Information in a Two-Stage Adaptive Optimal Design

Adam Lane[a], Ping Yao[b], Nancy Flournoy[a*]

[a]Department of Statistics,
146 Middlebush Hall
University of Missouri,
Columbia, MO, 65203 USA

[b]School of Nursing and Health Studies,
253 Wirtz Hall,
Northern Illinois University,
DeKalb, IL, 60115, USA

Abstract
In *adaptive optimal designs*, each stage uses an optimal design evaluated at maximum likelihood estimates that are derived using cumulative data from all prior stages. This dependency affects the properties of maximum likelihood estimates. To illuminate the effects, we assume for simplicity that there are only two stages and that the first stage design is fixed. The information measure most commonly used in the optimal design literature is compared with Fisher's information. To make the analysis explicit, responses are assumed to be normal with a one parameter exponential mean function. With this model, several estimates of information are compared and a procedure for selecting the proportion of subjects allocated to stage 1 is recommended.

Keywords: adaptive design, experimental design, Fisher's information, incremental information, observed information, nonlinear regression

[*]Principal corresponding author. Phone 573-882-6376. Fax 573-884-5524.
*Email addresses:*aclpp9@mail.missouri.edu (Adam Lane), pyao@niu.edu (Ping Yao), flournoyn@missouri.edu (Nancy Flournoy)

1 Introduction

The Fisher information from linear models does not depend on model parameters, and hence designs that maximize the information can be found and implemented directly. Information from nonlinear models is a function of model parameters, which complicates the implementation of efficient designs. Stein [1945] introduced two-stage procedures in which information from the first stage is used to improve the design of a second stage. Fisher [1947, Chapter 68] and Chernoff [1953] suggest that optimal designs be approximated by guessing the parameter values; however, this may be inefficient when the guess is far from the actual parameter value.

Adaptive optimal design is estimated from all prior data. This approach was endorsed by Box and Hunter [1965], Fedorov [1972] White [1975] and Silvey [1980] and used by many, including Dragalin, Fedorov, and Wu [2008]. Its appeal is that if an adaptive optimal design converges to the true optimal design, heuristically arguing, the overall experiment will become more efficient with additional stages.

In the adaptive optimal design literature, in place of constructing a likelihood from the joint density for responses and design points, responses have been treated as independent, conditional on treatment - both for selecting the next design point and for evaluating the design's efficiency. Silvey [1980, Chapter 7] and others point out that the information measure they employ is not, by definition, Fisher's information. While conditioning on treatment assignments is generally accepted for analysis, the role of conditioning in adaptive design construction has not been clarified.

This paper explores how dependency among treatments affects the properties of maximum likelihood estimates (MLEs). To illuminate the effects, we assume for simplicity that there are only two stages and that the first stage design is fixed. Responses are assumed to be normal with a one parameter exponential mean function. A procedure for selecting the proportion of subjects allocated to stage 1 is recommended. Measures of information and their estimates are analyzed and compared. The information measure most commonly used in the optimal design literature is compared with Fisher's information.

2 Information in a Two-Stage Experiment

Consider an experiment with two treatment groups. Suppose summary statistics $t_1 = t(\mathbf{y_1})$ and $t_2 = t(\mathbf{y_2})$ are obtained from subjects in group 1 and group 2, respectively, where $\mathbf{y_i}$ is a vector of independent observations from n_i subjects in group i. We use lower case letters to denote both random variables and their realization when there is no confusion.

Let x_1 be the treatment given to group 1, and assume that it is fixed. Suppose the treatment given to group 2 is an onto function of the stage 1 summary statistic, t_1, that is, $x_2 = x_2(t_1)$. Further suppose t_1 and t_2 are jointly sufficient for all observations $n = n_1 + n_2$, so that inference can be based on their joint density, which is assumed to be bounded and twice differentiable with respect to $\theta \in \Theta$. We also assume that the true value of θ is an internal point of Θ.

Let $\tilde{\theta}_n$ be an estimator of θ with finite expectation $\mathrm{E}\left(\tilde{\theta}_n\right) = \theta + b(\theta)$ and let $\hat{\theta}_n$ be the MLE based on the total sample. The derivation of the information inequality [cf. Cox and Hinkley [1974, p. 254], Hogg, McKean, and Craig [2005, p. 322]] applies in a straightforward manner for the two-stage experiment. Let $S = d\,ln\,f(t_1,t_2;\theta,x_1)/d\theta$ denote the score function. Then

$$\frac{d}{d\theta}\mathrm{E}\left[\tilde{\theta}_n\right] = 1 + b'(\theta) = \mathrm{Cov}\left[\tilde{\theta}_n, S\right]$$

and by the Cauchy-Schwartz inequality, $\left[\mathrm{Cov}\left[\tilde{\theta}_n, S\right]\right]^2 \leq \mathrm{Var}\left[\tilde{\theta}_n\right]\mathrm{Var}\left[S\right]$. Therefore, provided $\mathrm{Var}\left[S\right] > 0$,

$$\mathrm{Var}\left[\tilde{\theta}_n\right] \geq \frac{\left[1 + b'(\theta)\right]^2}{\mathrm{Var}\left[S\right]}. \tag{1}$$

Most estimators will not attain the lower bound, as equality in (1) requires $\tilde{\theta}_n$ to be perfectly correlated with S and to have constant bias with respect to θ. However, $(\mathrm{Var}\left[S\right])^{-1}$ is asymptotically equivalent to $\mathrm{Var}\left[\hat{\theta}_n\right]$, and as a result, $(\mathrm{Var}\left[S\right])^{-1}$ is an approximation commonly called Fisher's information.

When the sufficient statistics are sample means, i.e., $t_1 = \bar{y}_1$, $t_2 = \bar{y}_2$,

$$\begin{aligned}\frac{1}{n}\mathrm{Var}[S] &= \mathrm{Var}\left[d\,ln\,f\left(\bar{y}_1,\bar{y}_2;\theta,x_1\right)/d\theta\right] = \mathrm{E}\left[-d^2\,ln\,f\left(\bar{y}_1,\bar{y}_2;\theta,x_1\right)/d\theta^2\right]\\ &= \mathrm{E}\left[-d^2\,ln\,f\left(\bar{y}_2|\bar{y}_1;\theta,x_1\right)/d\theta^2 - d^2\,ln\,f\left(\bar{y}_1;\theta,x_1\right)/d\theta^2\right]\\ &= \mathrm{E}\left[-d^2 ln\,f\left(\bar{y}_2|x_2\left(\bar{y}_1\right);\theta,x_1\right)/d\theta^2 - d^2 ln\,f\left(\bar{y}_1;\theta,x_1\right)/d\theta^2\right]\\ &= w_2\mathrm{E}\left[\mu\left(\theta,x_2\left(\bar{y}_1\right)\right)\right] + w_1\mu\left(\theta,x_1\right),\end{aligned}$$

where $w_i = n_i/n$ and $\mu(x,\theta) = \mathrm{E}_{y|x}\left[-d^2 ln\, f\left(y|\,\theta,x\right)/d\theta^2\right]$ gives a measure of information on a single observation conditional of their treatment.

3 The Model: Normal Random Response Variables Having a One Parameter Exponential Mean Function

Consider n subjects with responses

$$y_j = \eta(x,\theta) + \varepsilon_j, \quad \varepsilon_j \sim \mathscr{N}(0,1), \quad j = 1,\ldots,n. \tag{2}$$

Here let $\eta = \eta(x,\theta) = e^{-\theta x}$, $\theta \in \Theta = (0,\infty)$ be a one parameter exponential mean function evaluated at the design point x. As is typically necessary, due to experimental or practical considerations, we bound the design space, i.e., $x \in \mathscr{X} = [a,b]$, $0 < a < b < \infty$.

Fisher [1947] used a variant of this model in which x indicates the number of serial dilutions in a laboratory experiment to illustrate the relationship between information on η and information on θ. With serial dilutions the response is expected to decrease with x. Cochran [1973] elaborated on the experiment that motivated Fisher and used the exponential mean function (as we do) to illustrate statistical complications with nonlinear regression more generally. Complications encountered with model (2) evaluated at a single point are likely to exist, or be exaggerated, with more complicated designs and/or more complicated mean functions.

In particular, complications for finite sample sizes are induced by the mean function being bounded on the interval $(0,1)$ while responses are on $(-\infty,\infty)$. Other functions with this property include, for example, the widely used E_{MAX} model discussed in Dragalin et al. [2007] and in Leonov and Miller [2009]. For the E_{MAX} model,

$$\ln\left[\frac{\theta_1-\eta}{\eta-\theta_2}\right] = \theta_3 + \theta_4 x \Longrightarrow \eta(x,\theta) = \frac{\theta_1+\theta_2 e^{\theta_3+\theta_4 x}}{1+e^{\theta_3+\theta_4 x}}.$$

So η also has a restricted range, yet $y \in (-\infty,\infty)$.

3.1 A Fixed One Point Design

In model (2) the sample mean, $\bar{y}$, is complete and sufficient for θ given x and n, and hence inference can be based on the likelihood

$$\mathscr{L}(\theta|x,\bar{y}) = f(\bar{y}|x,\theta) = \left(\frac{n}{2\pi}\right)^{1/2} \exp\left\{-\frac{n}{2}\left(\bar{y} - e^{-\theta x}\right)^2\right\}.$$

Since the mean function $e^{-\theta x}$ is bounded in $(0,1)$, the likelihood must be maximized separately for $\bar{y} < 0$, $\bar{y} \in (0,1)$ and $\bar{y} > 1$:

1. If $\bar{y} \in (0,1)$, then the MLE is the unique solution to
$$d\ln\ f(y_j|x,\theta)/d\theta = \left(\bar{y} - e^{-\theta x}\right) x e^{-\theta x} = 0. \tag{3}$$
2. If $\bar{y} > 1$, the left side of (3) is a decreasing function of θ and x. The MLE of θ is the solution to $\bar{y} = e^{-\theta x} = 1$, which is 0.
3. If $\bar{y} < 0$, the left side of (3) is a increasing function of θ and x. The MLE of θ is the solution to $\bar{y} = e^{-\theta x} = 0$; in other words, the MLE diverges to infinity. The divergence of the MLE to infinity necessitates the restriction of a search for the MLE to be less than some a predetermined constant $\overline{\theta}$.

In summary, for the one point design

$$\hat{\theta}_n = \begin{cases} \dfrac{-\ln\ \bar{y}}{x}, & \text{if } \bar{y} \in \left(e^{-\overline{\theta}x}, 1\right), \\ 0 & \text{if } \bar{y} \geq 1, \\ \overline{\theta} & \text{if } \bar{y} \leq e^{-\overline{\theta}x}. \end{cases}$$

It is common practice to use $\hat{\theta}_n$ to estimate θ and $[n\mu(x,\theta)]^{-1}$ to approximate $\text{Var}\left[\hat{\theta}_n\right]$. For small sample applications, this approximation should be used with caution because (1) $\hat{\theta}_n$ has large bias; (2) $[n\mu(x,\theta)]^{-1}$ can be quite far from $\text{Var}\left[\hat{\theta}_n\right]$; and (3) $[n\mu(x,\theta)]^{-1}$ must be estimated.

3.2 The Adaptive Stage 2 Treatment

Henceforth, let the subscript i represent stage i. The treatment that maximizes the increment in information in stage 2 is

$$x^* = \arg\max_{x\in\mathscr{X}} [\mu(x,\theta)] = \arg\max_{x\in\mathscr{X}} \left(\frac{d\eta(x,\theta)}{d\theta}\right)^2 = \begin{cases} \theta^{-1}, & \text{if } a \leq \frac{1}{\theta} \leq b \\ b, & \text{if } \frac{1}{\theta} \geq b \\ a, & \text{if } \frac{1}{\theta} \leq a \end{cases} \tag{4}$$

which we call the *locally optimal design point*. A two stage design is one way to deal with the fact that x^* depends on θ. That is, after treating an initial subset of subjects at x_1, use the first stage data to estimate θ in (4). However, the restriction that $x_2 \in [a,b]$ on model (2) imposes yet more boundary constraints. Namely,

$$x_2 = \begin{cases} \hat{\theta}_{n_1}^{-1}, & \text{if } \bar{y}_1 \in \left(e^{-a^{-1}x_1}, e^{-b^{-1}x_1}\right) \\ b, & \text{if } \bar{y}_1 \geq e^{-b^{-1}x_1} \\ a, & \text{if } \bar{y}_1 \leq e^{-a^{-1}x_1}. \end{cases} \tag{5}$$

Now, since a and $\overline{\theta}$ are predetermined constants, one can simplify the procedure and its analysis by selecting $\overline{\theta} = a^{-1}$. This is done henceforth in this paper.

We emphasize that x_2 depends on $\bar{y}_1$, that is, on $\{\varepsilon_{1j}\}_1^{n_1}$; but x_2 is independent of $\{\varepsilon_{2j}\}_1^{n_2}$. Let $\xi_A = \{x_i, w_i\}_{i=1}^2$ denote a two-stage design with x_2 adapted as in (5).

In later comparisons, we employ a locally optimal two-stage design, ξ^*, as a benchmark. This design has the same first stage design $\{x_1, w_1\}$ as the adaptive design, but the second stage design is $\{x^*, w_2\}$, i.e., the second stage uses the unknown optimal treatment.

Since $\bar{y}_1 | x_1$ is distributed $\Phi\left(\sqrt{n_1}\,(\bar{y}_1 - \eta_1)\right)$, where $\Phi(z)$ denotes the cumulative standard normal distribution function, the probabilities that x_2 will lie on the boundaries and in the interior of the parameter space are given by

$$\begin{aligned} \pi_1 &= P(x_2 = a) = P\left(\bar{y}_1 \leq e^{-a^{-1}x_1}\right) = \Phi\left(\sqrt{n_1}\left(e^{-a^{-1}x_1} - e^{-\theta x_1}\right)\right); \\ \pi_3 &= P(x_2 = b) = P\left(\bar{y}_1 \geq e^{-b^{-1}x_1}\right) = 1 - \Phi\left(\sqrt{n_1}\left(e^{-b^{-1}x_1} - e^{-\theta x_1}\right)\right); \\ \pi_2 &= P(a < x_2 < b) = 1 - \pi_1 - \pi_3. \end{aligned}$$

Note that π_1 and π_3 go to zero as $n \to \infty$.

Let $\mathrm{I}(\cdot)$ denote the indicator function. Then the per-subject information is

$$\begin{aligned} M(\xi_A, \theta) = \frac{1}{n}\mathrm{Var}[S] &= w_1 x_1^2 e^{-2\theta x_1} + w_2 \pi_1 a^2 e^{-2\theta a} + w_2 \pi_3 b^2 e^{-2\theta b} \\ &\quad + w_2 \mathrm{E}_{\bar{y}_1}\left[\left(\frac{-x_1}{\ln \bar{y}_1}\right)^2 e^{-2\theta\left(\frac{-x_1}{\ln \bar{y}_1}\right)} \cdot \mathrm{I}\left(e^{-a^{-1}x_1} < \bar{y}_1 < e^{-b^{-1}x_1}\right)\right]. \end{aligned} \tag{6}$$

For finite samples π_1 and π_3 can be large.

4 Selection of the Stage 1 Sample Size

When the stage 1 sample proportion, $w_1 = n_1/n$, is to be fixed, a natural question is how to determine n_1. We say a stage one sample size is good if it is close to

$$n_1^* = \arg \max_{n_1 \in \{1,\ldots,n\}} M(\xi_A, \theta). \tag{7}$$

Consider the following proposition (proof is in the appendix):

Proposition 1 *For model (2) if x^* an interior point of $\mathscr{X}$, then $n_1^* = O(\sqrt{n})$.*

This proposition indicates that it is best to select n_1 proportional to $\sqrt{n}$. This relationship was noted by Luc Pranzanto for a more general model (personal communication, 2012). However, practically this knowledge is of little value.

For finite sample sizes the following two points can be shown. First, there exists a neighborhood for x_1 around x^* such that $n_1^* = n$. This indicates that over an interval of x_1 around x^* the expected per observation information given the initial treatment is greater than that for observations treated at the adaptively selected treatment. Second, provided $\mathscr{X}$ is a sufficiently large interval, there exists a point $x' < x^*$ and $x'' > x^*$ such that for all $x_1 < x'$ and all $x_1 > x''$, $n_1^* < n$. These two points imply that, for a finite sample, the distance between x_1 and x^* may have greater influence on the value of n_1^* than does the value of $\sqrt{n}$. For proof of the existence of these neighborhoods and further discussion, see appendix 9.2.

Based on the preceding discussion, we suggest one assign θ a prior distribution, calculate n_1^* conditional on θ and then average $n_1^*|\theta$ with respect to the prior of θ. That is, we recommend using

$$\tilde{n}_1^* = \int (n_1^*|\theta) d\pi(\theta), \tag{8}$$

where $\pi(\theta)$ denotes a prior distribution of θ.

For example, consider model (2) with $\eta(x,\theta) = e^{-\theta x}$, true parameter value $\theta_t = 1$, $n = 100$, $x_1 = 1$, and $\mathscr{X} = [.25, 10]$. Then noting that $x_2^{-1}(.25) = 4$ and $x_2^{-1}(10) = 1$, let θ be Uniform(.1,4). Figure 1 shows $[n_1^*|\theta]/n$ for $\theta \in (.1,4)$. Note the interval around θ_t where $n_1^* = n$, and that outside this interval $n_1 \leq n$. For $n = 100$, (8) yields $\tilde{n}_1^* = 42$.

[Figure 1 about here.]

5 The Final MLE $\hat{\theta}$ of θ

Since responses are independent conditional on treatment and x_2 is an onto function of $\mathbf{y_1}$, the likelihood is

$$\mathscr{L}(\theta|\mathbf{y_1},\mathbf{y_2},x_1) \propto exp\left\{-\frac{1}{2}\left(n_1\left(\bar{y}_1-\eta\left(x_1,\theta\right)\right)^2+n_2\left(\bar{y}_2-\eta\left(x_2(\bar{y_1},x_1),\theta\right)\right)^2\right)\right\}$$

which has the same form as when x_2 is fixed except now the second stage mean is a random function of $\bar{y}_1$. Henceforth n_1 and n_2 are assumed to be fixed in advance.

After the second stage, the MLE $\hat{\theta}$ using all the data can be found by solving

$$\frac{1}{n}S = w_1\left(\bar{y}_1-e^{-x_1\theta}\right)x_1e^{-x_1\theta}+w_2\left(\bar{y}_2-e^{-x_2\theta}\right)x_2e^{-x_2\theta}=0 \tag{9}$$

subject to boundary conditions, i.e., if $\tilde{\theta}$ is the unique solution to (9), then

$$\hat{\theta}_n = \begin{cases} \tilde{\theta} & \text{if } \tilde{\theta} \in \left(0,a^{-1}\right) \\ 0 & \text{if } \tilde{\theta} \leq 0 \\ a^{-1} & \text{if } \tilde{\theta} \geq a^{-1}. \end{cases}$$

The upper bound a^{-1} is necessary to guarantee $\mathrm{E}[\hat{\theta}_n] < \infty$.

The distributions of $\bar{y}_1$ and $\bar{y}_2$ determine the distribution of $\hat{\theta}$ and hence inference on θ_n. Although, $\bar{y}_1$ is normally distributed, the boundaries of the design space and the adaptive selection of x_2 result in $\bar{y}_2$ following a mixture distribution:

$$f\left(\bar{y}_2\right)=\pi_1 f\left(\bar{y}_2|x_2=a\right)+\pi_3 f\left(\bar{y}_2|x_2=b\right)+\int_a^b f\left(\bar{y}_2|x_2\right)f\left(x_2\right)dx_2.$$

6 Staged Information

Ignoring the dependency induced by selecting x_2 adaptively and treating the responses as independent, one obtains the commonly used information measure

$$M_{ind}(\xi_A,\theta,\hat{\theta}_{n_1})=\sum_1^2 w_i\mu(x_i,\theta)=\sum_1^2 w_i x_i^2 e^{-2\theta x_i} \tag{10}$$

where $\hat{\theta}_{n_1}$ denote the MLE based on stage 1 data alone. The locally optimal design gives the benchmark information

$$M(\xi^*,\theta)=w_1x_1^2e^{-2\theta x_1}+w_2x^{*2}e^{-2\theta x^*}. \tag{11}$$

Since $x_2 \longrightarrow x^*$, $\mu(x_2,\theta) \to \mu(x^*,\theta)$ as $n \longrightarrow \infty$ and because the limit can be passed under the expectation in (6), it follows that

$$M(\xi_A,\theta) \xrightarrow{n\to\infty} w_1\mu(x_1,\theta) + w_2\mu(x^*,\theta) = M(\xi^*,\theta).$$

In general, one needs to verify that (6) and (10) have the same asymptotic limits. In particular, different limits are likely in models where the support of x_1 contains θ.

6.1 Comparison of $M(\xi_A,\theta)$, $M(\xi^*,\theta)$ and $\left[n\mathbf{Var}[\hat{\theta}_n]\right]^{-1}$.

Figure 2 compares the (simulated) main object of interest: $\left[n\mathrm{Var}[\hat{\theta}_n]\right]^{-1}$ with $M(\xi_A,\theta)$ and $M(\xi^*,\theta)$. $M(\xi^*,\theta)$ decreases linearly with w_1; it attains its maximum when all subjects are treated at $x^* = 1$, i.e., $w_1 = 0$, and its minimum when x^* is never used. $M(\xi_A,\theta)$ achieves a maximum when $w_1 \approx 0.45$ and only when $w_1 = 1$ does it equal $M(\xi^*,\theta)$. For values $w_1 \geq 0.20$ $M(\xi_A,\theta)$ has very little variability.

Note that $\left[n\mathrm{Var}[\hat{\theta}_n]\right]^{-1}$ is overestimated by both $M(\xi_A,\theta)$ and $M(\xi^*,\theta)$. Thus $M(\xi^*,\theta)$ is misleading as a benchmark. The difference between $M(\xi_A,\theta)$ and $\left[n\mathrm{Var}[\hat{\theta}_n]\right]^{-1}$ is small only when w_1 is very near 0. This difference increases dramatically as w_1 increases to 1.

[Figure 2 about here.]

6.2 Comparison of $M(\xi_A,\theta)$, $M_{ind}(\xi_A,\theta,\hat{\theta}_{n_1})$ and $\left[n\mathbf{Var}[\hat{\theta}_n]\right]^{-1}$.

A primary interest is the difference between $M(\xi_A,\theta)$ and the commonly used $M_{ind}(\xi_A,\theta,\hat{\theta}_{n_1})$. Since, for the one parameter exponential mean function, one can pass the limit through the expectation, it follows that $M(\xi_A,\theta) - M_{ind}(\xi_A,\theta,\hat{\theta}_{n_1}) \xrightarrow{n\to\infty} 0$. Even though the two measures are asymptotically equivalent, one would like to determine which is closer to $\left[n\mathrm{Var}[\hat{\theta}_n]\right]^{-1}$ when the sample size is finite. Unfortunately, because $\hat{\theta}_n$ is the solution to the likelihood equation subject to boundary conditions, its distribution and variance cannot be found explicitly for a fixed n. As a result, it is not possible to determine the distance of either measure from $\left[n\mathrm{Var}[\hat{\theta}_n]\right]^{-1}$ explicitly, for unknown θ. However, rewriting (1) for $\tilde{\theta}_n = \hat{\theta}_n$ as

$$nM(\xi_A,\theta) \geq (1 + b'(\theta))^2 \mathrm{Var}^{-1}[\hat{\theta}_n]$$

one sees that the lower bound of $\left[\mathrm{Var}[\hat{\theta}_n]\right]^{-1}$ is proportional to $nM(\xi_A,\theta)$ given θ. Furthermore, simulations for $n = 100$ suggest that

$$\arg\max_{x\in\mathscr{X}} M(\xi_A,\theta) = \arg\max_{x\in\mathscr{X}} \mu(x,\theta) \approx \arg\max_{x\in\mathscr{X}} Var^{-1}\left[\hat{\theta}_n\right].$$

This provides some small sample justification for procedure (4).

An important distinction between $M(\xi_A,\theta)$ and $M_{ind}(\xi_A,\theta,\hat{\theta}_{n_1})$ deals with their relationship to $\hat{\theta}_{n_1}$: in particular, $M(\xi_A,\theta)$ is constant given the design, where $M_{ind}(\xi_A,\theta,\hat{\theta}_{n_1})$ is a function of $\hat{\theta}_{n_1}$, and thus a function of the first stage data.

Figure 3 plots $M_{ind}(\xi_A,\theta,\hat{\theta}_{n_1})$ and $M(\xi_A,\theta)$ for $x_1 = 2$, $\theta_t = 1$, $w_1 = 0.20$ as functions of $\hat{\theta}_{n_1}$. Note that $M_{ind}(\xi_A,\theta,\hat{\theta}_{n_1})$ will not change as n increases, but the probability $M_{ind}(\xi_A,\theta,\hat{\theta}_{n_1})$ is close to $M(\xi_A,\theta)$ increases. For example when $n = 100$, $M(\xi_A,\theta) \approx 0.081$ and $\mathrm{P}\big(\left(M(\xi_A,\theta) - M_{ind}(\xi_A,\theta,\hat{\theta}_{n_1})\right)/M(\xi_A,\theta) < .1\big) \approx .61$. If the sample size is increased to 400, $M(\xi_A,\theta) \approx 0.103$ and the same probability would be approximately 0.806.

[Figure 3 about here.]

This discussion has been aimed at illuminating the appropriateness of using the information measure $M(\xi_A,\theta)$ to construct two-stage designs with finite sample sizes. Using $M_{ind}(\xi_A,\theta,\hat{\theta}_{n_1})$ can be justified only with large stage 1 sample sizes which are common in many environmental sampling schemes, but not in clinical trials. However, motivated by the Cramer-Rao lower bound and supported by simulations, $nM(\xi_A,\theta)$ and $Var^{-1}[\hat{\theta}_n]$ appear proportional to a constant except for extreme values of w_1. These findings support using $nM(\xi_A,\theta)$ to make design decisions.

7 Estimates of Staged Information

A number of estimates of the information are possible.

Plug in information estimates:

$$\begin{aligned}
M_{ind}(\xi_A,\hat{\theta}_n,\hat{\theta}_{n_1}) &= \left[w_1\mu(x_1,\theta)+w_2\mu(x_2(x_1,\bar{y}_1),\theta)\right]_{\theta=\hat{\theta}_n}\\
&= w_1x_1^2e^{-2\hat{\theta}_nx_1}+w_2(\hat{\theta}_{n_1}^{-2}e^{-2\hat{\theta}_n\hat{\theta}_{n_1}^{-1}})\\
M(\xi_A,\hat{\theta}_n) &= \left[w_1\mu\left(x_1,\theta\right)+w_2\mathrm{E}_{\bar{y}_1}\left[\mu\left(x_2(\bar{y}_1,x_1),\theta\right)\right]\right]_{\theta=\hat{\theta}_n}\\
&= w_1x_1^2e^{-2\hat{\theta}_nx_1}+w_2\left(\hat{\pi}_1a^2e^{-2\hat{\theta}_na}+\hat{\pi}_3b^2e^{-2\hat{\theta}b}\right.\\
&\quad \left.+\hat{E}_{\bar{y}_1}\left(\left(\frac{x_1}{\ln\bar{y}_1}\right)^2e^{2\theta\frac{x_1}{\ln\bar{y}_1}}|e^{-a^{-1}x_1}<\bar{y}_1<e^{-b^{-1}x_1}\right)\right).
\end{aligned}$$

observed information estimate:

$$\begin{aligned}
M_{obs}(\xi_A,\hat{\theta}_n,\hat{\theta}_{n_1}) &= -\left(d^2\ln f\left(y_1,y_2|\theta,x_1\right)/d\theta^2\right)|_{\theta=\hat{\theta}_n}\\
&= M_{ind}(\xi_A,\hat{\theta}_n,\hat{\theta}_{n_1})-\sum_{i=1}^{2}w_i(\bar{y}_i-e^{-\hat{\theta}_nx_i})x_i^2e^{-2\hat{\theta}_nx_i}.
\end{aligned}$$

To compare estimators, consider a simulation where $n=100$, $w_1=0.20$, $x_1=2.0$, $\theta_t=1.0$, $a=0.25$, and $b=10$. Figures 4(b), 4(a) and 4(c), plots the quartiles and mean values of $M_{ind}(\xi_A,\hat{\theta}_n,\hat{\theta}_{n_1})$, $M(\xi_A,\hat{\theta})$, and $M_{obs}(\xi_A,\theta_n,\hat{\theta}_{n_1})$, respectively, against $\left[n\hat{\mathrm{V}}\mathrm{ar}[\hat{\theta}]\right]^{-1}$. In these pictures, all three estimators appear to perform very similarly. However, a significant difference appears in Figure 5 which compares the the frequency with which

$$\left|M_i-\left[n\hat{\mathrm{V}}\mathrm{ar}[\hat{\theta}]\right]^{-1}\right|>\left|M_j-\left[n\hat{\mathrm{V}}\mathrm{ar}[\hat{\theta}]\right]^{-1}\right|,$$

$\{M_i,M_j\}\subset\{M_{ind}(\xi_A,\hat{\theta}_n,\hat{\theta}_{n_1}),M(\xi_A,\hat{\theta}_n),M_{obs}(\xi_A,\theta_n,\hat{\theta}_{n_1})\}$, $i\neq j$. Thus each line in Figure 5 represents a comparison of two measures, and it can be seen that $M(\xi_A,\hat{\theta}_n)$ dominates both $\hat{M}_{obs}(\xi_A,\theta_n,\hat{\theta}_{n_1})$ and $M_{ind}(\xi_A,\hat{\theta}_n,\hat{\theta}_{n_1})$ for nearly all values of w_1.

[Figure 4 about here.]

[Figure 5 about here.]

8 Discussion

We have explored the information in a two-stage adaptive optimal design in the context of a nonlinear regression model with standard normal errors and exponential mean function. We introduced a procedure for deciding on the first stage sample size. This procedure's usefulness is not restricted to model (2) with the exponential mean function. It is possible to improve upon our suggested method by attempting to select $n_1 = \arg\min_{n_1 \in \{1,\ldots,n\}} \mathrm{Var}\left[\hat{\theta}_n\right]$ which for finite sample could be significantly different that n_1^*, as can be seen in Figure 2. The procedure would remain almost the same except instead one would use Monte Carlo simulations to approximate $\mathrm{Var}[\hat{\theta}_n]$ for given values of θ.

From a theoretical perspective, we compared the variance of the score function with an analogous information measure derived under the incorrect assumption of independence. A numeric example demonstrated that the independent measure is of little use from a design perspective unless w_1 is very near 1. Potential uses for the true Fisher information measure were addressed.

Using a simulated example, information was evaluated from an analysis perspective. Three different estimators were examined; the plug in estimate of $\mathrm{Var}[S]$, the analogous estimate under independence and the observed information. The observed information is shown by Yao and Flournoy [2010] to fluctuate randomly, asymmetrically around $\mu(x^*, \theta)$, yet to converge to $\mu(x^*, \theta)$ as $n \to \infty$. Efron and Hinkley [1978] and Lindsay and Li [1997] argue that the observed information is to be preferred over Fisher's information for analysis, but our simulations call their argument into question. In fact, all three estimators perform almost imperceptibly similar with only a small advantage for the plug in estimator of the $\mathrm{Var}[S]$ in that it is slightly closer to the $\left[n\mathrm{Var}[\hat{\theta}_n]\right]^{-1}$. That the plug in estimator under independence is competitive with other estimators we examined should reassure practitioners who regularly use it due to its much simpler form. Of course, we have evaluated only one specific mean function and caution is recommended when using other mean functions.

We considered the case where θ is a fixed positive real value as in dilution assays. Exponential growth models, where $\theta \in (-\infty, 0)$, also have numerous applications, for example in nuclear chair reactions, numbers of micro-organisms, spread of a virus, compound interest. Results for growth models are analogous. The procedures can also be generalized for more complex mean functions.

This paper addresses w_i fixed, and as can be seen in the plots of the theoretical information measures, values of $w_i < 0.10$ correspond to small values of information. However, often in practice it is the case that 0.10 is the maximum stage one sample proportion considered. For this reason, it will be worthwhile to examine the case where n_1 is small.

9 Appendix

9.1 Proof of Proposition 1

Let $\bar{\varepsilon} = \bar{y} - \eta(x_1, \theta)$. Then

$$n_1^* = \arg \max_{n_1 \in \{1,\ldots,n\}} M(\xi_A, \theta) = \arg \max_{n_1 \in \{1,\ldots,n\}} \left[\frac{n_1}{n} \mu(x_1, \theta) + \left(1 - \frac{n_1}{n}\right) \mu(x_2, \theta) \right]$$

where

$$\mu(x_2, \theta) = \mathrm{E}_{\overline{\varepsilon}} \left(\left(\frac{d\eta(x_2(x_1, \eta(x_1, \theta) + \overline{\varepsilon}), \theta)}{d\theta} \right)^2 \right).$$

Taking a Taylor expansion of $[d\eta(x_2(x_1, \eta(x_1, \theta) + \overline{\varepsilon}), \theta)/d\theta]^2$ around $\overline{\varepsilon} = 0$,

$$\mathrm{E}_{\overline{\varepsilon}} \left(\left(\frac{d\eta(x_2(x_1, \eta(x_1, \theta) + \overline{\varepsilon}), \theta)}{d\theta} \right)^2 \right) = \left(\frac{d\eta(x^*, \theta)}{d\theta} \right)^2 + \frac{c_2}{n_1} + \sum_{k=2}^{\infty} \frac{c_{2k}}{n_1^k (2k)!},$$

where $c_l = \left[\frac{d^l}{d\overline{\varepsilon}^l} \left(\frac{d\eta(x_2(x_1, \eta(x_1, \theta) + \overline{\varepsilon}), \theta)}{d\theta} \right)^2 \right]_{\overline{\varepsilon}=0}$. Setting $a = \left(\frac{d\eta(x_1, \theta)}{d\theta} \right)^2 - \left(\frac{d\eta(x^*, \theta)}{d\theta} \right)^2$, note that $a < 0$ and $\sum_{k=2}^{\infty} \frac{c_{2k}}{n_1^k (2k)!} = O\left((nn_1)^{-1}\right)$. Therefore,

$$n_1^* = \arg \max_{n_1 \in \{1,\ldots,n\}} \left[\frac{n_1}{n} a + \frac{1}{n_1} c_2 + O\left((nn_1)^{-1}\right) \right].$$

It can be shown that since x^* is and interior point of $\mathscr{X}$, c_2 evaluated in the neighborhood of x^* is negative. Thus to approximate n_1^* set the derivative of the argument equal to 0 and solve to get

$$n_1 = \left(\frac{1}{c_2} \left(\frac{1}{n} a - O\left(n^{-1} n_1^{-2}\right) \right) \right)^{-\frac{1}{2}}$$

which implies the result.

9.2 Proof of neighborhoods described in section 4

Write

$$n_1^* = \arg \max_{n_1 \in \{1,\ldots,n\}} \left[\frac{n_1}{n} \mu(x_1, \theta) + \left(1 - \frac{n_1}{n}\right) \mathrm{E}_{\bar{y}_1} \left[\mu(x_2(\bar{y}_1), \theta)\right] \right].$$

It is sufficient to show that there is a neighborhood of x_1 around x^* such that

$$\mathrm{E}_{\bar{y}_1} \left[\mu(x_2(\bar{y}_1), \theta)\right] < \mu(x_1, \theta). \tag{12}$$

From Jensen's Inequality

$$\mathrm{E}_{\bar{y}_1} \left[\mu(x_2(\bar{y}_1), \theta)\right] < \mu\left(x_2\left(\mathrm{E}_{\bar{y}_1}[\bar{y}_1]\right), \theta\right) = \mu(x^*, \theta).$$

Thus if $x_1 = x^*$, (12) holds. Because $\mu(x, \theta)$ is continuous and the equality is strict the existence of the interval around x^* is confirmed.

To show that neighborhoods of x_1 exist such that a subject treated in those neighborhoods provides less expected information than does a subject treated adaptively, it is sufficient to show there exists an $x' > x^*$ and $x'' < x^*$ such that for all $x_1 > x'$ and for all $x_1 < x''$

$$\mathrm{E}_{\bar{y}_1} \left[\mu(x_2(\bar{y}_1), \theta)\right] < \mu(x_1, \theta). \tag{13}$$

Expand and rewrite (13) as

$$\pi_1 \left(\frac{a}{x_1}\right)^2 e^{2\theta(x_1 - a)} + \pi_3 \left(\frac{b}{x_1}\right)^2 e^{2\theta(x_1 - b)} + \mathrm{E}_{\bar{y}_1} \left[\left(\frac{1}{\ln \bar{y}_1}\right)^2 e^{2\theta x_1 \left(1 + \frac{1}{\ln \bar{y}_1}\right)} \right] > 1. \tag{14}$$

Note each term in the left side of (14) is strictly greater than 0, $x_1 \in [a, b]$ and $\theta \in [b^{-1}, a^{-1}]$. Consider two cases. Case 1: If $a < \frac{1}{\theta} < x_1 < b$, then π_1 increases to 1/2 and $\left(\frac{a}{x_1}\right)^2 e^{2\theta(x_1 - a)}$ strictly increases as x_1 increases. Then letting x' be the unique solution to

$$\left(\frac{a}{x_1}\right)^2 e^{2\theta(x_1 - a)} = 0,$$

(14) will be satisfied for all $x_1 > x'$, provided of course that b is sufficiently large. Case 2: If $a < x_1 < \frac{1}{\theta} < b$ then π_3 increases to 1/2 as x_1 increases and

$\left(\frac{b}{x_1}\right)^2 e^{2\theta(x_1-b)}$, strictly increases as x_1 decreases. Then letting x'' be the unique solution to

$$\left(\frac{b}{x_1}\right)^2 e^{2\theta(x_1-b)} = 0,$$

(14) will be satisfied for all $x_1 < x''$, provided of course that a is sufficiently small.

Remark. To show an interval of x_1 exists such that $n_1^* = n$ we used the fact that for our procedure $\mu(x_2(\bar{y}_1), \theta)$ is concave with respect to $\bar{y}_1$. However, noting that $\mu(x, \theta)$ is concave and with maximum at x^* one can use Jensen's inequality to argue that for any concave or convex $x_2(\bar{y}_1)$ (12) still holds. In fact (12) can only be an equality for $x_1 = x^*$ if $\mathrm{E}[x_2(\bar{y}_1)] = x^*$. So we conclude that in most practical examples such a neighborhood will exist.

Acknowledgements Ping Yao is grateful to have had a stimulating internship with Val Fedorov at GlaxoSmithKline where this research idea germinated. Nancy Flournoy and Adam Lane sincerely thank Rosemary Bailey and the Isaac Newton Institute For Mathematical Sciences for hosting them in Cambridge, UK during the fall 2011 and Val Fedorov for comments and suggestions during their stay in Cambridge.

Bibliography

Box, G., Hunter, W., 1965. Sequential design of experiments for nonlinear models. In: Korth, J. J. (Ed.), Proceedings of the Scientific Computing Symposium: Statistics. White Plains: IBM, pp. 113–137.

Chernoff, H., 1953. Locally optimal designs for estimating parameters. Annals of Mathematical Statistics 24, 586–602.

Cochran, W., 1973. Experiments for nonlinear functions. Journal of the American Statistical Association 68, 771–781.

Cox, D. R., Hinkley, D. V., 1974. Theoretical Statistics. Chapman and Hall, London, UK.

Dragalin, V., Fedorov, V., Wu, Y., 2008. Adaptive designs for selecting drug combinations based on efficacy-toxicity response. Journal of Statistical Planning and Inference 2, 352–373.

Dragalin, V., Hsuan, F., Padmanabhan, S. K., 2007. Adaptive designs for dose–finding studies based on sigmoid E_{MAX} model. Journal of Biopharmaceutical Statistics 17, 1051–1070.

Efron, B., Hinkley, D., 1978. Assessing the accuracy of the maximum likelihood estimate: observed versus Fisher information (with discussion). Biometrika 65, 457–483.

Fedorov, V., 1972. Theory of Optimal Experiments. Academic Press: New York.

Fisher, R. A., 1947. The Design of Experiments. Oliver and Boyd, Edinburgh, Scotland.

Hogg, R. V., McKean, J. W., Craig, A. T., 2005. Introduction to Mathematical Statistics. Pearson Education.

Leonov, S., Miller, S., 2009. An adaptive design for the $\mathrm{E}_{\mathrm{MAX}}$ model and its application in clinical trials. Journal of Biopharmaceutical Statistics 19, 360–385.

Lindsay, B., Li, B., 1997. On second order optimality of the observed Fisher information. Annals of Statistics 25, 2172–2199.

Silvey, S., 1980. Optimal Design: An Introduction to the Theory for Parameter Estimation. Chapman and Hall, London, UK.

Stein, C., 1945. A two-sample test for a linear hypothesis whose power is independent of the variance. The Annals of Mathematical Statistics 16 (3), 243–258.

White, L. V., 1975. The optimal design of experiments for estimation of nonlinear models. Dissertation. University of London, UK.

Yao, P., Flournoy, N., 2010. MoDa 9 – Advances in Model-Oriented Design and Analysis. Springer (eds. Giovagnoli, A., Atkinson, A.C., Torsney, B., May, C.), Ch. Information in a Two-stage Adaptive Optimal Design for Normal Random Variables having a One Parameter Exponential Mean Function, pp. 229–236.

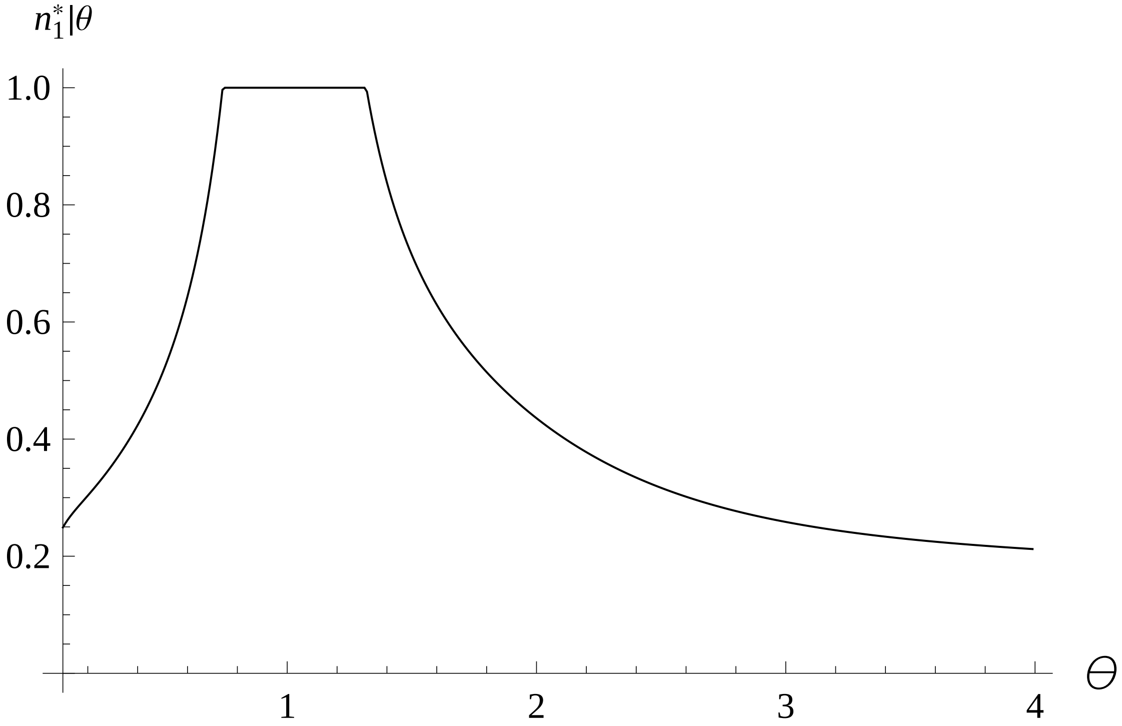


Figure 1: Optimal Sample Size Allocated to Stage 1 versus θ; $x_1 = 1$

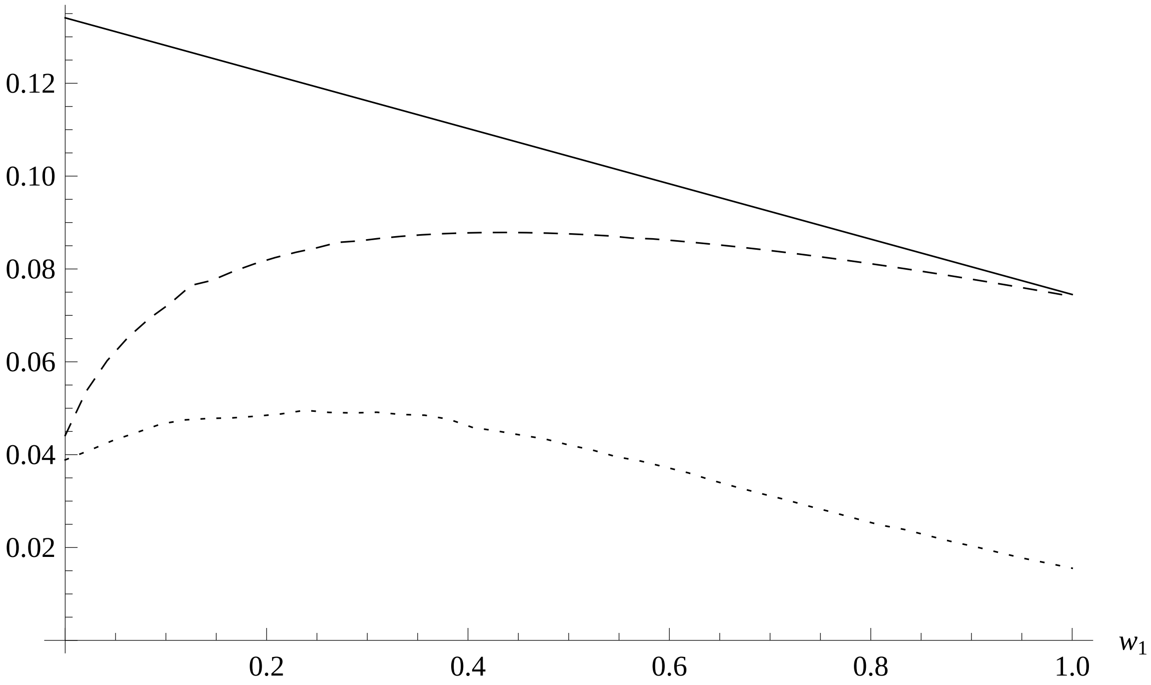


Figure 2: Information measures by w_1. The solid, dashed, and dotted lines represent $M(\xi^*,\theta)$, $M(\xi_A,\theta)$ and $\left[n\widehat{\text{Var}}\left(\hat{\theta}_n\right)\right]^{-1}$ by w_1, respectively

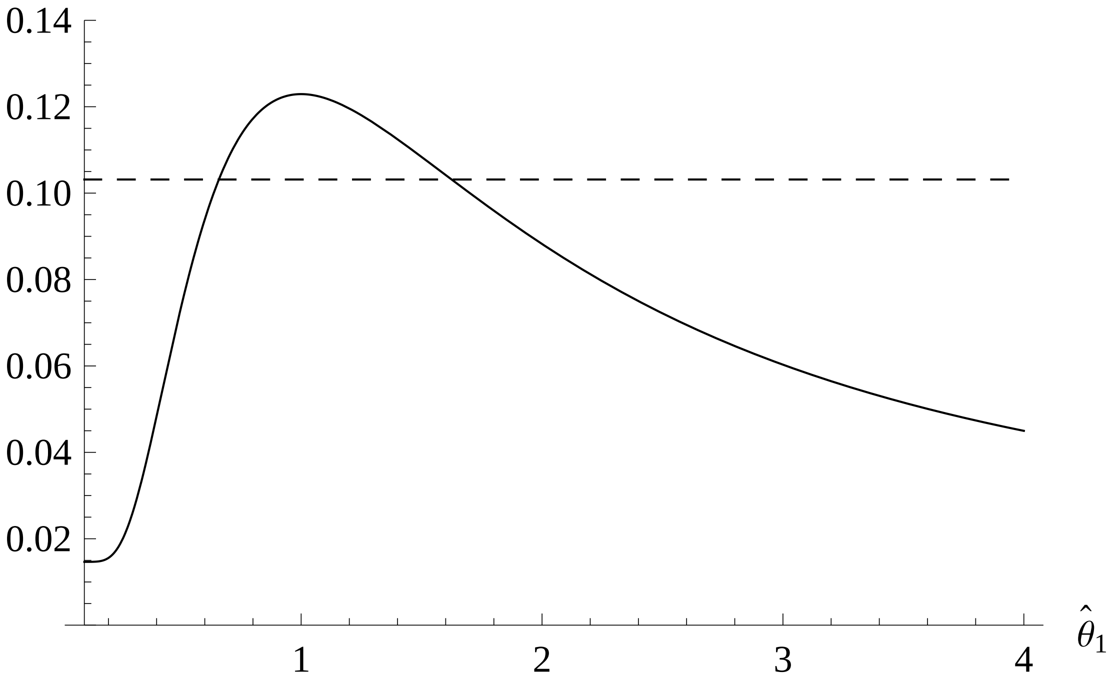
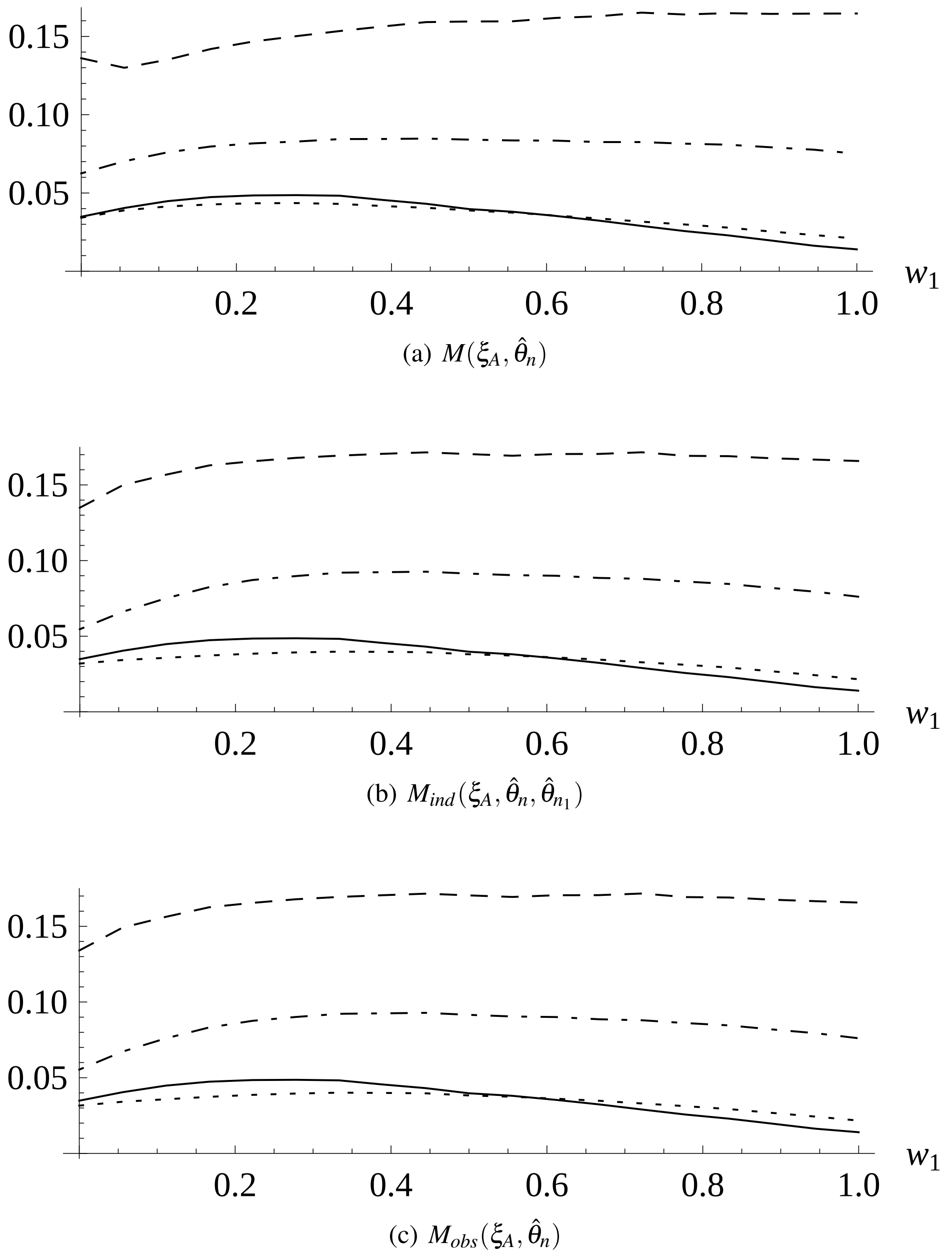


Figure 3: $M_{ind}(\xi_A, \theta, \hat{\theta}_{n_1})$, the solid line, and $M(\xi_A, \theta)$, the dashed line, plotted by $\hat{\theta}_{n_1}$ at values $x_1 = 2$, $\theta_t = 1$ and $w_1 = 0.20$.



(a) $M(\xi_A, \hat{\theta}_n)$

(b) $M_{ind}(\xi_A, \hat{\theta}_n, \hat{\theta}_{n_1})$

(c) $M_{obs}(\xi_A, \hat{\theta}_n)$

Figure 4: Estimates of Information Compared with Simulated $\left[n\widehat{\mathrm{Var}}\left[\hat{\theta}_n\right]\right]^{-1}$ as Functions of w_1. The solid lines represent $\left[n\widehat{\mathrm{Var}}\left[\hat{\theta}_n\right]\right]^{-1}$. The dotted,dot-dashed, and dashed lines are the 1^{st}, 2^{nd}, and 3^{rd} quartiles of the three information measure estimates.

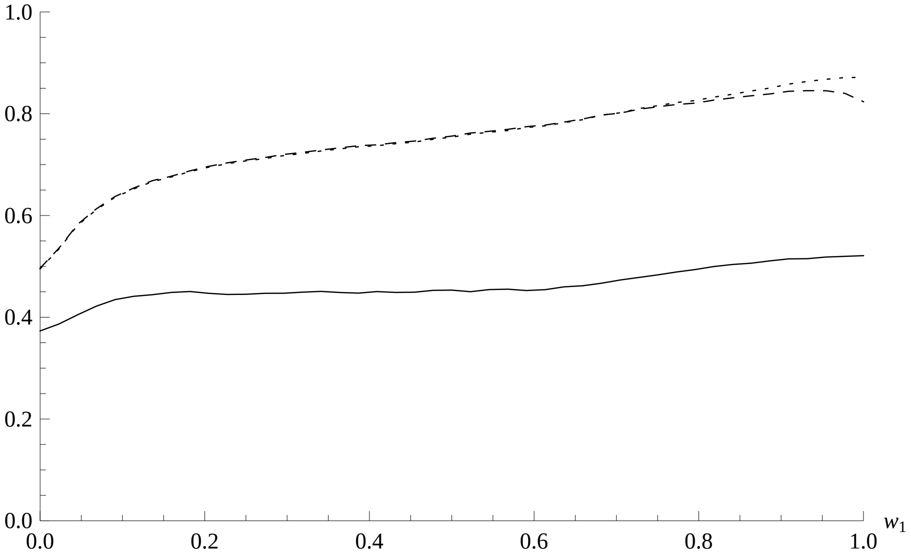


Figure 5: Proportion of times one information measure estimate is closer to $\left[n\widehat{\mathrm{Var}}\left(\hat{\theta}_n\right)\right]^{-1}$ than another. The dotted line, $M(\xi_A, \hat{\theta}_n, \theta)$ is closer than $M_{ind}(\xi_A, \hat{\theta}_n, \bar{y}_1, \theta)$, the dashed line, $M(\xi_A, \hat{\theta}_n, \theta)$ is closer than $M_{obs}(\xi_A, \hat{\theta}_n, \hat{\theta}_{n_1})$ and the solid line, $M_{obs}(\xi_A, \hat{\theta}_n, \hat{\theta}_{n_1})$ is closer than $M_{ind}(\xi_A, \hat{\theta}_n, \bar{y}_1, \theta)$.