A Monte Carlo exchange algorithm for finding near-optimal designs under model contamination

Dave Woods

School of Mathematics, University of Southampton, Southampton, UK, SO17 1BJ; D.C.Woods@maths.soton.ac.uk

Introduction

Consider a factorial experiment where, prior to experimentation, the relationship between a continuous response Y and m factors $\boldsymbol{x} = (x_1, \ldots, x_m)$ is thought to be approximated over the design region $\mathcal{R}$ of interest by the *assumed model*

$$Y(\boldsymbol{x}) = f(\boldsymbol{x}) + \varepsilon\,,$$

where $f(\cdot)$ is a function of the m factors, ε is a random error term with zero mean and variance σ^2 and the random error terms for different observations are assumed independent. Instead of assuming some known true functional form, we may represent our uncertainty about the true model by assuming a random *contamination* term $\Phi(\boldsymbol{x})$, giving a population of *true models*

$$Y(\boldsymbol{x}) = f(\boldsymbol{x}) + \Phi(\boldsymbol{x}) + \varepsilon\,. \tag{1}$$

This is often a realistic scenario for practical experiments in the physical sciences. An example is when a simple approximation to some complicated relationship will be fitted to the data from an experiment. This approximate model will then be used to make inferences which are subject to possible bias. One example is using the Brønsted-Debye-Huckel equation to estimate the relationship between the rate constant of a reaction and the ionic strength of the compounds in the reaction. The validity of this linear approximation has been tested extensively and found to be accurate at low ionic strength. However, curvature may appear for higher ionic strength, as in the schematic in figure 1.

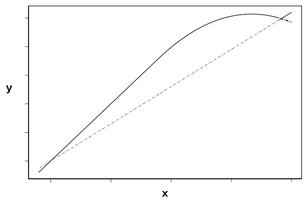

Figure 1: Schematic for an assumed model (dashed line) and true model (solid line).

Mean-squared error with random contamination

For *fixed* contamination $\phi(\boldsymbol{x})$, a design selection criterion which takes account of model misspecification is the *average mean squared error* (AMSE), which can be approximated over a grid of r evaluation points and can be split into *variance* (V) and *bias* (B) components as follows:

$$\begin{aligned} AMSE &= \frac{n}{r\sigma^2}\sum_{i=1}^{r} E\left\{[\hat{f}(\boldsymbol{x}) - E(\hat{f}(\boldsymbol{x}))]^2\right\} + \frac{n}{r\sigma^2}\sum_{i=1}^{r}\left\{E[\hat{f}(\boldsymbol{x})] - f(\boldsymbol{x}) - \phi(\boldsymbol{x})\right\}^2 \\ &= V + B\,, \end{aligned} \tag{2}$$

where n is the number of design points and $\hat{f}(\boldsymbol{x})$ is the least squares estimator of $f(\boldsymbol{x})$.

Under model (1), the bias for any given design and assumed model is a random variable, $B(\Phi)$, which can be shown to have the form

$$B(\Phi) = \frac{n}{r\sigma^2} tr\left\{[FAD - I]'[FAD - I]\boldsymbol{\Phi}\boldsymbol{\Phi}'\right\}\,.$$

where $A = (X'X)^{-1}X'$, F is an $r \times p$ matrix holding the values of the p assumed model terms for the r evaluation points, D is an $n \times r$ matrix with ijth entry non-zero and equal to 1 only when the ith design point and the jth evaluation point coincide and I is an $r \times r$ identity matrix.

Design selection criteria

Let $\mathcal{D}$ denote the set of all n point designs in $\mathcal{R}$. A design $\delta^* \in \mathcal{D}$ is

1. expected bias (EB-) optimal if

$$E\left[B(\Phi)|\delta^*\right] = \min_{\delta \in \mathcal{D}} E\left[B(\Phi)|\delta\right]\,.$$

2. expected mean squared error ($EMSE$-) optimal if

$$E(AMSE|\delta^*) = \min_{\delta \in \mathcal{D}} V(\delta) + E(B(\Phi)|\delta)\,.$$

Criterion 2 was also considered by Allen et al. (2003).

Implementation

By placing a prior distribution on Φ, we can find designs using these criteria which protect against a population of possible contaminations. This task is analytically intractable and so we use a search approach and have developed a modified Fedorov exchange algorithm (Cook and Nachtsheim, 1980) with embedded Monte Carlo simulation to find designs under these, and other, criteria.

An exchange algorithm swaps points between a candidate list of possible points and the points in the design in an attempt to find an optimal design. The criteria require the values of properties of the bias distribution. For each design search, a Monte Carlo simulation is performed in order to approximate these properties, with a sample of size $s = 10000$ simulated for each approximation.

Polynomial spline contamination

For a single factor, consider an assumed model

$$f(x) = \sum_{i=0}^{d} \beta_i x^i\,, \tag{3}$$

and contamination built from polynomial spline basis functions (Eubank, 1999, ch. 6) given by

$$\Phi(x) = \sum_{i=1}^{K} \Gamma_i (x - \Lambda_i)^d\,, \tag{4}$$

where the number of knots K, their locations Λ_i and coefficients Γ_i are random variables ($i = 1, \ldots, K$). This type of contamination allows different curvature over different sections of the range of x.

A theorem for spline contamination

For the important case when the expected value of each term of the contamination function is zero, analytical results can be derived which greatly increase the computational efficiency of the design search.

Theorem Suppose the true model is specified by (3) and the contamination has the form (4), consisting of independent random variables K, with mean μ_k and variance σ_k^2, and, for given $K = k$, random locations Λ_l and coefficients Γ_l, with mean μ_p and variance σ_p^2 ($l = 1, \ldots, k$). Then if $\mu_p = 0$

(i) $E\left[B(\Phi)\right] = n\mu_k\sigma_p^2 tr\left\{[F(X'X)^{-1}X'D - I]'[F(X'X)^{-1}X'D - I]E^{(1)}\right\}/(r\sigma^2)$.

where $E^{(1)}_{ij} = E\left[(x_i - \Lambda)^d_+ (x_j - \Lambda)^d_+\right]$.

(ii) the expected bias optimal design is invariant to μ_k, σ_k^2 and σ_p^2.

Corollary When $\mu_p = 0$,

$$E[AMSE] = \frac{n}{r} tr\left\{F'(X'X)^{-1}F + R[FAD - I]'[FAD - I]E^{(1)}\right\}\,,$$

where $R = \mu_k\sigma_p^2/\sigma^2$.

Example

Consider an assumed model (3) with $d = 2$ and contamination (4) where $K \sim$ Poisson(μ_k), $\Lambda_i \sim U(-0.2, 0.2)$ and $\Gamma_i \sim N(0, \sigma_p^2)$. Table 1 gives four point approximate $EMSE$-optimal designs for different values of the ratio $R = \mu_k\sigma_p^2/\sigma^2$.

R	$\frac{V}{E(B)}$	Design
$\to \infty$	$\to 0$	-0.8, -0.2, 0.2, 0.8
213	1.0	-0.85, -0.25, 0.25, 0.85
90	2.0	-0.9, -0.3, 0.3, 0.9
25	5.2	-0.95, -0.25, 0.25, 0.95
10	7.9	-1, -0.15, 0.15, 1
5	14.5	-1, 0, 0, 1

Table 1: Approximate $EMSE$-optimal designs for different values of the ratio R.

As R tends to infinity, the $EMSE$-optimal design tends towards the EB-optimal design, namely $\{-0.8, -0.2, 0.2, 0.8\}$. As R decreases, the design points slowly shift towards the V-optimal design for this assumed model, $\{-1, 0, 0, 1\}$. A V-optimal design minimises the variance component, V, in equation (2). However, as can be seen from the table, the V-optimal design is not achieved until V is over 14 times the size of $E(B)$. The limiting EB-optimal and V-optimal designs are determined by the distribution of the additional knot locations and the assumed model respectively.

Figure 2 shows the expected average mean squared error for the approximate $EMSE$- and EB-optimal designs and for the V-optimal design for a selection of values of R. It is clear that the gain in expected mean squared error from using the V-optimal design over the EB-optimal design for low R values is much smaller than the gain from using the EB-optimal design for higher values of R. The advantage of the $EMSE$-optimal design is clear, as for *any* value of R, it has the lowest possible value for $E[AMSE]$.

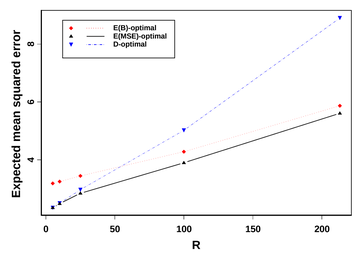

Figure 2: Expected mean squared error values for a selection of R values.

Acknowledgements

The author would like to thank Professor S.M.Lewis, his PhD supervisor, for her advice and guidance. This work was supported by EPSRC grant GR/R67729.

References

Allen, T. T., Yu, L. and Schmitz, J. (2003) An experimental design criterion for minimizing meta-model prediction errors applied to die casting process design. *J. Roy. Statist. Soc. C.*, **52**, 103–117.

Cook, R. D. and Nachtsheim, C. J. (1980) A comparison of algorithms for constructing exact D-optimal designs. *Technometrics*, **22**, 315–324.

Eubank, E. L. (1999) *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker.