

A Review On Optimal Experimental Design

Noha A. Youssef
London School Of Economics

1 Introduction

Finding an optimal experimental design is considered one of the most important topics in the context of the experimental design. Optimal design is the design that achieves some targets of our interest. The Bayesian and the Non-Bayesian approaches have introduced some criteria that coincide with the target of the experiment based on some specific utility or loss functions. The choice between the Bayesian and Non-Bayesian approaches depends on the availability of the prior information and also the computational difficulties that might be encountered when using any of them.

This report aims mainly to provide a short summary on optimal experimental design focusing more on Bayesian optimal one. Therefore a background about the Bayesian analysis was given in Section 2. Optimality criteria for non-Bayesian design of experiments are reviewed in Section 3. Section 4 illustrates how the Bayesian analysis is employed in the design of experiments. The remaining sections of this report give a brief view of the paper written by (Chalenor & Verdinelli 1995). An illustration for the Bayesian optimality criteria for normal linear model associated with different objectives is given in Section 5. Also, some problems related to Bayesian optimal designs for normal linear model are discussed in Section 5. Section 6 presents some ideas for Bayesian optimal design for one way and two way analysis of variance. Section 7 discusses the problems associated with nonlinear models and presents some ideas for solving these problems.

2 Bayesian Basic Definitions

Given a data set y_1, \dots, y_n , where n is the sample size and $\theta \in \Theta$ the vector of unknown parameters where Θ is the parameter space, we define the following;

Definition 1 *Prior distribution, $\pi(\theta)$, represents the prior information we have about the unknown parameter θ .*

Definition 2 *Posterior distribution, $p(\theta|y)$, is the conditional distribution of θ given the data y_1, \dots, y_n . It can be derived using the following relation*

$$p(\theta|y) = \text{constant} \times f(y|\theta) \times \pi(\theta) \quad (1)$$

where $f(y|\theta)$ is the sampling distribution and the constant is just $\frac{1}{f_y(y)}$ the marginal distribution of Y which is considered a constant with respect to θ .

Definition 3 Risk is the expected loss given by

$$R(\theta) = E_y L(\hat{\theta}, \theta) \quad (2)$$

where $L(\hat{\theta}, \theta)$ is a predetermined loss function. The loss function can be replaced by a utility function.

Definition 4 Bayes Risk is the expected risk where the integration is taken over θ ,

$$E(R(\theta)) = E_\theta E_{y|\theta} L(\hat{\theta}, \theta) \quad (3)$$

$$= E_y E_{\theta|y} L(\hat{\theta}, \theta) \quad (4)$$

and $E_{\theta|y} L(\hat{\theta}, \theta)$ refers to the posterior risk.

Definition 5 Bayes estimator is the estimator obtained by finding the minimum of the posterior expected risk or the maximum of the posterior expected loss.

3 Non-Bayesian Optimality Criteria For The Normal Linear Model (Atkinson & Donev 1992)

Assuming a normal linear model for the response $Y(x)$ at the point x , then we have

$$E(Y_x) = \sum_j f_j(x) \theta_j \quad (j = 1, \dots, p) \quad x \in \mathcal{X} \quad (5)$$

where $y \sim N(X\theta, \sigma^2 I)$, θ is a vector of unknown parameters, σ^2 is known and \mathcal{X} is the design space. Fisher information matrix is defined as $\sigma_2 n M$ or for simplicity nM where $M = \frac{1}{n} X^T X$.

The optimality criteria in the normal linear model can be divided in two main branches first of them is parameter based criteria and response based criteria. The optimality criteria for parameter based criteria can be summarized as follows;

Trace criterion is chosen when the aim of our experiment is to minimize the total variance of the L.S. estimates $\hat{\theta}$:

$$\text{var}\hat{\theta}_1 + \text{var}\hat{\theta}_2 + \dots + \text{var}\hat{\theta}_p.$$

Since

$$\text{cov}(\hat{\theta}) = \sigma^2 (X^T X)^{-1}$$

and

$$\text{var}\hat{\theta}_1 + \text{var}\hat{\theta}_2 + \dots + \text{var}\hat{\theta}_p = \sigma^2 \text{trace}(X^T X)^{-1}.$$

A Optimality is used when the aim of the experiment is to estimate more than one linear function of the parameters, e.g. $K^T \hat{\theta}$, since

$$\text{cov}(K^T \hat{\theta}) = \sigma^2 K^T (X^T X)^{-1} K,$$

then we minimize

$$\begin{aligned}\sigma^2 \text{trace}[K^T(X^T X)^{-1}K] &= \sigma^2 \text{trace}[(X^T X^{-1})KK^T] \\ &= \sigma^2 \text{trace}[(X^T X)^{-1}A]\end{aligned}\quad (6)$$

with $A = KK^T$, i.e. A any $p \times p$ symmetric non-negative definite

C Optimality is chosen when estimating one particular linear function of the parameters is of our interest, $c^T \theta$, this criterion is a special case of A optimality criterion. It is also called linear optimality. So we aim to minimize

$$\begin{aligned}\text{var}(c^T \hat{\theta}) &= \sigma^2 c^T (X^T X)^{-1} c \\ &= \sigma^2 \text{trace}[c^T (X^T X)^{-1} c] \\ &= \sigma^2 \text{trace}[(X^T X)^{-1} c c^T]\end{aligned}\quad (7)$$

D Optimality is used to minimize the covariance of of the estimators of $\hat{\theta}$,

$$\det \text{cov}(\hat{\theta}) = \sigma^2 \det(X^T X)^{-1} = \sigma^2 |X^T X|^{-1} = \sigma^2 \prod_j \lambda_j^{-1} \quad (8)$$

which is equivalent to maximize the determinant of the information matrix, $|X^T X|$.

E Optimality is used when we are interested in estimating a normalized linear function of the parameters. It can be considered a special case of C optimality. So we may want to minimize $\max \text{var}(c^T \hat{\theta}) \forall c$, such that, $\|c\| = 1$. By leaving σ^2 out we have

$$\begin{aligned}\max_{\|c\|=1} \text{var}(c^T \hat{\theta}) &= \max_{\|c\|=1} c^T (X^T X)^{-1} c \\ &= \max \text{eigenvalue of } (X^T X)^{-1} \\ &= [\lambda_{\max}](X^T X)^{-1}.\end{aligned}\quad (9)$$

This criterion is equivalent to maximizing $[\lambda_{\min}(X^T X)]$, the smallest eigenvalue of $X^T X$.

For the response based criteria, we aim at minimizing in some sense the variance of the expected response,

$$\hat{Y} = f(x)^T \hat{\theta} \quad (10)$$

$$\text{var} = \sigma^2 f(x)^T (X^T X)^{-1} f(x). \quad (11)$$

Another optimality criterion can be formulated for this purpose called **G optimality** defined as;

$$\min_{\xi} \sup f(x)^T (X^T X)^{-1} f(x).$$

It has been showed in the literature that G and D optimality are equivalent under certain conditions and the famous theory for that is called equivalence theorem (Kiefer & Wolfowitz 1959)

4 Bayesian Approach For Design of Experiments

Let ξ be a design chosen from Ξ the set of all possible designs and d the decision chosen from the set \mathcal{D} . Bayesian approach aims at finding the design that maximizes the expected utility of the best decision, i.e.,

$$U(\xi^*) = \max_{\xi} \int_{\Theta} \max_{d \in \mathcal{Y}} \int_{\Theta} U(d, \theta, \xi, y) \cdot p(\theta|y, \xi) p(y|\xi) d\theta dy \quad (12)$$

where $U(d, \theta, \xi, y)$ is the utility function chosen to satisfy certain goal, y is the observed data from a sample space \mathcal{Y} , d is the decision which consists of two parts: first selection of ξ and then the choice of a terminal decision d . By rearranging the integrand,

$$U(\xi^*) = \max_{\xi} \int_{\mathcal{Y}} \max_{d \in \mathcal{Y}} [\int_{\Theta} U(d, \theta, \xi, y) \cdot p(\theta|y, \xi) d\theta] p(y|\xi) dy, \quad (13)$$

we are able to say that we want to find the decision that maximize the expected posterior utility and then find the design that maximizes the expected pre-posterior utility of the best decision.

The utility function is specified according to the purpose of the experiments whether it is inference about the parameters or prediction of the response or both. The design that is optimal for estimation is not optimal for prediction, even for one purpose one can find several utility functions that lead to different designs. Many criteria have been obtained from these different utility functions. In the context of linear models we have alphabetical design criteria in the non-Bayesian approach (*A*-, *C*-, *D*-, *E*- optimality), which some of them can be extended to serve in the Bayesian one i.e. selecting appropriate utility functions can help deriving the alphabetical optimality of non-Bayesian criteria.

5 Bayesian Designs For The Normal Linear Model

Bayesian approach differs from the Non-Bayesian approach by assuming a known prior distribution for the parameters. By assuming σ^2 is known and normally distributed $\theta|\sigma^2$ with mean θ_0 and variance covariance matrix $\sigma^2 R^{-1}$, where the $k \times k$ is known matrix. The posterior distribution for $\theta|y$ is also normal with mean vector

$$\theta^* = (nM(\xi) + R)^{-1} (X^T y + R\theta_0) \quad (14)$$

and covariance matrix $\sigma^2 (nM(\xi) + R)^{-1}$. Bayesian optimality criteria can be derived from non-Bayesian optimality criteria. This section summarizes the Bayesian optimality criteria for normal linear models as follows;

Bayes A Optimality can be obtained by maximizing the following expected utility

$$U(\xi) = - \int_y \int_{\theta} (\theta - \hat{\theta})^T A (\theta - \hat{\theta}) \cdot p(y, \theta|\xi) d\theta dy \quad (15)$$

where A is a symmetric nonnegative definite matrix. So by integrating over θ the corresponding criterion is $\phi_2(\xi) = -\text{trace} A (nM(\xi) + R)^{-1}$ which is called Bayes *A*-optimality while the corresponding non-Bayes *A*-optimality is $\text{trace} A nM(\xi)^{-1}$. If rank A is 1 which means that $A = cc^T$ we have the *C*-optimality. If this linear combination is normalized this is called Bayes

E -optimality, this means to minimize the maximum posterior variance of all possible combinations of the parameters. So we are minimizing

$$\sup_{\|c\|=wc^T} (nM(\xi) + R)^{-1}c = w^2 \lambda_{\max}[(nM(\xi) + R)^{-1}], \quad (16)$$

but the Bayesian interpretation of this criterion is not clear.

Bayes D Optimality can be obtained when the gain in Shannon information is used as the utility function so that we choose the design that maximize the expected gain of Shannon information

$$U(\xi) = \int_y \int_{\theta} \log p(\theta|y, \xi) p(y, \theta|n) d\theta dy. \quad (17)$$

This function takes the following form in the normal linear model

$$U(\xi) = -\frac{k}{2} \log(2\pi) - \frac{k}{2} + \frac{1}{2} \log \det \sigma^2(nM(\xi) + R) \quad (18)$$

and this reduces to maximize $\det(nM(\xi) + R)$ which is Bayes D optimality, while non-Bayes D optimality is just $\det nM(\xi)$. Sometimes in the non-Bayes optimality this criterion is equivalent to Bayes D optimality when there is a previous design done before so we have to maximize $\det(nM + X_0^T X_0)$ and $X_0^T X_0$ is fixed. D optimality can be obtained using other utility functions that aim to different goals other than inference about θ , like prediction or discriminating between two models. Examples for that we have scoring functions

$$U(\theta, p(\cdot), \xi) = 2p(\theta) - \int p^2(\tilde{\theta}) d\tilde{\theta} \quad (19)$$

and

$$U(\tilde{\theta}, \theta, \xi) = \begin{cases} 0, & |\hat{\theta} - \theta| < a \\ -1, & |\hat{\theta} - \theta| > a \end{cases} \quad (20)$$

where a is an arbitrarily small positive constant.

G -optimality is chosen to minimize $\sup_{x \in \mathcal{X}} x^T (nM(\xi) + R)^{-1} x$. This criterion is very related to E -optimality since they don't represent any utility function.

Several issues have been discussed in the literature of optimal designs. In this report we just mention three main topics the first issue is the difference between Bayesian and non-Bayesian optimality, the second is when σ^2 is unknown and the third is when the target of the experiment is more than one goals.

First Issue is the main difference between Bayesian and non-Bayesian which is the dependence on sample size, which vanishes when the sample size is too large. This means that when the sample size is small, prior distribution has more effect on the design and posterior distribution. There is also a big difference between Bayesian and non-Bayesian is that when using C -optimality or D_s optimality, which is also a D optimality but concerns only with subset of the parameters, the choice of $nM(\xi)$ maybe singular while in the Bayesian R is nonsingular for a proper informative prior distribution, so $(nM(\xi) + R)$ is always nonsingular.

Second Issue discussed in the literature is when σ^2 is unknown. We can assume the conjugate priors for (θ, σ^2) is in normal inverted gamma so that both prior and posterior for θ are multivariate T. In this case evaluating these integrals are difficult, so numerical techniques are needed to find Bayesian designs. Regarding this problem, many papers have also discussed the alphabet optimality avoiding any distributional assumptions using the equivalence theory.

Third Issue is about getting a design that achieves two targets like estimation and prediction so we can form a utility function that is a sum of two utilities with different weights corresponding to the importance of each target. Verdinelli suggested the following utility function

$$\begin{aligned}
 U(\xi) &= \gamma \int \log p(y_{n+1}|y, \xi) \cdot p(y, y_{n+1}|\xi) dy dy_{n+1} \\
 &+ w \int \log p(\theta|y, \xi) p(y, \theta|\xi) dy d\theta
 \end{aligned} \tag{21}$$

for two purposes both inference and prediction.

In the case of normal linear model this is equivalent to minimizing

$$\sigma_{n+1}^2 [\det \sigma^2 (nM(\xi) + R)^{-1}]^{\frac{w}{\gamma}}. \tag{22}$$

6 Design For Analysis Of Variance Models

In the one way analysis of variance, the design consists of choosing the number of observations on each treatment, while in the two way case, the choice of the design is equivalent to choosing the number of observations taken on i^{th} treatment in the j^{th} block. This situation is solved in the non-Bayesian context using the following criterion

$$\Phi_p = \{k^{-1} \text{trace}[(nM(\xi) + R)^{-1}]^p\}^{1/p}. \tag{23}$$

Bayesian A optimality is a special case when $p = 1$, Bayesian D optimality is also a special case when $p \rightarrow 0$) and E optimality is when $p \rightarrow \infty$. Also there is the hierarchical form of the prior distribution which consists of three stages the first of them is the sampling distribution, the second and the third are used to model the prior distribution of θ .

7 Bayesian Nonlinear Design Problems

In the case of nonlinear models the problem can be formulated as maximizing expected utility but approximations must be used because it is often a complicated integral. Examples of nonlinear problems may include logistic regression and nonlinear regression(in X 's). Most of the approximations are done using normal approximation to the posterior distribution. These approximations involves Fisher information matrix or the second derivative of the log-likelihood function or the posterior distribution. The criteria in this case are just the integral over the prior distribution of θ . Hence, specifying a good prior is crucial

for nonlinear design problems. Local optimality can be used as a crude approximation to the prior distribution of θ by a one-point distribution considered as a best guess. Equivalence theorem plays a pivotal role in obtaining the Bayesian optimal criterion in such cases.

Support points concept is associated with approximated designs, which means the largest set of points with positive measure with respect to θ . Bayesian optimal designs have the support points as the number of unknown parameters which is not an appealing feature. Moreover, exact results can be found by numerical optimization.

In their paper (Chalenor & Verdinelli 1995) they claimed that sequential design does not provide any gain in terms of information whether in Bayesian or Non-Bayesian approaches in normal linear models but in nonlinear models the posterior utility clearly depends on the data y so there is a gain in information.

References

- Atkinson, A. C. & Donev, A. N. (1992), *Optimum experimental designs*, Vol. 8 of *Oxford Statistical Science Series*, Oxford University Press, Oxford.
- Chalenor, K. & Verdinelli, I. (1995), ‘Bayesian experimental design: A review’, *Statistical Science* **10**(3), 273–304.
- Kiefer, J. & Wolfowitz, J. (1959), ‘Optimum designs in regression problems’, *Ann. Math. Statist.* **30**, 271–294.