Optimal design of experiments and its potential application to high-dimensional data

Kirsten Schorning

Mathematical Statistics
TU Dortmund University

June 28, 2020

Motivating Example

Situation:

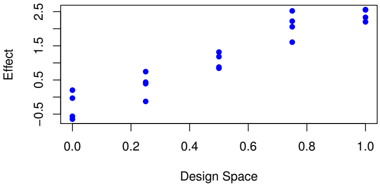
- We want to describe the relationship between observations Y_{ij} and measuring points x_i by a linear regression model:

$$Y_{ij} = \theta_1 + \theta_2 x_i + \varepsilon_{ij} \quad i = 1, \dots, k; j = 1, \dots, n_i$$

- We are able to define the positions of measuring points $x_1, \dots, x_k$ in advance.

Example:

$n = 20$ observations are taken in $\mathcal{X} = [0, 1]$ with $n_i = 4$ observations at $k = 5$ different points $x_1 = 0$, $x_2 = 0.25$, $x_3 = 0.5, x_4 = 0.75$, $x_5 = 1$



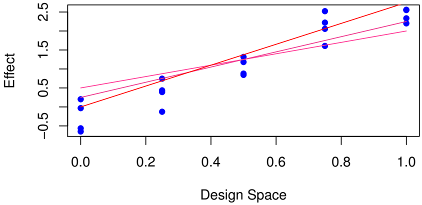
Question: Can the quality of the estimation be improved by choosing the measuring points in $\mathcal{X}$ appropriately?

The classical approach in optimal experimental design

We assume:

$$Y_{ij} = \eta(x_i, \theta) + \varepsilon_{ij} \ ; \quad i = 1, \ldots, k; \ j = 1, \ldots, n_i$$

- η is a regression function
- $x_i \in \mathcal{X} \subset \mathbb{R}^d$, $\mathcal{X}$ design space
- $\theta \in \mathbb{R}^p$ is unknown
- Today: ε_{ij} independent $\sim \mathcal{N}(0, \sigma^2)$



Interested in: Estimation of the parameter θ

Goal: Select the points $x_1, \ldots, x_k$ and $n_1, \ldots, n_k$ such that the estimator, $\hat{\theta}$, of the parameter is most precise.

Minimizing the Covariance of the Estimator $\hat{\theta}$

Assume a linear model:

$$Y_{ij} = f^T(x_i)\theta + \varepsilon_{ij} \ ; \quad i = 1, \dots, k;\ j = 1, \dots, n_i$$

Then the covariance matrix of the least-squares estimator $\hat{\theta}$ is given by:

$$\mathrm{Cov}(\hat{\theta}) = \left(\sum_{i=1}^{k} n_i f(x_i) f^T(x_i) \right)^{-1} \in \mathbb{R}^{p \times p}$$

Goal: Select the points $x_1, \dots, x_k$ and $n_1, \dots, n_k$ such that $\mathrm{Cov}(\hat{\theta})$ is small.

Approach: Minimize real-valued, convex functions of $\mathrm{Cov}(\hat{\theta})$ with respect to $x_1, \dots, x_k$ and $n_1, \dots, n_k$.

Commonly used criteria for minimization

$$\mathrm{Cov}(\hat{\theta}) = \left(\sum_{i=1}^{k} n_i f(x_i) f^T(x_i) \right)^{-1} \in \mathbb{R}^{p \times p}$$

D-optimality criterion: $\Phi_D(x_1, \ldots, x_k, n_1, \ldots, n_k) = \det\left(\mathrm{Cov}(\hat{\theta})\right)$

A-optimality criterion: $\Phi_A((x_1, \ldots, x_k, n_1, \ldots, n_k) = \mathrm{tr}\left(A * \mathrm{Cov}(\hat{\theta})\right)$

Popular *D*-optimal designs on $[0, 1]$:

Linear regression $f(x) = (1, x)^T$ $x_1 = 0, x_2 = 1$ $n_1 = \frac{n}{2}, n_2 = \frac{n}{2}$

- $\Phi_D(x_1 = 0, x_2 = 1, n_1 = n_2 = n/2) = 4/n$
- $\Phi_D(x_1 = 0, x_2 = 0.25, x_3 = 0.5, x_4 = 0.75, x_5 = 1, n_1 = \ldots = n_5 = n/5) = 8/n$

Optimal design of **BIG DATA** experiments

Situation: Huge data set is available.

Problem: The calculation of the LSE $\hat{\theta}$ based on the whole data set takes too much time.

Target: Efficient selection of an optimal subsample which results in a precise estimation in an acceptable amount of time.

Optimal design of **BIG DATA** experiments

Current approaches:

- **Wang et al. (2018)** derive subsamples for linear regression using the D-optimality criterion.
- **Wang (2019)** derives optimal subsamples for logistic regression using the A-optimality criterion and a newly developed estimator.

Optimal design of **high-dimensional** experiments

Situation: The data or parameter θ is high-dimensional with $d, p >> n$.

Problem: The classical least-square estimator is not feasible.
Other estimators (LASSO) and sparsity arguments have to be used.

Target: Efficient selection of an optimal sample which results in a precise estimation of the high-dimensional model.

Optimal design of **high-dimensional** experiments

Current approaches:

- **Hu and Lu (2019)**: Derive asymptotics and optimal designs of LASSO for sparse linear regression.
- **Candès and Sur (2020); Sur and Candès (2019)**: Derive the asymptotic bias and variance of the maximum-likelihood-estimator in high-dimensional logistic regression.

Conclusion and Outlook

- In principle, **optimal design of experiments** can be used whenever the experimenter can influence the positions of the measuring points.
- **Optimal design of experiments** can improve the quality of such experiments substantially.
- Methods of **optimal design of experiments** might also be applicable to the setting of big data and high-dimension.
- Do you already have an improvable experiment in mind?

Thank you very much for your attention!

References

Candès, E. J. and Sur, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Annals of Statistics*, 48(1).

Hu, H. and Lu, Y. M. (2019). Asymptotics and optimal designs of slope for sparse linear regression. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 375–379.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Wang, H. (2019). More efficient estimation for logistic regression with optimal subsamples. *Journal of Machine Learning Research*, 20(132):1–59.

Wang, H., Yang, M., and Stufken, J. (2018). Information-Based Optimal Subdata Selection for Big Data Linear Regression. *Journal of the American Statistical Association*, pages 1–13.