A short note on the experimental design

Yen-Chi Chen
University of Washington
May 14, 2020

Experimental design is a classical problem in statistics. Consider a simple linear regression where we want to investigate the linear relationship between a covariate $X \in \mathbb{R}^p$ and a response $Y \in \mathbb{R}$. In a linear model, this relationship is often expressed as the model

$$\mathbb{E}(Y|X) = \beta^T X.$$

The design occurs in a situation where we get to choose the covariate X (or its distribution). For simplicity, we assume that

$$Y_i = \beta^T X_i + \varepsilon_i,$$

where $\varepsilon_1, \cdots, \varepsilon_n$ are IID normal with variance σ^2.

This note is based on Chapter 3.7 of the following book

> Design of Experiments for Generalized Linear Models, Kenneth Russell, CRC Press, 2019

and the following classical review paper

> Chaloner, K., & Verdinelli, I. (1995). Bayesian Experimental Design: A Review. Statistical Science, 10(3), 273-304.

1 Frequentist experimental design

In practice, we choose the covariates $X_1, \cdots, X_n$ and then observe the corresponding response variables $Y_1, \cdots, Y_n$. Then we attempt to study the linear relationship between X and Y using this data. Namely, we want to estimate the slope β.

A regular approach (Frequentist) is to apply a least square estimate, which leads to

$$\widehat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y},$$

where $\mathbb{X} \in \mathbb{R}^{n\times(p+1)}$ is the design matrix (matrix of the covariate) that is constructed from $X_1, \cdots, X_n$ and $\mathbb{Y} = (Y_1, \cdots, Y_n)^T$ is the response vector. Under regularity conditions (linear model is correct and noise is homogenous), the variance (covariance) of $\widehat{\beta}$ is

$$\sigma^2 \cdot (\mathbb{X}^T\mathbb{X})^{-1} = \frac{\sigma^2}{n} \cdot \Sigma_X^{-1},$$

where $\sigma^2 = \mathsf{Var}(Y|X = x)$ is the noise level and $\Sigma_X = \frac{1}{n}\mathbb{X}^T\mathbb{X}$ is the design. Thus, in the Frequentist perspective, we want to choose the design $X_1, \cdots, X_n$ such that Σ_X is as large as possible.

In summary, a design problem consists of the following elements:

- A response variable Y or a response vector $\mathbb{Y}$.
- A design of covariates X or $\mathbb{X}$ or Σ_X that we can choose.
- A parameter of interest β; in the linear regression, β is the slope.
- An estimator $\widehat{\beta} = \eta(\mathbb{Y}, \mathbb{X})$.
- A measure of success $V(\Sigma_X) \in \mathbb{R}$ or $V(\mathbb{X}) \in \mathbb{R}$ that marginalizes out the sampling variability of $\widehat{\beta}$ so it depends only on Σ_X; in the above example, it is often some function of the variance of the estimator $\mathsf{Var}(\widehat{\beta})^{-1}$.

Here are some popular examples of the measure of success; they are often called the *alphabetical optimality*.

- **D-optimality.** The D-optimality corresponds to the volume of a confidence ellipse with minimal volume. So the measure of success is

$$V_D(\Sigma_X) = \mathsf{det}(\Sigma_X^{-1}).$$

  So we want to choose the design to maximize the determinant of the inverse design matrix Σ_X. Note that there is a D_s-optimality that refers to using the the determinant of a 'subset' of Σ_X^{-1}.

- **A-optimality.** The A-optimality is the average coordinate-wise variance of $\widehat{\beta}$. So it corresponds to

$$V_A(\Sigma_X) = \mathsf{Tr}(\Sigma_X^{-1}).$$

- **C-optimality.** The C-optimality corresponds to the prediction variance at a particular location $c \in \mathbb{R}^p$, i.e., we want to minimize $\mathsf{Var}(\widehat{\beta}^T c)$. This corresponds to

$$V_C(\Sigma_X) = c^T \Sigma_X^{-1} c.$$

- **E-optimality.** The E-optimality is the maximal variance under every direction. Namely, we want to minimize the variance $\sup_{c:\|c\|=1} \mathsf{Var}(c^T \widehat{\beta})$. Using linear algebra, one can show that this is essentially

$$V_E(\Sigma_X) = \lambda_{\max}(\Sigma_X^{-1}),$$

  the maximal eigenvalue of Σ_X^{-1}.

- **G-optimality.** Let $\mathcal{X}$ be the support of X. The G-optimality is to minimize the maximal predictive variance $\sup_{x \in \mathcal{X}} \mathsf{Var}(x^T \widehat{\beta})$. It is similar to the E-optimality but we focus on the support $\mathcal{X}$. It corresponds to

$$V_G(\Sigma_X) = \sup_{x \in \mathcal{X}} x^T \Sigma_X^{-1} x.$$

2 Bayesian experimental design via the posterior

Now we turn to the linear regression problem under the Bayesian setting. For simplicity, we assume that

$$Y_i = \beta^T X_i + \varepsilon_i,$$

where $\varepsilon_1, \cdots, \varepsilon_n$ are IID normal with variance σ^2. For the prior, we assume that $\beta \sim N(\beta_0, \sigma^2 \Omega^{-1})$. Under these assumptions, the posterior of β will be

$$\beta | \mathbb{Y}, \mathbb{X} \sim N\left((n\Sigma_X + \Omega)^{-1} (\mathbb{X}^T \mathbb{Y} + \Omega\beta_0), \sigma^2 (n\Sigma_X + \Omega)^{-1} \right).$$

The precision matrix (inverse of covariance matrix) Σ_X^{-1} in the Frequentist setting plays the key role in the optimality. Because the posterior distribution of β has a precision matrix proportional to $(\Sigma_X + \frac{1}{n}\Omega)^{-1}$, we can directly construct Bayesian alphabetical optimality by replacing Σ_X^{-1} with $(\Sigma_X + \frac{1}{n}\Omega)^{-1}$, which leads to the following Bayesian alphabetical optimality:

- **D-optimality.**
$$V_D(\Sigma_X) = \mathsf{det}\left(\left(\Sigma_X + \frac{1}{n}\Omega \right)^{-1} \right).$$
- **A-optimality.**
$$V_A(\Sigma_X) = \mathsf{Tr}\left(\left(\Sigma_X + \frac{1}{n}\Omega \right)^{-1} \right).$$
- **C-optimality.**
$$V_C(\Sigma_X) = c^T \left(\Sigma_X + \frac{1}{n}\Omega \right)^{-1} c.$$
- **E-optimality.**
$$V_E(\Sigma_X) = \lambda_{\max}\left(\left(\Sigma_X + \frac{1}{n}\Omega \right)^{-1} \right).$$
- **G-optimality.**
$$V_G(\Sigma_X) = \sup_{x \in X} x^T \left(\Sigma_X + \frac{1}{n}\Omega \right)^{-1} x.$$

3 Bayesian experimental design via a utility function

Here we introduce an alternative way to construct optimality in a Bayesian setting. We first introduce a utility function

$$U_0(d, \mathbb{Y}, \mathbb{X}, \beta),$$

where $d \in \mathcal{D}$ is the 'decision' that we are making. We *want to choose the design* $\mathbb{X}$ *by maximizing the utility after properly handling other inputs* $(d, \mathbb{Y}, \beta)$.

β is unobserved but given what we observed $\mathbb{Y}$ and $\mathbb{X}$, it follows from the posterior $\pi(\beta|\mathbb{X},\mathbb{Y})$. So a common way to 'marginalize' out the effect of an unobserved β is to integrate it over the underlying conditional density. Namely,

$$U_1(d,\mathbb{Y},\mathbb{X}) = \int U_0(d,\mathbb{Y},\mathbb{X},\beta)\pi(\beta|\mathbb{Y},\mathbb{X})d\beta.$$

Then the optimal decision given $\mathbb{Y}$ and $\mathbb{X}$, is

$$d^*(\mathbb{Y},\mathbb{X}) = \mathsf{argmax}_{d\in\mathcal{D}}U_1(d,\mathbb{Y},\mathbb{X}) = \mathsf{argmax}_{d\in\mathcal{D}}\int U_0(d,\mathbb{Y},\mathbb{X},\beta)\pi(\beta|\mathbb{Y},\mathbb{X})d\beta.$$

In estimation problem, the decision is often the estimator and the optimal decision often corresponds to some natural estimator.

Example 1: posterior mean. Suppose we choose the utility function to be negative squared L_2 distance

$$U_0(d,\mathbb{Y},\mathbb{X},\beta) = -\|d-\beta\|_2^2.$$

Then the optimal decision $d^*(\mathbb{Y},\mathbb{X})$ will be the posterior mean.

Example 2: MAP (Maximum a posteriori estimation; posterior mode). If we choose

$$U_0(d,\mathbb{Y},\mathbb{X},\beta) = I(d=\beta),$$

then the optimal decision $d^*(\mathbb{Y},\mathbb{X})$ will be the MAP.

Example 3: posterior median. In the case of the choice negative L_1 norm:

$$U_0(d,\mathbb{Y},\mathbb{X},\beta) = -\|d-\beta\|_1,$$

the optimal decision $d^*(\mathbb{Y},\mathbb{X})$ will be the posterior median.

With the optimal decision $d^*(\mathbb{Y},\mathbb{X})$, we can further express the utility in terms of $\mathbb{Y}$ and $\mathbb{X}$:

$$U_2(\mathbb{Y},\mathbb{X}) = U_1(d^*(\mathbb{Y},\mathbb{X}),\mathbb{Y},\mathbb{X}).$$

The final objective function will be the utility after adjusting for $\mathbb{Y}$, i.e.,

$$U_3(\mathbb{X}) = \int U_2(\mathbb{Y},\mathbb{X})p(\mathbb{Y}|\mathbb{X})d\mathbb{Y}.$$

This quantity, $U_3(\mathbb{X})$, is the Bayesian version of the measure of success $V(\mathbb{X})$ in the Frequentist setting. So the Bayesian design problem can be written as finding $\mathbb{X}$ or Σ_X such that $U_3(\mathbb{X})$ is maximized. Namely, the optimal design is

$$\begin{aligned}\Sigma_X^* &= \mathsf{argmax}_{\mathbb{X}}U_3(\mathbb{X}) \\ &= \mathsf{argmax}_{\mathbb{X}}\int \max_{d\in\mathcal{D}}\int U_0(d,\mathbb{Y},\mathbb{X},\beta)\pi(\beta|\mathbb{Y},\mathbb{X})p(\mathbb{Y}|\mathbb{X})d\beta d\mathbb{Y}.\end{aligned}$$

3.1 Bayesian alphabetical optimality as utility function

We come back to the linear regression and show that some alphabetical optimality can be written in terms of utility functions. Again, we assume that

$$Y_i = \beta^T X_i + \varepsilon_i,$$

where $\varepsilon_1, \cdots, \varepsilon_n$ are IID normal with variance σ^2. For the prior, we assume that $\beta \sim N(\beta_0, \sigma^2\Omega^{-1})$. Under these assumptions, the posterior of β will be

$$\beta|\mathbb{Y},\mathbb{X} \sim N\left((n\Sigma_X + \Omega)^{-1}(\mathbb{X}^T\mathbb{Y} + \Omega\beta_0), \sigma^2(n\Sigma_X + \Omega)^{-1}\right).$$

3.1.1 D-optimality

A popular utility function is to formulate the problem as *the expected information gain* from the prior to the posterior, which corresponds to the utility function

$$U_0(d, \mathbb{Y}, \mathbb{X}, \beta) = \log\left(\frac{\pi(\beta|\mathbb{Y},\mathbb{X})}{\pi(\beta)}\right).$$

Note that in this case, the decision is not important so we can ignore it. This leads to

$$\begin{aligned} U_2(\mathbb{Y},\mathbb{X}) &= \int \log\left(\frac{\pi(\beta|\mathbb{Y},\mathbb{X})}{\pi(\beta)}\right)\pi(\beta|\mathbb{Y},\mathbb{X})d\beta \\ &= \mathsf{KL}(\pi(\cdot|\mathbb{Y},\mathbb{X})||\pi(\cdot)), \end{aligned}$$

which is the Kullback-Leibler (KL) divergence between the posterior and the prior.

The final objective function will then be

$$\begin{aligned} U_3(\mathbb{X}) &= \int \mathsf{KL}(\pi(\cdot|\mathbb{Y},\mathbb{X})||\pi(\cdot)))p(\mathbb{Y}|\mathbb{X})d\beta d\mathbb{Y} \\ &= \mathbb{E}\left(\mathsf{KL}(\pi(\cdot|\mathbb{Y},\mathbb{X})||\pi(\cdot))|\mathbb{X}\right), \end{aligned}$$

which is the expected KL divergence between the posterior and prior.

Interestingly, the prior distribution $\pi(\cdot)$ in $U_2(\mathbb{Y},\mathbb{X})$ is independent of $\mathbb{X}$. Thus, maximizing $U_3(\mathbb{X})$ is equivalent to maximizing

$$U_4(\mathbb{X}) = \int \left(\log\pi(\beta|\mathbb{Y},\mathbb{X})\right)\pi(\beta|\mathbb{Y},\mathbb{X})p(\mathbb{Y}|\mathbb{X})d\beta d\mathbb{Y}.$$

Now plugging the normal model into the posterior distribution, we obtain a simple form

$$U_4(\mathbb{X}) = -\frac{p}{2}\log(2\pi) - \frac{p}{2} + \frac{1}{2}\log\det(\sigma^2(n\Sigma_X + \Omega)^{-1}).$$

Thus, the optimal choice of $\mathbb{X}$ is to maximize $\det(\sigma^2(n\Sigma_X + \Omega)^{-1})$, i.e.,

$$\mathbb{X}_D^* = \mathsf{argmax}_{\mathbb{X}}\det(\sigma^2(n\Sigma_X + \Omega)^{-1}),$$

which recovers the Bayesian D-optimality. Note that there are other utility functions of obtaining the D-optimality design.

3.1.2 A-optimality

Now consider the utility function $U_0(d, \mathbb{Y}, \mathbb{X}, \beta) = \|d - \beta\|_2^2$. From example 1, we know that the optimal $d^*(\mathbb{X}, \mathbb{Y})$ will be the posterior mean, i.e.,

$$d^*(\mathbb{X}, \mathbb{Y}) = \widehat{\beta}_{pm} = (n\Sigma_X + \Omega)^{-1}(\mathbb{X}^T\mathbb{Y} + \Omega\beta_0).$$

Thus,

$$U_2(\mathbb{Y}, \mathbb{X}) = \int \|\widehat{\beta}_{pm} - \beta\|_2^2 \pi(\beta|\mathbb{Y}, \mathbb{X}) d\beta = \mathsf{Tr}(\sigma^2(n\Sigma_X + \Omega)^{-1}),$$

which is independent of $\mathbb{Y}$ so we immediately obtain

$$U_3(\mathbb{X}) = \mathsf{Tr}(\sigma^2(n\Sigma_X + \Omega)^{-1}),$$

the same result as the Bayesian A-optimality.

3.1.3 C-optimality

The C-optimality can also be derived using the utility function. Let $c \in \mathbb{R}^p$ and consider

$$U_0(d, \mathbb{Y}, \mathbb{X}, \beta) = |c^T(d - \beta)|^2.$$

The optimal decision d^* will again be the posterior mean, i.e. $d^*(\mathbb{X}, \mathbb{Y}) = \widehat{\beta}_{pm}$. As a result,

$$U_2(\mathbb{Y}, \mathbb{X}) = \int |c^T(\widehat{\beta}_{pm} - \beta)|^2 \pi(\beta|\mathbb{Y}, \mathbb{X}) d\beta = \sigma^2 c^T(n\Sigma_X + \Omega)^{-1})c.$$

Again, it is independent of $\mathbb{Y}$ so

$$U_3(\mathbb{X}) = \sigma^2 c^T(n\Sigma_X + \Omega)^{-1})c,$$

which recovers the Bayesian C-optimality.

Remark on E-optimality and G-optimality. Although many alphabetical optimality can be written as utility function form, it is still unclear how to write the E-optimality and G-optimality as a utility function problem. A major issue is that the two optimality involves the supreme operator of the posterior covariance matrix. It is unclear how do we adjust the utility function so that this supreme operator can be included.