The Theory of Designed Experiments

6. Optimal Choice of Treatments

UNESP-Botucatu, August 2010

Optimal Choice of Treatments

For linear models we can find optimal designs directly, e.g. the D-optimal design for a quadratic model with $x \in \{-1, 0, 1\}$. Assume that the proportions of points at the three levels are p_-, p_0 and p_+ respectively.

$$\begin{aligned}
\mathbf{X}'\mathbf{X} &= \begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \\
&= \begin{bmatrix} n & n(p_+ - p_-) & n(p_+ + p_-) \\ n(p_+ - p_-) & n(p_+ + p_-) & n(p_+ - p_-) \\ n(p_+ + p_-) & n(p_+ - p_-) & n(p_+ + p_-) \end{bmatrix}
\end{aligned}$$

Optimal Choice of Treatments

$$\begin{aligned} \Rightarrow |\mathbf{X}'\mathbf{X}| &= 4n^3 p_- p_+ (1 - p_- + p_+) \\ &\propto p_- p_+ (1 - p_- + p_+). \end{aligned}$$

Differentiating with respect to p_- and equating to zero gives

$$p_+ = 0 \text{ or } 1 - 2p_- - p_+ = 0$$

and differentiating with respect to p_+ and equating to zero gives

$$p_- = 0 \text{ or } 1 - 2p_+ - p_- = 0.$$

Hence maximum is at

$$p_+ = 1 - 2p_- \Rightarrow 3p_- - 1 = 0 \Rightarrow p_- = \frac{1}{3}$$

$$\Rightarrow p_- = p_0 = p_+ = \frac{1}{3}.$$

Notes

- It can be shown using the *General Equivalence Theorem* that this design is in fact D-optimal for $x \in [-1, 1]$. This beautiful theorem makes finding D-optimal designs even simpler, but cannot be used for many other criteria.
- The D-optimal design does not allow for detection of lack of fit. There is no reason why it should, since the D criterion assumes that the model is correct.
- This design is also weighted-A-optimal, with weights 0, 0.8 and 0.2 for β_0, β_1 and β_{11} respectively. These are very sensible weights, since we are interested in *comparing* levels and the scale of $2\beta_1$ corresponds to the scale of β_{11}.

Design for Nonlinear Models

Example: A food scientist is studying a reaction which causes the gelatinization of starch. The rate of reaction is known to be first order, i.e. if temperature, pH, etc. are held constant, the rate at which the substrate molecules are converted to molecules of product is constant. Chemists write $\mathrm{S} \xrightarrow{\mathrm{k}} \mathrm{P}$.

This means

$$\begin{aligned} \frac{dS}{dt} &= -kS \\ \Rightarrow \int \frac{dS}{S} = -\int k dt &\Rightarrow \log S = -kt + b \\ \Rightarrow S &= Ae^{-kt} \end{aligned}$$

An experiment is to be performed to determine the rate of the reaction. Which times should the reaction be run for?

Design for Nonlinear Models

The first thing to consider in dealing with nonlinear models is what is a reasonable error structure to assume?

Consider two possibilities for the example:

- errors are multiplicative and approximately log-Normal;
- errors are additive and approximately Normal.

Sometimes consideration of the error structure will allow linearisation of the model, but linearisation should be done only for this reason, *not just for convenience*.

Design for Nonlinear Models

With linear models, we can study the variances of parameter estimates for different designs, but *for nonlinear models the variances of parameter estimates depend on the (unknown) values of the parameters*.

Two possible ways round the problem:

- compare designs for a range of specific values of parameters and find one which is fairly good across the range;
- use a prior distribution for the parameters and find a design which maximises the expected value of some design criterion over the prior.

The first method is computationally simpler, but might not allow us to find a good design.

Design for Nonlinear Models

Assume that we will estimate the parameters, $\boldsymbol{\theta}$, using maximum likelihood. Approximately,

$$\hat{\boldsymbol{\theta}} \sim N\left[\boldsymbol{\theta}, \{\mathcal{I}(\boldsymbol{\theta}, \mathbf{X})\}^{-1}\right],$$

where

$$\mathcal{I}(\boldsymbol{\theta}, \mathbf{X}) = \frac{1}{\sigma^2}\mathbf{M}(\mathbf{X}, \boldsymbol{\theta})$$

is the expected Fisher information matrix.

Design for Nonlinear Models

Most criteria are related to $\{\mathbf{M}(\mathbf{X}, \boldsymbol{\theta})\}^{-1}$. Criteria corresponding to the common criteria for linear models can be developed.

- Pseudo-Bayesian (or average) weighted-A-efficiency chooses a design $\mathbf{X}$ to minimise

$$\phi(\mathbf{X}) = \int tr\left[\mathbf{A}\{\mathbf{M}(\mathbf{X}, \boldsymbol{\theta})\}^{-1}\right] p(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

- Pseudo-Bayesian D-efficiency maximises

$$\phi(\mathbf{X}) = \int \log\{|\mathbf{M}(\mathbf{X}, \boldsymbol{\theta})|\} - p(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Notes

- For linear models $\mathbf{M}(\mathbf{X}, \boldsymbol{\theta}) = \mathbf{X}'\mathbf{X}$ does not depend on $\boldsymbol{\beta}$ and so the pseudo-Bayesian criteria reduce to weighted-A-efficiency, D-efficiency, etc.
- If the parameters are estimated using nonlinear least squares, the usual first order approximation to the covariance matrix of the parameter estimates is also $\{\mathcal{I}(\boldsymbol{\theta}, \mathbf{X})\}^{-1}$.

It is rarely possible to find the optimal design analytically; a general grid search is usually used. This is computationally intensive.

The pseudo-Bayesian optimal design rarely turns out to have simple rational proportions of points at a few values of x.

Rounding of the design will be necessary. This will be acceptable if the number of experimental units is considerably greater than the number of distinct levels of x.

For small designs it might be better to use an *exchange algorithm* - see next chapter.

Bayesian Design

Since an informative prior distribution is being used for the design, it seems logical to also use it for a fully Bayesian analysis.

It is difficult to make general progress. Usually we assume that the posterior variance matrix will be well approximated by

$$\left\{ \frac{1}{\sigma^2}\mathbf{R} + \mathcal{I}(\boldsymbol{\theta}, \mathbf{X}) \right\}^{-1},$$

where the prior distribution of $\boldsymbol{\theta}$ has covariance matrix $\sigma^2\mathbf{R}^{-1}$.

Hence criteria corresponding to the standard optimality criteria will be based on $\{\mathbf{R} + \mathbf{M}(\mathbf{X}, \boldsymbol{\theta})\}^{-1}$.

Bayesian Design

In a Bayesian context it is natural to derive design criteria from a decision theoretic viewpoint.

A quadratic loss utility function,

$$U(\mathbf{X}) = -(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\mathbf{L}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

gives expected utility

$$E\{U(\mathbf{X})\} = -E\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\mathbf{L}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\}.$$

If $\hat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$

$$E\{U(\mathbf{X})\} \propto -\int tr\{\mathbf{L}(\boldsymbol{\theta})\{\mathbf{R} + \mathbf{M}(\mathbf{X}, \boldsymbol{\theta})\}^{-1}\}p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

and maximising expected utility is the same as maximising Bayesian L-efficiency. If $\mathbf{L}(\boldsymbol{\theta})$ is diagonal, we have Bayesian weighted-A-efficiency.

Bayesian Design

Another utility function is the gain in Shannon information or, equivalently, the Kullback-Leibler distance between the posterior and prior

$$U(\mathbf{X}) = \log\left\{\frac{p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X})}{p(\boldsymbol{\theta})}\right\}.$$

Since the prior does not depend on the design, the expected utility can be written as

$$E\{U(\mathbf{X})\} \propto \int \log\{p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X})\}p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

It can be shown that this can be approximated as

$$E\{U(\mathbf{X})\} \propto \int \log\left\{|\mathbf{R} + \mathbf{M}(\mathbf{X},\boldsymbol{\theta})|\right\} p(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

We will refer to this as Bayesian D efficiency.

Bayesian Design

Even from a Bayesian perspective, it is not immediately obvious that these methods are correct:

- The prior used for the design can be purely personal to the experimenter and need not be defended in public; indeed, nobody but the experimenter need ever know that it was used.
- The analysis might have to be done using several different priors, perhaps including vague priors, as the results might be published, presented to decision makers, or otherwise used to convince other people than those involved in planning the experiment.

Thus using an informative, personal, prior for the design, but assuming a vague prior for the analysis might be more appropriate.

With a vague prior $\mathbf{R} + \mathbf{M}(\mathbf{X}, \theta) \approx \mathbf{M}(\mathbf{X}, \theta)$ and the design problem becomes equivalent to that for maximum likelihood estimation.