



**CHALMERS**  
UNIVERSITY OF TECHNOLOGY



UNIVERSITY OF GOTHENBURG

# Optimal Subsampling Designs

---

Henrik Imberg <henrik.imberg@stat-grp.se>

Joint work with Marina-Axelson Fisk and Johan Jonasson

Department of Mathematical Sciences

Chalmers University of Technology and University of Gothenburg, Sweden

Presentation at the LiU Seminar Series in Statistics and Mathematical Statistics, 10 October 2023

I: Optimal subsampling designs

II: Expected-distance-minimising designs

III: Active sampling and martingale CLT

IV: Application and experiments

# I: Optimal subsampling designs

---

Consider dataset  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$  and a parameter  $\theta_0$  defined as the solution to an estimating equation

$$\sum_{i=1}^N \psi_i(\theta) = \mathbf{0}.$$

$\psi_i(\theta)$  is a function of the parameter  $\theta$  and data vector  $(\mathbf{x}_i, \mathbf{y}_i)$ .

**Examples:** least squares regression, generalised linear models, maximum likelihood estimation, quasi-likelihood methods, M-estimation.

Assume that full data  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$  cannot be utilised (e.g., some components of  $(\mathbf{x}_i, \mathbf{y}_i)$  too expensive to observe)  $\rightarrow \theta_0$  cannot be evaluated.

This is what we call a *measurement-constrained experiment*.

Related problem in Big Data: parameter  $\theta_0$  computationally too expensive to calculate.

We assume that complete data can be observed and utilised for a subsample of size  $n \ll N$ .

Subsampling method:

- ▶ Select subsample using unequal probability sampling.
- ▶ Each instance has a unique and strictly positive probability of selection.

Estimation using sample weighting / importance weighting / inverse probability weighting:

$$\text{find } \hat{\theta} \text{ such that } \sum_i w_i \psi_i(\hat{\theta}) = \mathbf{0}.$$

$$w_i = S_i / \mu_i,$$

$S_i$  is the number of times an instance  $i$  is selected by the sampling mechanism, and

$\mu_i = \mathbb{E}[S_i]$  the corresponding expected number of selections.

Under suitable conditions, the estimator  $\hat{\theta}$  converges (at parametric rate) to a multivariate normal distribution with mean  $\theta_0$  covariance matrix

$$\Gamma(\mu; \theta_0) = \mathbf{H}(\theta_0)^{-1} \mathbf{V}(\mu; \theta_0) \mathbf{H}(\theta_0)^{-1} \quad (\text{sandwich formula})$$

as the sample size increases.

Result derived by first-order Taylor expansion around  $\theta_0$  (Binder, 1983).

- ▶ This holds under mild assumptions on the sampling mechanism (selection probabilities suitably bounded away from zero).
- ▶ No assumptions on the data distribution (only mild moment conditions).

The statistical properties of the estimator  $\hat{\theta}$  are determined by choice of sampling design and mean inclusion vector (sampling scheme)  $\mu = (\mu_1, \dots, \mu_N)$ .

Choose

$$\mu^* = \arg \min_{\mu} \Phi(\Gamma(\mu; \theta_0))$$

for some suitable function  $\Phi$ . Optimisation over all (feasible) values  $\mu$  such that  $\sum_{i=1}^N \mu_i = n$ , where  $n$  is the desired sample size.

**Examples:**

Optimality criterion	Description	Objective function $\Phi(\Gamma)$
A-optimality	Minimise average variance	$\text{tr}(\Gamma)$
D-optimality	Minimise generalised variance (determinant)	$\det(\Gamma)$
E-optimality	Minimise largest eigenvalue	$\lambda_{\max}(\Gamma)$
L-optimality	Minimise average variance of lin. comb. $\mathbf{L}^T \hat{\theta}$ .	$\text{tr}(\Gamma \mathbf{L} \mathbf{L}^T)$

Assume that solutions with  $\mu_i > 1$  are allowed (i.e., sampling is with replacement).

The constrained stationary points of  $\Phi(\Gamma(\mu; \theta_0))$  are obtained as the critical points of the Lagrangian

$$\Lambda(\mu, \lambda) = \Phi(\Gamma(\mu; \theta_0)) + \lambda g(\mu), \quad g(\mu) = \sum_{i \in \mathcal{D}} \mu_i - n.$$

Taking the derivatives with respect to  $\mu$  and  $\lambda$ , we obtain the system of equations

$$\nabla \Lambda(\mu, \lambda) = \mathbf{0} \quad \Leftrightarrow \quad \begin{cases} g(\mu) = 0 & (\text{feasibility}) \\ -\nabla_{\mu} \Phi(\Gamma(\mu; \theta_0)) = \lambda \nabla g(\mu) & (\text{stationarity}). \end{cases}$$

If sampling is without replacement we have  $N$  additional inequality constraints  $0 < \mu_i \leq 1$ . Lagrangian slightly more involved. Solution must satisfy the Karush-Kuhn-Tucker conditions.

Need the derivatives of  $\Phi(\Gamma(\mu; \theta_0))$  with respect to  $\mu$ .



## Lemma 1

Whenever it exists, the partial derivative of  $\Phi(\mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0))$  with respect to  $\mu_i$  is given by

$$\frac{\partial \Phi(\mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0))}{\partial \mu_i} = \text{tr} \left( \boldsymbol{\phi}(\mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0))^\top \frac{\partial \mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)}{\partial \mu_i} \right),$$

where  $\boldsymbol{\phi}(\mathbf{U}) = \frac{\partial \Phi(\mathbf{U})}{\partial \mathbf{U}}$  is the  $p \times p$  matrix derivative of  $\Phi$  with respect to its matrix argument, and  $\frac{\partial \mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)}{\partial \mu_i}$  is the elementwise derivative of  $\mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$  with respect to  $\mu_i$ .

This is the chain rule in matrix differential calculus (see, e.g., Petersen and Pedersen, 2012).

## Lemma 2

Assume that

1.  $\Gamma(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$  decreases monotonically with  $\mu_1, \dots, \mu_N$  in the Loewner order sense, i.e.,  $\Gamma(\boldsymbol{\mu}_1; \boldsymbol{\theta}_0) - \Gamma(\boldsymbol{\mu}_2; \boldsymbol{\theta}_0)$  is positive semi-definite (PSD) for every pair of vectors  $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}_{>0}^N$  such that  $\mu_1 \leq \mu_2$  (elementwise), and
2.  $\Phi$  is monotone for Loewner's ordering, i.e., that

$\Phi(\mathbf{U}) - \Phi(\mathbf{V})$  is PSD for all PSD matrices  $\mathbf{U}, \mathbf{V}$  such that  $\mathbf{U} - \mathbf{V}$  is PSD.

Then the derivative matrix  $\phi(\Gamma(\boldsymbol{\mu}; \boldsymbol{\theta}_0))$  is PSD and there exists a real matrix  $\mathbf{L}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$  such that  $\mathbf{L}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)\mathbf{L}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)^\top = \phi(\Gamma(\boldsymbol{\mu}; \boldsymbol{\theta}_0))$ .

1. D-optimality criterion:  $\log \det(\mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0))$  is differentiable with respect to  $\boldsymbol{\mu}$  and  $\phi(\mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)) = \mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)^{-1}$ , provided that  $\mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$  is of full rank.
2. L-optimality criterion:  $\text{tr}(\mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)\mathbf{L}\mathbf{L}^T)$  is differentiable with respect to  $\boldsymbol{\mu}$ , and  $\phi(\mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)) = \mathbf{L}\mathbf{L}^T$ .
3.  $\mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$  is differentiable with respect to  $\boldsymbol{\mu}$  and  $\frac{\partial \mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)}{\partial \mu_i} = -\mu_i^{-2} \mathbf{H}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{\psi}_i(\boldsymbol{\theta}_0) \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^T \mathbf{H}(\boldsymbol{\theta}_0)^{-1}$ , provided that  $\mu_i > 0$  for all  $i$ .
4. With  $\mathbf{L}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$  as on the previous slide,

$$\frac{\partial \Phi(\mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0))}{\partial \mu_i} = -\mu_i^{-2} c_i^2, \quad c_i = \|\mathbf{L}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)^T \mathbf{H}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\|_2^2$$

whenever the derivative exists.

See, e.g., Petersen and Pedersen (2012).

Assume that

- ▶ the derivative matrix  $\phi(\mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0))$  is constant with respect to  $\boldsymbol{\mu}$  (this is true for linear optimality criteria),
- ▶ sampling is with replacement (e.g., according to a Multinomial sampling design).

Then the optimal sampling scheme for a given size  $n$  is obtained as

$$\mu_i^* = n \frac{\sqrt{c_i}}{\sum_{j=1}^N \sqrt{c_j}} \quad i = 1, \dots, N.$$

For sampling with replacement: may need simple correction to ensure that  $\mu_i^* \leq 1$ .

For non-linear optimality criteria: apply iterative procedure by local approximation as a linear optimality criterion until convergence to a fixed-point (local optimum).

## II: Expected-distance-minimising designs

---

Desirable properties of an optimal subsampling design:

- ▶ Tractability: should be computationally and analytically tractable.
- ▶ Invariance: should not depend on the choice of parameterisation, nor on the measurement-scale of the data or coding of the data prior to modelling.
- ▶ Appropriateness: should address the scientific question of interest and/or be an appropriate measure for the overall performance of the estimator.

	A-optimality	D-optimality	E-optimality	L-optimality	???
Tractable?	✓	✗	✗	✓	✓
Invariant?	✗	✓	✗	?	✓

D-optimality is often considered the gold standard in traditional experimental design problems.

We evaluate the performance of optimal subsampling under the D-optimality criterion for a quasi-binomial logistic regression model with  $\sim 200$  parameters and  $\sim 40,000$  observations.

- ▶ Fitting the model to the full dataset takes 8.46 sec on a desktop computer.
- ▶ Computation time for finding a D-optimal subsampling design is 27.69 sec. 3-fold increase!
- ▶ Finding an A-optimal subsampling design takes only 1.16 sec.

Can we find a class of optimality criteria with

- ▶ Computational complexity like the A-optimality criterion,
- ▶ Performance and invariance properties like the D-optimality criterion?

Aim of data subsampling: obtain estimate  $\hat{\theta}$  for unknown full-data parameter  $\theta_0$ .

Natural target for optimal design: minimise the expected distance  $E[d(\hat{\theta})]$  of  $\hat{\theta}$  from  $\theta_0$ .

**Examples:** Kullback-Leibler divergence, empirical risk distance (deviance), Mahalanobis distance.

By linearisation, minimising the expected distance corresponds a linear optimality criterion.

- ▶ Convex optimisation problem with a simple, closed-form solution.

For the statistical distance functions above, the optimal sampling scheme is also **invariant**:

- ▶ Does not depend on the choice of parameterisation, nor on the measurement-scale of the data or coding of the data prior to modelling.

In experiments: performance like D-optimality with 10–20 times lower computational demand.



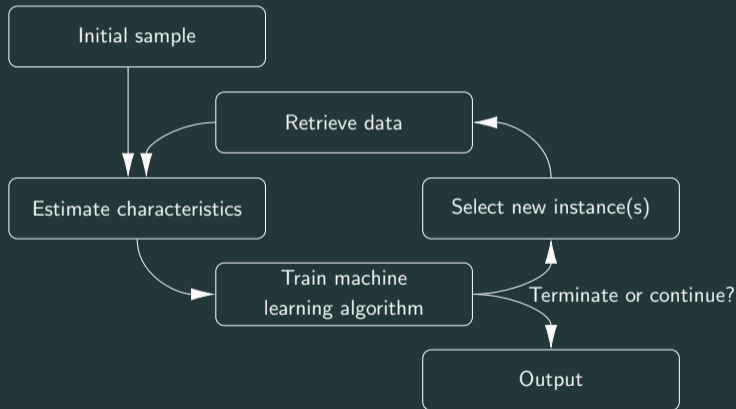
## III: Active sampling and martingale CLT

---

**Problem:** optimal design depends on unknown full-data characteristics  $(\mathbf{x}_i, \mathbf{y}_i, \theta_0)$ .

**Solution:**

**Active Sampling**



Asymptotic properties (consistency, asymptotic normality, variance estimators) are established using martingale central limit theory.

**Theorem 3 (Brown, 1971)**

Consider a sequence  $\{X_j\}_{j=1}^{\infty}$  of random variables such that  $E[X_j] = E[X_j|X_1, \dots, X_{j-1}] = 0$  and  $E[X_j^2] < \infty$  (i.e.,  $\{X_j\}$  is a zero-mean martingale with finite variance).

Let  $\sigma_j^2 = E[X_j^2|X_1, \dots, X_{j-1}]$ ,  $U_k = \sum_{j=1}^k X_j$ ,  $V_k^2 = \sum_{j=1}^k \sigma_j^2$ , and  $u_k^2 = E[U_k^2] = E[V_k^2]$ .

Assume that

1.  $V_k^2 u_k^{-2} \xrightarrow{P} 1$  as  $k \rightarrow \infty$   
(ratio of total conditional variance to total variance converges in probability to 1),
2. The Lindeberg-Feller condition holds:

$$u_k^{-2} \sum_{j=1}^k E[X_j^2 I(|X_j| > \varepsilon u_k)] \rightarrow 0 \quad \text{as } k \rightarrow \infty \quad \text{for all } \varepsilon > 0.$$

(uniform asymptotic negligibility).

Then  $U_k/u_k \xrightarrow{d} \mathcal{N}(0, 1)$  as  $k \rightarrow \infty$ .

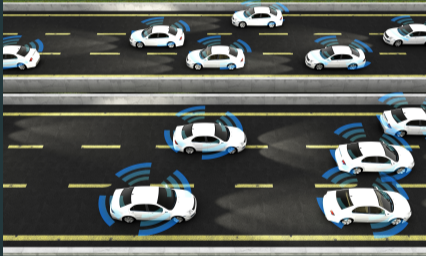
In our case:

- ▶ Construct sequence of conditionally unbiased estimators (martingale).
- ▶ Assume total variance tends to infinity (variance in each sampling stage not allowed to approach zero too fast).
- ▶ Assume conditional variance approaches total variance (correlation between sampling stages is asymptotically negligible).
- ▶ Show that Lindeberg-Feller condition holds if  $S_i/\mu_i$  have uniformly bounded second moments (sampling probabilities properly bounded away from zero).

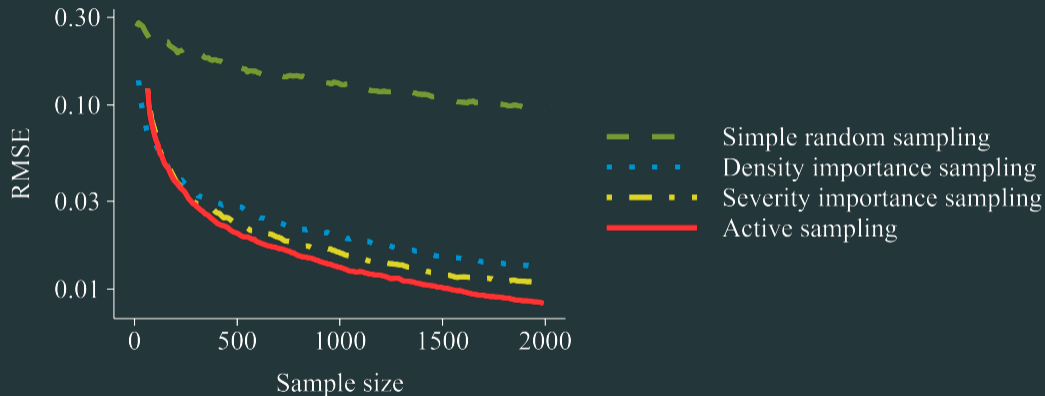
Variance estimation using martingale estimator, pooling of classical survey sampling variance estimators, or bootstrap. Requires some additional moment conditions to ensure consistency.

## **IV: Application and experiments**

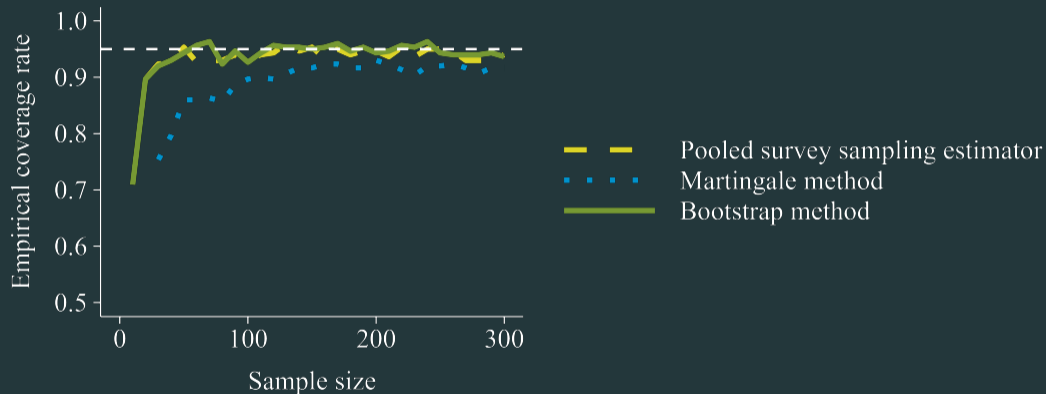
---



- ▶ Safety benefit evaluation of an advanced driver assistance system (e.g., an automatic emergency braking system) using virtual simulations.
- ▶ Computer simulation tool takes input  $x_i$  (e.g., vehicle kinematics, driver behaviour) and returns output  $y_i$  (e.g., collision impact speed).
- ▶ Want to understand characteristics of  $y_i$  (given  $x_i$ ) for a large number of experiments.
- ▶ Observing  $y_i$  comes with high computational cost.
- ▶ Can we reduce computational load through subsampling?



Sample size reductions of 7–48% compared to traditional importance sampling methods.



Confidence interval coverage rates approach the nominal 95% confidence level quickly.



- ▶ Subsampling methods have seen a huge increase in popularity over the past few years. Our work contributes to this development by presenting a framework for optimal design in general subsampling problems using active sampling and sequential optimal design.
- ▶ Asymptotic properties are established using a martingale central limit theorem, assuming that the number of sampling stages tends to infinity. Properties when the number of sampling stages is fixed requires further attention.
- ▶ The method is limited to regular asymptotically linear estimators. It would be interesting to study optimal subsampling methods in high-dimensional and non-parametric settings, for instance for kernel generalised linear models.

## Supervisors

Marina Axelson-Fisk and Johan Jonasson

Department of Mathematical Sciences, Chalmers University of Technology

## Collaborators

Xiaomi Yang and Jonas Bärgrman

Division of Vehicle Safety, Chalmers University of Technology


Carol Flannagan


University of Michigan Transportation Research Institute


## Industry partners

Malin Svärd and Simon Lundell at Volvo Car Corporation for sharing data.



 D. A. Binder (1983).  
**On the Variances of Asymptotically Normal Estimators from Complex Surveys.**  
*International Statistical Review*, 51:279–292.

 B. M. Brown (1971).  
**Martingale Central Limit Theorems.**  
*The Annals of Mathematical Statistics*, 42:59–66.

 K. B. Petersen and M. S. Pedersen (2012).  
**The Matrix Cookbook.**  
Technical University of Denmark.



H. Imberg (2023).

**Optimal Subsampling Designs Under Measurement Constraints.**

PhD thesis, Chalmers University of Technology, Gothenburg.



H. Imberg, M. Axelson-Fisk, and J. Jonasson (2023).

**Optimal subsampling designs.**

arXiv:2304.03019 [math.ST].



H. Imberg, X. Yang, C. Flannagan, and J. Bärnman (2022).

**Active sampling: A machine-learning-assisted framework for finite population inference with optimal subsamples.**

arXiv:2212.10024 [stat.ME].



H. Imberg, J. Jonasson, and M. Axelson-Fisk (2020).

**Optimal sampling in unbiased active learning.**

Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research, 108:559–569.

# Thank you for listening!



UNIVERSITY OF  
GOTHENBURG

---



**CHALMERS**  
UNIVERSITY OF TECHNOLOGY