Sequential estimation with adaptive designs and its applications

Yuan-chin Ivan Chang
Institute of Statistical Science
Academia Sinica, Taipei, Taiwan

1. Application to CARA Designs

From an ethical viewpoint, it is desirable to minimize the number of subjects allocated to inferior treatments in the course of a clinical trial without jeopardizing the generation of useful and meaningful statistical inferences. The response adaptive (RA) design in clinical trials (Zelen and Wei 1995 and Hu and Rosenberger 2006) is dedicated to this purpose. Due to innovation in genomic technologies and the nature of developing targeted drugs (Simon and Maitournam 2005), it is natural to incorporate the information available on individual covariates that have a strong influence on responses to a model, since they may be associated with the efficacy of treatments. The sequential method is a natural choice for a CARA design based clinical trial (Hu and Rosenberger (2006)); however, it is rare to find literature regarding the application of stopping rules for the sequential estimation procedure based on CARA designs.

Let \( N_{m,k} \) be the number of subjects assigned to treatment \( k \) during the first \( m \) assignments and \( N_m = (N_{m,1}, \ldots, N_{m,K}) \). Suppose that \( \{Y_{m,k}, m = 1, 2, \ldots, k = 1, \ldots, K\} \) denotes responses of the \( m \)-th subject to the \( k \)-th treatment and \( Y_m = (Y_{m,1}, \ldots, Y_{m,K}) \). Let \( \xi_m \) be the covariates of the \( m \)-th subject. Suppose that \( X_1, X_2, \ldots \) is the sequence of random assignment treatments, and \( X_m = (X_{m,1}, \ldots, X_{m,K}) \), \( X_{m,k} \in \{0, 1\} \), denotes assignment of treatment \( k \) to the \( m \)-th subject. Then \( X_{m,k} = 1 \) for some \( k \) and \( \sum_{k=1}^{K} X_{m,k} = 1 \). That is, each subject is allocated to one treatment only. Hence, it follows that the response of subject \( m \) to the treatment \( k \), \( Y_{m,k} \), is observed only if \( X_{m,k} = 1 \). (Note that this implies that \( N_m = \sum_{i=1}^{m} X_i \).)

Define \( X_m = \sigma(X_1, \ldots, X_m) \), \( Y_m = \sigma(Y_1, \ldots, Y_m) \), and \( Z_m = \sigma(\xi_1, \ldots, \xi_m) \), \( \xi_i \in \mathbb{R}^p \), be the corresponding \( \sigma \)-fields. Let \( F_m = \sigma(X_m, Y_m, Z_m) \), then a general CARA design is defined as \( \psi_m = E[X_m|F_{m-1}, \xi_m] = E[X_m|X_{m-1}, Y_{m-1}, Z_m] \), where \( \psi_m \) is actually a vector of randomization probabilities for treatments \( 1, \ldots, K \). Suppose that for each \( m \geq 1 \), the responses and covariate vector satisfy

\[
E[Y_{m,k}|\xi_m] = \mu_k(\theta_k, \xi_m),
\]

where \( \mu_k(\cdot, \cdot) \) are known functions, \( V_k \) denotes the covariance matrix based on Equation (1) and \( \theta_k \in \mathbb{R}^p \) for \( k = 1, \ldots, K \). The asymptotic properties of the estimate of \( \theta = (\theta_1, \ldots, \theta_K) \) and allocation function under such a general CARA design has been discussed in Zhang et al. (2007). The estimation of \( \theta \) is the primary goal in a clinical trial. Thus, it will be beneficial if treatment effects can be estimated with a certain accuracy using a minimum required sample size whilst simultaneously still retaining the good allocation properties. Since, in a CARA design, the design at the current stage depends on the past history, sequential analysis is the statistical tool of choice.

Sequential estimation of treatment effects Suppose no prior information about the effects of treatments is available. In order to estimate the treatment effects, at the beginning, we need to assign \( m_0(> 0) \) subjects to each treatment using restricted randomization. Hence, when we allocate the \( m \)-th subject \( m > Km_0 \), there are already \( m-1 \) observations, \( \{(X_1, Y_1, \xi_1), \ldots, (X_{m-1}, Y_{m-1}, \xi_{m-1})\} \), collected. Thus, we assign the \( m \)-th subject to the treatment \( k \) with probability \( \psi_k = P(X_{m,k} = 1|F_{m-1}, \xi_m) = \pi_k(\theta_{m-1}, \xi_m) \), where \( \pi_{m-1} \) is the maximum quasi-likelihood estimate of \( \theta \) based on the previous \( m-1 \) observations and \( \pi_k(\cdot, \cdot) \) is the true allocation probability for treatment \( k \) and the given covariate. Assume further that \( \mu_k(\theta_k, \xi_m) = \mu_k(\xi_m, \theta_k) \) for each \( m \geq 1 \). Hence, it follows from Equation (1) and assuming the existence of variance that the method of generalized linear models (quasi-likelihood)
can be applied (McCullagh and Nelder 1989). Assume that \( \theta_k \in \Theta_k \subseteq \mathbb{R}^p \) is bounded for \( k = 1, \ldots, K \), and let the parameter space \( \Theta = \prod_{k=1}^{K} \Theta_k \).

Under the above assumptions (see also Condition A of Zhang et al. (2007), Theorem 2.1), it is proved that as \( \min(N_{m,k}, k = 1, \ldots, K) \) goes to infinity, \( \sqrt{n}(\hat{\theta} - \theta) \to_{L^2} N(0, \mathbf{V}) \), where \( \mathbf{V} = \text{diag}\{V_1, \ldots, V_K\} \).

Based on the asymptotic normality of \( \hat{\theta} \), the sequential method is employed for estimating the confidence set of \( \theta = (\theta_1, \ldots, \theta_K) \). Define \( R = \{ \theta \in \Theta : n(\hat{\theta} - \theta)(\mathbf{V}^{-1}(\hat{\theta} - \theta) \leq C_\alpha^2 \} \), where \( C_\alpha^2 \) is the constant such that \( P(\chi^2(p \cdot K) \geq C_\alpha^2) \leq \alpha \). The asymptotic normality of \( \hat{\theta} \) implies that \( P(\theta \in R) \approx 1 - \alpha \) as the sample size becomes large.

Although large sample results guarantee the performance of estimates and some asymptotic properties of CARA designs, we want to know just how large a sample size is needed to guarantee a satisfactory performance in a practical sense. Moreover, no matter how high the coverage probability is, the confidence set becomes less useful if the size of the confidence set becomes too large. Now, suppose we further require that the maximum axis of \( R \) is no larger than \( 2\delta \) for some \( \delta > 0 \), then the minimum sample size to achieve this goal is \( n \Lambda_{\min}(\mathbf{V}^{-1}) \geq \frac{C_\alpha^2}{\delta^2} \). Equivalently, the above inequality can be re-written as

\[
n \geq \frac{C_\alpha^2 \Lambda_{\max}(\mathbf{V})}{\delta^2},
\]

where notations \( \Lambda_{\max}(A) \) and \( \Lambda_{\min}(A) \) denote the maximum and minimum eigenvalues of matrix \( A \), respectively. Let \( \mathcal{R}_\delta \) denote the corresponding confidence ellipsoid for given \( \delta \). So, once \( \delta > 0 \) is specified, the maximum axis of confidence ellipsoid \( \mathcal{R}_\delta \) is no greater than \( 2\delta \). The constant \( \delta \) here is used as a measure of precision of the confidence ellipsoid \( \mathcal{R}_\delta \). Please refer to Siegmund (1985), Albert (1966) and Ghosh and Sen (1991) for other measures of confidence sets.

If \( \mathbf{V} \) is known, then the optimal sample size required to construct a confidence ellipsoid \( \mathcal{R}_\delta \) with the required maximum axis no greater than \( 2\delta \) is \( n_{\text{opt}} = \text{first n such that } n \geq \frac{C_\alpha \Lambda_{\max}(\mathbf{V})}{\delta^2} \). Since the variance matrix \( \mathbf{V} \) is usually unknown, the above optimal sample size is not available. Replacing the unknown \( \mathbf{V} \) in Equation (2) with its consistent estimate \( \hat{\mathbf{V}} \) (to be defined later), a stopping rule to construct such a fixed size confidence ellipsoid is suggested:

\[
\tau_\delta = \inf\{n \geq n_0 : n \geq \frac{C_\alpha^2 \Lambda_{\max}(\hat{\mathbf{V}})}{\delta^2}\},
\]

where \( n_0 \geq K n_0 \) is the minimum initial sample size and \( n_0 \) is the initial sample size for each treatment. Similarly, we then define \( \mathcal{R}_\delta = \{ \theta \in \Theta : n(\hat{\theta} - \theta)(\hat{\mathbf{V}}^{-1}(\hat{\theta} - \theta) \leq C_\alpha^2 \} \). It follows from the strong consistency of \( \hat{\theta} \), if \( \hat{\mathbf{V}} \) is also a strongly consistent estimate of \( \mathbf{V} \), then \( \lim_{n \to \infty} P(\theta \in \mathcal{R}_\delta) = 1 - \alpha \). That is, \( \mathcal{R}_\delta \) is a confidence ellipsoid of \( \theta \) with coverage probability \( 1 - \alpha \), asymptotically.

It follows from the definition of \( \tau_\delta \) that, when the sequential sampling stops, the confidence ellipsoid will have its maximum axis no greater than \( 2\delta \). However, it is also known that there is no guarantee that \( \hat{\theta} \) will have the same asymptotic distribution if we replace the fixed sample size with a random sample size \( \tau_\delta \). Although the sequential estimation procedure provides a way to control the size of the confidence set by utilizing a stopping rule, it is interesting to know whether the asymptotic properties in Zhang et al. (2007) are still adhered to under such a randomly stopped criterion.

Suppose that allocation function \( \pi(\cdot, \cdot) = (\pi_1(\cdot, \cdot), \ldots, \pi_K(\cdot, \cdot)) \) and satisfies the following conditions:

\[\sum_{k=1}^{K} \pi_k = 1 \quad \text{and} \quad 0 < \pi_k = E[z_k|\pi_k(\theta, z)] < 1, \quad k = 1, \ldots, K.\]

\[\text{(C2) For fixed } \xi, \pi_k(\theta, \xi) > 0 \text{ is a continuous function of } \theta \text{ and is differentiable with respect to } \theta \text{ such that } v_k(\hat{\theta}) = v_k(\theta) + (\hat{\theta} - \theta)(\partial v_k/\partial \theta)' + o(||\hat{\theta} - \theta||^{1+\varepsilon}) \text{ for some } \varepsilon > 0.\]

The condition \( \pi_k > 0 \) for each \( k = 1, \ldots, K \) on the allocation function guarantees that subjects will be allocated to individual treatments, eventually. Thus, this condition also affirms that with probability one the design matrix is non-singular, and the \( \Lambda_{\min}(\mathbf{V}^{-1}) > 0 \), asymptotically. Under these conditions, we prove Theorem 1 below:

**Theorem 1** Under some regularity conditions on the link function \( \mu_k \) and Conditions (C1) and (C2) for the allocation function \( v_k \), for each \( k \), if \( \sup_n \|z_k\| < \infty \), then the proposed sequential estimation with...
the stopping rule defined in (3) has the following properties: (i) \( P(\tau_5 < \infty) = 1 \) and \( \lim_{\delta \to 0} \tau_5/n_{opt} = 1 \) almost surely. When the sampling stops, the estimate of \( \theta \) satisfies that (ii) \( \theta_{\tau_5} \to \theta \) almost surely as \( \delta \to 0 \), \( \sqrt{\hat{\tau}_5(\theta_{\tau_5} - \theta)} \to \mathcal{N}(0,V) \), and \( \lim_{\delta \to 0} P(\theta \in \hat{R}_5) = 1 - \alpha \). Then, in addition, the average of the stopping rule satisfies that (iii) \( \lim_{\delta \to 0} E \left[ \frac{\tau_{\delta}}{n_{opt}} \right] = 1 \). Moreover, for a given allocation function, it is shown that (iv) \( \lim_{\delta \to 0} \frac{N_{\tau_5}}{\hat{\tau}_5} = \nu \) almost surely, (v) \( \frac{N_{\tau_5,k}}{\hat{\tau}_5} = \pi_k(\theta, \xi) \) as \( \delta \to 0 \), \( k = 1, \ldots, K \), and (vi) \( \sqrt{\hat{\tau}_5(N_{\tau_5}/\tau_5 - \nu)} \to \mathcal{N}(0,\Sigma) \), where \( N_{\tau_5,k} \) is the number of subjects assigned to treatment \( k \) with covariate \( \xi \) up to \( \tau_5 \)th subject and \( N_{\tau_5} \) is the total number of subjects with covariate \( \xi \) up to \( \tau_5 \)th subject. Here \( \nu = (\nu_1, \ldots, \nu_K)' \) and \( \pi_k \), \( k = 1, \ldots, K \), depend on the allocation function, and \( \Sigma = \Sigma_1 + 2\Sigma_2 \) where \( \Sigma_1 = \mathrm{diag}(\nu) - \nu\nu' \) and \( \Sigma_2 = \sum_{k=1}^K \frac{\partial \nu}{\partial \theta_k}V_k\left(\frac{\partial \nu}{\partial \theta_k}'\right) \).

It can also be viewed as a multi-armed bandit problem. For further details, please see Chang and Park (2012).

2. Other Applications

2.1 Computerized Adaptive Testing

The 3 parameter logistic model, as defined below, is one of the most popular statistical models used for the item response theory (IRT) based standardized tests:

\[
P(\theta) = P(Y = 1|\theta, a, b, c) = c + (1 - c)\{1 + \exp(-a(\theta - b))\}^{-1},
\]

where \( \theta \) is the unknown latent trait of an examinee to be estimated, \( Y = 1(0) \) denotes the answer of the examinee, with ability level \( \theta \), is correct (incorrect) to an item with parameters \( a, b \) and \( c \). The idea of an adaptive testing is to administer a new item to the examinee based his/her responses to the previous items. Following the IRT model above, the Fisher information of \( \theta \) up to \( n \) items is \( I_n(\theta) = \sum_{i=1}^n (\partial^2 P_i(\theta)/\partial \theta^2)/P_i(1 - P_i(\theta)) \). Since our goal is to estimate \( \theta \), a natural and most efficient way to select a new item at the \( (n + 1) \)-st stage is to select an item that maximizes the \( I_n(\hat{\theta}_n) \); Because \( \theta \) is the unknown parameter to be estimated, hence we can only choose new item that maximizes the "estimated" Fisher information up to current stage. Hence, it is naturally a sequential estimation with adaptive designs (items). See Chang and Ying (2004) for details.

2.2 ROC Curve Estimation

The receiver operating characteristic (ROC) curve is a popular tool for evaluating the performance of classifiers; especially for the two-class classification problems. The receiver operating characteristic (ROC) curve is a statistical method to evaluate the performance of a diagnostic test, \( Y \), which classifies observations into two groups. Assume there are \( n_p \) cases and \( n_D \) controls in a case-control study, and let \( Y_D \) with a survival function \( S_D(c) = P_r(Y_D > c) \) and density \( f_D(c) \) and \( Y_D \) with a survival function \( S_D(c) = P_r(Y_D > c) \) and density \( f_D(c) \) be their diagnostic values respectively. The ROC curve is then defined as a plot of the sensitivity versus 1-specificity; that is, \( R(p) = S_D(S_D^{-1}(p)), p \in [0,1] \).

The area under ROC curve (AUC), a popular numerical summary index for ROC curve may be used to measure how accurately a given diagnostic test differentiates two populations. The area under any ROC curve \( R(S_D, S_D) \) between 0 and 1 is the probability that \( Y_D > Y_D \) is positive, supposing \( F_0 \) and \( F_1 \) are absolutely continuous. The definition of AUC is \( A = \int_0^1 R(p)dp = Pr(Y_D > Y_D) \). Due to \( \hat{R}(p) \) deriving from smoothed distribution functions \( \hat{S}_D \) and \( \hat{S}_D \), which are integrated, it follows that AUC is

\[
\hat{A} = \int_0^1 \hat{R}(p)dp = Pr(Y_D > Y_D).
\]

Its variance is

\[
V_A = V(\hat{A}) = \frac{V[S_D(Y_D)]}{n_D} + \frac{V[S_D(Y_D)]}{n_D},
\]
By plugging in the estimators, $\hat{S}_D(Y_D)$ and $\hat{S}_D(Y_{\bar{D}})$, the $V_A$ can be strongly estimated by

$$\hat{V}_A = \frac{\mathbf{V}[\hat{S}_D(Y_D)]}{n_D} + \frac{\mathbf{V}[\hat{S}_D(Y_{\bar{D}})]}{n_{\bar{D}}}.$$  \hspace{1cm} (7)

Then the optimal ratio, which is minimized $V_A$ at sampling without coat, is given by

$$\rho_A = \frac{n_D}{n_{\bar{D}}} = \sqrt{\frac{\mathbf{V}[\hat{S}_D(Y_D)]}{\mathbf{V}[\hat{S}_D(Y_{\bar{D}})]}}.$$ \hspace{1cm} (8)

If the samples are collected sequentially, then what is the most efficient way to estimate the AUC? Suppose that there are initial samples with $n_D$ cases and $n_{\bar{D}}$ controls is observed. After that, we sample one subject either from the case group or from the control group in order to obtain a smaller $\hat{V}_{A_{n+1}}$.

Assume that $\hat{V}_A = \frac{\mathbf{V}[\hat{S}_D(Y_D)]}{n_D} + \frac{\mathbf{V}[\hat{S}_D(Y_{\bar{D}})]}{n_{\bar{D}}}$ are calculated by initial sample, and

$$\hat{V}_{A_{n+1}} = \begin{cases} 
\frac{\mathbf{V}[\hat{S}_D(Y_D)]}{n_D} + \frac{\mathbf{V}[\hat{S}_D(Y_{\bar{D}})]}{n_{\bar{D}}} & \text{if select a new case} \\
\frac{\mathbf{V}[\hat{S}_D(Y_D)]}{n_D} + \frac{\mathbf{V}[\hat{S}_D(Y_{\bar{D}})]}{n_{\bar{D}+1}} & \text{if select a new control.}
\end{cases}$$

Then,

$$\hat{V}_A - \hat{V}_{A_{n+1}} = \begin{cases} 
\frac{\mathbf{V}[\hat{S}_D(Y_D)]}{n_D} - \frac{\mathbf{V}[\hat{S}_D(Y_D)]}{n_{D+1}} & \text{if select a new case} \\
\frac{\mathbf{V}[\hat{S}_D(Y_D)]}{n_D} - \frac{\mathbf{V}[\hat{S}_D(Y_{\bar{D}})]}{n_{D+1}} & \text{if select a new control.}
\end{cases}$$ \hspace{1cm} (I) \hspace{1cm} \hspace{1cm} (II)

If (I)$>(II)$, then we sample a new subject from the case group; otherwise, we sample a new subject from the control group. Again, this problem can be viewed as a two-armed bandit problem. See Chen, Wang, and Chang (2011) for details.

References


