Torsten Reuter

**mODa13**

July 10, 2023

# $D$-Optimal Subsampling Design for Polynomial Regression In One Covariate

Joint work with Rainer Schwabe

Otto-von-Guericke-Universität Magdeburg

DFG-Graduiertenkolleg
MATHEMATISCHE
KOMPLEXITÄTSREDUKTION

# Quadratic Regression

We consider the quadratic regression model in one covariate

$$\begin{aligned} Y_i &= \mathbf{f}(X_i)^\top \boldsymbol{\beta} + \varepsilon_i \\ &= \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i, \quad i = 1, \ldots, n. \end{aligned}$$

- $X_i$ are iid random variables in $\mathbb{R}$.
- $\mathbf{f}(x) = (1, x, x^2)^\top$.
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$ is the parameter vector.
- $\varepsilon_i$ are independent, homoscedastic random errors with $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2 > 0$.

# Massive Data Setting

Density $f_X(x)$ of $X_i$ known. We want to find a design $\xi$ with density $f_\xi(x)$ such that

# Massive Data Setting

Density $f_X(x)$ of $X_i$ known. We want to find a design $\xi$ with density $f_\xi(x)$ such that

- $f_\xi(x) \leq f_X(x)$ so that $\xi$ generates a subsample of the $X_i$.

# Massive Data Setting

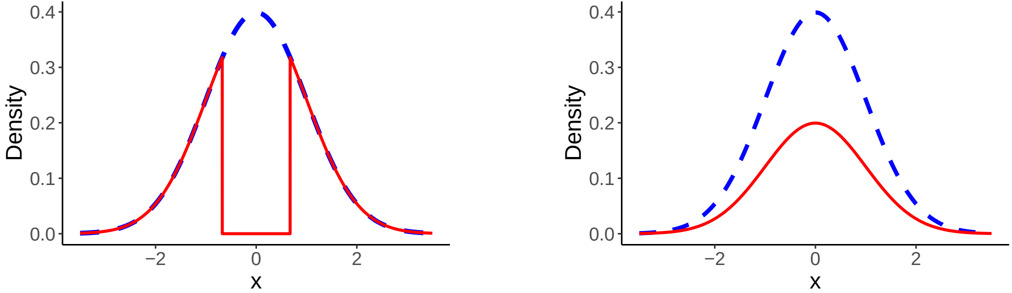
Density $f_X(x)$ of $X_i$ known. We want to find a design $\xi$ with density $f_\xi(x)$ such that

- $f_\xi(x) \leq f_X(x)$ so that $\xi$ generates a subsample of the $X_i$.
- $\int f_\xi(x)\,\mathrm{d}x = \alpha$, $\alpha$ is the percentage of the full data to be selected.

# Massive Data Setting

Density $f_X(x)$ of $X_i$ known. We want to find a design $\xi$ with density $f_\xi(x)$ such that

- $f_\xi(x) \leq f_X(x)$ so that $\xi$ generates a subsample of the $X_i$.
- $\int f_\xi(x)\,\mathrm{d}x = \alpha$, $\alpha$ is the percentage of the full data to be selected.




$f_X(x)$ (blue) for $X_i \sim \mathcal{N}(0,1)$ and two possible $f_\xi(x)$ (red) for $\alpha = 0.5$.

# $D$-optimality

We define the information matrix as

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

We want to find the $D$-optimal design, i.e. $\xi^*$ that maximizes

$$\Psi(\xi) = \det(\mathbf{M}(\xi)).$$

# Sensitivity Function

**Sensitivity function** $\psi(x,\xi)$ from $\xi$ to a single point measure $\xi_x$ at point $x$

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

$\psi(x,\xi)$ is a polynomial of degree 4 in $x$.

# Equivalence Theorem for Quadratic Regression

Assume $f_X(x)$ is symmetric.

**Theorem**

*The design $\xi^*$ is D-optimal if and only if there exist $\mathcal{X}^* \subset \mathbb{R}$ and a threshold $s^*$ such that*

(i) *the D-optimal design $\xi^*$ is given by*

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

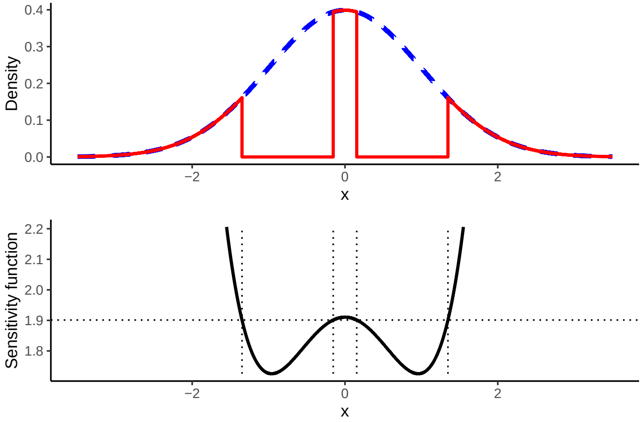
(ii) $\psi(x, \xi^*) \geq s^*$ *for* $x \in \mathcal{X}^*$, *and*

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

*where $\mathcal{X}^*$ is the union of at most three symmetrically placed intervals.*

E.g. $\mathcal{X}^* = (-\infty, -a] \cup [-b, b] \cup [a, \infty)$, $a > b > 0$.

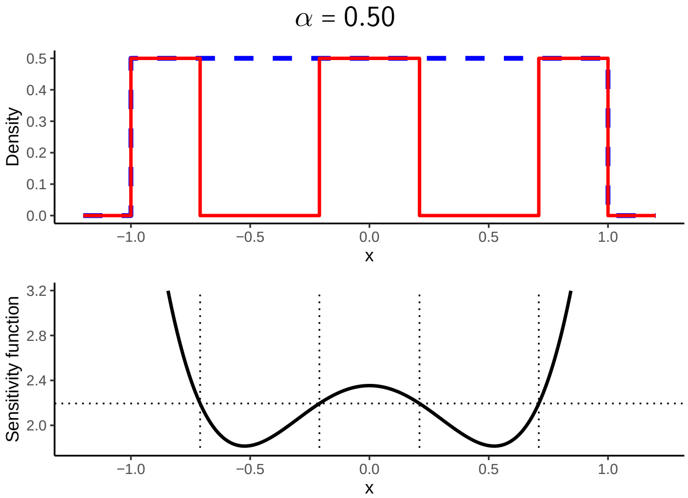
# Equivalence Theorem for Quadratic Regression




Density of the optimal design (red) and the standard normal distribution (blue) and sensitivity function (lower panel) for $\alpha = 0.3$.

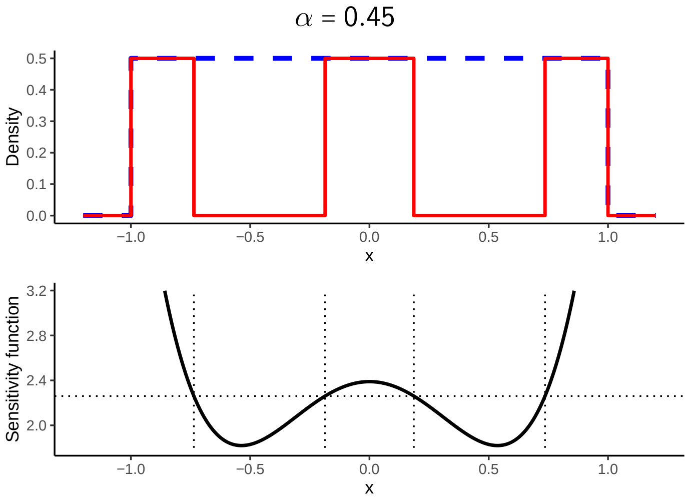
Application of [Sahm and Schwabe, 2001]; see also: [Pronzato and Wang, 2021].

# Uniform Distribution - no surprise



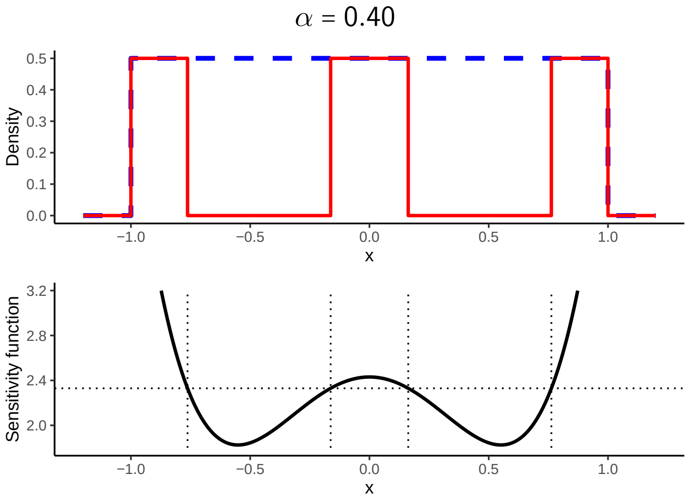

Density of the optimal design (red) and the uniform distribution (blue) and sensitivity function (lower panel).

# Uniform Distribution - no surprise



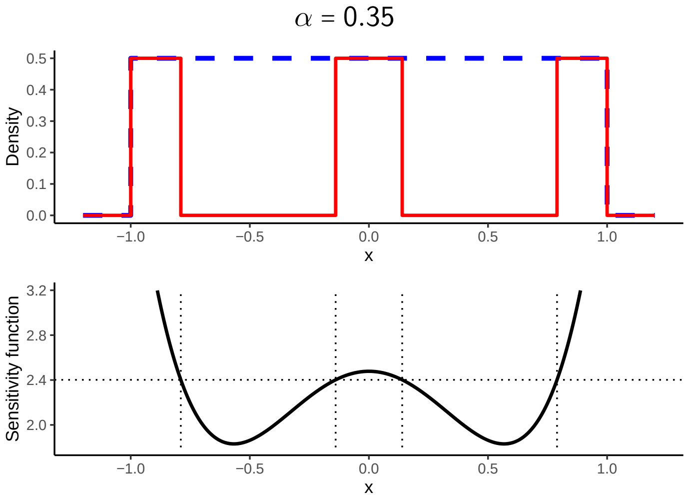

Density of the optimal design (red) and the uniform distribution (blue) and sensitivity function (lower panel).

# Uniform Distribution - no surprise



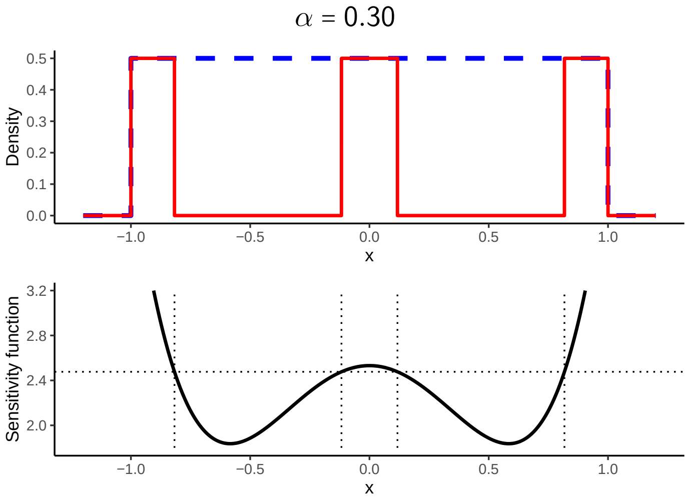

Density of the optimal design (red) and the uniform distribution (blue) and sensitivity function (lower panel).

# Uniform Distribution - no surprise



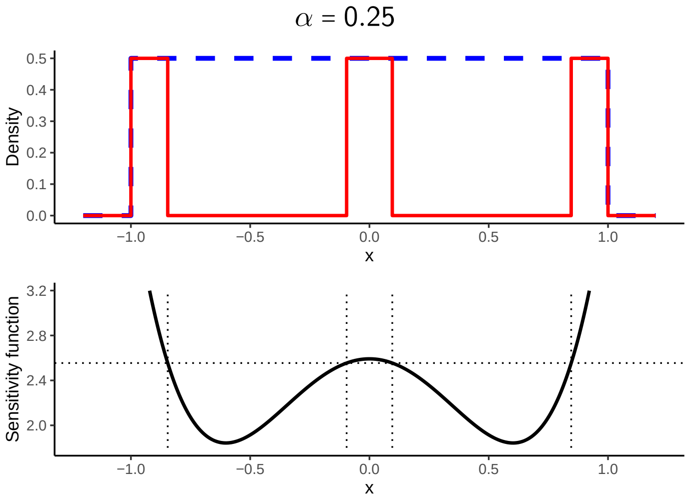

Density of the optimal design (red) and the uniform distribution (blue) and sensitivity function (lower panel).

# Uniform Distribution - no surprise



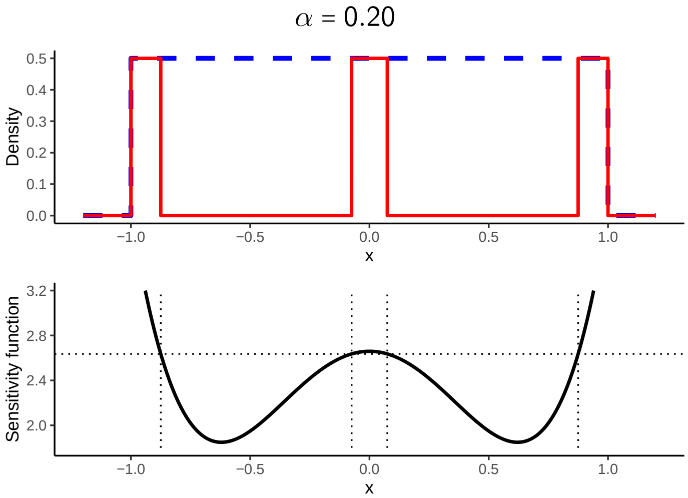

Density of the optimal design (red) and the uniform distribution (blue) and sensitivity function (lower panel).

# Uniform Distribution - no surprise



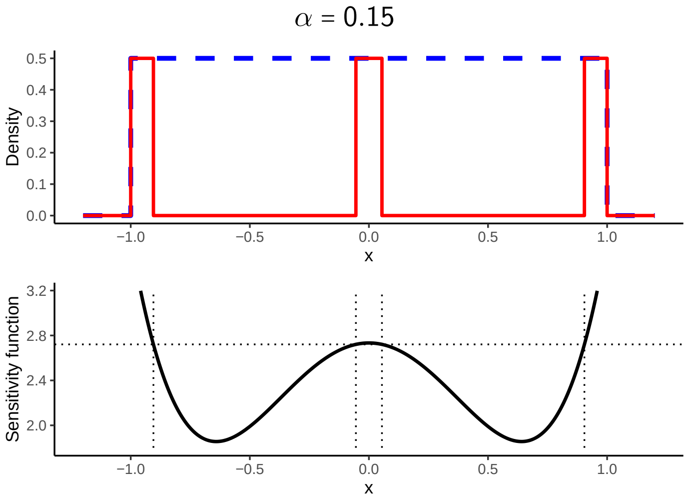

Density of the optimal design (red) and the uniform distribution (blue) and sensitivity function (lower panel).

# Uniform Distribution - no surprise




Density of the optimal design (red) and the uniform distribution (blue) and sensitivity function (lower panel).

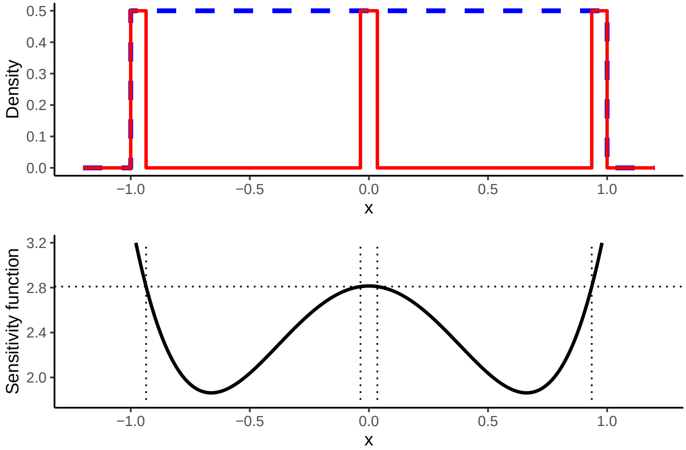
# Uniform Distribution - no surprise




Density of the optimal design (red) and the uniform distribution (blue) and sensitivity function (lower panel).

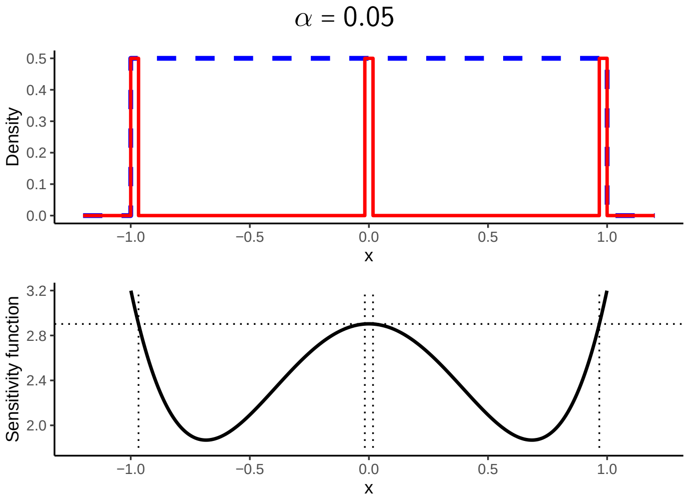
# Uniform Distribution - no surprise

$\alpha = 0.10$



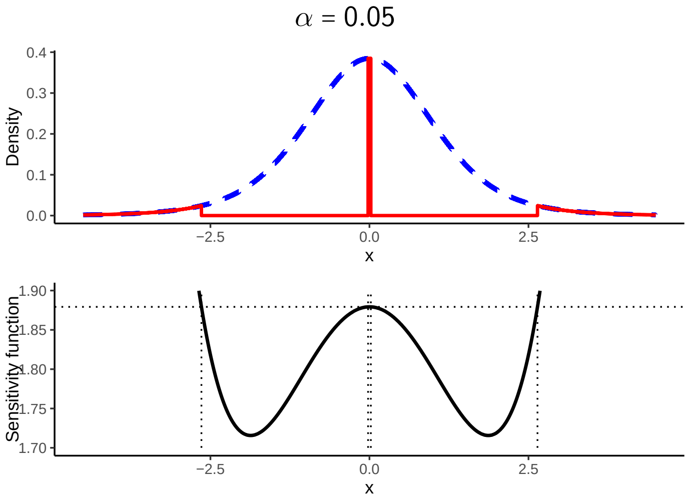

Density of the optimal design (red) and the uniform distribution (blue) and sensitivity function (lower panel).

# Uniform Distribution - no surprise



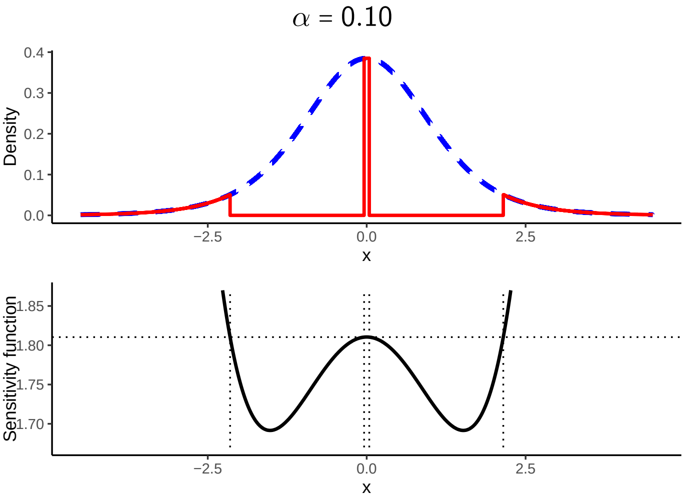

Density of the optimal design (red) and the uniform distribution (blue) and sensitivity function (lower panel).

# $t$-distribution - very surprising!



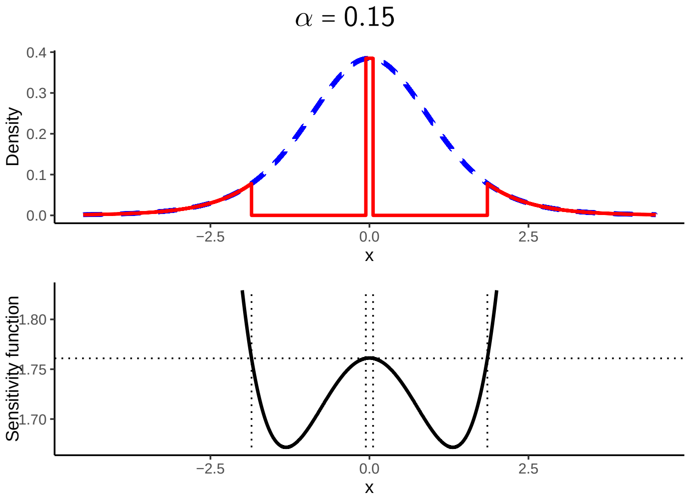

Density of the optimal design (red) and the $t_7$-distribution (blue) and sensitivity function (lower panel).

# $t$-distribution - very surprising!



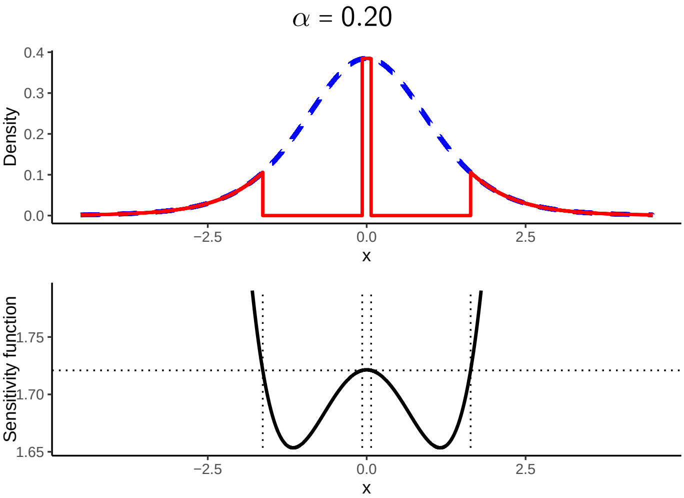

Density of the optimal design (red) and the $t_7$-distribution (blue) and sensitivity function (lower panel).

# $t$-distribution - very surprising!



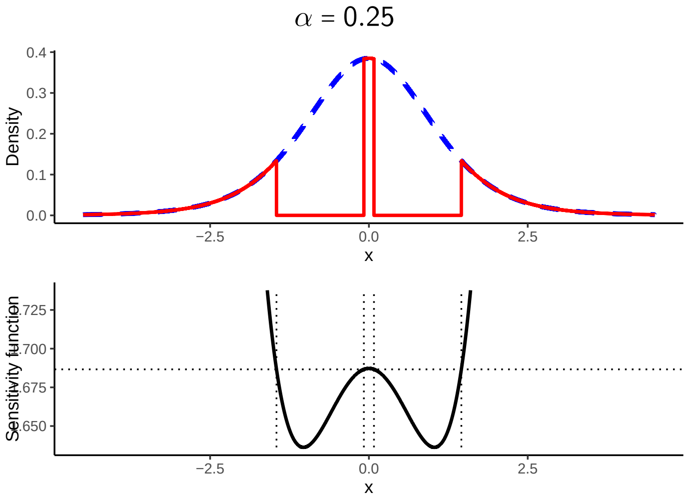

Density of the optimal design (red) and the $t_7$-distribution (blue) and sensitivity function (lower panel).

# $t$-distribution - very surprising!



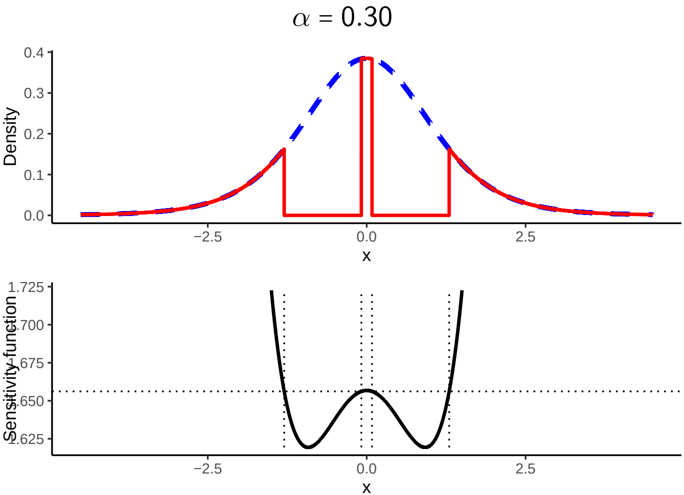

Density of the optimal design (red) and the $t_7$-distribution (blue) and sensitivity function (lower panel).

# $t$-distribution - very surprising!



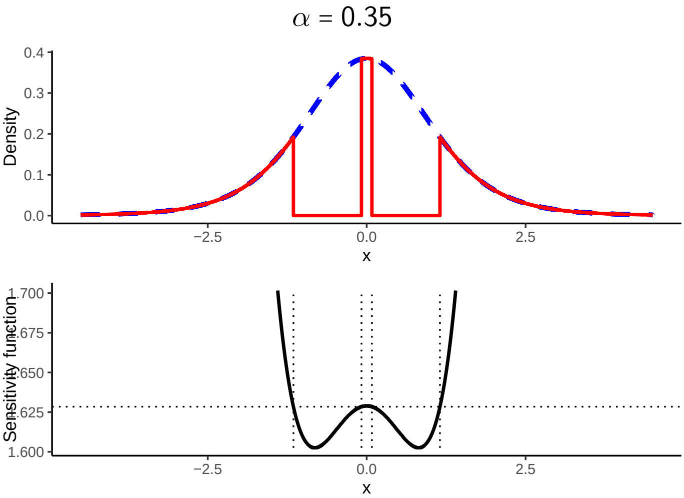

Density of the optimal design (red) and the $t_7$-distribution (blue) and sensitivity function (lower panel).

# $t$-distribution - very surprising!



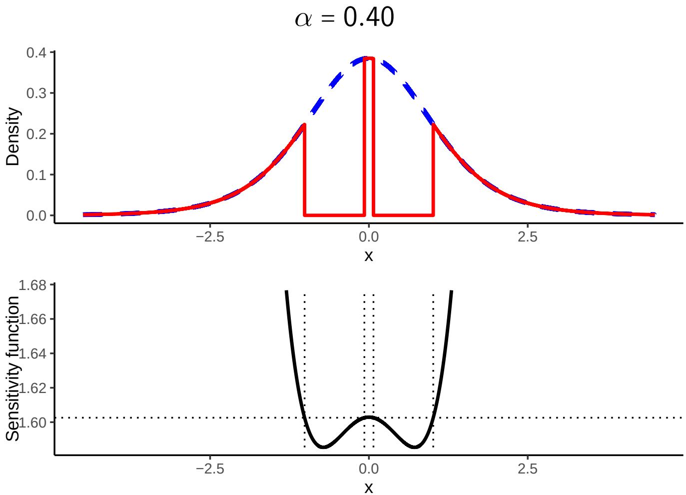

Density of the optimal design (red) and the $t_7$-distribution (blue) and sensitivity function (lower panel).

# $t$-distribution - very surprising!



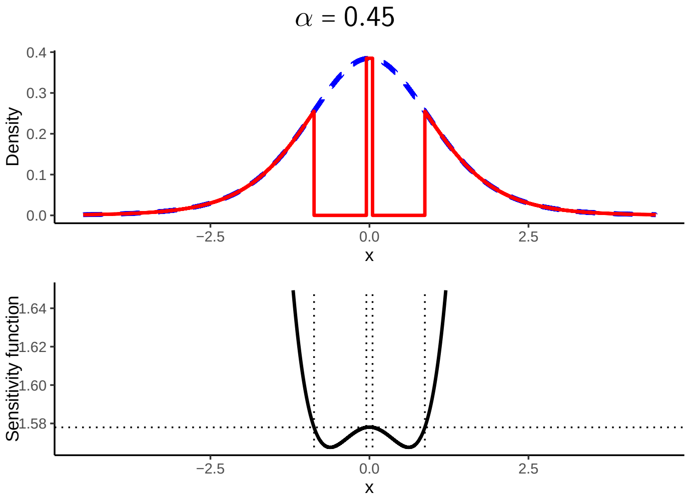

Density of the optimal design (red) and the $t_7$-distribution (blue) and sensitivity function (lower panel).

# $t$-distribution - very surprising!



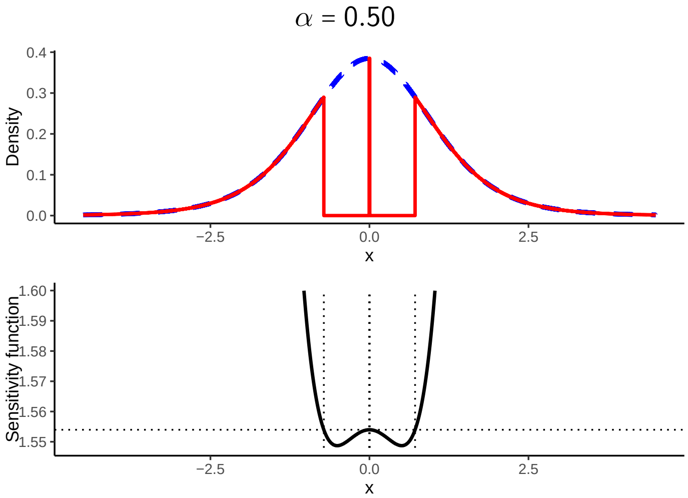

Density of the optimal design (red) and the $t_7$-distribution (blue) and sensitivity function (lower panel).

# $t$-distribution - very surprising!



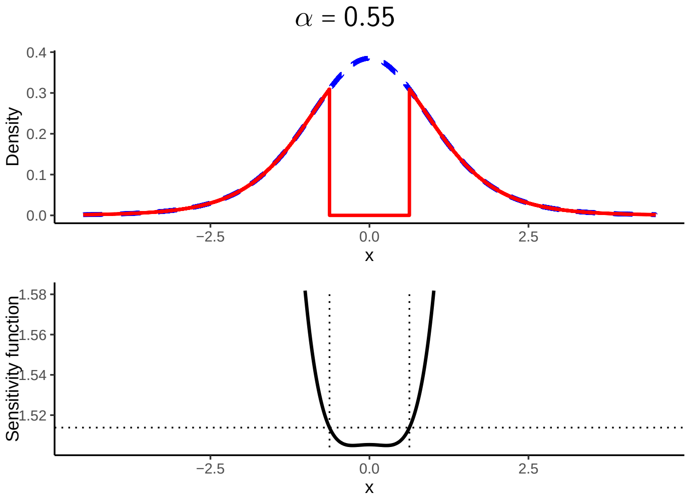

Density of the optimal design (red) and the $t_7$-distribution (blue) and sensitivity function (lower panel).

# $t$-distribution - very surprising!



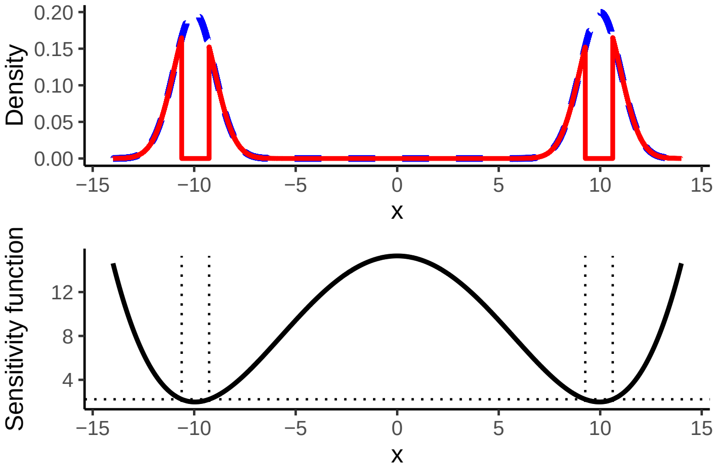

Density of the optimal design (red) and the $t_7$-distribution (blue) and sensitivity function (lower panel).

# $t$-distribution - very surprising!




Density of the optimal design (red) and the $t_7$-distribution (blue) and sensitivity function (lower panel).

# Mixture of Gaussians




Density of the optimal design (red) and a mixture of two normal distributions (blue) and sensitivity function (lower panel) for $\alpha = 0.5$.

# IBOSS-like Designs

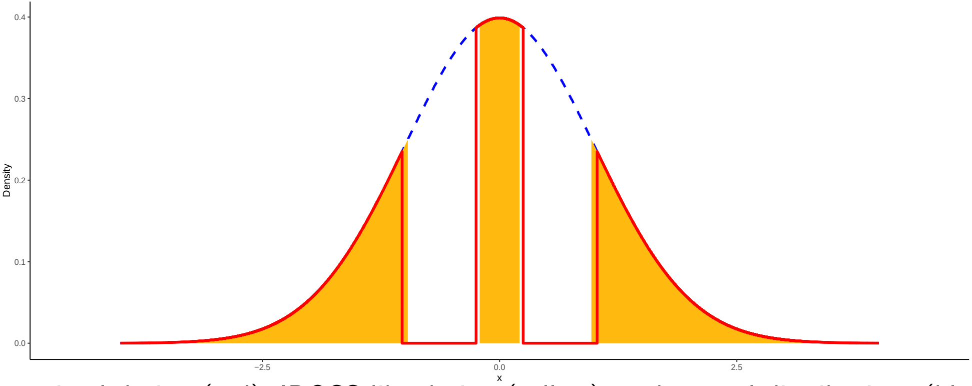
**Goal:** Find a subsampling strategy independent of the distribution of the $X_i$.

**Idea:** Subsampling design with three symmetrically placed intervals of measure $\alpha/3$.
Like IBOSS [Wang et al., 2019] for linear regression.

# IBOSS-like Designs

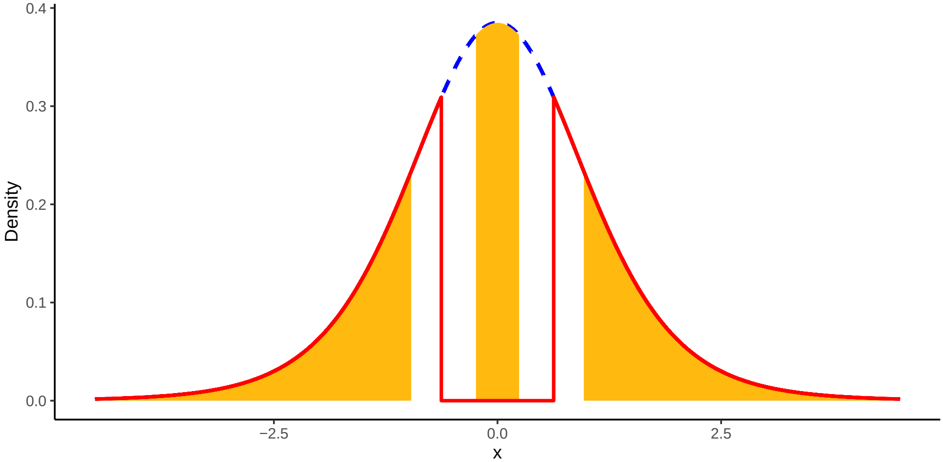
**Goal:** Find a subsampling strategy independent of the distribution of the $X_i$.

**Idea:** Subsampling design with three symmetrically placed intervals of measure $\alpha/3$.
Like IBOSS [Wang et al., 2019] for linear regression.




D-optimal design (red), IBOSS-like design (yellow), and normal distributions (blue) for $\alpha = 0.5$.

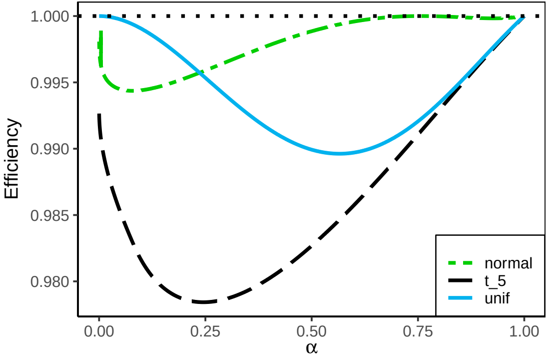
# IBOSS-like Designs




D-optimal design (red), IBOSS-like design (yellow), and $t_7$-distributions (blue) for $\alpha = 0.55$.

# Efficiency of IBOSS-like designs

$$\mathrm{eff}_{D,\alpha}(\xi_\alpha^{IBOSS}) = \left( \frac{\det(\mathbf{M}(\xi_\alpha^{IBOSS}))}{\det(\mathbf{M}(\xi_\alpha^{*}))} \right)^{1/3},$$




Efficiencies of IBOSS-like designs for standard normal (green), $t_5$ (black), and uniform (blue) distributions.

# Multiple Linear Regression

We consider the multiple linear regression model

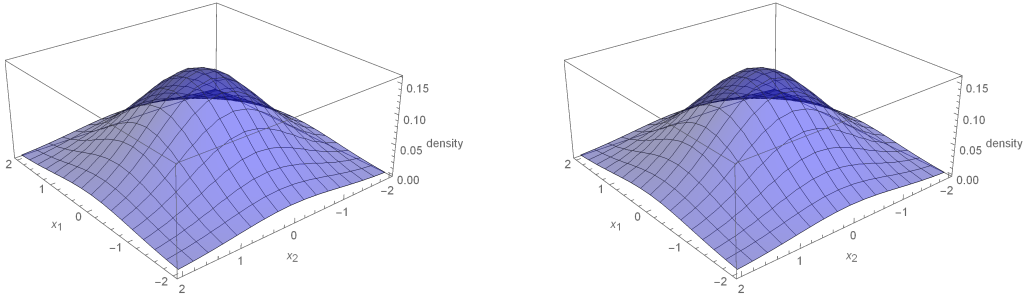
$$\begin{aligned} Y_i &= \mathbf{f}(\boldsymbol{X}_i)^\top \boldsymbol{\beta} + \varepsilon_i \\ &= \beta_0 + \beta_1 X_{i1} + \cdots + \beta_d X_{id} + \varepsilon_i, \quad i = 1, \ldots, n. \end{aligned}$$

- $\boldsymbol{X}_i$ are iid random vectors on $\mathbb{R}^d$.
- $\mathbf{f}(\mathbf{x}) = (1, \mathbf{x}^\top)^\top$.
- $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)^\top$ is the $d + 1$-dimensional parameter vector.
- $\varepsilon_i$ are uncorrelated and homoscedastic errors with $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2 > 0$.

# Big Data Setting

Density $f_{\boldsymbol{X}}(\mathbf{x})$ of $\boldsymbol{X}_i$ known. We want to find a design $\xi$ with density $f_\xi(\mathbf{x})$ such that

- $f_\xi(\mathbf{x}) \leq f_{\boldsymbol{X}}(\mathbf{x})$ so that $\xi$ generates a subsample of the $\boldsymbol{X}_i$.
- $\int f_\xi(\mathbf{x})\,\mathrm{d}\mathbf{x} = \alpha$, $\alpha$ is the percentage of the full data to be selected.

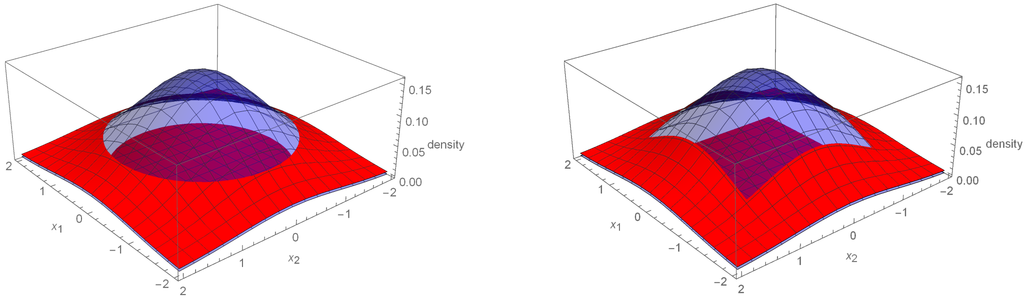



$f_{\boldsymbol{X}}(\mathbf{x})$ (blue) for $\boldsymbol{X}_i \sim \mathcal{N}_2(\mathbf{0}, \mathbb{I}_2)$ and two possible $f_\xi(\mathbf{x})$ (red) for $\alpha = 0.5$.

# Big Data Setting

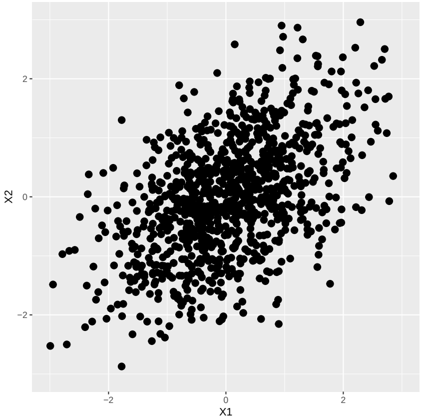
Density $f_{\boldsymbol{X}}(\mathbf{x})$ of $\boldsymbol{X}_i$ known. We want to find a design $\xi$ with density $f_\xi(\mathbf{x})$ such that

- $f_\xi(\mathbf{x}) \leq f_{\boldsymbol{X}}(\mathbf{x})$ so that $\xi$ generates a subsample of the $\boldsymbol{X}_i$.
- $\int f_\xi(\mathbf{x})\,\mathrm{d}\mathbf{x} = \alpha$, $\alpha$ is the percentage of the full data to be selected.



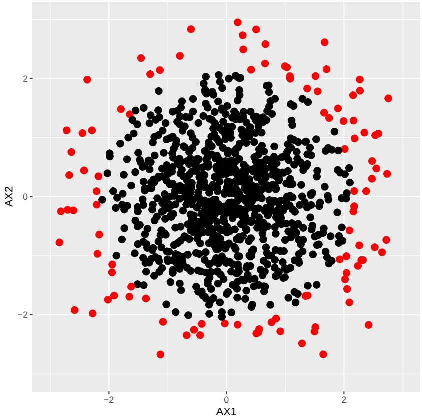

$f_{\boldsymbol{X}}(\mathbf{x})$ (blue) for $\boldsymbol{X}_i \sim \mathcal{N}_2(\mathbf{0}, \mathbb{I}_2)$ and two possible $f_\xi(\mathbf{x})$ (red) for $\alpha = 0.5$.

**Example:** $\boldsymbol{X}_i \sim \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\sigma_j^2 = 1$ and $\rho = 0.5$.

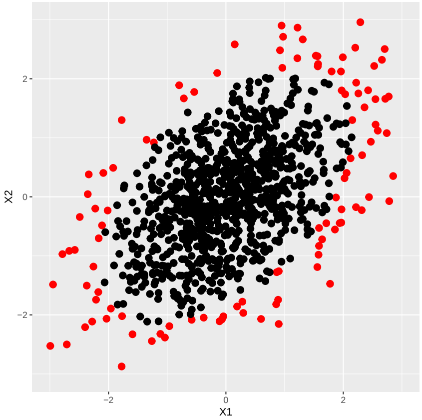


**Example:** $\boldsymbol{X}_i \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$, where $\sigma_j^2 = 1$ and $\rho = 0.5$.

Transform $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i \rightarrow$ select units with largest euclidean distance.



**Example:** $\boldsymbol{X}_i \sim \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma})$, where $\sigma_j^2 = 1$ and $\rho = 0.5$.



# Simplified Method

Instead of transforming the covariates with a root of the full covariance matrix $\mathbf{\Sigma}$ we only scale the variance by

$$\tilde{\mathbf{\Sigma}} = \begin{pmatrix} \sigma_1^2 & & & \mathbf{0} \\ & \sigma_2^2 & & \\ & & \ddots & \\ \mathbf{0} & & & \sigma_d^2 \end{pmatrix} \text{ and its root } \tilde{\mathbf{\Sigma}}^{1/2} = \begin{pmatrix} \sigma_1 & & & \mathbf{0} \\ & \sigma_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \sigma_d \end{pmatrix}.$$

# Simplified Method

Instead of transforming the covariates with a root of the full covariance matrix $\boldsymbol{\Sigma}$ we only scale the variance by

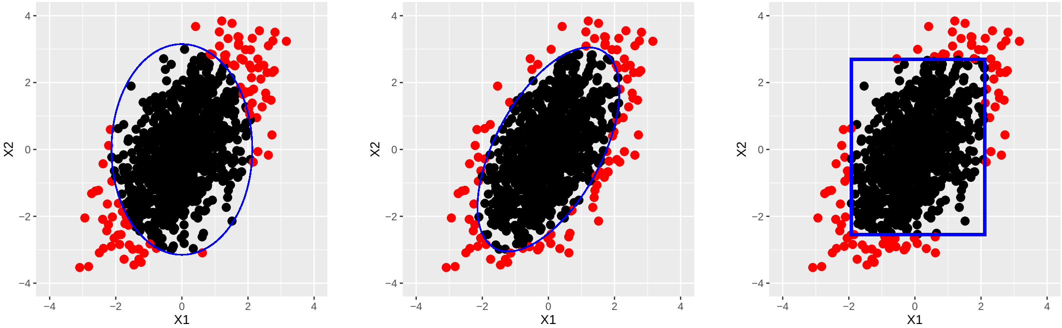
$$\tilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \sigma_1^2 & & & \mathbf{0} \\ & \sigma_2^2 & & \\ & & \ddots & \\ \mathbf{0} & & & \sigma_d^2 \end{pmatrix} \text{ and its root } \tilde{\boldsymbol{\Sigma}}^{1/2} = \begin{pmatrix} \sigma_1 & & & \mathbf{0} \\ & \sigma_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \sigma_d \end{pmatrix}.$$

**Advantages:**

- Lower computing time $\mathcal{O}(nd)$.
- $\tilde{\boldsymbol{\Sigma}}$ is easier to estimate than $\boldsymbol{\Sigma}$.

# Simplified Method

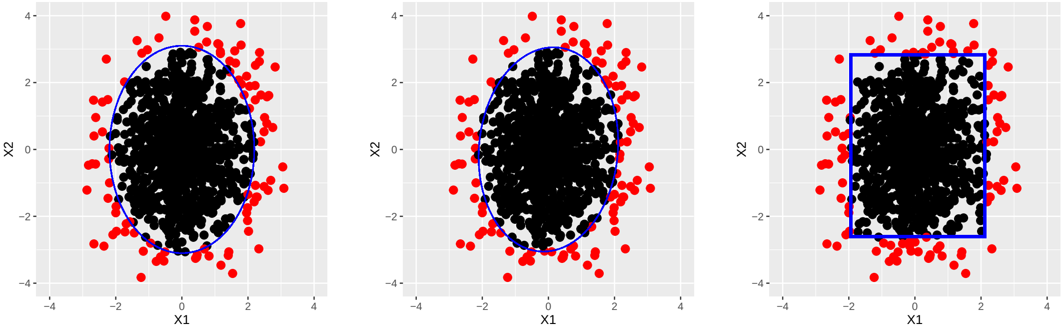
**Example:** $\boldsymbol{X}_i \sim \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\sigma_1^2 = 1$, $\sigma_2^2 = 2$ and $\rho = 0.56$.




Simplified Method (left), optimal subsampling design (middle), IBOSS (right).

# Simplified Method

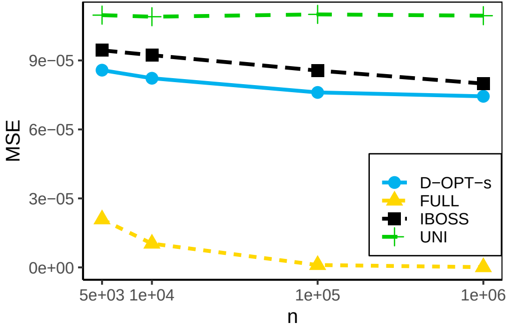
**Example:** $\boldsymbol{X}_i \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$, where $\sigma_1^2 = 1$, $\sigma_2^2 = 2$ and $\rho = 0.07$ .



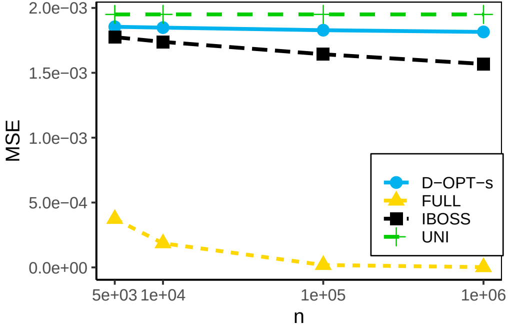

Simplified Method (left), optimal subsampling design (middle), IBOSS (right).

# Simulation Study - Simplified Method - Normal Distribution

$\boldsymbol{X}_i \sim \mathcal{N}_{50}(\mathbf{0}, \boldsymbol{\Sigma})$ with compound symmetry.




$\rho = 0.05.$




$\rho = 0.5.$

# Discussion & Outlook

**Quadratic Regression**

- Interior interval may vanish for heavy-tailed distributions.
- Subsampling design with three symmetrically placed intervals of measure $\alpha/3$ is highly efficient w.r.t. the $D$-optimal subsampling design.

**Multiple Linear Regression**

- $D$-optimal subsample by transforming data by $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$. Then select units with largest euclidean distance.
- Simplified method (only transforming w.r.t. the variances) can be a preferred alternative to IBOSS when correlations are small.

# References I

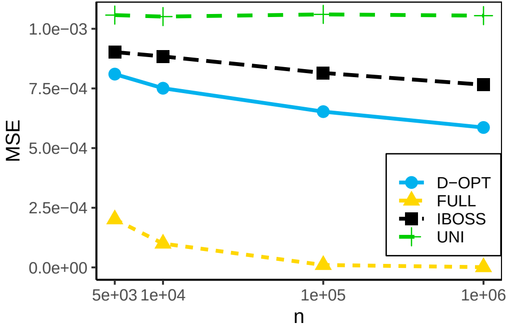

Pronzato, L. and Wang, H. (2021).
Sequential online subsampling for thinning experimental designs.
*Journal of Statistical Planning and Inference*, 212:169–193.

Sahm, M. and Schwabe, R. (2001).
A note on optimal bounded designs.
In Atkinson, A., Bogacka, B., and Zhigljavsky, A., editors, *Optimum Design 2000*, pages 131–140. Kluwer, Dordrecht, The Netherlands.

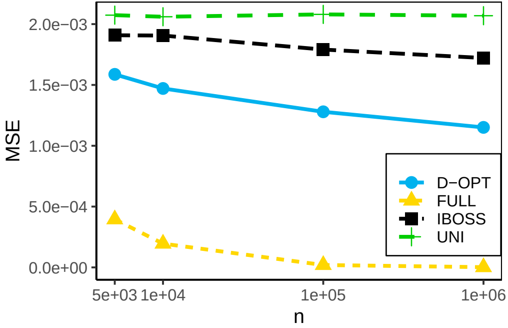
Wang, H., Yang, M., and Stufken, J. (2019).
Information-based optimal subdata selection for big data linear regression.
*Journal of the American Statistical Association*, 114(525):393–405.

# Simulation Study - Normal Distribution

$\boldsymbol{X}_i \sim \mathcal{N}_{50}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$



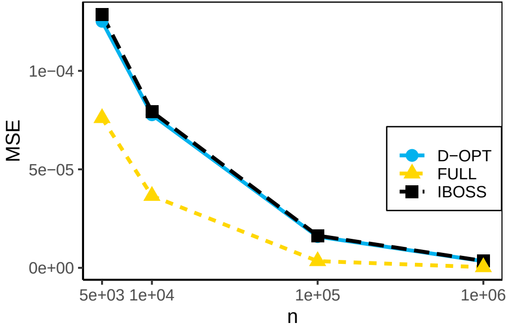

$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \mathbb{I}_{50}$ (uncorrelated).



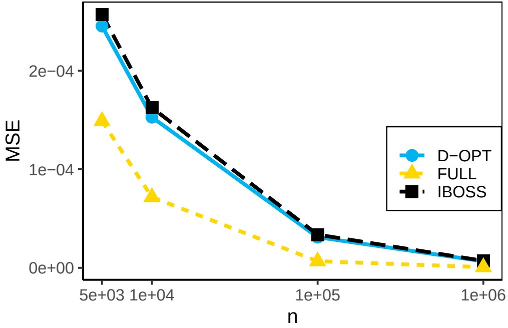

$\rho = 0.5$ (compound symmetry).

# Simulation Study - $t_3$-Distribution

$\boldsymbol{X}_i \sim t_3(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$




$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \mathbb{I}_{50}$ (uncorrelated).




$\rho = 0.5$ (compound symmetry).

# $D$-optimal Subsampling Design

Let the covariates $\boldsymbol{X}_i$ be distributed on $\mathbb{R}^d$ with density $f_{\boldsymbol{X}}(\mathbf{x})$, $\mathrm{E}(\boldsymbol{X}_i) = \boldsymbol{\mu}$ and non-singular covariance matrix $\boldsymbol{\Sigma}$, such that the distribution of $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_i - \boldsymbol{\mu})$ is rotationally invariant.

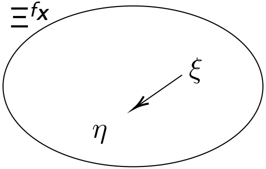
**Theorem**

*Then the density of the D-optimal subsampling design $\xi^*$ is*

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

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

# Sensitivity Function




Directional derivative $F_\Psi(\xi, \eta)$ from $\xi$ to $\eta$

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

**Sensitivity function** $\psi(x, \xi)$ from $\xi$ to a single point measure at point $x$

$$\psi(x, \xi) = (q+1) - F_D(\xi, \xi_x) = \alpha \mathbf{f}(x)^\top \mathbf{M}(\xi)^{-1} \mathbf{f}(x).$$