Optimality of Observed Information Adaptive Designs in Linear Models

Adam Lane

October 17, 2019
OUTLINE

INTRODUCTION

WEIGHTED LEAST SQUARES

DESIGN

OPTIMALITY

EXAMPLES

DISCUSSION
LINEAR MODEL

1. It is assumed the responses are observed from a linear regression model

\[ y = \beta^T f_x(x) + \varepsilon. \]

2. \( \beta \) is a \( p \) dimensional vector within the parameter space \( B \).

3. \( f_x(x) \) is a function of the design points \( x \in \mathcal{X} \) and \( \mathcal{X} \) is the design space.

4. The distribution of the errors satisfies the location family constraint

\[ f_\beta(y) = f_0[y - \beta^T f_x(x)] = f_0(\varepsilon). \]

5. The errors, from a sample of size \( n \), are independent and identical distributed.
**Exact Design**

1. An exact design is comprised of a set of $d$ design points, $x_i$, with corresponding weights, $w_i = n_i/n$ at each design point, $i = 1, \ldots, d$, and is denoted

$$\xi_n = \left\{ \begin{array}{c} x_1 \quad x_2 \quad \cdots \quad x_d \\ w_1 \quad w_2 \quad \cdots \quad w_d \end{array} \right\},$$

where the total sample size $n = \sum n_i$.

2. Let $y_i = (y_{i1}, \ldots, y_{in_i})^T$ represent the responses observed at the $i$th design point.

3. Let $y = (y_1^T, \ldots, y_d)^T$ be the complete response vector.

4. $\eta_i = \beta^T f_x(x_i)$ is the mean function of the $i$th design point.
THE ARGUMENT FOR CONDITIONAL INFERENCE

1. Fisher (1934) showed that, for the location family, \((\hat{\eta}, a)\) is a sufficient statistic, where \(\hat{\eta}\) is the MLE and \(a\) is an ancillary configuration statistic.

2. A random variable is *ancillary* if its distribution is independent of the model parameters.

3. Since \(\hat{\eta}\) is not sufficient, inference based on the MLE alone represents a loss of information.

4. Conditional on \(a\) the MLE is the minimal sufficient statistic and the full information available in the data is recovered in the conditional distribution \(\hat{\eta}|a\).

THE ARGUMENT FOR OBSERVED INFORMATION

1. In practice the conditional distribution, $\hat{\eta}|a$, is often untractable or unknown.
2. The joint distribution of $(\hat{\eta}, a)$ is

$$f(\hat{\eta}, a) = c(a)g_a(\hat{\eta} - \eta).$$

The expected and observed Fisher information are

$$\mathcal{F} = E \left[ \left\{ \frac{\partial \log g_a(\hat{\eta} - \eta)}{\partial \eta} \right\}^2 \right] \quad \text{and} \quad i_a = \left[ \frac{\partial^2 \log g_a(\hat{\eta} - \eta))}{\partial \eta^2} \right]_{\eta=\hat{\eta}},$$

respectively.
THE ARGUMENT FOR OBSERVED INFORMATION (CONT.)

1. In a seminal paper Efron and Hinkley (1978) establish that

\[
\text{Var}[\hat{\eta} \mid a] = i_a^{-1} + O_p(n^{-1}) \\
\text{Var}[\hat{\eta} \mid a] = \mathcal{F}^{-1} + O_p(n^{-1/2}).
\]

2. The conditional variance is the appropriate measure of the variance of the MLE.

3. The observed Fisher information is a better approximation of the conditional variance than the expected Fisher information.

4. Conclusion: Inference should be conducted using the observed Fisher information.
SEQUENTIAL EXPERIMENT

1. Expected Fisher information is known *a priori* and thus can be used to design an experiment completely in advance of the data collection.

2. Observed Fisher information is a function of the design points and responses and is thus not available in advance.

3. If an experiment is observed sequentially then it is possible to develop adaptive designs that use observed Fisher information from the previous runs to inform design decisions of the current and future runs.

4. Specifically, it is possible to apply well developed tools of optimal design to an adaptive procedure in an effort to ”minimize” the inverse of observed Fisher information.
OUTLINE

INTRODUCTION

WEIGHTED LEAST SQUARES

DESIGN

OPTIMALITY

EXAMPLES

DISCUSSION
**Information in the Regression Model**

1. Let $a_i$ be the ancillary configuration statistic for the $i$th design point and $\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_d)^T$.
2. The log likelihood, with respect to $\beta$, is
   \[
   l(\beta|\hat{\eta}) = \sum_{i=1}^{d} \log g_{a_i}[\hat{\eta}_i - \beta^T f(x_i)]
   \]
3. Assuming derivatives and integrals can be exchanged the per-subject expected Fisher information can be written
   \[
   M(\xi_n) = \nu \sum_{i=1}^{d} w_i f(x_i)f_x^T(x_i) = \nu F^T W F,
   \]
4. $\nu = E [(\partial^2 / \partial \eta^2) \log g_{a_i}(\hat{\eta}_i - \eta_i)]$
5. $F$ is a $n \times p$ matrix with the $i$th row $f(x_i)^T$ and $W = \text{diag}(w_1, \ldots, w_d)$.
Information in the Regression Model (cont)

1. Let $A = (a_1, \ldots, a_d)^T$.
2. The observed Fisher information is

$$J_A(x) = \frac{1}{n} \sum_{i=1}^{d} i_{a_i} f_x(x_i) f_x^T(x_i) = \frac{1}{n} F^T I_AF.$$

3. $I_A = \text{diag}(i_{a_1}, \ldots, i_{a_d})$.
4. The observed Fisher information as defined in $J_A(x)$ is not a standard one used in linear regression.
From the independence of responses it is clear that

$$\text{Var}[\hat{\eta}|a] = I_A^{-1}[1 + O_p(n^{-1})].$$

The above representation suggests that from the conditional perspective a weighted least squares approach represents a straightforward alternative to the MLE.

Specifically, the WLSE of $\beta$ is

$$\tilde{\beta} = [F^T I_A F]^{-1} F^T I_A \hat{\eta}. $$
Weighted Least Squares Estimate (cont.)

1. For the WLSE estimate it is straightforward to show that the results of Efron and Hinkley (1978) extend to the regression setting, i.e.,

\[ n \text{Var}[\tilde{\beta}|A] = \{J_A(x)\}^{-1}[1 + O_p(n^{-1})]. \]

and

\[ n \text{Var}[\tilde{\beta}|A] = \{M(\xi_n)\}^{-1}[1 + O_p(n^{-1/2})]. \]

2. In the regression setting \( \text{Var}[\tilde{\beta}|A] \) is a more appropriate measure of the variance of the WLSE than \( \text{Var}[\tilde{\beta}] \).

3. The inverse of observed Fisher information, \( \{J_A(x)\}^{-1} \), is a more precise approximation of the conditional variance that the inverse of expected Fisher information, \( \{M(\xi_n)\}^{-1} \).
OUTLINE

INTRODUCTION

WEIGHTED LEAST SQUARES

DESIGN

OPTIMALITY

EXAMPLES

DISCUSSION

References
Fixed Optimal Design (FOD)

1. Let $\Psi(\cdot)$ be a convex criterion function. The optimal design problem is to find the design $\xi^*_\Psi(\theta)$ defined as

$$\xi^*_\Psi = \arg\min_{\xi \in \Xi} \Psi\{M(\xi_n)\},$$

where $\Xi$ represents the set of all permissible designs.

2. We refer to this as the fixed (not adaptive) optimal design (FOD).
SEQUENTIAL CONSTRUCTION OF THE FIXED OPTIMAL DESIGN

① In the FOD setting sequential means that the design points (and not responses) for the preceding observations are available to determine the design for the current observation.

② Once the algorithm has terminated there is no reason for the experiment to be conducted sequentially.

③ Let $\phi(x, \xi)$ denote the sensitivity function for the criterion function $\Psi$.

④ Suppose, the design for the first $j - 1$ observations $\xi(j - 1)$ is a non-optimal design such that $M[\xi(j - 1)]$ is non-singular.
Sequential Construction of the Fixed Optimal Design (cont.)

1. Let $\bar{\xi}_x$ as a design with design point $x$ and allocation 1 and
   
   $$\xi(j) = (1 - \alpha_{j-1})\xi(j - 1) + \alpha_{j-1}\bar{\xi}_x,$$

   where $\alpha_{j-1} = 1/j$ is the step size of the search algorithm.

2. For the design $\xi(j)$ we can write the expected Fisher information as
   
   $$M\{\xi(j)\} = (1 - \alpha_{j-1})M\{\xi(j - 1)\} + \alpha_{j-1}M\{\bar{\xi}_x\}.$$

3. The objective of a sequential algorithm is to minimize
   $$\Psi\{M[\xi(j)]\}$$
   for each iteration in the algorithm.

4. The $j$th observation in the search algorithm is placed at the design point
   
   $$\bar{x}_j = \min_{x \in X} \phi[x, \xi(j)].$$
In the sequential setting the sample is an ordered set of experimental runs.

In the fully sequentially setting each run is comprised of a single observation. Specifically, the data for observations 1, \ldots, j − 1 are known prior to the design assignment of the jth observation, for j = 2, \ldots, n.

To construct the proposed adaptive design it is assumed that the continuous FOD with respect to Ψ, denoted ξΨ, is known or can be computed. The optimal design points are denoted \( x^* = (x_1^*, \ldots, x_d^*)^T \).

The ancillary configuration vector for the ith design point from the first j runs is \( a_i(j) \) and the corresponding ancillary configuration matrix is \( A(j) = [a_1(j), \ldots, a_d(j)] \).
Let \( Q_A(j) = \sum_i i a_i(j) \), \( \omega_a(i(j)) = i a_i(j)/Q_A(j) \) and

\[
\tau_A(j) = \begin{bmatrix}
    x_1^* \\
    \omega_a_1(j) \\
    x_2^* \\
    \omega_a_2(j) \\
    \vdots \\
    \omega_a_d(j) \\
    \vdots \\
    x_d^*
\end{bmatrix}.
\]

Remark: \( \tau_A(j) \in \Xi_\Delta \). In other \( \tau_A(j) \in \Xi_\Delta \) is a design.

The observed Fisher information from the first \( j \) observations can be written as

\[
J_A(j)(\mathbf{x}^*) = \frac{1}{j} Q_A(j) M(\tau_A(j)).
\]

The importance of this is that \( J_A(j) \) is proportional, up to a known constant, to the expected Fisher information evaluated at the design \( \tau_A(j) \).
Observed Information Adaptive Design (cont.)

1. Now the previously described sequential algorithm can be adapted to incorporate observed Fisher information.

2. Let

   \[ \tau(j) = (1 - \alpha_j)\tau_A(j-1) + \alpha_j\bar{\xi}_x. \]

3. In the same way that \( j - 1 \) is weight the expected Fisher information from the first \( j - 1 \) observations, \( Q_A(j-1) \) is the weight of the observed Fisher information after the first \( j - 1 \) subjects.

4. For \( \alpha_j = [1 + Q_A(j-1)]^{-1} \) we can write

   \[ M\{\tau(j)\} = (1 - \alpha_{j-1})M\{\tau_A(j-1)\} + \alpha_{j-1}M\{\bar{\xi}_i\}. \]
**Observed Information Adaptive Design (OAD) Algorithm**

1. For observations $j = 1, \ldots, kd$ initiate the design by setting $k$ observations at each of the optimal design points $x_1^*, \ldots, x_d^*$.

2. Compute $i_{ai}(j-1)$, $Q_A(j-1)$, $\omega_{ai}(j-1)$, and $\tau_A(j-1)$, where $(j - 1)$ indicates that it the relevant quantity is computed using the ancillary statistics from the first $j - 1$ observations.

3. The design point for observation $j = kd + 1, \ldots, n$ is

$$
\hat{x}_j = \min_{x \in x^*} \phi[x, \tau(j)],
$$

where $x^*$ is the set of optimal design points of the FOD.
OUTLINE

INTRODUCTION

WEIGHTED LEAST SQUARES

DESIGN

OPTIMALITY

EXAMPLES

DISCUSSION
Inference for the OAD

There are two challenges associated with inference following adaptive designs.

1. After the completion of an adaptive procedure the design is not ancillary. As a result, it is argued that an analysis done conditionally on the design will result in information loss.

   The OAD uses only the ancillary information contained in $A(j - 1)$ to select the design for observation $j$.

2. Therefore, the design remains ancillary and there is no information lost by conditioning on the observed design.

   The induced dependence of the response of the $j$th observation on the responses from the preceding $j - 1$ runs, which can lead to bias and non-normality of the approximate distribution of the estimates.
Inference for the OAD

Theorem
Let \( \tilde{y}, \tilde{A} \) and \( \tilde{X} \) represent the responses, the ancillary configuration matrix and the observed design matrix from OAD, respectively. If \( A = \tilde{A} \) and \( X = \tilde{X} \) then \( \tilde{X}|\tilde{A} = \tilde{X} \) and

\[
f(\tilde{y}|\tilde{A}) = f(y|X, A).
\]

1. The theorem states that conditional on the ancillary configuration matrix, \( A \), the distribution of responses is the same regardless of whether they were observed adaptively or not.

2. The implication is that with respect to conditional inference the adaptive nature of the design can be ignored.
The conditions for optimality are as follows:

1. The distribution of $y$ is a member of the location family.
2. The residual vector, $\varepsilon$, is a vector of i.i.d random variables.
3. $E[(\hat{\eta}_i - \eta^2)] < \infty$.
4. For $k = 1, \ldots, 4$, $l^{(k)}(y - \eta)$ exists and $E[|l^{(k)}(y - \eta)|] < \infty$ with $\tilde{l}_j > 0$ and $E[\tilde{l}_j^2] < \infty$.
5. If $|\eta - \eta'| < \delta$ then $|l^{(5)}(y - \eta) - l^{(5)}(y - \eta')| < K_\delta(y, \eta)$, where $\lim_{\delta \to 0} E[K_\delta(y, \eta)] = 0$.

These conditions are the same as Efron and Hinkley (1978) except the additional restriction on the 5th derivative.
Inference of the OAD

Let $\tilde{A}$ and $A^*$ represent the responses and the ancillary configuration matrix for the OAD and FOD, respectively.

Corollary

Under the stated conditions

$$E[\tilde{\beta} - \beta | A] = O_p(n^{-1})$$

$$n \text{Var}[\tilde{\beta} | A] = \{J_A(x)\}^{-1} \left\{1 + O_p(n^{-1})\right\}.$$  

for $i = 1, \ldots, n$ and

$$n(\tilde{\beta} - \beta)J_A(x)(\tilde{\beta} - \beta)|A \rightarrow \chi^2_d + O_p(n^{-1})$$  

where $A$ can be either $\tilde{A}$ or $A^*$.  

Optimal Inference for the OAD

If it is accepted that entries of $J_A$ should be used for inference then it is implied that inference is optimized by minimizing $\Psi\{J_A\}$.

Theorem

Under the stated conditions

$$nE[\Psi\{J_A^*(x^*)\} - \Psi\{J_{\tilde{A}}(x^*)\}] \to \frac{\gamma^2}{2} tr[HV^*] \geq 0$$

as $n \to \infty$, where $H^* = \nabla^2 \Psi\{M(x^*)\}$, $V^* = \text{diag}[\omega^*_{(d-1)}] - \omega^*_{(d-1)}\omega^*_{(d-1)}^T$, $\omega^*_{(d-1)} = (\omega^*_1, \ldots, \omega^*_d)\),

$$\gamma^2 = \frac{\nu_{02}\nu_{20} - \nu_{11}}{\nu_{20}^3}$$

and $\nu_{jk} = E[\ddot{\psi}(\ddot{\psi} + E[\dot{i}]^2)^k]$. 

Adam Lane Observed Information Adaptive Designs 27 of 41
Optimal Conditional Mean Square Error (MSE) of the OAD

There a design that lead to the smallest $\Psi\{MSE[\hat{\beta}|A]^{-1}\}$ is considered optimal.

**Theorem**

*Under the stated conditions*

$$nE \left[ \Psi\{MSE[\hat{\beta}|A^*]^{-1}\} - \Psi\{MSE[\tilde{\beta}|\tilde{A}]\} \right] \rightarrow \frac{\gamma^2}{2} tr[H^*V^*] \geq 0$$

*as $n \rightarrow \infty$.***
Optimal Power of the OAD

1. The corollary established that under the null hypothesis that $E[c^T \beta] = C_0$,

$$[c^T J_A^{-1} c]^{-1} (c^T \beta - C_0)^2 | A = \chi_1^2 + O_p(n^{-1}).$$

2. Under an alternative hypothesis $E[c^T \beta] = C_1$ power of a $\chi^2$ test is calculated as

$$\text{Power} = 1 - P\{\chi_1^2(\lambda) \geq \chi_1^{2(1-\alpha)}\},$$

where $\chi_1^2(\lambda)$ is a non-central $\chi^2$ distribution, $\chi_1^{2(1-\alpha)}$ is the $1 - \alpha$ quantile of the $\chi^2$ distribution and $\lambda$ is the non-centrality parameter.
In the current context the non-centrality parameter is expressed as

$$\lambda_A = n [c^T J_A^{-1} c]^{-1} \delta^2,$$

where $\delta = C_1 - C_0$.

Power is a nondecreasing function of the non-centrality parameter.

The optimal design maximizes the power of the $\chi^2$ test.
Corollary

Under the stated conditions

\[ E[\lambda_{A^*}] \to \delta^2 h_n \{c^T M_c^{-1} c\}^{-1} [n - \gamma^2 \text{tr}[V^* D^*]] \]
\[ E[\lambda_{\tilde{A}}] \to \delta^2 h_n \{c^T M_c^{-1} c\}^{-1} [n], \]

where \( h_n = 1 - (\tau_3^2 / \tau + \tau_4) / (2n\tau^2) \) and \( D^* \) is a symmetric matrix with the ikth entry equal to
\[ [f(x_i) + f(x_d)]^T (F^T W^* F)^{-1} [f(x_k) + f(x_d)]. \]
OUTLINE

INTRODUCTION

WEIGHTED LEAST SQUARES

DESIGN

OPTIMALITY

EXAMPLES

DISCUSSION
**POWER EXAMPLE**

1. Suppose, we have an experiment with \(d\) treatments, indexed by 0, \ldots, \(d - 1\). Each observation can receive only 1 treatment.

2. This can be written as a linear regression model with \(f_x(x) = (1, x_1, \ldots, x_d)^T\), where \(x_i = 1\) for the treatment \(i\) and 0 otherwise for \(i = 1, \ldots, d - 1\).

3. Treatment 0 represents a standard treatment and the treatments 1, \ldots, \(d\) are a collection experimental treatments.

4. The null hypothesis of this hypothetical experiment is \(H_0 : \sum_i \beta_i = 0\).

5. The most common setting for this particular set up is for \(d = 2\). Then the null hypothesis is simply \(H_0 : \beta_1 = 0\).
Power Example (cont.)

① A $c$ optimal design, with $c = (0, 1_{d-1}^T)^T$, will maximize the power of the $\chi^2$ test.

② It can be shown that the $c$ optimal design allocates 1/2 of the observations to treatment 0 and divides the remain 1/2 evenly among the remaining $(d - 1)$ treatments.

③ The corollary implies, after some basic algebra, that

$$E[\lambda_A^*] \to 4\tau^{-1}(d - 1)^2\delta^2h_n[n - (d - 1)\gamma^2]$$
$$E[\tilde{\lambda}_A] \to 4\tau^{-1}(d - 1)^2\delta^2h_n[n].$$

④ Written this way and it is clear that a FOD with sample size $n$ has approximately the same expected power as an OAD with as sample size of $n - (d - 1)\gamma^2$. 
Power Example (cont.)

1. Suppose that \( \varepsilon \) has a T-distribution with parameter \( v \), i.e.

\[
f(\varepsilon) = \left[\sqrt{B(1/2, v/2)}\right]^{-1}(1 + \varepsilon^2/v)^{-(1+v)/2},
\]

where \( B(a, b) \) is the Beta function.

2. For the T-distribution

\[
\gamma^2 = \frac{6[19 + 3v(6 + v)]}{v(v + 1)(v + 5)(v + 7)}.
\]

3. In this example \( \gamma^2 \to \infty \) as \( v \to 0 \) and is a decreasing as a function of \( v \).

4. For a T-distribution with 1/2, 1 and 2 degree of freedom \( \gamma^2 \approx 5.6, 2.5 \) and 1.06, respectively.
POWER EXAMPLE: SIMULATION RESULTS

As an example a T-distribution with $v = 1$ has been used. This corresponds to a Cauchy distribution. In the top figure $d = 1$ and in the bottom figure $d = 4$. The figures present $\Psi\{J_A(x^*)\}$, $\Psi\{MSE[\tilde{\beta}]^{-1}\}$ and Power (left to right). In each figure the solid line represents the results of the OAD. The dashed line corresponds to the FOD.
**Example: D Optimal Design**

1. For a second example $D$ optimal design for a quadratic model is considered for $f_x(x) = (1, x_1, x_1^2, \ldots, x_r, x_r^2)$.
2. The design space is $[-1, 1]^r$.
3. A $D$ optimal design places equal weight the points that represent a $3^r$ factorial with levels $-1, 0, 1$.
4. Again, a T-distribution with $\nu = 1$ has been used (Cauchy distribution).
5. The efficiency of the OAD relative to the FOD, with respect to the observed information and conditional MSE, is defined as

$$REff_{Obs} = \frac{\Psi \{ J_{A^*}(x^*) \}}{\Psi \{ J_{\tilde{A}}(x^*) \}} \quad \text{and} \quad REff_{MSE} = \frac{\Psi \{ MSE[\tilde{\beta}|A^*]^{-1} \}}{\Psi \{ MSE[\tilde{\beta}|\tilde{A}]^{-1} \}},$$

respectively.
6. A $REff > 1$ indicates the OAD is more efficient than the FOD.
SIMULATION RESULTS

The figures present $\Psi\{J_A(x^*)\}$, $\Psi\{MSE[\hat{\beta}|A^*]^{-1}\}$ and $REff$ (left to right). Values $r = 1$ (top) and $r = 2$ (bottom) were used. For columns 1 and 2 the solid line represents the results of the OAD. The dashed line corresponds to the FOD. For the third column the solid line is the $REff_{Obs}$ and the dashed line is the $REff_{MSE}$.
OUTLINE

INTRODUCTION

WEIGHTED LEAST SQUARES DESIGN

OPTIMALITY EXAMPLES

DISCUSSION

References
SUMMARY

① An adaptive design that incorporates observed information was developed and discussed.

② The proposed design is optimal with respect to inference and conditional MSE.

③ This optimality was confirmed via simulation study where the OAD was uniformly better than the FOD.


