LECTURE 11

Optimal Design

The issue of how to optimally design experiments has been around for a long time, extending back to at least 1918 (Smith).

Before we start discussing the highlights of this topic we need to set up the problem. Note that we are going to set it up for the types of situations that we have encountered, although in fact the problem can be set up in more general ways.

We will define $X^{(n \times p)}$ as our design matrix, $\beta^{(p \times 1)}$ our vector of regression parameters, $y^{(n \times 1)}$ our vector of observations, and $\epsilon^{(n \times 1)}$ our error vector giving

$$y = X\beta + \epsilon.$$

We assume $\epsilon$ will be iid with mean zero and $\text{Cov}(\epsilon) = \sigma^2 I$. As before, we have

$$\hat{\beta} = (X'X)^{-1}X'y$$
$$\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$$
$$\hat{y}_x = x\hat{\beta}$$
$$\text{Var}(\hat{y}_x) = \sigma^2x(X'X)^{-1}x'.$$

The design problem consists of selecting row vectors $x^{(1 \times p)}$, $i = 1, 2, \ldots, n$ from the design space $X$ such that the design defined by these $n$ vectors is, in some defined sense, optimal. We are assuming that $n$ is fixed. By and large, solutions to this problem consist of developing some sensible criterion based on the above model and using it to obtain optimal designs.

One of the first to state a criterion and obtain optimal designs for regression problems was Smith (1918). The criterion she proposed was: minimize the maximum variance of any predicted value (obtained by using the regression function) over the experimental space. I.e.,

$$\min_{x_i, i=1, \ldots, n} \max_{x \in X} \text{Var}(\hat{y}_x).$$

This criterion was later called global, or $G$-optimality by Kiefer and Wolfowitz (1959).

A second criterion, proposed by Wald (1943), puts the emphasis on the quality of the parameter estimates. The criterion is to maximize the determinant of $X'X$. That is

$$\max_{x_i, i=1, \ldots, n} |X'X|.$$

©Steven Buyske and Richard Trout.
This was called $D$-optimality by Kiefer and Wolfowitz (1959). Comment on confidence ellipsoids: the determinant is the product of the eigenvalues, which is inversely proportional to the product of the axes of the confidence ellipsoid around $\hat{\beta}$, so maximizing $|X'X|$ is equivalent to minimizing the volume of the confidence ellipsoid.

In their General Equivalence Theorem, the equivalence of $D$ and $G$ optimality was established under certain conditions (to be discussed shortly).

While these criteria are the ones which have received the most attention in the literature, others have also been used. For example, so called $A$-optimality,

$$\min_{x_i,i=1,\ldots,n} \text{trace}(X'X^{-1}),$$

minimizes the average variance of the parameter estimates (Chernoff, 1953). Another criterion, $E$-optimality, finds the design which maximizes the minimum eigenvalue of $X'X$ (Ehrenfeld, 1955). A conceptually attractive criterion is called $V$-optimality (sometimes $IV$-optimality or $Q$-optimality). Here the criterion is to minimize the integrated prediction variance over the region of interest.

Of all these designs, only $D$-optimality is invariant under reparametrization.

Since $D$ and $G$ optimality are the criteria receiving the most attention in the applied literature, we will have a more detailed discussion of these criteria. Before doing so, we need to make a distinction between what is called the exact theory and the approximate theory. Suppose you had a problem involving maximizing a function over the integers. Standard calculus techniques don’t apply. A common technique would be to extend the function definition to the real numbers, use calculus to find the number where the maximum occurs, and then argue that the maximum over the integers will occur at an adjacent integer. The analogous design problem distinguishes the exact theory (like the integers) from the easier approximate theory (like the reals).

All of the design criteria just discussed have the property that

$$\phi(aX'X) = \text{positive constant} \times \phi(X'X),$$

so a design that maximizes $\phi(aX'X)$ also maximizes $\phi(X'X)$. Suppose that we have an $n$-point design with $n_i$ observations at $x_i$, so that $\sum n_i = n$. This, or any, design can be viewed as a measure $\xi$ on the design space $\mathcal{X}$. Let $\xi$ be a probability measure on $\mathcal{X}$ such that

- $\xi(x_i) = 0$ if there are to be no observations at $x_i$, and
- $\xi(x_i) = n_i/n$ if there are to be $n_i > 0$ observations at $x_i$.

For a discrete $n$-point design $\xi$ takes on values which are multiples of $1/n$, and defines an exact design on $\mathcal{X}$.

If we remove the restriction that $\xi$ be a multiple of $1/n$, we can extend this idea to a design measure which satisfies

$$\xi(x) \geq 0, \quad x \in \mathcal{X}$$

$$\int_{\mathcal{X}} \xi(dx) = 1.$$
Now let
\[ m_{ij}(\xi) = \int_{\mathcal{X}} x_i x_j \xi(dx), \quad \text{for all } i, j = 1, \ldots, p, \]
where \( m_{ij}(\xi) \) is the \( ij \) element of the matrix \( M(\xi) \). Note, for an exact design (this is called the moment matrix),
\[ M(\xi) = \frac{1}{n} X'X. \]
Similarly, a normalized generalization relating to \( \text{Var}(\hat{y}_x) \) is
\[ d(x, \xi) = x(M(\xi))^{-1}x', \]
again for an exact design
\[ d = d(x, \xi) = nx(X'X)^{-1}x'. \]
Using this notation we have the following definitions,
- \( \xi^* \) is \( D \)-optimal if and only if \( M(\xi^*) \) is nonsingular and
\[ \max_{\xi} |M(\xi)| = |M(\xi^*)|. \]
- \( \xi^* \) is \( G \)-optimal if and only if
\[ \min_{\xi} \max_{x \in \mathcal{X}} d(x, \xi) = \max_{x \in \mathcal{X}} d(x, \xi^*). \]
It turns out that a sufficient condition for \( \xi^* \) to satisfy the \( G \)-optimality criterion is
\[ \max_{x \in \mathcal{X}} d(x, \xi^*) = p, \]
where \( p \) is the dimension of \( M(\xi^*) \), or equivalently, the number of parameters in the model. To see that \( p \) is a lower bound for \( \max_{x \in \mathcal{X}} d(x, \xi) \), consider
\[ p = \text{trace} \quad MM^{-1} \]
\[ = \frac{1}{n} \text{trace} \quad X'X \quad M^{-1} \]
\[ = \frac{1}{n} \sum_i \text{trace} (x_i' x_i) M^{-1} \]
\[ = \sum_i \text{trace} (x_i M^{-1} x_i') \]
\[ \leq \max_{x \in \mathcal{X}} x M^{-1} x'. \]
How would one show that a specific design is the best there is? The key is to look at the derivative. For a given design \( M \) and a given optimality criterion \( \phi \) to maximize, like log det, the Fréchet derivative is defined as
\[ F_\phi(M, x'x) = \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \left[ \phi \left( (1 - \epsilon)M + \epsilon x'x \right) - \phi(M) \right]. \]
For example, let’s consider \( D \)-optimality. We want to maximize the determinant of \( M \). This is equivalent to maximizing the log of the determinant of \( M \). The function
$\phi = \log \det$ has the advantage that it is convex on the space of information matrices $M$, so that a local maximum will in fact be a global maximum. At any rate

$$F_\phi(M, x^t x) = \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \left[ \phi \left( (1 - \epsilon)M + \epsilon x^t x \right) - \phi(M) \right]$$

$$= \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \left[ \log \det \left( (1 - \epsilon)M + \epsilon x^t x \right) - \log \det(M) \right]$$

$$= \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \left[ \log \frac{\det \left( (1 - \epsilon)M + \epsilon x^t x \right)}{\det M} \right]$$

$$= \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \left[ \log \det \left( (1 - \epsilon)I + \epsilon x^t x M^{-1} \right) \right]$$

$$= \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \left[ \log \left( 1 - \epsilon \right)^p \det \left( I + \frac{\epsilon}{(1 - \epsilon)} x^t x M^{-1} \right) \right]$$

$$= \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \left[ \log \left( 1 - \epsilon \right)^p \left( 1 + \frac{\epsilon}{(1 - \epsilon)} \text{trace}(x^t x M^{-1}) + \mathcal{O}(\epsilon^2) \right) \right]$$

$$= \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \left[ \log \left( 1 - \epsilon \right)^p + \log \left( 1 + \frac{\epsilon}{(1 - \epsilon)} \text{trace}(x^t x M^{-1}) + \mathcal{O}(\epsilon^2) \right) \right]$$

$$= \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \left[ p \log \left( 1 - \epsilon \right) + \frac{\epsilon}{(1 - \epsilon)} \text{trace}(x^t x M^{-1}) + \mathcal{O}(\epsilon^2) \right]$$

$$= x^t M^{-1} x - p.$$ (Remember that $\log(1 + t) = t + \mathcal{O}(t^2)$.)

Why have we only considered the derivative of $M$ in the direction of a matrix of the form $x^t x$ and not something more general? As a consequence of Carathéodory’s Theorem, every element of the design space can be expressed as a convex combination of no more than $p(p + 1)/2 + 1$ elements of the form $x^t x$.

Finally, the equivalence of $D$- and $G$-optimality is established in the General Equivalence Theorem of Kiefer and Wolfowitz. The General Equivalence Theorem says: If $\phi$ is concave on $\mathcal{M}$, the space of design information matrices, and differentiable at $M(\xi^*)$, then the following are equivalent

1. The measure $\xi^*$ is $\phi$-optimal
2. The Fréchet derivative $F_\phi(M(\xi^*), x^t x) \leq 0$ for all $x \in \mathcal{X}$.
3. The following equality holds

$$\max_{x \in \mathcal{X}} F_\phi(M(\xi^*), x^t x) = \min_{\xi} \max_{x \in \mathcal{X}} F_\phi(M(\xi), x^t x).$$

This last is what gives the equality of $D$-and $G$-optimality.

The implication of this result is that we can use the sufficient condition for $G$-optimality to verify whether or not a specific design is $D$-optimal. That is, if

$$\max_{x \in \mathcal{X}} d(x, \xi^*) = p,$$

where as before $p$ is the number of parameters in the the model, including the intercept, then the design is $D$-optimal.
Note that $D$-optimality is essentially a parameter estimation criterion, whereas $G$-optimality is a response estimation criterion. The Equivalence Theorem says that these two design criteria are identical when the design is expressed as a measure on $A'$.

Note that when dealing with exact designs the equivalence of the two criteria does not hold.

In practice, the design problem consists of selecting an exact design to define the experimental runs. The design measure that we just discussed gives us an approximate design.

Measure designs are of interest primarily because the $D$-optimal measure design provides the reference against which exact designs can be evaluated, and also because the points in an optimal exact design will often correspond to the points of support (points of positive measure) of the $D$-optimal measure design.

For the practical problem we will let $\xi$ be an $n$-point design. The moment matrix

$$M(\xi) = \frac{1}{n} X' X$$

and

$$|M(\xi)| = \frac{1}{n^p} |X'X|.$$ 

Again a normalized measure of the variance of the prediction at $x$ is

$$d = x(M(\xi))^{-1} x' = nx(X'X)^{-1} x'.$$

Note, with our usual assumption,

$$\text{Var}(\hat{y}_x) = \frac{1}{n} \sigma^2 d.$$ 

The values of $|M(\xi)|$ and $d$ give an indication of the information per point for a design, so designs having differing numbers of points can be compared. Then designs can be compared based on their $D$- and $G$-efficiencies.

For a given design, call it $\xi'$, define the $D$-efficiency to be

$$\left[ \frac{|M(\xi')|}{\max_{\xi} |M(\xi)|} \right]^{\frac{1}{n}}.$$ 

Similarly, $G$-efficiency is defined as

$$\frac{p}{\max_{x \in X} d(x, \xi)} = \frac{p}{nd'}.$$ 

One should note that for finite designs, especially small ones, the efficiency is likely to be quite a bit less than 1.

Now let’s briefly consider a few examples of $D$-optimal designs. First note in the above presentation, that two pieces of information must be supplied prior to obtaining an optimal design. These are:

1. model to be used
2. Number of data points
We could also, as Kiefer did, talk about the location of points and proportion of points at that location, rather than the number of points, but that is not the usual practice in designing experiments.

Some examples:

1. Consider the model

\[ E(y) = \beta_0 + \sum \beta_i x_i + \sum_{i \leq j} \beta_{ij} x_i x_j + \beta_{123} x_1 x_2 x_3, \]

with \( i, j, k = 1, 2, 3 \), so \( p = 8 \). Consider all possible designs with eight points (so \( n = 8 \)) with the restriction \(|x_i| \leq 1\).

It can be shown that, for fixed diagonal terms, \(|X'X|\) is largest when all off-diagonal terms are zero. This can be achieved by adopting a 2⁴ design with points \((\pm a, \pm a, \pm a)\) with \(|a| \leq 1\). For this design it can be shown that

\[ |X'X| = npa^2 (3 + 3 \cdot 2 + 3) = 88 a^4. \]

Clearly this is maximized if \(|a| = 1\). That, in turn, implies that the optimal design is 2³ with \( a = \pm 1 \).

2. \( E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3. \)

Here \( p = 4 \), \( n = 4 \). Using the same type of argument as in example 1, the optimum design is 2³−¹ with \( I = 123 \).

3. \( E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \beta_{112} x_1 x_2 x_3 + \beta_{122} x_1 x_2 x_3 + \beta_{111} x_1 x_2 x_3. \)

Here \( p = 9 \) and \( n = 9 \). The optimal design is the 3² factorial with levels \((-1,0,1)\).

4. \( E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2, \)

the full second-order model in two variables. Here \( p = 6 \) and suppose \( n = 6 \). The optimal design is

\[
\begin{array}{c|c}
 x_1 & x_2 \\
-1 & -1 \\
1 & -1 \\
-1 & 1 \\
-\alpha & -\alpha \\
1 & 3\alpha \\
3\alpha & 1 \\
\end{array}
\]

where \( \alpha = 0.1315 \). These designs may be rotated by 90° or a multiple of that.

5. Using the same model with \( n = 9 \) the optimal design is 3².

6. Using a full second-order model and the number of points for a central composite for \( p \geq 2 \), the \( D \)-optimal design places \( 2^p \) points at ±1 and the star points are placed on the face of the hypercube.
Note that in the last two designs, the design space was a hypercube, not a hypersphere.

A theory, no matter how beautiful, won’t be used much in practice if it’s not practical. For optimal designs, this comes down to a question of having algorithms to find optimal designs. These were first worked out for $D$-optimality in the early 1970s. The basic idea is quite simple. Suppose you have a design with information matrix $M_n$. Find $x_{n+1}$ to maximize $F(M_n, xx^t)$. Unless $M_n$ is optimal, this value will be positive, and you can increase the design criterion measure by moving from $M_n$ in the direction of $x_{n+1} x_{n+1}'$. Thus for some $\alpha_{n+1}$ we define a new measure by

$$\eta_{n+1} = (1 - \alpha_{n+1})\eta_n + \alpha_{n+1}\eta_{n+1},$$

with corresponding information matrix

$$M_{n+1} = (1 - \alpha_{n+1})M_n + \alpha_{n+1}x_{n+1}x_{n+1}' .$$

The step-length $\alpha_n$ can either be chosen to maximize the criterion along that ray, as was done by Federov (1972), or as a sequence converging to zero but with divergent partial sums, as was done by Wynn (1970).

In particular, for $D$-optimality,

$$F(M, xx^t) = x'M^{-1}x - p.$$ 

Writing $d_n = x'M_n^{-1}x$ and $\tilde{d}_n = \max x'M_n^{-1}x = x_{n+1}'M_n^{-1}x_{n+1}$, the optimal step-size turns out to be

$$\alpha_{n+1} = \frac{\tilde{d}_n - p}{p(d_n - 1)} .$$ 

Further comments on algorithms.

SAS has proc optex to generate optimal designs. Typically one gives it a set of candidate points and a model. For example, suppose you’ve done a resolution IV fractional factorial design on seven factors. You’d like to augment the design so that you can estimate all two-factor interactions. First generate a set of candidate points:

```
proc factex;
factors x1-x7;
output out=can;
run;
```

and now the resolution IV design:

```
proc factex;
factors x1-x7;
model resolution=4;
size design=min;
output out=aug;
run;
```

Finally, find the augmented design totalling 30 points, according to the $D$-optimality criterion:
You will note that the criteria, and therefore the optimal designs that are developed, are model dependent. A question that arises is how well do these designs perform if the model is incorrect. We discussed that earlier in the lecture about bias and variance.

In fact, when using a computer to find an optimal design, one should always keep in mind that

1. The designs are model dependent, and may not be particularly good for other models.
2. Exact designs for $D$-optimality do not address prediction variance; the equivalence of $D$- and $G$-optimality does not hold for exact designs.
3. $D$-optimal designs do not allow for many center runs.

In general, for a situation that can be handled with a Central Composite Design or Box-Behnken, those are better designs in an overall sense. Nonetheless, computer-aided optimal designs can be invaluable for

1. Mixture designs
2. Constrained designs
3. Trying to salvage a “botched” experiment

**References** Although the literature on optimality is vast (over 600 articles), I’ll stick to three books, in increasing order of difficulty:

