf{0}$. The submatrix $\mathcal{C}_1$ is of dimension $d \times d$. Note that $\mathcal{C}$ and $\mathcal{C}_1$ are the simulated covariance matrices of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{\text{slope}}$, respectively. The mean of the covariance matrices is taken instead of the mean of the information matrices, which has been the target quantity for asymptotic behavior. Note that the inverse of the mean information matrix is a lower bound for the mean covariance matrix by Jensen's inequality. Then, we calculate the determinant of $\mathcal{C}_1$ and scale it to homogeneity, i. e. $\det(\mathcal{C}_1)^{(1/d)}$. Alternatively to using $\det(\mathcal{C}_1)^{(1/d)}$ to compare the different methods, we have also used the MSE of $\hat{\boldsymbol{\beta}}_{\text{slope}}$, i. e. $\text{MSE}_{\hat{\boldsymbol{\beta}}_{\text{slope}}} = S^{-1} \sum_{s=1}^{S} \|\hat{\boldsymbol{\beta}}_{\text{slope}}^{(s)} - \boldsymbol{\beta}_{\text{slope}}\|_2^2$, where $\hat{\boldsymbol{\beta}}_{\text{slope}}^{(s)}$ is the estimator of the s-th iteration. Results were very similar in all cases and, importantly, the comparison between them does not change. In particular note that the trace of $\mathcal{C}_1$ is equal to $\text{MSE}_{\hat{\boldsymbol{\beta}}_{\text{slope}}}$.

Consider the special case of homoscedastic covariates. Then all diagonal elements of the theoretical counterpart of $\text{Cov}(\hat{\boldsymbol{\beta}}_{\text{slope}})$ are equal and all off-diagonal entries are equal to zero. Thus in theory we have the MSE divided by d is equal to the term of interest

$\det(\text{Cov}(\hat{\boldsymbol{\beta}}_{\text{slope}}))^{1/d}$ in our simulations. In this situation, the D-optimal subsampling design is equal to the A-optimal subsampling design for the slope parameters, which minimizes $\text{trace}(\text{Cov}(\hat{\boldsymbol{\beta}}_{\text{slope}}))$ and thus the MSE. As A- and D-optimal subsampling designs are not equal in other cases we recommend using an A-optimal subsampling design as a benchmark for other methods when the MSE is used as the measure of comparison, but we will not follow this line further here. All simulations are performed using R Statistical Software (R Core Team, 2023, v4.2.2).

3.5.1 Optimal Subsampling Design $(\mathbf{x}_{1:n}, \ldots, \mathbf{x}_{k:n})$

Here we use the subsampling design $(\mathbf{x}_{1:n}, \ldots, \mathbf{x}_{k:n})$ from Algorithm 2 with fixed k. Let $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \mathbb{I}_{50}$ or $\boldsymbol{\Sigma}_{0.5}$, where $\boldsymbol{\Sigma}_{0.5} = (\mathbb{I}_{50} + \mathbf{1}_{50}\mathbf{1}_{50}^{\top})/2$ represents compound symmetry with correlation $\rho = 0.5$. Figure 3.3 shows the results for normally distributed covariates $\boldsymbol{X}_i$ with $\mathbb{I}_{50}$ and $\boldsymbol{\Sigma}_{0.5}$ as the covariance matrix respectively. Figure 3.4 shows the results for the t-distribution with three degrees of freedom where $\mathbb{I}_{50}$ and $\boldsymbol{\Sigma}_{0.5}$ are the respective scale matrices, so again we have compound symmetry with correlation $\rho = 0.5$ in the latter case. Here, we omit the uniformly selected subsample for better visibility because uniform subsampling performs substantially worse. For uniform random subsampling, the determinant is close to constant at around 4.6×10^{-4} for all four values of n in the case of no correlation and similarly around 8.5×10^{-4} in the case with correlation.

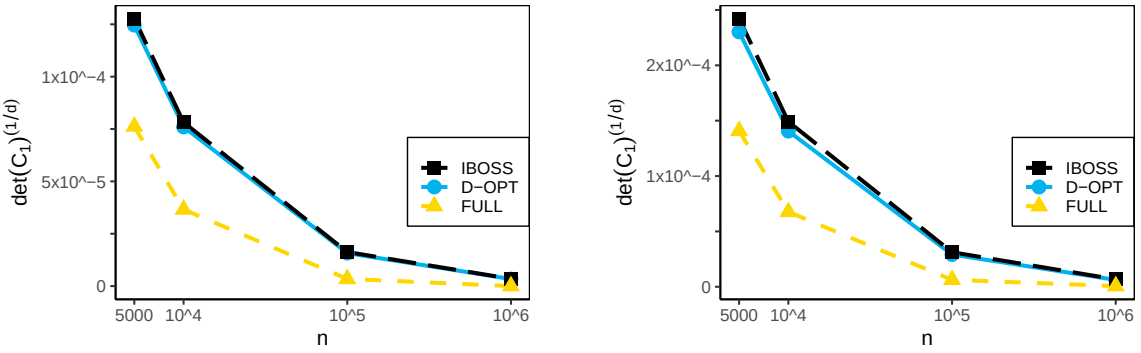


(a) $\boldsymbol{X}_i \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_{50})$.

(b) $\boldsymbol{X}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{0.5})$.

Figure 3.3: Standardized determinant of the inverse information matrix for normally distributed covariates given different covariance matrices.

As expected, regardless of the distribution of the covariates, for uniform random subsampling the full sample size n has no impact on $\det(\mathcal{C}_1)^{(1/d)}$, which is equal to n/k times that value of the full data.



(a) $\boldsymbol{X}_i \sim t_3(\mathbf{0}, \mathbb{I}_{50})$ (b) $\boldsymbol{X}_i \sim t_3(\mathbf{0}, \boldsymbol{\Sigma}_{0.5})$.

Figure 3.4: Standardized determinant of the inverse information matrix for t-distributed covariates with three degrees of freedom given different scale matrices

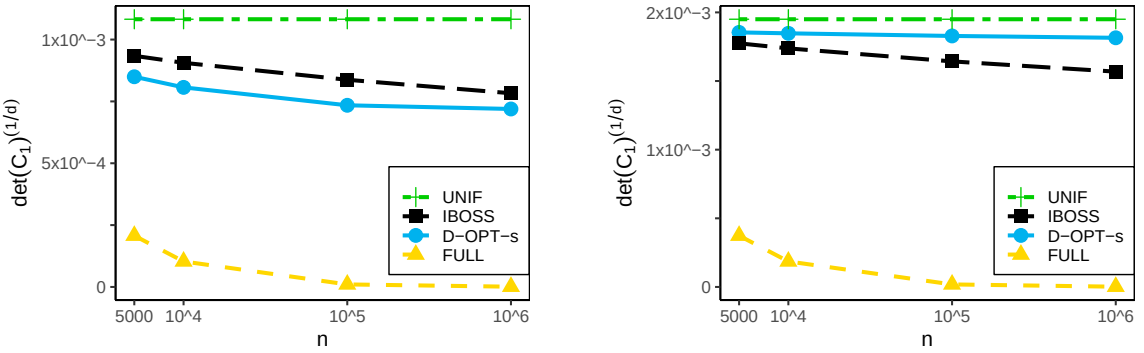
With the prior knowledge of the distribution of the covariates, our method is able to outperform the IBOSS method. As is to be expected, our approach can increase its advantage over the IBOSS method when there is correlation between the covariates. The advantage is however minor for the heavy-tailed t-distribution, where both methods perform much closer to the full data. In particular, for large n both perform basically as good as the full data. For reference, in the case of positive correlations the relative efficiency of the IBOSS method with respect to the D-OPT method, i.e. the ratio of the corresponding values of D-OPT divided by IBOSS, ranges from approximately 0.951 to 0.928 for the different values in full sample size n.

3.5.2 Simplified Method $(\tilde{\mathbf{x}}_{1:n}, \ldots, \tilde{\mathbf{x}}_{k:n})$

Finally, we want to study the simplified design $(\tilde{\mathbf{x}}_{1:n}, \ldots, \tilde{\mathbf{x}}_{k:n})$ of the D-OPT method that only scales by standard deviations and can be performed in $\mathcal{O}(nd)$. In this method, we ignore the correlations between the covariates. We use the diagonal matrix $\tilde{\boldsymbol{\Sigma}} = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ containing only the diagonal entries of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ for transformation of the data, such that the entire procedure has computational complexity $\mathcal{O}(nd)$. We examine this method in the case of normally distributed covariates and refer to the simplified D-OPT method as "D-OPT-s" in the figures.

Note that in the case of no correlation between the covariates the simplified method is equal to the D-OPT method of the previous section. Thus results for this scenario can be inherited from Figure 3.3(A). Further, we consider compound symmetry with $\rho = 0.05$ and $\rho = 0.5$. The 50×50 covariance matrices of the $\boldsymbol{X}_i$ are $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{\Sigma}_{0.05}$ or $\boldsymbol{\Sigma}_{0.5}$, with $\sigma_{0.05,ii} = 1$

and $\sigma_{0.05,ij} = 0.05$ for $i \neq j$ and $\mathbf{\Sigma}_{0.5}$ as before. Figure 3.5 shows the results for normally distributed covariates $\boldsymbol{X}_i$ with $\mathbf{\Sigma}_{0.05}$ and $\mathbf{\Sigma}_{0.5}$ as the covariance matrix, respectively. While the advantage of the D-optimal subsampling design over the IBOSS method is reduced, there are still scenarios where it can outperform the IBOSS method such as the one of covariance matrix $\mathbf{\Sigma}_{0.05}$ with small correlations. However, if correlations are particularly large as in the case of covariance matrix $\mathbf{\Sigma}_{0.5}$, the simplified method D-OPT-s seems to perform much worse and only slightly outperforms uniform subsampling.



(a) $\boldsymbol{X}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{0.05})$. (b) $\boldsymbol{X}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{0.5})$.

Figure 3.5: Standardized determinant of the inverse information matrix for normally distributed covariates given different covariance matrices for the simplified D-OPT-s method

3.6 Concluding Remarks

We have constructed optimal subsampling designs ξ^* for multiple linear regression, first for centered spherical distributions, then for distributions that can be generated from such a distribution via location-scale transformation. We have given two methods of implementation and discussed that the computational complexity of the D-optimal method, that selects the k data points with the largest Mahalanobis distance from the mean of the data, is $\mathcal{O}(nd^2)$, whereas the simplified version can be performed in $\mathcal{O}(nd)$. We have compared these implementations to the IBOSS method of Wang et al. (2019) in simulation studies with the expected result that the full method outperforms IBOSS as well as the simplified method outperforms the IBOSS method in certain settings with small correlations between the covariates. Besides applications where the covariance matrix of the covariates is known, our method can be used as a benchmark for other methods that do not require prior knowledge of the distribution of the covariates. Note, that the proposed subsampling designs depend both

on the distribution of the covariates and the model. If either is incorrect, the subsampling designs will no longer be optimal. Recent work on subsampling for model discrimination is done by Yu and Wang (2022).

3.A Technical Details

Lemma 3.16. *The essential part of the directional derivative*

$$F_\Psi(\xi, \xi_{\mathbf{x}}) = \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \left(-\log(\det(\mathbf{M}((1-\epsilon)\xi + \epsilon\xi_{\mathbf{x}}))) + \log(\det(\mathbf{M}(\xi)))\right)$$

at a design ξ in the direction of a one-point measure $\xi_{\mathbf{x}}$ with total measure α is the sensitivity function $\psi(\mathbf{x}, \xi) = (d+1) - F_\Psi(\xi, \xi_{\mathbf{x}}) = \alpha \mathbf{f}(\mathbf{x})^\top \mathbf{M}(\xi)^{-1} \mathbf{f}(\mathbf{x})$.

Proof of Lemma 3.16. The directional derivative $F_\Psi(\xi, \xi_{\mathbf{x}})$ can be calculated as $F_\Psi(\xi, \xi_{\mathbf{x}}) = (d+1) - \mathrm{Tr}(\mathbf{M}(\xi)^{-1}\mathbf{M}(\xi_{\mathbf{x}}))$ (see Silvey, 1980, Example 3.8) which reduces to $F_\Psi(\xi, \xi_{\mathbf{x}}) = (d+1) - \alpha \mathbf{f}(\mathbf{x})^\top \mathbf{M}(\xi)^{-1} \mathbf{f}(\mathbf{x})$ for a one-point measure $\xi_{\mathbf{x}}$. Equivalently, we consider the sensitivity function $\psi(\mathbf{x}, \xi) = \alpha \mathbf{f}(\mathbf{x})^\top \mathbf{M}(\xi)^{-1} \mathbf{f}(\mathbf{x})$, which incorporates the essential part of the directional derivative ($\psi(\mathbf{x}, \xi) = (d+1) - F_\Psi(\xi, \xi_{\mathbf{x}})$). □

Remark 3.17. For the representation of a design ξ in hyperspherical coordinates we make use of the transformation $T : [0, \infty) \times [0, \pi]^{d-2} \times [0, 2\pi) \to \mathbb{R}^d, T(r, \boldsymbol{\theta}) = \mathbf{x}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{d-1})^\top$, $x_k = r \cos(\theta_k) \prod_{j=1}^{k-1} \sin(\theta_j)$ for $k = 1, \ldots, d-1$, and $x_d = r \prod_{j=1}^{d-1} \sin(\theta_j)$. We identify all points with radius zero with the origin and denote the inverse of the transformation T by $S = T^{-1}$. Then, for a subsampling design $\xi \in \Xi^{f\boldsymbol{x}}$ on $\mathbb{R}^d$, the induced subsampling design ξ^S is the same subsampling design in hyperspherical coordinates, i. e. on $[0, \infty) \times \mathbb{B}$, where $\mathbb{B} = [0, \pi]^{d-2} \times [0, 2\pi)$.

Proof of Theorem 3.1. The result follows immediately from Corollary 1 (c) in Sahm and Schwabe (2001) as in Theorem 3.1. in Reuter and Schwabe (2023a). □

Proof of Lemma 3.3. We define $\mathbf{h}$ as a mapping from $\mathbb{R}_{>0} \times \mathbb{B}$ to itself. Let $\mathbf{h} = (h_0, \mathbf{h}_1^\top)^\top$, where $h_0(r) = r$ is the identity on the radius and $\mathbf{h}_1 \in SO(d)$ acts on the angle $\boldsymbol{\theta}$. First note that for any $B = B_R \times B_{\boldsymbol{\Theta}}$ with $B_R \in \mathcal{B}(\mathbb{R}_{>0})$ and $B_{\boldsymbol{\Theta}} \in \mathcal{B}(\mathbb{B})$ and any $\mathbf{h}_1 \in SO(d)$ the mapping $\mathbf{h}$ only affects the set $B_{\boldsymbol{\Theta}}$ on the angle. Since μ is a left Haar measure w.r.t $SO(d)$, it holds that $\mu(\mathbf{h}_1^{-1}(B_{\boldsymbol{\Theta}})) = \mu(B_{\boldsymbol{\Theta}})$ for any $\mathbf{h}_1 \in SO(d)$ and any $B_{\boldsymbol{\Theta}} \in \mathcal{B}(\mathbb{B})$. We first prove that the composition $\xi = \xi_R \otimes \mu$ of a measure ξ_R on the radius and the Haar

measure μ implies invariance. For any $B = B_R \times B_{\boldsymbol{\Theta}}$ and any $\mathbf{h}_1 \in SO(d)$, we have

$$\begin{aligned}\xi^{\mathbf{h}}(B) &= (\xi_R \otimes \mu)(\mathbf{h}^{-1}(B_R \times B_{\boldsymbol{\Theta}})) \\ &= \xi_R(B_R)\mu(\mathbf{h}_1^{-1}(B_{\boldsymbol{\Theta}})) \\ &= \xi_R(B_R)\mu(B_{\boldsymbol{\Theta}}) \\ &= \xi(B).\end{aligned}$$

Because the σ-algebra $\mathcal{B}(\mathbb{R}_{>0}) \otimes \mathcal{B}(\mathbb{B})$ is generated by $\mathcal{B}(\mathbb{R}_{>0}) \times \mathcal{B}(\mathbb{B})$, we conclude that ξ is invariant w.r.t. $SO(d)$.

Conversely, let us assume $\xi = \xi_R \otimes \xi_{\boldsymbol{\Theta}|R}$ (this decomposition exists by the Radon-Nikodym Theorem) is invariant w.r.t. $SO(d)$ and there exist sets $B_R \in \mathcal{B}(\mathbb{R}_{>0})$ with $\xi_R(B_R) > 0$ and $B_{\boldsymbol{\Theta}} \in \mathcal{B}(\mathbb{B})$ such that $\xi_{\boldsymbol{\Theta}|R\in B_R}(B_{\boldsymbol{\Theta}}) \neq \mu(B_{\boldsymbol{\Theta}}) > 0$. Then there exists a rotation $\mathbf{h}_1 \in SO(d)$ such that $\xi_{\boldsymbol{\Theta}|R\in B_R}(B_{\boldsymbol{\Theta}}) \neq \xi_{\boldsymbol{\Theta}|R\in B_R}(\mathbf{h}_1^{-1}(B_{\boldsymbol{\Theta}}))$, and subsequently we have $\xi^{\mathbf{h}}(B_R \times B_{\boldsymbol{\Theta}}) \neq \xi(B_R \times B_{\boldsymbol{\Theta}})$. This contradicts the invariance assumption and we derive that invariance of ξ w.r.t. $SO(d)$ implies that $\xi_{\boldsymbol{\Theta}|R}(B_{\boldsymbol{\Theta}}) = \mu(B_{\boldsymbol{\Theta}})$ almost everywhere w.r.t. ξ_R. This concludes the proof. □

Proof of Lemma 3.4. The $\boldsymbol{X}_i$ are invariant w.r.t. $SO(d)$ and thus we can write the density of the $\boldsymbol{X}_i$ as $f_{\boldsymbol{X}}(\mathbf{x}) = f_{R(\boldsymbol{X})}(r)f_\mu(\boldsymbol{\theta})$ by Lemma 3.3. We can decompose the density of ξ into $f_\xi(\mathbf{x}) = f_{R(\xi)}(r)f_{\boldsymbol{\Theta}|R}(\boldsymbol{\theta})$. We have $f_{R(\xi)}(r) \leq f_{R(\boldsymbol{X})}(r)$ because $\xi \in \Xi^{f_{\boldsymbol{X}}}$. As a result $f_{\overline{\xi}}(\mathbf{x}) = f_{R(\xi)}(r)f_\mu(\boldsymbol{\theta}) \leq f_{R(\boldsymbol{X})}(r)f_\mu(\boldsymbol{\theta})$ and thus $\overline{\xi} \in \Xi^{f_{\boldsymbol{X}}}$. □

Proof of Lemma 3.5. Consider the information matrix of a subsampling design ξ in hyperspherical coordinates, i. e. with the transformation T and its inverse S. The Jacobi matrix of T is denoted by $\boldsymbol{J}_T(r, \boldsymbol{\theta})$. Then

$$\begin{aligned}\mathbf{M}(\xi) &= \int_{\mathbb{R}^d} \mathbf{f}(\mathbf{x})\mathbf{f}(\mathbf{x})^\top \xi(\mathrm{d}\mathbf{x}) \\ &= \int_{S(\mathbb{R}^d)} \mathbf{f}(T(r,\boldsymbol{\theta}))\mathbf{f}(T(r,\boldsymbol{\theta}))^\top \xi^S(\mathrm{d}(r,\boldsymbol{\theta})) \\ &= \int_{[0,\infty)} \int_{\mathbb{B}} \mathbf{f}(T(r,\boldsymbol{\theta}))\mathbf{f}(T(r,\boldsymbol{\theta}))^\top |\det(\boldsymbol{J}_T(r,\boldsymbol{\theta}))| \xi_{\boldsymbol{\Theta}|R=r}(\mathrm{d}\boldsymbol{\theta})\xi_R(\mathrm{d}r).\end{aligned}$$

Now we study the sum of information matrices of rotated subsampling designs.

$$
\begin{aligned}
&\frac{1}{|G|}\sum_{g\in G}\mathbf{M}(\xi^g)\\
&\quad=\frac{1}{|G|}\sum_{g\in G}\int_{[0,\infty)}\int_{\mathbb{B}}\mathbf{f}(T(r,\boldsymbol{\theta}))\mathbf{f}(T(r,\boldsymbol{\theta}))^\top|\det(\boldsymbol{J}_T(r,\boldsymbol{\theta}))|\xi^g_{\boldsymbol{\Theta}|R=r}(\mathrm{d}\boldsymbol{\theta})\xi_R(\mathrm{d}r)\\
&\quad=\frac{1}{|G|}\sum_{g\in G}\int_{[0,\infty)}\int_{\mathbb{B}}\mathbf{f}(g^{-1}(T(r,\boldsymbol{\theta})))\mathbf{f}(g^{-1}(T(r,\boldsymbol{\theta})))^\top|\det(\boldsymbol{J}_T(r,\boldsymbol{\theta}))|\xi_{\boldsymbol{\Theta}|R=r}(\mathrm{d}\boldsymbol{\theta})\xi_R(\mathrm{d}r)\\
&\quad=\int_{[0,\infty)}\int_{\mathbb{B}}\sum_{g\in G}\frac{1}{|G|}\mathbf{f}(g^{-1}(T(r,\boldsymbol{\theta})))\mathbf{f}(g^{-1}(T(r,\boldsymbol{\theta})))^\top|\det(\boldsymbol{J}_T(r,\boldsymbol{\theta}))|\xi_{\boldsymbol{\Theta}|R=r}(\mathrm{d}\boldsymbol{\theta})\xi_R(\mathrm{d}r).
\end{aligned}
$$

The inner sum can be regarded as the information matrix of a design putting equal weight on the vertices of a rotated d-dimensional cross-polytope. This is equal to the information matrix of the uniform distribution on the d-sphere, see Pukelsheim (1993, Chapter 15.18.) or Gaffke and Heiligers (1996, Lemma 4.9.). Then

$$
\begin{aligned}
&\frac{1}{|G|}\sum_{g\in G}\mathbf{M}(\xi^g)\\
&\quad=\int_{[0,\infty)}\int_{\mathbb{B}}\int_{\mathbb{B}}\mathbf{f}(T(r,\boldsymbol{\gamma}))\mathbf{f}(T(r,\boldsymbol{\gamma}))^\top|\det(\boldsymbol{J}_T(r,\boldsymbol{\gamma}))|\mu(\mathrm{d}\boldsymbol{\gamma})\xi_{\boldsymbol{\Theta}|R=r}(\mathrm{d}\boldsymbol{\theta})\xi_R(\mathrm{d}r)\\
&\quad=\int_{[0,\infty)}\int_{\mathbb{B}}\mathbf{f}(T(r,\boldsymbol{\gamma}))\mathbf{f}(T(r,\boldsymbol{\gamma}))^\top|\det(\boldsymbol{J}_T(r,\boldsymbol{\gamma}))|\mu(\mathrm{d}\boldsymbol{\gamma})\int_{\mathbb{B}}\xi_{\boldsymbol{\Theta}|R=r}(\mathrm{d}\boldsymbol{\theta})\xi_R(\mathrm{d}r)\\
&\quad=\int_{[0,\infty)}\int_{\mathbb{B}}\mathbf{f}(T(r,\boldsymbol{\gamma}))\mathbf{f}(T(r,\boldsymbol{\gamma}))^\top|\det(\boldsymbol{J}_T(r,\boldsymbol{\gamma}))|\mu(\mathrm{d}\boldsymbol{\gamma})\xi_R(\mathrm{d}r)\\
&\quad=\mathbf{M}(\xi_R\otimes\mu).
\end{aligned}
$$

In the third equality we used $\int_{\mathbb{B}}\xi_{\boldsymbol{\Theta}|R=r}(\mathrm{d}\boldsymbol{\theta})=1$. □

Proof of Theorem 3.6. By the result $\mathbf{M}(\bar{\xi})=\frac{1}{|G|}\sum_{g\in G}\mathbf{M}(\xi^g)$ of Lemma 3.5 we have

$$\Phi(\mathbf{M}(\bar{\xi}))\le\frac{1}{|G|}\sum_{g\in G}\Phi(\mathbf{M}(\xi^g))$$

by the convexity of Φ. Note that $g\in G\subset SO(d)$. We then utilize that Φ is invariant w.r.t. $SO(d)$, i. e. $\Phi(\mathbf{M}(\xi^g))=\Phi(\mathbf{M}(\xi))$, and obtain

$$\Phi(\mathbf{M}(\bar{\xi}))\le\Phi(\mathbf{M}(\xi)).$$

□

Proof of Theorem 3.7. We apply Theorem 3.1. ξ^* is defined such that it is equal to the bounding distribution on $\mathcal{X}^* = \{\mathbf{x} \in \mathcal{X} : \|\mathbf{x}\|_2^2 \geq q_{1-\alpha}\}$ and equal to zero on $\{\mathbf{x} \in \mathcal{X} : \|\mathbf{x}\|_2^2 < q_{1-\alpha}\}$. The sensitivity function from equation (3.1) is

$$\psi(\mathbf{x}, \xi^*) = 1 + \frac{\alpha}{m}\|\mathbf{x}\|_2^2.$$

Then we can immediately see that $s^* = \inf_{x \in \mathcal{X}^*} \psi(\mathbf{x}, \xi^*) = \sup_{x \notin \mathcal{X}^*} \psi(\mathbf{x}, \xi^*) = \frac{\alpha}{m} q_{1-\alpha}$. and thus conditions (i)-(iii) are satisfied, which concludes the proof. □

Proof of Lemma 3.9. Note that for any $d \times d$ matrix $\boldsymbol{A}$ and any vector $\boldsymbol{\mu} \in \mathbb{R}^d$, there exists a bijection $\Xi^{f\boldsymbol{Z}} \to \Xi^{f\boldsymbol{X}}$, where every subsampling design $\zeta \in \Xi^{f\boldsymbol{Z}}$ is mapped to $\xi \in \Xi^{f\boldsymbol{X}}$, which is defined as $\xi(B) = \zeta(\boldsymbol{A}^{-1}(B - \boldsymbol{\mu}))$ for any measurable set $B \subset \mathbb{R}^d$. Let $\zeta \in \Xi^{f\boldsymbol{Z}}$. Consider the information matrix of the subsampling design ξ,

$$\begin{aligned}
\mathbf{M}(\xi) &= \int \mathbf{f}(\mathbf{x})\mathbf{f}(\mathbf{x})^\top \xi(\mathrm{d}\mathbf{x}) \\
&= \int \mathbf{f}(\boldsymbol{A}\mathbf{z} + \boldsymbol{\mu})\mathbf{f}(\boldsymbol{A}\mathbf{z} + \boldsymbol{\mu})^\top \zeta(\mathrm{d}\mathbf{z}) \\
&= \int \begin{pmatrix} 1 & \mathbf{0}^\top \\ \boldsymbol{\mu} & \boldsymbol{A} \end{pmatrix} \mathbf{f}(\mathbf{z})\mathbf{f}(\mathbf{z})^\top \begin{pmatrix} 1 & \boldsymbol{\mu}^\top \\ \mathbf{0} & \boldsymbol{A}^\top \end{pmatrix} \zeta(\mathrm{d}\mathbf{z}) \\
&= \begin{pmatrix} 1 & \mathbf{0}^\top \\ \boldsymbol{\mu} & \boldsymbol{A} \end{pmatrix} \mathbf{M}(\zeta) \begin{pmatrix} 1 & \boldsymbol{\mu}^\top \\ \mathbf{0} & \boldsymbol{A}^\top \end{pmatrix}.
\end{aligned}$$

The determinant of the information matrix can be calculated as follows.

$$\det\left(\mathbf{M}(\xi)\right) = \det(\boldsymbol{A}\boldsymbol{A}^\top) \det\left(\mathbf{M}(\zeta)\right).$$

Thus ξ^* minimizes $\Psi(\xi)$ in $\Xi^{f\boldsymbol{X}}$, if ζ^* minimizes $\Psi(\zeta)$ in $\Xi^{f\boldsymbol{Z}}$. □

Proof of Theorem 3.10. From Lemma 3.9 we have that $\xi^*(B) = \zeta^*(\boldsymbol{A}^{-1}(B - \boldsymbol{\mu}))$ for any measurable set $B \subset \mathbb{R}^d$, where ζ^* is the optimal subsampling design for covariates $\boldsymbol{Z}_i$ in the setting of Theorem 3.7. We inherit the desired result by applying Theorem 3.7. □

Proof of equation (3.2). Let $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,d})^\top \sim \mathcal{N}_d(\mathbf{0}, \mathbb{I}_d)$ with density $f_{\boldsymbol{X}}$. From Theorem 3.7, we know that the support of ξ^* is $\mathcal{X}^* = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2^2 \geq \chi^2_{d,1-\alpha}\}$ on which it is equal to the d-dimensional standard normal distribution. By definition, the information matrix of ξ^* is $\mathbf{M}(\xi^*) = \int \mathbf{f}(\mathbf{x})\mathbf{f}(\mathbf{x})^\top \xi^*(\mathrm{d}\mathbf{x})$. Any off-diagonal entries $\int x_j \xi^*(\mathrm{d}\mathbf{x})$ and

$\int x_j x_{j'} \xi^*(\mathrm{d}\mathbf{x})$, $j, j' = 1, \ldots, d, j \neq j'$ are equal to zero. The upper left element of the matrix is $\xi^*(\mathbb{R}^d) = \alpha$ by the definition of a subsampling design. The other elements on the main diagonal are equal because ξ^* is invariant w.r.t. $SO(d)$ and thus $\int x_j^2 \xi^*(\mathrm{d}\mathbf{x}) = \int x_{j'}^2 \xi^*(\mathrm{d}\mathbf{x})$ for any $j, j' = 1, \ldots, d$. Note that $W = \|\boldsymbol{X}_i\|_2^2$ follows a χ^2-distribution with d degrees of freedom. We start to calculate m by formulating it as the expected value of W.

$$\begin{aligned} m &= \int_{\mathcal{X}^*} x_1^2 f_{\boldsymbol{X}}(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ &= \mathrm{E}\left[X_{i,1}^2 \mathbb{1}_{\{\|\boldsymbol{X}_i\|_2^2 \geq \chi^2_{d,1-\alpha}\}} \right] \\ &= \frac{1}{d} \mathrm{E}\left[\|\boldsymbol{X}_i\|_2^2 \mathbb{1}_{\{\|\boldsymbol{X}_i\|_2^2 \geq \chi^2_{d,1-\alpha}\}} \right] \\ &= \frac{1}{d} \mathrm{E}\left[W \mathbb{1}_{\{W \geq \chi^2_{d,1-\alpha}\}} \right]. \end{aligned}$$

We write the expected value in its integral form and insert the density $f_{\chi^2_d}$ of the χ^2-distribution with d degrees of freedom. Then

$$\begin{aligned} m &= \frac{1}{d} \int_{\chi^2_{d,1-\alpha}}^{\infty} w f_{\chi^2_d}(w) \, \mathrm{d}w \\ &= \frac{1}{d 2^{d/2} \Gamma(d/2)} \int_{\chi^2_{d,1-\alpha}}^{\infty} w^{d/2} e^{-w/2} \, \mathrm{d}w. \end{aligned}$$

Integration by parts yields

$$\begin{aligned} m &= \frac{1}{d 2^{d/2} \Gamma(d/2)} \left(2(\chi^2_{d,1-\alpha})^{d/2} e^{-\chi^2_{d,1-\alpha}/2} + d \int_{\chi^2_{d,1-\alpha}}^{\infty} w^{(d/2)-1} e^{-w/2} \, \mathrm{d}w \right) \\ &= \frac{(\chi^2_{d,1-\alpha})^{d/2} e^{-\chi^2_{d,1-\alpha}/2}}{d 2^{(d/2)-1} \Gamma(d/2)} + \int_{\chi^2_{d,1-\alpha}}^{\infty} \frac{w^{(d/2)-1} e^{-w/2}}{2^{d/2} \Gamma(d/2)} \, \mathrm{d}w. \end{aligned}$$

The latter term simplifies to α because the integrand is the density of the χ^2 distribution with d degrees of freedom. Then

$$m = \frac{2\chi^2_{d,1-\alpha}}{d} f_{\chi^2_d}(\chi^2_{d,1-\alpha}) + \alpha.$$

□

Proof of equation (3.4). We get from equation (3.3) that the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{Cov}[\hat{\boldsymbol{\beta}}; (\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})] \approx \sigma_\varepsilon^2 (\mathbf{M}_n(\xi_n^*))^{-1}.$$

As in the proof of Lemma 3.9,

$$\mathbf{M}_n(\xi_n^*) = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \boldsymbol{\mu}_{\boldsymbol{X}} & \boldsymbol{A} \end{pmatrix} \mathbf{M}_n(\zeta_n^*) \begin{pmatrix} 1 & \boldsymbol{\mu}_{\boldsymbol{X}}^\top \\ \mathbf{0} & \boldsymbol{A}^\top \end{pmatrix}.$$

Recall that $\mathbf{M}_n(\zeta_n^*) = n\,\mathrm{diag}(k/n, m_n, \ldots, m_n)$, where $m_n = \int x_1^2 \zeta_n^*(\mathrm{d}\mathbf{x})$. We get for the covariance matrix of the asymptotic distribution of the parameter estimator

$$\begin{aligned}&\mathrm{Cov}[\hat{\boldsymbol{\beta}}; (\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})] \\ &\quad\approx \sigma_\varepsilon^2 \begin{pmatrix} 1 & -(\boldsymbol{A}^{-1}\boldsymbol{\mu}_{\boldsymbol{X}})^\top \\ \mathbf{0} & (\boldsymbol{A}^{-1})^\top \end{pmatrix} n^{-1}\,\mathrm{diag}((k/n), m_n, \ldots, m_n)^{-1} \begin{pmatrix} 1 & \mathbf{0}^\top \\ -\boldsymbol{A}^{-1}\boldsymbol{\mu}_{\boldsymbol{X}} & \boldsymbol{A}^{-1} \end{pmatrix}.\end{aligned}$$

The approximation of the covariance matrix of the slope parameters estimators $\hat{\boldsymbol{\beta}}_{\mathrm{slope}}$ is given by the lower right block of the matrix above.

$$\begin{aligned}\mathrm{Cov}[\hat{\boldsymbol{\beta}}_{\mathrm{slope}}; (\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})] &\approx \sigma_\varepsilon^2 (\boldsymbol{A}^{-1})^\top (nm_n)^{-1} \mathbb{I}_d \boldsymbol{A}^{-1} \\ &= \sigma_\varepsilon^2 (nm_n)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}.\end{aligned}$$

□

Proof of Corollary (3.14). For the covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{slope}}$ under the uniform random subsampling design ξ_n^u we find

$$\det(\mathrm{Cov}[\hat{\boldsymbol{\beta}}_{\mathrm{slope}}; \xi_n^u]) \approx \sigma_\varepsilon^{2d} k^{-d} \det(\boldsymbol{\Sigma}_{\boldsymbol{X}})^{-1}.$$

For the covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{slope}}$ given $(\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})$ we have with the approximation in equation (3.4)

$$\det(\mathrm{Cov}[\hat{\boldsymbol{\beta}}_{\mathrm{slope}}; (\mathbf{x}_{1:n}, ..., \mathbf{x}_{k:n})]) \approx \sigma_\varepsilon^{2d} (nm_n)^{-d} \det(\boldsymbol{\Sigma}_{\boldsymbol{X}})^{-1},$$

where m_n are the diagonal entries of $\mathbf{M}_n(\zeta_n^*) = n\,\mathrm{diag}(k/n, m_n, \ldots, m_n)$. Thus $\mathrm{eff}_{D_{\mathrm{slope}}}(\xi_n^u) \approx k/(nm_n)$. □

Chapter 4

Poisson Regression in one Covariate on Massive Data

In this chapter we present the work titled "Poisson Regression in one Covariate on Massive Data" (Reuter and Schwabe, 2024) published as an electronic preprint.

4.1 Introduction

Progress in technology has lead to the collection of increasingly large data sets. The field of subsampling or subdata selection has gained popularity in recent years, where the aim is to decrease the number of observations in the data set while maintaining as much information as possible. To illuminate fundamental features of the concept, we solely focus on the reduction of observations in massive data for a single covariate, rather than reduction in covariates of high-dimensional data. Subdata selection for massive data can be done via a probabilistic subsampling scheme or through deterministic rules. Earlier works on subsampling for generalized linear models (GLMs) focus on probabilistic methods, in particular on subsampling for logistic regression, see e.g. Wang et al. (2018). More recently there are more works on GLMs, including Poisson regression: For probabilistic subsampling under the A and L-optimality criteria see Ai et al. (2021) and Yu et al. (2022). After Wang et al. (2019) introduced information-based optimal subdata selection (IBOSS) for linear regression, Cheng et al. (2020) proposed IBOSS for logistic regression, a deterministic subsampling technique with a probabilistic initial subsample to estimate the unknown parameter. This is necessary because, as is well known, the optimal design depends on the unknown parameter for GLMs.

In the present paper on Poisson regression we derive locally D-optimal continuous subsampling designs directly bounded by the density of the covariate. Such directly bounded designs were first studied by Wynn (1977) and Fedorov (1989). Recently, Ul Hassan and Miller (2019) derived such bounded optimal subsampling designs for logistic regression in the context of optimal item calibration similarly to our approach. Such subsampling designs can then easily be used for subdata selection by including all observations that lie in the support of the optimal subsampling design and exclude all others. Though an initial step to estimate the parameter is necessary when it is unknown. When there are no constraints on the design, literature on Poisson regression includes Rodríguez-Torreblanca and Rodríguez-Díaz (2007) and Russell et al. (2009).

In Section 4.2 we introduce the Poisson regression model to be used in this paper. Then, we present a theorem on the support of a locally D-optimal continuous subsampling design as well as a theorem concerning scale-location shifts of the covariate in Section 4.3. Further, we give examples when the covariate has an exponential or a uniform distribution. In Section 4.4 we study the efficiency of uniform random subsampling and some heuristic designs in comparison to the optimal subsampling designs. In addition, we consider the loss in efficiency when the regression parameter is misspecified. We add closing remarks in Section 4.5. Proofs are deferred to an appendix.

4.2 Model Specification

We consider pairs $(x_i, y_i), i = 1, \ldots, n$, of data, where y_i is the value of the response variable Y_i. x_i is a realization of the random variable X_i. The covariate X_i has probability density function f_X. We suppose that the dependence of the response variable on the covariate X_i is given by a Poisson regression model.

(A1) Conditionally on the covariate X_i, the response Y_i is Poisson distributed with conditional mean $\mathrm{E}(Y_i|X_i) = \exp(\beta_0 + \beta_1 X_i)$.

Model (A1) constitutes a generalized linear model with random covariate and log link. The aim is to estimate the regression parameter $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$. $\mathbf{f}(x) = (1, x)^\top$ denotes the regression function in the linear component $\mathbf{f}(X_i)^\top \boldsymbol{\beta}$ such that $\mathrm{E}(Y_i|X_i) = \exp(\mathbf{f}(X_i)^\top \boldsymbol{\beta})$.

We will further assume that the covariate X_i has a continuous distribution satisfying some moment conditions.

(A2) The covariate X_i has density f_X and $\mathrm{E}(X_i^2 \exp(\beta_1 X_i)) < \infty$.

4.3 Subsampling Design

We assume that the number of observations n is very large. However, we encounter the challenge of dealing with responses, denoted by Y_i, which are either costly or difficult to observe. Meanwhile, the values x_i of all units X_i of the covariate are readily available. To tackle this problem, we consider a scenario in which the responses Y_i will only be observed for a specific subsampling proportion α of the units, $0 < \alpha < 1$. The selection of these units is based on the knowledge of the covariate values x_i for all units. Our objective is to identify a subsample of pairs (x_i, y_i) that provides the most accurate estimation of the parameter vector $\boldsymbol{\beta}$ by means of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$. As the covariate X_i has a continuous distribution, we are going to identify a subsample from this distribution that maximizes information, but only covers a percentage α of the distribution. Therefore, we consider continuous designs ξ as measures of mass α on $\mathbb{R}$ with density f_ξ bounded by the density f_X of X_i ensuring $\int f_\xi(x)\,\mathrm{d}x = \alpha$ and $f_\xi(x) \leq f_X(x)$ for all $x \in \mathbb{R}$. A subsample can then be generated according to such a bounded continuous design ξ by accepting units i with probability $f_\xi(x_i)/f_X(x_i)$. To obtain analytical results, we assume that the distribution of the covariate X_i and, hence, its density f_X is known.

The information arising for a single observation at covariate value x is defined by the elemental information $\mathbf{M}(x, \boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x)\mathbf{f}(x)\mathbf{f}(x)^\top$ (see Russell et al., 2009). For a continuous design ξ, the information matrix $\mathbf{M}(\xi, \boldsymbol{\beta})$ is defined by

$$\mathbf{M}(\xi, \boldsymbol{\beta}) = \int \mathbf{M}(x, \boldsymbol{\beta})\xi(\mathrm{d}x) = \exp(\beta_0) \begin{pmatrix} m_0(\xi, \beta_1) & m_1(\xi, \beta_1) \\ m_1(\xi, \beta_1) & m_2(\xi, \beta_1) \end{pmatrix},$$

where $m_k(\xi, \beta_1) = \int x^k \exp(\beta_1 x) f_\xi(x)\,\mathrm{d}x$. The moment condition $\mathrm{E}(X_i^2 \exp(\beta_1 X_i)) < \infty$ stated in assumption (A2) for the distribution of the covariates X_i ensures that the entries $m_k(\xi, \beta_1)$ in the information matrix are finite for any bounded continuous design ξ. Otherwise no meaningful optimization would be possible. The moment condition is obviously satisfied when the distribution of X_i has a finite support. It also holds for other not heavy-tailed distributions like the normal distribution. In the case of an exponentially distributed covariate X_i considered below, the additional condition $\beta_1 < \lambda$ on the slope parameter β_1 is required where λ is the rate parameter of the exponential distribution.

The information matrix $\mathbf{M}(\xi, \boldsymbol{\beta})$ serves as a measure for evaluating the performance of the design ξ. Note that $\mathbf{M}(\xi, \boldsymbol{\beta})$ has full rank for any continuous design ξ. This ensures the

existence of the inverse

$$\mathbf{M}(\xi, \boldsymbol{\beta})^{-1} = \frac{1}{\exp(\beta_0) d(\xi, \beta_1)} \begin{pmatrix} m_2(\xi, \beta_1) & -m_1(\xi, \beta_1) \\ -m_1(\xi, \beta_1) & m_0(\xi, \beta_1) \end{pmatrix}.$$

where $d(\xi, \beta_1) = m_0(\xi, \beta_1) m_2(\xi, \beta_1) - m_1(\xi, \beta_1)^2$ is the standardized determinant of $\mathbf{M}(\xi, \boldsymbol{\beta})$, $d(\xi, \beta_1) = \exp(-2\beta_0) \det(\mathbf{M}(\xi, \boldsymbol{\beta}))$. Then, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically normal with mean zero and covariance matrix $\mathbf{M}(\xi, \boldsymbol{\beta})^{-1}$ for the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$.

Maximization of the information matrix in the Loewner sense of nonnegative definiteness will not be possible, in general. Therefore, we have to consider some one-dimensional information functional. We will focus here on the most popular design criterion, the D-criterion, in its widely used form, $\log(\det(\mathbf{M}(\xi, \boldsymbol{\beta})))$, to be maximized. A subsampling design ξ^* with density f_{ξ^*} that maximizes the D-criterion for a given parameter value $\boldsymbol{\beta}$ will be called locally D-optimal at $\boldsymbol{\beta}$. Maximization of the D-criterion can be interpreted in terms of the covariance matrix as minimization of the volume of the asymptotic confidence ellipsoid for the parameter vector $\boldsymbol{\beta}$.

Remark 4.1. Note that β_0 appears in the information matrix only by the multiplicative factor $\exp(\beta_0)$. Thus, a locally D-optimal subsampling design ξ^* only depends on the slope β_1.

For the characterization of a locally D-optimal design, we will make use of an equivalence theorem based on constrained convex optimization (see e. g. Sahm and Schwabe, 2001). For this, we have to distinguish between cases related to the sign of the slope β_1. In applications, the slope will often be negative ($\beta_1 < 0$). We will focus on that case and establish a representation of the locally D-optimal subsampling designs for $\beta_1 < 0$ first.

Denote by F_X and q_α the cumulative distribution function and the α-quantile of X_i. Let $\mathbb{1}_A$ the indicator function of a set A, i. e. $\mathbb{1}_A(x) = 1$, if $x \in A$ and $\mathbb{1}_A(x) = 0$ otherwise. Further, denote by

$$\psi(x, \xi, \beta_1) = \frac{1}{d(\xi, \beta_1)} \exp(\beta_1 x)(m_0(\xi, \beta_1) x^2 - 2 m_1(\xi, \beta_1) x + m_2(\xi, \beta_1))$$

the sensitivity function of a design ξ (see Theorem 4.11). Note that the sensitivity function $\psi(x, \xi, \beta_1)$ does not depend on β_0.

Theorem 4.2. *Let assumptions* (A1) *and* (A2) *be satisfied and let* $\beta_1 < 0$. *Then the subsampling design* ξ^* *is locally D-optimal at* $\boldsymbol{\beta}$ *if and only if* ξ^* *has density* $f_{\xi^*}(x) = f_X(x)\mathbb{1}_{\mathcal{X}^*}(x)$ *and either*

(i) there exist $a_1 < a_2 < a_3$ *such that*

$$\mathcal{X}^* = (-\infty, a_1] \cup [a_2, a_3],$$

$$F_X(a_1) + F_X(a_3) - F_X(a_2) = \alpha, \text{ and} \tag{4.2a}$$

$$\psi(a_1, \xi^*, \beta_1) = \psi(a_2, \xi^*, \beta_1) = \psi(a_3, \xi^*, \beta_1), \tag{4.2b}$$

or

(ii) $\mathcal{X}^* = (-\infty, q_\alpha]$, (4.2a')

$$\psi(x, \xi^*, \beta_1) > \psi(q_\alpha, \xi^*, \beta_1) \text{ for } x < q_\alpha, \text{ and}$$

$$\psi(x, \xi^*, \beta_1) < \psi(q_\alpha, \xi^*, \beta_1) \text{ for } x > q_\alpha. \tag{4.2b'}$$

Conditions (4.2a) and (4.2a') correspond to the subsampling percentage α while (4.2b) and (4.2b') are related to the conditions on the sensitivity function in the general equivalence theorem for bounded designs (Theorem 4.11) reproduced in the Appendix.

In view of the shape $f_{\xi^*}(x) = f_X(x)\mathbb{1}_{\mathcal{X}^*}(x)$ of the density of the continuous optimal subsampling designs ξ^* in Theorem 4.2, the subsampling mechanism becomes deterministic for the optimal design: The subsample can be generated by accepting all units i for which $x_i \in \mathcal{X}^*$ and by rejecting all others.

According to Theorem 4.2, there are two different scenarios for the locally D-optimal design ξ^*. Either the supporting set $\mathcal{X}^*$ consists of two separate intervals $(-\infty, a_1]$ and $[a_2, a_3]$ (scenario (i)) or these intervals will be merged into a single one (scenario (ii)).

Remark 4.3. The optimal subsampling design ξ^* is unique because of the strict concavity of the D-criterion and the shape of the sensitivity function.

For the construction of a locally D-optimal subsampling design by Theorem 4.2, first the conditions of scenario (ii) for an optimal design supported on a single interval can be checked. If scenario (ii) does not apply, the boundary points $a_1 < a_2 < a_3$ for the support $\mathcal{X}^*$ have to be calculated by solving the system of (nonlinear) equations (4.2a) and (4.2b). In the latter case, the rightmost boundary point a_3 of $\mathcal{X}^*$ may lie outside the support of X_i, i. e. $a_3 > x_{\max}$, when the support of the covariate X_i is bounded from above, i. e. $x_{\max} = \operatorname{ess\,sup}(X_i) < \infty$, where ess sup denotes the essential supremum (see, e. g., Example 4.9 for the uniform distribution below). Then, in scenario (ii), explicit calculation of the rightmost boundary point c is not necessary. Instead, it is sufficient for (4.2b) to verify that $\psi(x_{\max}, \xi^*, \beta_1) \geq \psi(a_1, \xi^*, \beta_1) = \psi(a_2, \xi^*, \beta_1)$.

Remark 4.4. The leftmost boundary point a_1 of a D-optimal subsampling design ξ^* cannot lie outside the range of X_i, i. e. $a_1 > x_{\min}$, where $x_{\min} = \operatorname{ess\,inf}(X_i)$ the essential infimum of the distribution of X_i.

Remark 4.5. When $\beta_1 = 0$, the information matrix $\mathbf{M}(\xi, \boldsymbol{\beta})$ is, up to the multiplicative constant $\exp(\beta_0)$, equal to the information matrix $\mathbf{M}(\xi) = \int \mathbf{f}(x)\mathbf{f}(x)^\top \xi(\mathrm{d}x)$ in the linear model (treated in Reuter and Schwabe, 2023a). Therefore, the D-optimal subsampling design for ordinary linear regression is also locally D-optimal in the Poisson regression model. Hence, according to (Reuter and Schwabe, 2023a, Section 4), the subsampling design ξ^* is locally D-optimal for $\beta_1 = 0$ if and only if there exist $a_1 < a_2$ such that

$f_{\xi^*}(x) = f_X(x)\mathbb{1}_{(-\infty,a_1]\cup[a_2,\infty)}(x)$,
$F_X(a_2) - F_X(a_1) = 1 - \alpha$, and
$\psi(a_1, \xi^*, \beta_1) = \psi(a_2, \xi^*, \beta_1)$.

By means of equivariance considerations, we may transfer a locally D-optimal subsampling design ξ^* for a covariate X_i to a location-scale transformed covariate $Z_i = aX_i + b$, $a \neq 0$. However, the transformation of a locally D-optimal subsampling design is not as straightforward as in polynomial regression (see Reuter and Schwabe, 2023a), but requires a simultaneous transformation of the slope parameter β_1. This kind of simultaneous transformation typically has to be used in generalized linear models where the elemental information depends on β_1 by the linear component $\mathbf{f}(x)^\top\beta_1$, see e. g. Radloff and Schwabe (2016).

Theorem 4.6. *Let ξ^* be a locally D-optimal subsampling design at β_1 for a covariate X_i with density f_X. Then, for a covariate Z_i with density $f_Z(z) = \frac{1}{|a|} f_X(\frac{z-b}{a})$, the design ζ^* with density $f_{\zeta^*}(z) = \frac{1}{|a|} f_{\xi^*}(\frac{z-b}{a})$ is locally D-optimal at the transformed parameter β_1/a.*

For $a = -1$, Theorem 4.6 covers sign change. Then we can transfer the characterization of a locally D-optimal subsampling design in the equivalence theorem (Theorem 4.2) to positive values for the slope β_1.

Corollary 4.7. *Let $\beta_1 > 0$. Then the subsampling design ξ^* is locally D-optimal at $\boldsymbol{\beta}$ if and only if $f_{\xi^*} = f_X\mathbb{1}_{\mathcal{X}^*}$ and either*

(i) there exist $a_1 < a_2 < a_3$ such that
$\mathcal{X}^* = [a_1, a_2] \cup [a_3, \infty)$,
$F_X(a_1) + F_X(a_3) - F_X(a_2) = 1 - \alpha$, *and*
$\psi(a_1, \xi^*, \beta_1) = \psi(a_2, \xi^*, \beta_1) = \psi(a_3, \xi^*, \beta_1)$,

or

(ii) $\mathcal{X}^* = [q_{1-\alpha}, \infty)$,
$\psi(x, \xi^*, \beta_1) < \psi(q_{1-\alpha}, \xi^*, \beta_1)$ *for* $x < q_\alpha$, *and* $\psi(x, \xi^*, \beta_1) > \psi(q_\alpha, \xi^*, \beta_1)$ *for* $x > q_\alpha$.

To illustrate how the equivalence theorem (Theorem 4.2) can be used to construct locally D-optimal subsampling designs, we consider $\beta_1 < 0$ in the situation of an exponentially and of a uniformly distributed covariate in the following two examples.

Example 4.8 (exponential distribution)**.** We assume the covariate X_i to follow an exponential distribution with rate λ, i. e. X_i has density $f_X(x) = \lambda \exp(-\lambda x)$ for $x \geq 0$. The condition of finite moments $m_k(\xi, \beta_1)$ is satisfied for $\beta_1 < \lambda$ and hence, in particular, for $\beta_1 \leq 0$. For $\beta_1 < 0$, let

$$g_0(t) = \frac{\lambda}{\lambda - \beta_1} \exp(-(\lambda - \beta_1)t),\ g_1(t) = \left(t + \frac{1}{\lambda - \beta_1}\right) g_0(t) \text{ and } g_2(t) = t^2 g_0(t) + \frac{2}{\lambda - \beta_1} g_1(t)$$

such that $g_k(t) = \int_t^\infty x^k \exp(\beta_1 x) f_X(x)\, \mathrm{d}x,\ t \geq 0$. Then, in scenario (i), the entries in $\mathbf{M}(\xi^*, \boldsymbol{\beta})$ are

$$m_k(\xi^*, \beta_1) = g_k(0) - g_k(a_1) + g_k(a_2) - g_k(a_3)\,, \quad k = 0, 1, 2,$$

while they reduce to $m_k(\xi^*, \beta_1) = g_k(0) - g_k(q_\alpha)$ in scenario (ii) when there is only one interval, where $q_\alpha = -\log(1 - \alpha)/\lambda$.

In scenario (i), we obtain numerical results for the boundary points a_1 to a_3 solving the system of equations (4.2a) and (4.2b) using the Newton method implemented in the **R** package *nleqslv* by Hasselman (2018). Note that here $a_3 < x_{\max} = \infty$. For the case of a standard exponential distribution ($\lambda = 1$), results are given in Table 4.1 for selected values of β_1 and α. In addition, we give the values for the amount $F_X(a_1)$ as well as the percentage of mass the design ξ^* places on the left interval $[0, a_1]$. We also add the result for $\beta_1 = 0$ for reference (see Reuter and Schwabe, 2023a).

For other values of the rate λ, results can be derived from the case of a standard exponentially distributed covariate via equivariance (Theorem 4.6) by letting $a = 1/\lambda$ and $b = 0$: If we seek a locally D-optimal subsampling design at $\beta_1 < 0$ when the rate is λ, we can first construct a locally D-optimal design at β_1/λ for a standard exponentially distributed covariate and then divide the obtained boundary points by λ. For example, when $\lambda = 2$, $\beta_1 = -1$ and the subsampling proportion is $\alpha = 0.10$, we get the boundary points $0.05181/2$, $2.92225/2$, and $5.44835/2$ from the second line highlighted in the second block of Table 4.1 such that the locally D-optimal subsampling design wanted is supported on the two intervals $[0, 0.0259]$ and $[1.4611, 2.7242]$.

Table 4.1: Numerical values for the boundary points a_1, a_2, a_3, and q_α, respectively, for selected values of the subsampling proportion α and slope parameter β_1 in the case of a standard exponentially distributed covariate ($\lambda = 1$)

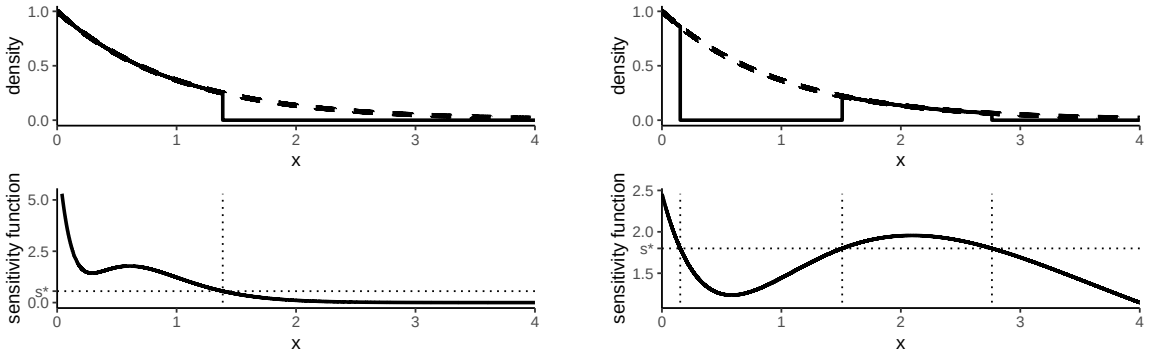
α	β_1	a_1	a_2	a_3, q_α	$F_X(a_1)$	% of mass on $[0, a_1]$
0.01	0.0	0.00579	5.46588	-	0.00577	57.71
	-0.5	0.00501	3.86767	4.14130	0.00500	49.95
	-1.0	0.00500	1.98399	2.02112	0.00499	49.88
	-4.0	0.00496	0.49830	0.50665	0.00495	49.51
0.10	0.0	0.06343	3.25596	-	0.06146	61.46
	-0.5	**0.05181**	**2.92225**	**5.44835**	**0.05049**	**50.49**
	-1.0	0.05011	1.83717	2.22435	0.04887	48.87
	-4.0	0.04680	0.47740	0.56896	0.04572	45.72
0.30	0.0	0.21398	2.23153	-	0.19264	64.21
	-0.5	0.17225	1.95006	7.60885	0.15823	52.74
	-1.0	0.15317	1.50902	2.76234	0.14202	47.34
	-4.0	0.12876	0.40855	0.72273	0.12081	40.27
0.75	0.0	0.67278	1.34596	-	0.48971	65.29
	-0.5	0.52804	1.07947	10.89214	0.41024	54.70
	-1.0	0.43176	0.88401	4.28609	0.35063	46.75
	-4.0	-	-	1.38629	-	-

When the subsampling proportion α goes to zero, the locally D-optimal subsampling design apparently tends to its counterpart in classical optimal design theory which assigns equal weight $1/2$ to two support points $x_1^* = 0$ and $x_2^* = -2/\beta_1$ (see e. g. Rodríguez-Torreblanca and Rodríguez-Díaz, 2007). In particular, we observe $a_2 < x_2^* < a_3$ for all numerically obtained values of a_2 and a_3.

On the contrary, we find that scenario (ii) appears for large values of α. This happens when the slope β_1 is strongly negative. More precisely, given α, there is a crossover point β_1^* such that the single interval design with density $f_{\xi^*} = f_X \mathbb{1}_{[0,q_\alpha]}$ is locally D-optimal at β_1 for all $\beta_1 \geq \beta_1^*$ This crossover point becomes stronger negative when α gets smaller and apparently tends to $-\infty$ as $\alpha \to 0$. On the other hand, when α gets larger, the crossover point apparently tends to zero. In Table 4.2, we give numerical results for the crossover point β_1^*/λ for selected values of α together with the quantile q_α, the setting x_2^* of the locally D-optimal unbounded design and their ratio. This shows that, for scenario (ii) to apply, the quantile q_α has to be substantially larger than x_2^*. Vice versa, for given slope $\beta_1 < 0$,

there is a critical subsampling proportion α^* such that the single interval design is locally D-optimal for larger subsampling proportions $\alpha \geq \alpha^*$. In particular, when $\beta_1 = 0$, only scenario (i) applies (see Reuter and Schwabe, 2023a) and, hence, $\alpha^* = 1$.

We further notice that the percentage of mass on the left interval $[0, a_1]$ is generally larger than 50% for β_1 closer to zero which coincides with what we have seen in Reuter and Schwabe (2023a) for the case $\beta_1 = 0$. There, observations from the right tail are more informative and thus more observations are needed on the left tail. Conversely, the percentage of mass on $[0, a_1]$ is smaller than 50% for strongly negative β_1. Figure 4.1 depicts the locally D-optimal subsampling designs for $\alpha = 0.75, 0.3$ and $\beta_1 = -1$ along with the corresponding sensitivity functions. The horizontal dotted line represents the threshold s^* from Theorem 4.11. The vertical dotted lines depict the boundary points. While smaller subsampling proportions $\alpha \leq 0.1$ are typically of interest in the context of subsampling, our selection of larger subsampling proportions α has been made for the sake of clarity and visibility in the tables and figures.



(a) $\beta_1 = -4$, $\alpha = 0.75$ (b) $\beta_1 = -1$, $\alpha = 0.3$

Figure 4.1: Density of the locally optimal design (solid) at β_1 and the standard exponential distribution (dashed, upper panels), and corresponding sensitivity functions (lower panels) for $\beta_1 = -4$, $\alpha = 0.75$ (left) and $\beta_1 = -1$, $\alpha = 0.3$ (right)

Example 4.9 (uniform distribution)**.** We assume the covariate to be uniform random on an interval $[x_{\min}, x_{\max}]$ with density $f_X(x) = \frac{1}{x_{\max} - x_{\min}} \mathbb{1}_{[x_{\min}, x_{\max}]}(x)$. The condition of finite moments $m_k(\xi, \beta_1)$ is satisfied for all β_1.

For $\beta_1 < 0$, let

$$g_0(t) = \frac{\exp(\beta_1 t)}{|\beta_1|(x_{\max} - x_{\min})},\ g_1(t) = \left(t + \frac{1}{|\beta_1|}\right) g_0(t) \text{ and } g_2(t) = t^2 g_0(t) + \frac{2}{|\beta_1|} g_1(t)\,.$$

Table 4.2: Numerical values for the standardized crossover point β_1^*/λ for an exponentially distributed covariate

α	β_1^*/λ	λq_α	λx_2^*	q_α/x_2^*
0.01	-360.34840	0.01005	0.00556	1.81081
0.10	-34.60684	0.10536	0.05779	1.82310
0.30	-10.41165	0.35667	0.19209	1.85679
0.50	-5.49454	0.69314	0.36400	1.90426
0.75	-2.89534	1.38629	0.69077	2.00690
0.90	-1.86128	2.30259	1.07453	2.14288

In scenario (i), unlike in Example 4.8, the support of the covariate is bounded from above and thus the rightmost boundary point a_3 may be larger than $x_{\max}$. We denote the essential supremum of ξ^* by $\tilde{a}_3 = \min(a_3, x_{\max})$. Then, in scenario (i), the entries in $\mathbf{M}(\xi^*, \boldsymbol{\beta})$ are

$$m_k(\xi^*, \beta_1) = g_k(x_{\min}) - g_k(a_1) + g_k(a_2) - g_k(\tilde{a}_3)\,, \quad k = 0, 1, 2,$$

while in scenario (ii), when there is only one interval, they reduce to $m_k(\xi^*, \beta_1) = g_k(x_{\min}) - g_k(q_\alpha)$ where $q_\alpha = (1-\alpha)x_{\min} + \alpha x_{\max}$.

For the case of a uniform distribution on the unit interval ($x_{\min} = 0$ and $x_{\max} = 1$), optimal boundary points are given in Table 4.3 for selected values of α and $\beta_1 < 0$. In addition, we give the values for the amount $F_X(a_1)$ as well as the percentage of mass the design ξ^* places on the left interval $[0, a_1]$. We also add formally the result for $\beta_1 = 0$ for reference (see Reuter and Schwabe, 2023a).

Apart from the situation that $a_3 > x_{\max}$ indicated by a hyphen ($-$) in the table when $\alpha = 0.5$ and $\beta_1 = -2$, the results are similar to those in Example 4.8: More weight is given to the left interval $[0, a]$ when β_1 is closer to zero. When the subsampling proportion α becomes small, the locally D-optimal subsampling design approaches the locally D-optimal unbounded design equally supported on $x_1^* = 0$ and $x_2^* = -2/\beta_1$. For large values of α, the two intervals are merged into one (e. g. for $\alpha = 0.50$ and $\beta_1 = -8$). Figure 4.2 depicts the locally D-optimal subsampling designs along the corresponding sensitivity functions in scenario (ii) of a single supporting interval for ξ^* in the left panel. The right panel exhibits scenario (i) of ξ^* supported on two proper intervals. The horizontal dotted line depicts the threshold s^*. The vertical dotted lines represent the boundary points a_1, a_2, and a_3. The situation when $a_3 > x_{\max}$ is displayed in Figure 4.3.

Because of the symmetry of the uniform distribution, locally D-optimal subsampling

Table 4.3: Numerical values for the boundary points a_1, a_2, a_3 and q_α, respectively, for selected values of the subsampling proportion α and slope parameter β_1 in the case of a uniformly distributed covariate on $[0, 1]$

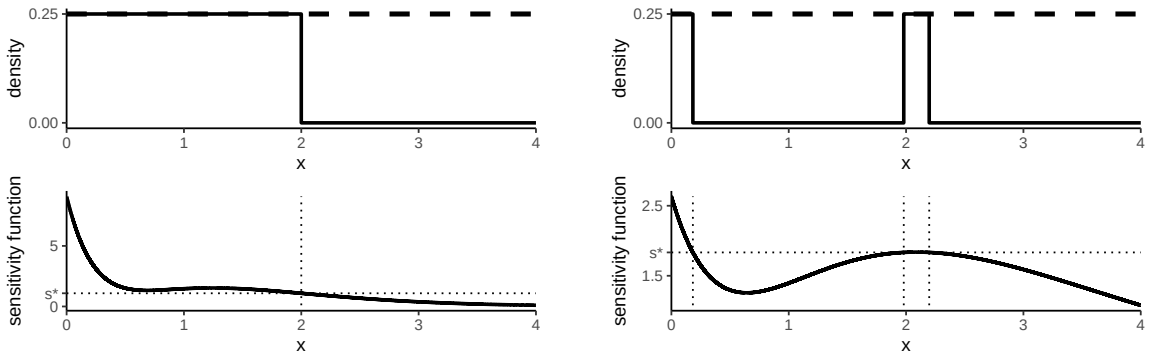
α	β_1	a_1	a_2	a_3, q_α	$F_X(a_1)$	% of mass on $[0, a_1]$
0.01	0	0.00500	0.99500	-	0.00500	50.00
	-2	0.00498	0.99498	-	0.00498	49.75
	-4	0.00495	0.49994	0.50499	0.00495	49.51
	-8	0.00490	0.24989	0.25498	0.00490	49.04
0.10	0	0.05000	0.9500	-	0.05000	50.00
	-2	0.04772	0.94772	-	0.04772	47.72
	-4	**0.04578**	**0.49506**	**0.54928**	**0.04578**	**45.78**
	-8	0.04269	0.24155	0.29887	0.04269	42.69
0.30	0	0.15000	0.8500	-	0.15000	50.00
	-2	0.13271	0.83271	-	0.13271	44.24
	-4	0.12102	0.46678	0.64577	0.12102	40.34
	-8	0.10847	0.20165	0.39318	0.10847	36.16
0.50	0	0.25000	0.7500	-	0.25000	50.00
	-2	0.20993	0.70993	-	0.20993	41.99
	-4	0.18578	0.42624	0.74046	0.18578	37.16
	-8	-	-	0.50000	-	-

designs can be derived for positive values of the slope β_1 via equivariance with respect to sign change by letting $a = -1$ and $b = 1$ in Theorem 4.6. For example, when $\beta_1 = 4$ and $\alpha = 0.10$, the optimal boundary points can be obtained from the third line highlighted in the second block of Table 4.3 as $1 - 0.04578$, $1 - 0.49506$, and $1 - 0.54928$ such that the locally D-optimal subsampling design is then supported on the two intervals $[0.45072, 0.50494]$ and $[0.95422, 1]$.

Further, for other ranges $[x_{\min}, x_{\max}]$ of the uniform covariate, optimal subsampling designs can be obtained by equivariance (Theorem 4.6) as well by letting $a = x_{\max} - x_{\min}$ and $b = x_{\min}$.

4.4 Efficiency

We want to study the performance of random subsampling as well as some heuristic subsampling designs in the style of IBOSS (see Wang et al., 2019) to quantify the gain in using a locally D-optimal subsampling design. Besides, we are interested in the quality of



(a) $\beta_1 = -8$, $\alpha = 0.5$ (b) $\beta_1 = -4$, $\alpha = 0.1$

Figure 4.2: Density of the locally optimal design (solid) at β_1 for a uniformly distributed covariate on $[0, 1]$ (dashed, upper panels), and sensitivity functions (lower panels) for $\beta_1 = -8$, $\alpha = 0.5$ (left) and $\beta_1 = -4$, $\alpha = 0.1$ (right)

the heuristic designs and how they compare to random subsampling. Further, we want to investigate the performance of designs when the parameter is misspecified. Specifically, a subsampling design $\xi^*(\boldsymbol{\beta}') = \arg\max \det(\mathbf{M}(\xi, \boldsymbol{\beta}'))$ that is locally D-optimal at $\boldsymbol{\beta}'$ is studied when the true parameter is $\boldsymbol{\beta}$. The performance of a design ξ may be compared to the locally D-optimal subsampling design $\xi^*(\boldsymbol{\beta})$ using D-efficiency. The D-efficiency of a subsampling design ξ with mass α is defined as

$$\text{eff}_{D,\alpha}(\xi, \boldsymbol{\beta}) = \left(\frac{\det(\mathbf{M}(\xi, \boldsymbol{\beta}))}{\det(\mathbf{M}(\xi^*(\boldsymbol{\beta}), \boldsymbol{\beta}))} \right)^{1/2}.$$

For this definition the homogeneous version $(\det(\mathbf{M}(\xi, \boldsymbol{\beta})))^{1/2}$ of the D-criterion is used which satisfies the homogeneity condition $(\det(\nu\mathbf{M}))^{1/2} = \nu(\det(\mathbf{M}))^{1/2}$ for all $\nu > 0$ (see Pukelsheim, 1993, Chapter 6.2). Note that by Remark 4.1, the efficiency $\text{eff}_{D,\alpha}(\xi, \boldsymbol{\beta})$ does not depend on β_0.

As uniform random subsampling we define the design ξ_α of size α, which has density $f_{\xi_\alpha}(x) = \alpha f_X(x)$. The information matrix of ξ_α is given by $\mathbf{M}(\xi_\alpha, \boldsymbol{\beta}) = \alpha\mathbf{M}(\xi_1, \boldsymbol{\beta})$. Here, ξ_1 represents the full sample with information matrix

$$\mathbf{M}(\xi_1, \boldsymbol{\beta}) = \int \exp(\beta_0 + \beta_1 x)\mathbf{f}(x)\mathbf{f}(x)^\top f_X(x)\, \mathrm{d}x.$$

Thus, the D-efficiency $\text{eff}_{D,\alpha}(\xi_\alpha, \boldsymbol{\beta})$ of uniform random subsampling can be nicely interpreted as noted in Reuter and Schwabe (2023a): for a fixed full sample size n, the required subsample

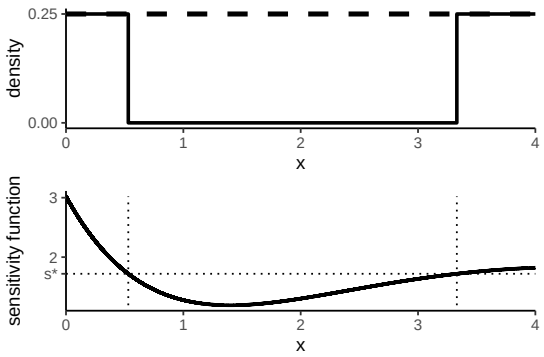


Figure 4.3: Density of the locally optimal design (solid) at β_1 for a uniformly distributed covariate on $[0, 1]$ (dashed, upper panel), and sensitivity functions (lower panel) for $\beta_1 = -2$, $\alpha = 0.3$

size (mass) $\tilde{\alpha}$ needed to achieve the same precision (measured by the D-criterion), compared to utilizing a locally D-optimal subsampling design ξ^* with mass α, is given by the inverse of the efficiency, $\text{eff}_{D,\alpha}(\xi_\alpha, \boldsymbol{\beta})^{-1}$, multiplied by α, i. e. $\tilde{\alpha} = \alpha / \text{eff}_{D,\alpha}(\xi_\alpha, \boldsymbol{\beta})$. For instance, if the efficiency $\text{eff}_{D,\alpha}(\xi_\alpha, \boldsymbol{\beta})$ equals 0.5, then twice the number of observations would be needed under uniform random sampling compared to a locally D-optimal subsampling design of mass α. Naturally, the full sample has higher information than any proper subsample such that, for uniform random subsampling, $\text{eff}_{D,\alpha}(\xi_\alpha, \boldsymbol{\beta}) \geq \alpha$ holds for all α.

Further, we analyze the efficiency of two heuristic designs. Again we only consider the case $\beta_1 < 1$. Let the α-quantile of the covariate X_i be denoted by q_α. First, we consider the one-sided design ξ_{os} with density $f_{\xi_{os}}(x) = f_X(x)\mathbb{1}_{(-\infty, q_\alpha]}(x)$ that assigns all of its mass on the left tail of the distribution of the covariate motivated by its optimality for large α. Second, we study the two-sided design ξ_{ts} with density $f_{\xi_{ts}}(x) = f_X(x)\mathbb{1}_{(-\infty, q_{\alpha/2}]\cup[q_{1-\alpha/2}, \infty)}(x)$ that allocates equal mass $\alpha/2$ on both tails of the distribution in the style of the IBOSS method (see Wang et al., 2019).

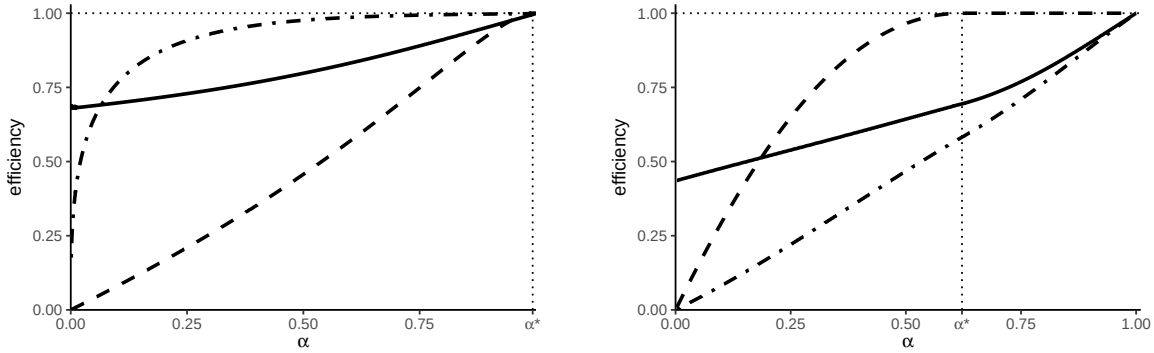
Example 4.10 (exponential distribution)**.** As in Example 4.8, we assume that the covariate X_i is exponentially distributed with rate λ.

Because uniform random subsampling ξ_α as well as the one- and two-sided designs ξ_{os} and ξ_{ts} are equivariant under location-scale transformations, their efficiency depends only on the slope and the rate by the ratio β_1/λ. In Figure 4.4, we depict the efficiency of these designs for $\beta_1/\lambda = -1$ and -4 in dependence on the subsampling proportion α. The efficiency of uniform random subsampling is quite low for reasonable proportions $\alpha \leq 0.1$ and, hence, the gain in using the D-optimal subsampling design is substantial. Similarly, the

efficiency of the one- and the two-sided design is small for $\alpha \leq 0.1$ and apparently tends to zero for $\alpha \to 0$ which may be explained by the fact that these designs miss observations close to the location x_2^* of the locally D-optimal unbounded design. This feature does not apply to uniform random subsampling such that, for very small subsampling proportions, both the one- and the two-sided design is severely less efficient than uniform random subsampling.

As is to be expected, the two-sided IBOSS-like design ξ_{ts} performs much better for β_1 near zero. In particular, for $\beta_1 = 0$, the two-sided design ξ_{ts} only differs slightly from the locally D-optimal subsampling design is ξ^* and has a high efficiency throughout (see Reuter and Schwabe, 2023a). Conversely, the one-sided design ξ_{os} performs better for strongly negative β_1. The vertical dotted line in Figure 4.4 displays the crossover point α^*. For all $\alpha > \alpha^*$, the one-sided design is the D-optimal subsampling design.

We observe similar behavior in Figure 4.5. Predictably, the one-sided design performs better for strongly negative β_1 and the two-sided design is better for β_1 closer to zero. Notably, the two-sided design exhibits a nonmonotonic behavior: It performs worst for $\beta_1/\lambda \approx -3.64$ ($\text{eff}_{D,\alpha}(\xi_{ts}, \boldsymbol{\beta}) = 0.07974506$) and attains a local maximum at $\beta_1/\lambda \approx -0.40$ ($\text{eff}_{D,\alpha}(\xi_{ts}, \boldsymbol{\beta}) = 0.9988009$). Further, we again see that uniform subsampling generally performs better for β_1 closer to zero, though it performs best for $\beta_1/\lambda \approx -1.05$ ($\text{eff}_{D,\alpha}(\xi_{ts}, \boldsymbol{\beta}) = 0.6978610$).



(a) $\beta_1/\lambda = -1$ (b) $\beta_1/\lambda = -4$.

Figure 4.4: D-efficiency of uniform random subsampling (solid), one-sided (dashed), and two-sided (dot-dashed) subsampling design in dependence on the subsampling proportion α for slope-rate ratio $\beta_1/\lambda = -1$ (left) and -4 (right) for an exponentially distributed covariate

For strongly negative β_1, the behavior of the efficiency of the three designs in Figure 4.5 gives additional insight. As $\beta_1 \to -\infty$, the efficiency of uniform random subsampling converges to its lower bound α whereas the efficiency of both one- and two-sided design converge to one. Most of the information is concentrated on the covariate values close to

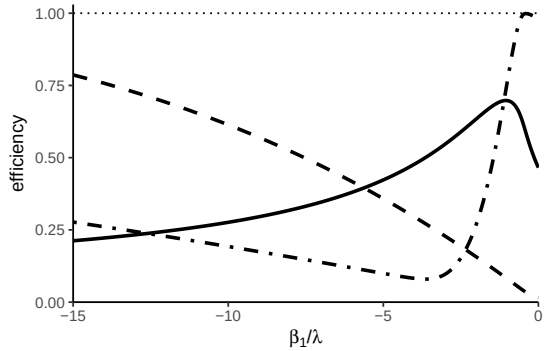


Figure 4.5: D-efficiency of uniform random subsampling (solid), one-sided (dashed), and two-sided (dot-dashed) subsampling design in dependence on the slope-rate ratio β_1/λ for subsampling proportion $\alpha = 0.1$ and an exponentially distributed covariate

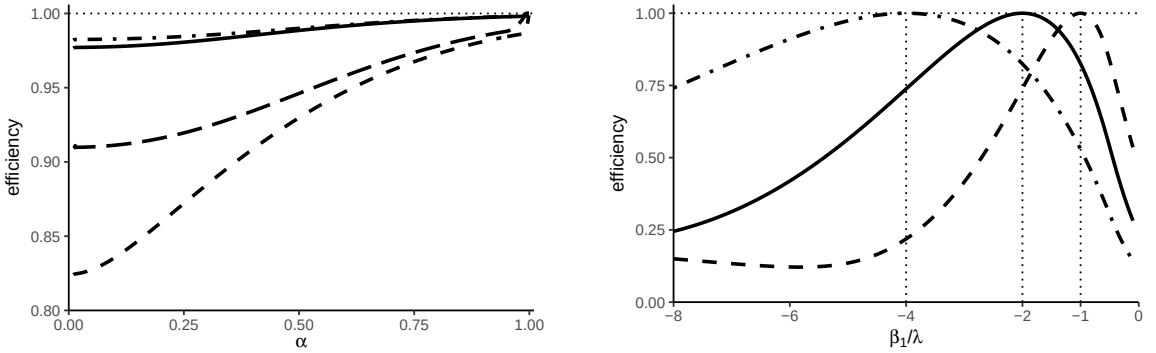
zero. Thus, for strongly negative β_1 the two heuristic designs as well as the D-optimal subsampling design have almost all the information of the full sample. This limiting behavior is not presented in Figure 4.5 in order to preserve intelligibility for β_1 closer to zero.

Finally, we consider the efficiency of locally D-optimal subsampling designs $\xi^*(\boldsymbol{\beta}')$, when the nominal value β_1' is misspecified and differs from the true slope β_1. The left panel of Figure 4.6 illustrates the efficiency of $\xi^*(\boldsymbol{\beta}')$ in dependence on the subsampling proportion α for selected values of the true ratio β_1/λ, when the nominal value is $\beta_1'/\lambda = -1$. For all values we find that the efficiency of the design $\xi^*(\boldsymbol{\beta}')$ under misspecification declines with decreasing α. When the deviation of the parameter is rather small, $\beta_1/\lambda = -0.8$ and $\beta_1/\lambda = -1.2$, the designs under misspecification are still very efficient, with efficiency above 0.98 for $\alpha = 0.01$. For larger deviations however, the efficiency can drop drastically. In particular, when β_1/λ is closer to 0, the efficiency is more strongly negatively affected than when the deviation of β_1/λ is away from zero. In the right panel of Figure 4.6, we exhibit the efficiency for various values of the nominal slope-rate ratio in dependence on the true value when the subsampling proportion is $\alpha = 0.1$. The nominal values are indicated by vertical dotted lines.

It can be seen that the efficiency decreases faster for β_1/λ towards zero than for stronger negative values. In particular, the efficiency increases again when β_1/λ goes to $-\infty$.

4.5 Concluding Remarks

Our investigation centers on a theoretical approach to evaluate subsampling designs under distributional assumptions on the covariate in the case of Poisson regression on a single



(a) True parameter $\beta_1/\lambda = -0.5$ (dashed), -0.8 (solid), -1.2 (dot-dashed), and -1.5 (long dashed)

(b) Locally D-optimal subsampling designs for $\beta_1'/\lambda = -1$ (dashed), -2 (solid), and -4 (dot-dashed)

Figure 4.6: Efficiency of the locally D-optimal subsampling design for $\beta_1'/\lambda = -1$ and various true slope-rate ratios β_1/λ in dependence on α (left) and for $\alpha = 0.1$ and various values of the nominal slope-rate ratio β_1'/λ in dependence on the true slope-rate ratio β_1/λ (right) for an exponentially distributed covariate

covariate. We adjust a standard equivalence theorem to Poisson regression, given a general distribution of the covariate and negative slope parameter β_1. This equivalence theorem also characterizes the support of the locally D-optimal subsampling design and allows us to derive such designs for a given covariate and slope parameter. Then, we establish a theorem to identify locally D-optimal subsampling designs under a scale-location transformation of the covariate and simultaneous rescaling of the slope parameter. We make use of this to give a corollary to the equivalence theorem for $\beta_1 > 0$. It is worthwhile noting that many of the results can be extended from D-optimality to other criteria within Kiefer's class of Φ_p-optimality criteria, including, in particular, linear criteria. The derivation relies mostly on the fact that the sensitivity function can be factorized into the exponential function and a quadratic polynomial, rather than its specific form. Our efficiency analysis shows, among other things, that heuristic one- or two-sided designs can be highly efficient under certain circumstances, however, they display substantial loss in efficiency for the most relevant small subsampling proportions. Addressing uncertainty about the parameter β_1 and the covariate distribution may involve an initial random subsampling step, before deploying the locally D-optimal subsampling design. Lastly, note that the results presented here may be extended to polynomial Poisson regression, where the linear predictor is a polynomial of degree q in the covariate X_i. Then, the equation $\psi(x, \xi, \boldsymbol{\beta}) = s$ has at most $2q + 1$ solutions and the support of ξ^* is the union of at most $q + 1$ intervals.

4.A Proofs

Before we establish the equivalence theorem (Theorem 4.2), we introduce some technical tools: The directional derivative of the D-criterion at design ξ in the direction of a design η is $\Psi(\xi, \eta, \boldsymbol{\beta}) = \text{trace}(\mathbf{M}(\xi, \boldsymbol{\beta})^{-1}\mathbf{M}(\eta, \boldsymbol{\beta})) - 2$. Here, η may be any design of total mass α which is not necessarily required to have a density bounded by f_X. The sensitivity function $\psi(x, \xi, \boldsymbol{\beta}) = \text{trace}(\mathbf{M}(\xi, \boldsymbol{\beta})^{-1}\mathbf{M}(\xi_x, \boldsymbol{\beta}))$ is the essential part of the directional derivative at ξ in the direction of a single point design ξ_x with all mass α at point x. Then

$$\begin{aligned}\psi(x, \xi, \boldsymbol{\beta}) &= \alpha \exp(\beta_0 + \beta_1 x)\mathbf{f}(x)^\top \mathbf{M}(\xi, \boldsymbol{\beta})^{-1}\mathbf{f}(x) \\ &= \frac{\alpha}{d(\xi, \beta_1)} \exp(\beta_1 x)(m_0(\xi, \beta_1)x^2 - 2m_1(\xi, \beta_1)x + m_2(\xi, \beta_1))\end{aligned}$$

does not depend on β_0 and will be denoted by $\psi(x, \xi, \beta_1)$, for short. Note that, for any continuous subsampling design ξ, the information matrix $\mathbf{M}(\xi, \boldsymbol{\beta})$ is positive definite and, hence, $\psi(x, \xi, \beta_1)$ is well-defined.

For convenience, we reproduce an equivalence theorem for subsampling designs in a general model context which follows from Corollary 1(c) in Sahm and Schwabe (2001).

Theorem 4.11. *Let condition*
(A) $\mathrm{P}(\psi(X_i, \xi, \beta_1) = s) = 0$ *for any* ξ *and* s
be satisfied. Then the subsampling design ξ^* *is locally* D*-optimal at* $\boldsymbol{\beta}$ *if and only if there exist a set* $\mathcal{X}^*$ *and a threshold* s^* *such that*

(i) ξ^* *has density* $f_{\xi^*}(x) = f_X(x)\mathbb{1}_{\mathcal{X}^*}(x)$

(ii) $\psi(x, \xi^*, \beta_1) \geq s^*$ *for* $x \in \mathcal{X}^*$*, and*

(iii) $\psi(x, \xi^*, \beta_1) < s^*$ *for* $x \notin \mathcal{X}^*$*.*

Next we establish that condition (A) holds for the Poisson regression model.

Lemma 4.12. *Given* ξ *and* s*, the equation* $\psi(x, \xi, \beta_1) = s$ *has, at most, three different solutions in* x*.*

Proof. For $\beta_1 = 0$, the sensitivity function is a quadratic polynomial in x. Hence, there are, at most, two solutions.

For $\beta_1 \neq 0$, the sensitivity function $\psi(x, \xi, \boldsymbol{\beta}) = \exp(\beta_1 x)q(x)$ factorizes into the exponential function ($\exp(\beta_1 x)$) and a quadratic polynomial q with positive leading term. Because $\psi(x, \xi, \beta_1)$ is positive, only $s > 0$ has to be considered. Let $v(x) = q(x) - s\exp(-\beta_1 x)$.

The third derivative $v^{(3)}(x) = s\beta_1^3 \exp(-\beta_1 x)$ has no roots. By iterative application of the mean value theorem, we see that v has, at most, three roots. Because the solutions of $\psi(x, \xi, \beta_1) = s$ are the roots of v, this completes the proof. □

Condition (A) follows from the continuous distribution of the covariate X_i.

Proof of Theorem 4.2. If ξ^* is locally D-optimal, then, by Theorem 4.11, its density has the shape $f_\xi = f_X \mathbb{1}_{\mathcal{X}}$ and $\mathcal{X}^* = \{x;\, \psi(x, \xi^*, \beta_1) \geq s^*\}$ for some $s^* > 0$. Because $\beta_1 < 0$, the sensitivity function $\psi(x, \xi^*, \beta_1)$ ranges from ∞ for $x \to -\infty$ to 0 for $x \to \infty$ with $\psi(x, \xi^*, \beta_1) > 0$ throughout. Thus, the number of sign changes in $\psi(x, \xi^*, \beta_1) - s^*$ is odd and, by Lemma 4.12, equal to one or three. Hence, $\mathcal{X}^*$ consists of one or two intervals including a left open interval $(-\infty, a_1]$, say, and potentially a second finite interval $[a_2, a_3]$. Conditions (4.2a) and (4.2a'), respectively, follow from the subsampling percentage α. If there are two intervals, then $\psi(a_k, \xi^*, \beta_1) = s^*$, $k = 1, 2, 3$, by continuity of the sensitivity function and we get condition (4.2b) in scenario (i). If there is only one interval, then condition (4.2b') follows from (ii) and (iii) in Theorem 4.11 which completes the proof that the locally D-optimal subsampling design satisfies the properties stated in Theorem 4.2.

Conversely, by the shape of the sensitivity function, the properties stated in Theorem 4.2 imply the equivalence conditions in Theorem 4.11 which proves the reverse statement. □

Proof of Remark 4.4. Assume $a_1 \leq x_{\min}$. Then

$$m_1(\xi^*, \beta_1) = \int_{a_2}^{a_3} x \exp(\beta_1 x) f_X(x)\, \mathrm{d}x > a_2 \int_{a_2}^{a_3} \exp(\beta_1 x) f_X(x)\, \mathrm{d}x = a_2 m_0(\xi^*, \beta_1)$$

and q attains its minimum at $m_1(\xi^*, \beta_1)/m_0(\xi^*, \beta_1) > a_2$. Hence, the sensitivity function $\psi(x, \xi^*, \beta_1) = \exp(\beta_1 x) q(x)$ is strictly decreasing on $(-\infty, a_2]$ such that $\psi(a_1, \xi^*, \beta_1) > \psi(a_2, \xi^*, \beta_1)$ which leads to a contradiction to the optimality condition (4.2b). □

Proof of Theorem 4.6. The proof goes along the same lines as in Radloff and Schwabe (2016). Denote by g the location-scale transformation $g(x) = ax + b$. Let $Z_i = g(X_i)$. Note that only the distribution of the covariate plays a role, but not the covariate itself. The transformation g is conformable with the regression function $\mathbf{f}(x)$, i. e. there exists a nonsingular matrix $\mathbf{Q} = \begin{pmatrix} 1 & 0 \\ b & a \end{pmatrix}$ such that $\mathbf{f}(ax + b) = \mathbf{Q}\mathbf{f}(x)$ for all x. For a design ξ bounded by f_X, we define the transformed design $\zeta = \xi^g$ which has density $f_\zeta(z) = \frac{1}{|a|} f_\xi(\frac{z-b}{a})$ and is bounded by the density $f_Z(z) = \frac{1}{|a|} f_X(\frac{z-b}{a})$ of Z_i. Further, let $\tilde{\boldsymbol{\beta}} = (\mathbf{Q}^\top)^{-1}\boldsymbol{\beta} = (\beta_0 - \beta_1 b/a, \beta_1/a)^\top$.

By the transformation theorem for measure integrals,

$$\begin{aligned}\mathbf{M}(\zeta, \tilde{\boldsymbol{\beta}}) &= \int \exp(\beta_0 + \beta_1(z-b)/a)\mathbf{f}(z)\mathbf{f}(z)^\top \zeta(\mathrm{d}z) \\ &= \int \exp(\beta_0 + \beta_1 x)\mathbf{Q}\mathbf{f}(x)\mathbf{f}(x)^\top \mathbf{Q}^\top \xi(\mathrm{d}x) \\ &= \mathbf{Q}\mathbf{M}(\xi, \boldsymbol{\beta})\mathbf{Q}^\top.\end{aligned}$$

Therefore $\det(\mathbf{M}(\zeta, \tilde{\boldsymbol{\beta}})) = \det(\mathbf{Q})^2 \det(\mathbf{M}(\xi, \boldsymbol{\beta}))$. Thus ξ^* maximizes the D-criterion over the set of subsampling designs bounded by f_X for β_1 if and only if ζ^* maximizes the D-criterion over the set of subsampling designs bounded by f_Z for β_1/a. □

Bibliography

[1] Mingyao Ai, Jun Yu, Huiming Zhang, and HaiYing Wang. Optimal subsampling algorithms for big data regressions. *Statistica Sinica*, 31:749–772, 2021.

[2] Martijn P. F. Berger and Weng Kee Wong. *Applied Optimal Designs*. Wiley, Chichester, 2005.

[3] Qianshun Cheng, HaiYing Wang, and Min Yang. Information-based optimal subdata selection for big data logistic regression. *Journal of Statistical Planning and Inference*, 209:112–122, 2020.

[4] Donald L. Cohn. *Measure theory*. Birkhäuser, Boston, 2013.

[5] Angela Dean, Max Morris, John Stufken, and Derek Bingham. *Handbook of design and analysis of experiments*. CRC Press, Boca Raton, 2015.

[6] Laura Deldossi and Chiara Tommasi. Optimal design subsampling from big datasets. *Journal of Quality Technology*, 54:93–101, 2021.

[7] Michał Dereziński and Manfred K. Warmuth. Reverse iterative volume sampling for linear regression. *The Journal of Machine Learning Research*, 19:853–891, 2018.

[8] Petros Drineas, Michael W. Mahoney, and Shan Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136, 2006.

[9] Valerii V. Fedorov. *Theory of optimal experiments*. Acedemic Press, New York and London, 1972.

[10] Valerii V. Fedorov. Optimal design with bounded density: optimization algorithms of the exchange type. *Journal of Statistical Planning and Inference*, 22:1–13, 1989.

[11] Ronald A. Fisher. *The design of experiments*. Oliver and Boyd, 1935.

[12] Norbert Gaffke and Berthold Heiligers. Approximate designs for polynomial regression: Invariance, admissibility, and optimality. In S. Ghosh and C.R. Rao, editors, *Handbook of Statistics 13*, pages 1149–1199. Elsevier, Amsterdam, 1996.

[13] Allan Gut. *Stopped random walks*. Springer, New York, second edition, 2009.

[14] Berend Hasselman. *nleqslv: Solve Systems of Nonlinear Equations*, 2018. URL `https://CRAN.R-project.org/package=nleqslv`. R package version 3.3.2.

[15] Li He, William Li, Difan Song, and Min Yang. A systematic view of information-based optimal subdata selection: Algorithm development, performance evaluation, and application in financial data. *Statistica Sinica*, 34:611–636, 2024.

[16] Berthold Heiligers and Klaus Schneider. Invariant admissible and optimal designs in cubic regression on the v-ball. *Journal of Statistical Planning and Inference*, 31:113–125, 1992.

[17] T. H. Jones and N. B. Willms. Inverse eigenvalue problems for checkerboard toeplitz matrices. *Journal of Physics: Conference Series*, 1047:012016, 2018.

[18] V. Roshan Joseph and Simon Mak. Supervised compression of big data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14:217–229, 2021.

[19] Jack Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society: Series B*, 21:272–319, 1959.

[20] Jack Kiefer. Optimum experimental designs V, with applications to systematic and rotatable designs. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob.*, pages 381–405. Univ. California Press, Berkeley, 1960.

[21] Jack Kiefer. General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*, 2:849–879, 1974.

[22] Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

[23] Nan Lin and Ruibin Xi. Aggregated estimating equation estimation. *Statistics and its Interface*, 4:73–83, 2011.

[24] Yanxi Liu, Min Yang, and John Stufken. Information-based optimal subdata selection for clusterwise linear regression. *Statistica Sinica*, 36, 2026, forthcoming.

[25] Ping Ma, Michael W. Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. In *Proceedings of the 31st International Conference on Machine Learning*, pages 91–99, 2014.

[26] Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3:123–224, 2011.

[27] Conrado Martínez. Partial quicksort. In *Proc. 6th ACMSIAM Workshop on Algorithm Engineering and Experiments and 1st ACM-SIAM Workshop on Analytic Algorithmics and Combinatorics*, pages 224–228, 2004.

[28] Frank Miller. *Optimale Versuchspläne bei Einschränkungen in der Versuchspunktwahl*. PhD thesis, Universität Karlsruhe, 2002.

[29] Luc Pronzato. A minimax equivalence theorem for optimum bounded design measures. *Statistics & Probability Letters*, 68:325–331, 2004.

[30] Luc Pronzato. A delimitation of the support of optimal designs for Kiefer's ϕ_p-class of criteria. *Statistics & Probability Letters*, 83:2721–2728, 2013.

[31] Luc Pronzato and HaiYing Wang. Sequential online subsampling for thinning experimental designs. *Journal of Statistical Planning and Inference*, 212:169–193, 2021.

[32] Friedrich Pukelsheim. *Optimal Design of Experiments*. Wiley, New York, 1993.

[33] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

[34] Martin Radloff and Rainer Schwabe. Invariance and equivariance in experimental design for nonlinear models. In J. Kunert, C.H. Müller, and A.C. Atkinson, editors, *mODa 11-Advances in Model-Oriented Design and Analysis*, pages 217–224. Springer, 2016.

[35] Torsten Reuter and Rainer Schwabe. Optimal subsampling design for polynomial regression in one covariate. *Statistical Papers*, 64:1095–1117, 2023a.

[36] Torsten Reuter and Rainer Schwabe. D-optimal subsampling design for massive data linear regression. *arXiv preprint arXiv:2307.02236*, 2023b.

[37] Torsten Reuter and Rainer Schwabe. Poisson regression in one covariate on massive data. *arXiv preprint arXiv:2403.18432*, 2024.

[38] C. Rodríguez-Torreblanca and J.M. Rodríguez-Díaz. Locally D-and c-optimal designs for Poisson and negative binomial regression models. *Metrika*, 66:161–172, 2007.

[39] Kenneth G. Russell, David C. Woods, Susan M. Lewis, and John A. Eccleston. D-optimal designs for Poisson regression models. *Statistica Sinica*, 19:721–730, 2009.

[40] Michael Sahm and Rainer Schwabe. A note on optimal bounded designs. In A. Atkinson, B. Bogacka, and A. Zhigljavsky, editors, *Optimum Design 2000*, pages 131–140. Kluwer, Dordrecht, 2001.

[41] Rainer Schwabe. *Optimum designs for multi-factor models*. Springer, New York, 1996.

[42] Chenlu Shi and Boxin Tang. Model-robust subdata selection for big data. *Journal of Statistical Theory and Practice*, 15:1–17, 2021.

[43] Samuel D. Silvey. *Optimal design*. Chapman and Hall, London, 1980.

[44] Rakhi Singh. A model-free subdata selection method for classification. *arXiv preprint arXiv:2404.19127*, 2024.

[45] Rakhi Singh and John Stufken. Subdata selection with a large number of variables. *The New England Journal of Statistics in Data Science*, 1:426–438, 2023.

[46] Kirstine Smith. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12:1–85, 1918.

[47] Miaomiao Su, Ruoyu Wang, and Qihua Wang. A two-stage optimal subsampling estimation for missing data problems with large-scale data. *Computational Statistics & Data Analysis*, 173:107505, 2022.

[48] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996.

[49] Mahmood Ul Hassan and Frank Miller. Optimal item calibration for computerized achievement tests. *Psychometrika*, 84:1101–1128, 2019.

[50] HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113:829–844, 2018.

[51] HaiYing Wang, Min Yang, and John Stufken. Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114:393–405, 2019.

[52] Lin Wang, Jake Elmstedt, Weng Kee Wong, and Hongquan Xu. Orthogonal subsampling for big data linear regression. *The Annals of Applied Statistics*, 15:1273–1290, 2021.

[53] Henry P. Wynn. Optimum designs for finite populations sampling. In S.S. Gupta and D.S. Moore, editors, *Statistical Decision Theory and Related Topics II*, pages 471–478. Academic Press, New York, 1977.

[54] Yaqiong Yao and HaiYing Wang. A review on optimal subsampling methods for massive datasets. *Journal of Data Science*, 19:151–172, 2021.

[55] Jun Yu and HaiYing Wang. Subdata selection algorithm for linear model discrimination. *Statistical Papers*, 63:1883–1906, 2022.

[56] Jun Yu, HaiYing Wang, Mingyao Ai, and Huiming Zhang. Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association*, 117:265–276, 2022.

[57] Jun Yu, Mingyao Ai, and Zhiqiang Ye. A review on design inspired subsampling for big data. *Statistical Papers*, 65:467–510, 2024.

[58] Haixiang Zhang and HaiYing Wang. Distributed subdata selection for big data via sampling-based approach. *Computational Statistics & Data Analysis*, 153:107072, 2021.

[59] Tao Zhang, Yang Ning, and David Ruppert. Optimal sampling for generalized linear models under measurement constraints. *Journal of Computational and Graphical Statistics*, 30:106–114, 2021.

List of Symbols

$\mathbf{1}_m$	$m \times 1$ vector with all entries equal to one
$\mathbb{1}_A(\cdot)$	indicator function of a subset A
d	dimension of the covariate
$\det(\cdot)$	determinant of a matrix
$\mathrm{diag}(\cdot)$	diagonal matrix
$\mathrm{eff}_\Phi(\xi, \boldsymbol{\beta})$	Φ-efficiency of a (subsampling) design ξ for parameter $\boldsymbol{\beta}$
$\mathbf{f}(\cdot)$	regression function
$f_{\boldsymbol{X}}(\cdot)$	density of a random covariate $\boldsymbol{X}_i$
$f_\xi(\cdot)$	density of a subsampling design ξ
$F_\Phi(\xi, \eta, \boldsymbol{\beta})$	directional derivative of Φ at (subsampling) design ξ in the direction of (subsampling) design η for parameter $\boldsymbol{\beta}$
$\mathbb{I}_m$	$m \times m$ identity matrix
k	size of the subsample
$\mathbf{M}(\xi, \boldsymbol{\beta})$	information matrix of a (subsampling) design ξ for parameter $\boldsymbol{\beta}$
n	size of the full sample
$\mathrm{NND}(p)$	closed cone of nonnegative definite $p \times p$ matrices
p	dimension of $\boldsymbol{\beta}$
$SO(d)$	special orthogonal group acting on $\mathbb{R}^d$
$\mathrm{trace}(\cdot)$	trace of a matrix
$\boldsymbol{X}_i$	random covariate
$\mathcal{X}$	design region
$\mathcal{X}^*$	support of the (locally) Φ-optimal (subsampling) design
Y_i	response variable
α	subsampling proportion
$\boldsymbol{\beta}$	vector of unknown parameters
$\boldsymbol{\beta}_1$	vector of unknown parameters without intercept
$\boldsymbol{\mu}_{\boldsymbol{X}}$	mean vector of a random covariate $\boldsymbol{X}_i$

ξ	(subsampling) design
$\xi_{\mathbf{x}}$	one-point design at $\mathbf{x}$
ξ^*	(locally) Φ-optimal (subsampling) design
$\overline{\xi}$	symmetrized subsampling design
$\Xi^{f\boldsymbol{X}}$	set of subsampling designs bounded by $f_{\boldsymbol{X}}$
σ^2_ε	variance of the random errors ε_i
$\boldsymbol{\Sigma}_{\boldsymbol{X}}$	covariance matrix of a random covariate $\boldsymbol{X}_i$
$\psi(\mathbf{x}, \xi, \boldsymbol{\beta})$	sensitivity function at point $\mathbf{x}$ for a (subsampling) design ξ and parameter $\boldsymbol{\beta}$