

On the efficiency of adaptive designs

Holger Dette

Ruhr-Universität Bochum

Fakultät für Mathematik

44780 Bochum, Germany

e-mail: holger.dette@ruhr-uni-bochum.de

Björn Bornkamp, Frank Bretz

Statistical Methodology

Novartis Pharma AG

4002 Basel, Switzerland

e-mail: bjoern.bornkamp@novartis.com

e-mail: frank.bretz@novartis.com

October 11, 2010

Abstract

In this paper we develop a method to investigate the efficiency of two-stage adaptive designs from a theoretical point of view. Our approach is based on an explicit expansion of the information matrix for an adaptive design. The results enables one to compare the performance of adaptive designs with non-adaptive designs, without having to rely on extensive simulation studies. We demonstrate that their relative efficiency depends sensitively on the statistical problem under investigation and derive some general conclusions when to prefer an adaptive or a non-adaptive design. In particular, we show that in nonlinear regression models with moderate or large variances the first stage sample size of an adaptive design should be chosen sufficiently large in order to address variability in the interim parameter estimates. We illustrate the methodology with several examples.

Keywords and phrases: mean squared error, optimal design, maximum likelihood estimation, nonlinear regression

1 Introduction

It is well known that optimal designs can substantially improve the efficiency of statistical analyses. Numerous authors have worked on the problem of constructing optimal designs for regressions models. However, optimal designs for nonlinear regression models usually depend on the unknown parameter, leading to so-called locally optimal designs [Chernoff (1953); see also Ford et al. (1992), He et al. (1996), Fang and Hedayat (2008) or Yang (2010) among many others]. Locally optimal designs require

an upfront specification of the unknown parameter and might thus be sensitive with respect to an initial misspecification of that parameter. More advanced design strategies have been developed instead to overcome this sensitivity, such as Bayesian or other robust designs [see Chaloner and Verdinelli (1995), Dette (1997), Müller and Pázman (1998) or Wiens (2009) among others].

An attractive alternative are sequential designs, which update the information about the unknown parameter sequentially after each observation [see e.g. Ford and Silvey (1980) for an early reference]. Several authors have proved efficiency of sequentially optimal designs in the sense that sequential designs converge asymptotically to the locally optimal designs and the corresponding parameter estimates are asymptotically efficient [see Roy et al. (2008) or Chang and Ying (2009) among others]. However, these results usually refer to specific models and fully sequential designs are often not feasible in practice due to logistic restrictions.

Response-adaptive designs with several cohorts of subjects (adaptive designs, in short) are often used instead: After each stage the accumulated data of the ongoing study are used to update the initial guess of the underlying model parameters [Zhu and Wong (1998), Haines et al. (2003), Montepiedra and Yeh (2004)]. These designs continue to gain popularity in biopharmaceutical applications. For example, in clinical studies addressing dose finding objectives, trial designs that enable adaptations based on accrued data of an ongoing trial are perceived to be more efficient than non-adaptive designs (Bornkamp et al., 2007). Several adaptive designs have been introduced in the recent past; see, for example, Miller et al. (2007), Dragalin et al. (2007), or Bornkamp et al. (2010) for approaches in the context of dose finding clinical trials. In order to investigate the operating characteristics of a given adaptive design, in particular in comparison with a traditional, non-adaptive design, extensive simulations are typically necessary [see Dragalin et al. (2011)]. Theoretical comparisons often fail short because of the complicated structure in the data generating process of adaptive (or sequential) designs. In this paper we propose a new method to investigate the asymptotic efficiency of adaptive designs compared to non-adaptive designs from a theoretical point of view. We focus on two-stage adaptive designs and derive an explicit expression for the (asymptotic) Fisher information of these designs. These results are used for a theoretical comparison of the variances of the maximum likelihood estimates (MLE) obtained from adaptive and non-adaptive designs. We illustrate the methodology with several examples and demonstrate that the approximations derived by the asymptotic theory are accurate for realistic samples sizes. Moreover, we show that in nonlinear regression models with a moderate variance of the responses the first stage sample size of an adaptive design should be chosen sufficiently large in order to address variability in the interim parameter estimate. In particular, we demonstrate that the superiority of an adaptive or a non-adaptive design depends sensitively on the statistical problem under investigation.

2 Theoretical background

2.1 Notation

We consider models where a random variable Y with density $f(y, x, \theta)$ (with respect to the Lebesgue or counting measure) is observed at experimental condition $x \in \mathcal{X}$. Here, $\theta \in \Theta \subset \mathbb{R}^d$ denotes the unknown parameter vector and the explanatory variable x varies in the design space \mathcal{X} . We assume that N independent observations Y_1, \dots, Y_N at experimental conditions $x_1, \dots, x_N \in \mathcal{X}$ are available and denote by $\hat{\theta}$ the MLE based on the full sample $(x_1, Y_1), \dots, (x_N, Y_N)$. A design is defined as a probability measure ξ on the design space \mathcal{X} with finite support which specifies the different experimental conditions (Pukelsheim, 2006). That is, if N observations can be made in total and the design ξ has mass w_i at x_i , $i = 1, \dots, k$, the quantities $w_i N$ are rounded to integers n_i , such that $\sum_{i=1}^k n_i = N$, and the experimenter takes n_i observations at each condition x_i , $i = 1, \dots, k$. It is well known (see also the derivations in the Appendix) that – under certain regularity assumptions – the statistic $\sqrt{N}(\hat{\theta} - \theta)$ is asymptotically normal distributed with mean 0 and covariance matrix $M^{-1}(\xi, \theta)$, where

$$M(\xi, \theta) = \int \int \left(\frac{\partial}{\partial \theta} \log f(y, x, \theta) \right)^T \left(\frac{\partial}{\partial \theta} \log f(y, x, \theta) \right) f(y, x, \theta) dy d\xi(x) \quad (2.1)$$

denotes the information matrix of the given design ξ . A locally optimal design maximizes an appropriate functional of this matrix, the so-called optimality criterion [see Atkinson et al. (2007) or Pukelsheim (2006) among others]. Throughout this paper let $\xi_\theta = \{x_i(\theta), w_i(\theta)\}_{i=1}^k$ denote a locally optimal design and assume that the weights $w_i(\theta)$ and support points $x_i(\theta)$ are continuously differentiable functions of the parameter θ . We define the mapping

$$I : \begin{cases} \Theta \times \Theta & \longrightarrow \mathbb{R}^{d \times d} \\ (\theta, \tau) & \longrightarrow I(\theta, \tau) := M(\xi_\tau, \theta). \end{cases} \quad (2.2)$$

and note that $I(\theta, \tau)$ is the information matrix of the locally optimal design ξ_τ for the parameter $\tau \in \Theta$, if the “true” parameter is given by θ . In the following we consider two standard examples, a nonlinear regression and a binary response model, that will be referred to in Section 4.

Example 2.1 Consider the common nonlinear regression model $Y = \eta(x, \theta) + \varepsilon$, where the expectation of a real valued response Y under experimental condition x is given by $E[Y|x] = \eta(x, \theta)$ with a normally distributed homoscedastic error such that $E[\varepsilon] = 0$ and $\text{Var}(\varepsilon) = \sigma^2$, i.e.

$$f(y, x, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \eta(x, \theta))^2\right).$$

From the identities

$$\begin{aligned} \frac{\partial}{\partial \theta} \log f(y, x, \theta) &= \frac{1}{\sigma^2} \frac{\partial}{\partial \theta} \eta(x, \theta) (y - \eta(x, \theta)) = \frac{1}{\sigma^2} \frac{\partial}{\partial \theta} \eta(x, \theta) \varepsilon, \\ \frac{\partial^2}{\partial^2 \theta} \log f(y, x, \theta) &= \frac{1}{\sigma^2} \left\{ \frac{\partial^2}{\partial^2 \theta} \eta(x, \theta) (y - \eta(x, \theta)) - \left(\frac{\partial}{\partial \theta} \eta(x, \theta) \right)^T \frac{\partial}{\partial \theta} \eta(x, \theta) \right\}, \end{aligned}$$

it follows by straightforward calculation that

$$M(\xi, \theta) = \frac{1}{\sigma^2} \int \left(\frac{\partial}{\partial \theta} \eta(x, \theta) \right)^T \left(\frac{\partial}{\partial \theta} \eta(x, \theta) \right) d\xi(x). \quad (2.3)$$

Example 2.2 Consider a binary response model, where the probability of success is given by $P(Y = 1|x) = p(x, \theta)$. In this case we obtain

$$f(y, x, \theta) = (p(x, \theta))^y (1 - p(x, \theta))^{1-y},$$

and the Fisher information matrix of a design ξ is given by

$$M(\xi, \theta) = \int \frac{\left(\frac{\partial}{\partial \theta} p(x, \theta) \right)^T \left(\frac{\partial}{\partial \theta} p(x, \theta) \right)}{p(x, \theta)(1 - p(x, \theta))} d\xi(x). \quad (2.4)$$

2.2 Main results

We now introduce two design strategies: A traditional, non-adaptive design ξ_A , where observations are taken at fixed experimental conditions, and a two-stage adaptive design ξ_B , where after the first stage the accrued data is used to determine the second stage design. In the following, let θ_0 denote a preliminary guess for the unknown parameter θ . In many biopharmaceutical applications, such as the dose finding clinical trials mentioned in Section 1, preliminary information from previous trials (animal studies, proof-of-concept studies, etc.) is available to generate a best guess θ_0 .

- (A) *Non-adaptive design* ξ_A : Take all N observations according to the locally optimal design ξ_{θ_0} based on the best guess θ_0 . The resulting estimate of θ is denoted by $\hat{\theta}_A$.
- (B) *Two-stage adaptive design* ξ_B : Split the total sample N in two parts and proceed as follows.
 - Take N_0 observations according to the locally optimal design ξ_{θ_0} . For the asymptotic considerations below we assume that $p_0 = \frac{N_0}{N}$, where $\lim_{N \rightarrow \infty} N_0/N \in (0, 1)$ is a fixed constant.
 - Estimate the parameter θ by MLE from these N_0 observations. The resulting estimate of θ is denoted by $\hat{\theta}_1$.
 - Take $N_1 = N - N_0$ observations according to the locally optimal design $\xi_{\hat{\theta}_1}$ and estimate the parameter θ by MLE from all $N = N_0 + N_1$ observations. The final estimate is denoted by $\hat{\theta}_B$. We let $p_1 = \frac{N_1}{N}$ and note that $p_0 + p_1 = 1$.

In the following we provide an analytical comparison of the two design strategies. Note that under standard assumptions in nonlinear regression (see the Appendix for details) the variance of the MLE is of order $O(1/N)$ while the squared bias is of order $O(1/N^2)$ [see Box (1971)], which implies that the

mean squared error is dominated by the variance. Therefore we approximate the mean squared error of the MLE by its variance, i.e.

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] \approx \text{Var}(\hat{\theta}). \quad (2.5)$$

We aim at deriving asymptotic expansions for the variances $\text{Var}(\hat{\theta}_A)$ and $\text{Var}(\hat{\theta}_B)$ in order to compare the two design options ξ_A and ξ_B for a given statistical problem. Our first result to this end specifies the asymptotic properties of the MLE $\hat{\theta}_1$ and $\hat{\theta}_B$ obtained with the two-stage adaptive design. Throughout this paper let $I_d \in \mathbb{R}^{d \times d}$ denote the identity matrix and $\mathcal{N}_d(0, I_d)$ the d -variate standard normal distribution. The proof of the following result is in the Appendix.

Theorem 2.1 *Assume that the mapping (2.2) is differentiable with respect to τ and that the regularity conditions specified in the Appendix are satisfied. Then*

$$\hat{\theta}_1 = \tilde{\theta}_1 + O_p\left(\frac{1}{N_0}\right),$$

where

$$\tilde{\theta}_1 = \theta + \frac{1}{\sqrt{N_0}} I^{-1/2}(\theta, \theta_0) Z_{0, N_0} \quad (2.6)$$

and

$$Z_{0, N_0} = \frac{1}{\sqrt{N_0}} I^{-1/2}(\theta, \theta_0) \sum_{i=1}^{N_0} \frac{\partial}{\partial \theta} \log f(Y_i, x_i(\theta_0), \theta) \xrightarrow{\mathcal{D}} \mathcal{N}_d(0, I_d). \quad (2.7)$$

Moreover,

$$\sqrt{N}(\hat{\theta}_B - \theta) = \hat{\gamma}_B + O_p\left(\frac{1}{\sqrt{N}}\right),$$

where

$$\hat{\gamma}_B = \left(p_0 I(\theta, \theta_0) + p_1 I(\theta, \tilde{\theta}_1) \right)^{-1} \left(\sqrt{p_0} I^{1/2}(\theta, \theta_0) Z_{0, N_0} + \sqrt{p_1} I^{1/2}(\theta, \tilde{\theta}_1) Z_{1, N_1} \right),$$

and the random variable Z_{1, N_1} is defined by

$$Z_{1, N_1} = \frac{1}{\sqrt{N_1}} I^{-1/2}(\theta, \tilde{\theta}_1) \sum_{i=N_0+1}^{N_0+N_1} \frac{\partial}{\partial \theta} \log f(Y_i, x_i(\tilde{\theta}_1), \theta). \quad (2.8)$$

By Theorem 2.1, it is reasonable to approximate the variance of the estimate $\hat{\theta}_B$ by the variance of the random variable $\hat{\gamma}_B$, which can be calculated using the variance decomposition formula

$$\text{Var}(\hat{\gamma}_B) = E[\text{Var}(\hat{\gamma}_B | Y_1, \dots, Y_{N_0})] + \text{Var}(E[\hat{\gamma}_B | Y_1, \dots, Y_{N_0}]). \quad (2.9)$$

From Theorem 2.1 we obtain for the conditional expectation and variance of $\hat{\gamma}_B$ given Y_1, \dots, Y_{N_0}

$$E[\hat{\gamma}_B | Y_1, \dots, Y_{N_0}] = \left(p_0 I(\theta, \theta_0) + p_1 I(\theta, \tilde{\theta}_1) \right)^{-1} \sqrt{p_0} I^{1/2}(\theta, \theta_0) Z_{0, N_0}, \quad (2.10)$$

$$\text{Var}(\hat{\gamma}_B | Y_1, \dots, Y_{N_0}) = p_1 \left(p_0 I(\theta, \theta_0) + p_1 I(\theta, \tilde{\theta}_1) \right)^{-1} I(\theta, \tilde{\theta}_1) \left(p_0 I(\theta, \theta_0) + p_1 I(\theta, \tilde{\theta}_1) \right)^{-1}. \quad (2.11)$$

Here, we used the fact that $E[Z_{1,N_1} | Y_1, \dots, Y_{N_0}] = 0$, $\text{Var}(Z_{1,N_1} | Y_1, \dots, Y_{N_0}) = I_d$ and that $\tilde{\theta}_1$ depends only on Y_1, \dots, Y_{N_0} . The following result provides a further simplification by eliminating $\tilde{\theta}_1$ from the expressions (2.10) and (2.11).

Corollary 2.1 *Under the assumptions of Theorem 2.1 we obtain for the conditional expectation and variance of $\hat{\gamma}_B = \sqrt{N}(\hat{\theta}_B - \theta) + O_p(1/\sqrt{N})$*

$$\begin{aligned} E[\hat{\gamma}_B | Y_1, \dots, Y_{N_0}] &= H^{-1}(\theta, \theta_0) \{ I_d - p_1 R(\theta, Z_{0,N_0}) H^{-1}(\theta, \theta_0) + p_1^2 S(\theta, Z_{0,N}) \} \\ &\quad \times \sqrt{p_0} I^{1/2}(\theta, \theta_0) Z_{0,N_0} + o_p\left(\frac{1}{N}\right), \end{aligned} \quad (2.12)$$

$$\begin{aligned} \text{Var} [\hat{\gamma}_B | Y_1, \dots, Y_{N_0}] &= p_1 H^{-1}(\theta, \theta_0) \{ I_d - p_1 R(\theta, Z_{0,N_0}) H^{-1}(\theta, \theta_0) + p_1^2 S(\theta, Z_{0,N}) \} \{ I(\theta, \theta) + R(\theta, Z_{0,N_0}) \} \\ &\quad \times \{ I_d - p_1 R(\theta, Z_{0,N_0}) H^{-1}(\theta, \theta_0) + p_1^2 S(\theta, Z_{0,N}) \}^T H^{-1}(\theta, \theta_0) + o_p\left(\frac{1}{N}\right), \end{aligned} \quad (2.13)$$

respectively. Here, we have introduced the notation

$$H(\theta, \theta_0) = p_0 I(\theta, \theta_0) + p_1 I(\theta, \theta), \quad (2.14)$$

the matrices $R(\theta, Z_{0,N_0})$ and $S(\theta, Z_{0,N_0})$ are defined by

$$R(\theta, Z_{0,N_0}) = \frac{1}{\sqrt{N_0}} D_1(\theta, Z_{0,N_0}) + \frac{1}{2N_0} D_2(\theta, Z_{0,N_0}), \quad (2.15)$$

$$S(\theta, Z_{0,N_0}) = \frac{1}{N_0} D_1(\theta, Z_{0,N_0}) H^{-1}(\theta, \theta_0) D_1(\theta, Z_{0,N_0}) H^{-1}(\theta, \theta_0), \quad (2.16)$$

respectively, the matrices D_1 and D_2 are given by

$$D_1(\theta, Z_{0,N_0}) = (\nabla(I(\theta, \theta)))_{ij} I^{-1/2}(\theta, \theta_0) Z_{0,N_0} \Big|_{i,j=1}^d, \quad ,$$

$$D_2(\theta, Z_{0,N_0}) = (Z_{0,N_0}^T I^{-1/2}(\theta, \theta_0) \nabla^2(I(\theta, \theta))_{ij} I^{-1/2}(\theta, \theta_0) Z_{0,N_0}) \Big|_{i,j=1}^d, \quad (2.17)$$

respectively, ∇g and $\nabla^2 g$ denote the gradient and the Hessian matrix of a real valued function g . Finally, the random variable Z_{0,N_0} is defined in (2.7).

In general, the explicit calculation of the dominating terms of the variance of $\hat{\gamma}_B$ using (2.9) and Corollary 2.1 for a given non-linear model is very cumbersome. However, its general structure becomes

clear with the arguments above and can be calculated using computer algebra systems as Mathematica or Matlab. Roughly speaking, we obtain from (2.12) and (2.13)

$$\begin{aligned}\text{Var}(E[\hat{\gamma}_B | Y_1, \dots, Y_{N_0}]) &\approx p_0 H^{-1}(\theta, \theta_0) I(\theta, \theta_0) H^{-1}(\theta, \theta_0) + \frac{1}{\sqrt{N_0}} A_1(\theta, \theta_0) + \frac{1}{N_0} A_2(\theta, \theta_0), \\ E[\text{Var}(\hat{\gamma}_B | Y_1, \dots, Y_{N_0})] &\approx p_1 H^{-1}(\theta, \theta_0) I(\theta, \theta) H^{-1}(\theta, \theta_0) + \frac{1}{\sqrt{N_0}} A_3(\theta, \theta_0) + \frac{1}{N_0} A_4(\theta, \theta_0),\end{aligned}$$

which implies

$$\text{Var}(\hat{\gamma}_B) \approx H^{-1}(\theta, \theta_0) + \frac{1}{\sqrt{N_0}} A_5(\theta, \theta_0) + \frac{1}{N_0} A_6(\theta, \theta_0)$$

with appropriate matrices $A_i(\theta, \theta_0), i = 1, \dots, 6$. These matrices depend on the specific model under investigation. In particular, we obtain for the information matrix $M(\xi_B, \theta)$ of the design ξ_B

$$M(\xi_B, \theta) \approx H(\theta, \theta_0) + \frac{1}{\sqrt{N_0}} K(\theta, \theta_0) + \frac{1}{N_0} L(\theta, \theta_0) + o\left(\frac{1}{N_0}\right), \quad (2.18)$$

where the matrix $H(\theta, \theta_0)$ is defined in (2.14), and the matrices K and L depend on the specific model under consideration as well as the initial guess θ_0 for the unknown parameter θ . Note that the matrix $H(\theta, \theta_0)$ is a weighted average of the information matrices corresponding to the locally optimal designs ξ_{θ_0} and ξ_θ . Therefore, this matrix can be interpreted as a mixture of information matrices corresponding to two locally optimal designs: One for the ‘‘true’’ parameter θ and another one for the preliminary guess θ_0 . The weights p_0 and p_1 in this mixture correspond to the relative proportions of subjects treated in the first and second stage, respectively. Consequently, for ‘‘small’’ values of p_0 the dominating term in (2.18) becomes ‘‘close’’ to the Fisher information matrix of the locally optimal design ξ_θ . Similarly, the adaptive design ξ_B is approximately given by $\xi_B \approx p_0 \xi_{\theta_0} + p_1 \xi_\theta$ and the remainder corresponds to the error in these approximations.

Note that the expansion (2.18) refers to an asymptotic analysis where we assume the first stage sample size N_0 to be of the same order as the total sample size $N \rightarrow \infty$, i.e. $\lim_{N \rightarrow \infty} N_0/N \in (0, 1)$. The information matrix of the non-adaptive design ξ_A using the locally optimal design ξ_{θ_0} is given by

$$M(\xi_A, \theta) \approx \frac{1}{I(\theta, \theta_0)} + \frac{1}{\sqrt{N_0}} \bar{L}(\theta, \theta_0) + \frac{1}{N_0} \bar{K}(\theta, \theta_0) + o\left(\frac{1}{N_0}\right) \quad (2.19)$$

with appropriate matrices \bar{L}, \bar{K} . In general, the matrices K, \bar{K} and L, \bar{L} in (2.18) and (2.19) are neither positive nor negative definite and therefore it is not clear whether for finite sample sizes the matrix $M(\xi_B, \theta)$ is smaller (with respect to the Loewner ordering) than $M(\xi_A, \theta)$ corresponding to the locally optimal design ξ_θ . Because $H(\theta, \theta_0) \geq I(\theta, \theta_0)$, however, it follows that asymptotically the adaptive design ξ_B is always better than the non-adaptive design ξ_A . For finite sample sizes the correction terms of order $1/\sqrt{N_0}$ and $1/N_0$ have to be factored in and the relationship is not obvious anymore.

The arguments above remain valid for any differentiable optimality criterion ϕ . To be precise, assume that a (locally) ϕ -optimal design minimizes $\phi(M(\xi, \theta))$ in the class of all designs. When comparing the

efficiency of the two designs ξ_A and ξ_B from Section 2.2, this gives

$$\text{eff}_\phi(\xi_A, \xi_B) = \frac{\phi(M(\xi_A, \theta))}{\phi(M(\xi_B, \theta))} \approx \frac{\phi(I(\theta, \theta_0))}{\phi(p_0 I(\theta, \theta_0) + p_1 I(\theta, \theta))} + \frac{c}{\sqrt{N_0}} + \frac{d}{N_0} \quad (2.20)$$

for the ϕ -efficiency of the non-adaptive design ξ_A with respect to the adaptive design ξ_B , where no information regarding the sign of the constants c and d is available in general. A common application is the problem of estimating a function of the unknown parameter θ , say $\Psi(\theta)$. For example, a frequent problem in dose response analyzes is the estimation of relevant target doses as a function of the parameters of a regression model [see Dette et al. (2008, 2010) among others]. In such situations the asymptotic variance of the canonical estimate $\hat{\psi} = \psi(\hat{\theta}_B)$ from a two-stage design ξ_B is given by

$$\begin{aligned} \text{Var}(\hat{\psi}) &\approx \nabla \Psi(\theta) M^{-1}(\xi_B, \theta) (\nabla \psi(\theta))^T \\ &\approx \nabla \psi(\theta) H^{-1}(\theta, \theta_0) (\nabla \Psi(\theta))^T + \frac{1}{\sqrt{N_0}} \tilde{K}(\theta, \theta_0) + \frac{1}{N_0} \tilde{L}(\theta, \theta_0) \end{aligned}$$

with appropriate constants \tilde{K} , \tilde{L} and (2.20) can be extended accordingly.

3 Asymptotic variances in one-parameter models

If the parameter θ in the nonlinear regression model from Section 2.1 satisfies $\theta \in \Theta \subset \mathbb{R}$, the information matrix of a given design ξ is one-dimensional and an optimal design maximizes this matrix (or minimizes its inverse). Assume that for each $\theta \in \Theta$ a one-point design, say ξ_θ , maximizes $M(\xi, \theta)$ in the class of all designs on the design space \mathcal{X} . Let $x(\theta)$ denote the corresponding support point of the locally optimal design ξ_θ , which is assumed to be an interior point of the design space $\mathcal{X} \subset \mathbb{R}$. Consequently, it follows from (2.1) that for each $\theta \in \Theta$ the point $x(\theta)$ is a solution of the equation

$$\frac{\partial}{\partial x} \int f(y, x, \theta) \left(\frac{\partial}{\partial \theta} \log f(y, x, \theta) \right)^2 dy = 0. \quad (3.1)$$

We are now in a position to give an explicit expression for the asymptotic variance of the MLE $\hat{\theta}_B$ obtained from the two-stage adaptive design ξ_B .

Theorem 3.1 *Assume $d = 1$ and that for each $\theta \in \Theta$ the locally optimal design is a one-point design. Under the assumptions of Theorem 2.1, it follows*

$$\text{Var}(\hat{\gamma}_B) \approx \frac{1}{H(\theta, \theta_0)} - \frac{g(\theta) p_1 (5p_0 I(\theta, \theta_0) + p_1 I(\theta, \theta))}{2N_0 H^3(\theta, \theta_0) I(\theta, \theta_0)}.$$

With the result from Theorem 3.1 we can now express the efficiency of a non-adaptive design ξ_A compared to an adaptive design ξ_B as the ratio of the asymptotic variances $\text{Var}(\hat{\theta}_A)$ and $\text{Var}(\hat{\theta}_B)$.

Because the mean squared error is dominated by the variance, it follows from (2.5) and Theorem 3.1 that

$$\text{eff}(\xi_A, \xi_B) = \frac{\text{MSE}(\hat{\theta}_A)}{\text{MSE}(\hat{\theta}_B)} \approx \frac{\text{Var}(\hat{\theta}_A)}{\text{Var}(\hat{\theta}_B)} \approx \left\{ \frac{I(\theta, \theta_0)}{H(\theta, \theta_0)} - p_1 \frac{g(\theta)(5p_0I(\theta, \theta_0) + p_1I(\theta, \theta))}{2N_0H^3(\theta, \theta_0)} \right\}^{-1}. \quad (3.2)$$

If $\text{eff}(\xi_A, \xi_B) < 1$, the design ξ_A is preferable as it yields smaller MSEs for the MLE. If $\text{eff}(\xi_A, \xi_B) > 1$, the opposite is true and the design ξ_B is preferable. In general, a conclusion about the superiority of a design depends on the underlying regression model, see Section 4 for examples. Note again that in (3.2) the dominating term $I(\theta, \theta_0)/H(\theta, \theta_0) < 1$, because $H(\theta, \theta_0) = p_0I(\theta, \theta_0) + p_1I(\theta, \theta) \geq I(\theta, \theta_0)$. Therefore, for a large first stage sample sizes N_0 , we have $\text{eff}(\xi_A, \xi_B) > 1$ and expect the adaptive design ξ_B to be more efficient than the non-adaptive design ξ_A . However, the second term in (3.2) is positive [note that $g(\theta)$ is negative because it is the second derivative at the optimal point maximizing $M(\xi, \theta)$] and this contribution may be substantial for finite sample sizes as illustrated with examples in the following section.

4 Examples

In this section we illustrate the asymptotic theory with three examples by considering an exponential, a logistic and a Poisson regression model.

4.1 Exponential regression model

We consider the one-parameter exponential regression model with homoscedastic errors, that is

$$E[Y|x] = \eta(x, \theta) = e^{-\theta x}, \quad \text{Var}(Y|x) = \sigma^2 > 0, \quad (4.1)$$

where $\mathcal{X} = [0, \infty)$ and $\theta > 0$. In this case, we have $\frac{\partial}{\partial \theta} \eta(x, \theta) = -xe^{-\theta x}$ and the Fisher information matrix at the point x is obtained from (2.3). Optimal design problems for this model have been considered by numerous authors (Dette and Neugebauer, 1996). In particular, the local D -optimal design is a one-point design with $x(\theta) = 1/\theta$. Consequently,

$$I(\theta, \theta) = \frac{1}{\sigma^2}(e\theta)^{-2}, \quad I(\theta, \theta_0) = \frac{1}{\sigma^2}(e^{\theta/\theta_0} \theta_0)^{-2}, \quad \text{and} \quad g(\theta) = -\frac{2}{\sigma^2}(\theta^2 e)^{-2}.$$

Therefore, it follows from (3.2) that

$$\text{eff}(\xi_A, \xi_B) \approx r(\xi_A, \xi_B) = \left\{ \frac{1}{p_0 + p_1 \{(\theta e)^2 (e^{\theta/\theta_0} \theta_0)^{-4} \}^{-1}} + \frac{\sigma^2 p_1}{e^2 N_0 \theta^4} \frac{5p_0(e^{\theta/\theta_0} \theta_0)^{-4} + \frac{p_1}{(\theta e)^2}}{[p_0(e^{\theta/\theta_0} \theta_0)^{-4} + p_1 \{(\theta e)^2 \}^{-1}]^3} \right\}^{-1} \quad (4.2)$$

In the following we investigate the accuracy of this approximation in several concrete finite sample scenarios. In Figure 1 we plot the approximation $r(\xi_A, \xi_B)$ from (4.2) as function of p_0 (solid line) for

different configurations of σ, θ_0 and θ , where the total sample size is $N = 100$. A ratio larger (smaller) than 1 means that the adaptive design ξ_B yields smaller (larger) MSEs and is therefore better (worse) than the non-adaptive design ξ_A .

The first row in Figure 1 shows the results when the true parameter value is $\theta = 1$, while the parameter used for the calculation of the non-adaptive design (and also for the initial design of the adaptive design) was misspecified by 100%, i.e. $\theta_0 = 2$. In the second row we display the results for a stronger misspecification of 200%, i.e. $\theta_0 = 3$. Finally, the third row shows the results when $\theta = 2$ and $\theta_0 = 3$. The three columns correspond to three different variances in the error distribution, i.e. $\sigma^2 = 1, 0.1, 0.01$. In order to investigate the accuracy of the asymptotic results for finite sample sizes we have also performed simulations to calculate the MSE ratio for the MLEs obtained from the designs ξ_A and ξ_B (dashed lines in Figure 1; based on 20000 simulation runs). We observe a rather precise approximation of the simulated MSE ratios by the asymptotic theory. Note that the asymptotic approximation is better for smaller variances, because this variance also appears in the remainder of (4.2).

Comparing the two designs ξ_A and ξ_B reveals that for large variances the non-adaptive design ξ_A has a competitive, if not even better performance for a broad range of p_0 values. This observation can be nicely explained by the fact that the term of order $1/N_0$ in the approximation (4.2) is increasing with σ^2 and decreasing with p_0 . Heuristically, a large error variance leads to a highly variable first stage estimate $\hat{\theta}_1$ if the initial sample size N_0 is not sufficiently large. Therefore the corresponding design $\xi_{\hat{\theta}_1}$ used in the second stage may not be efficient in some cases. On the other hand, for small variances or large first stage sample sizes, the parameter θ can be estimated rather precisely from the data collected in the first stage. Consequently, updating the initial parameter guess θ_0 based on the first stage data will lead to a better second stage design and to an overall better performance for most p_0 values. Note also that the degree of initial misspecification of the parameter θ (through θ_0) has only little impact when the variance is large. Overall, the differences between the designs ξ_A and ξ_B are small for the situations considered here, except in the case of a very small variances σ^2 and where the initial guess θ_0 deviates substantially from θ .

4.2 Logistic regression

Consider a logistic regression model, where the responses are independent Bernoulli random variables with probability of success

$$p(x, \theta) = E[Y|x] = \frac{1}{1 + e^{x-\theta}},$$

and $\mathcal{X} = \mathbb{R}$. This model is sometimes called one-parameter Rasch model and is used to model the item characteristic curve in item response theory (Rasch, 1960). Sequential optimal designs for the Rasch model have recently been discussed by Chang and Ying (2009). It follows from (2.4) that the Fisher information matrix for a one-point design δ_x at the point x is given by

$$M(\delta_x, \theta) = \frac{e^{x-\theta}}{(1 + e^{x-\theta})^2}.$$

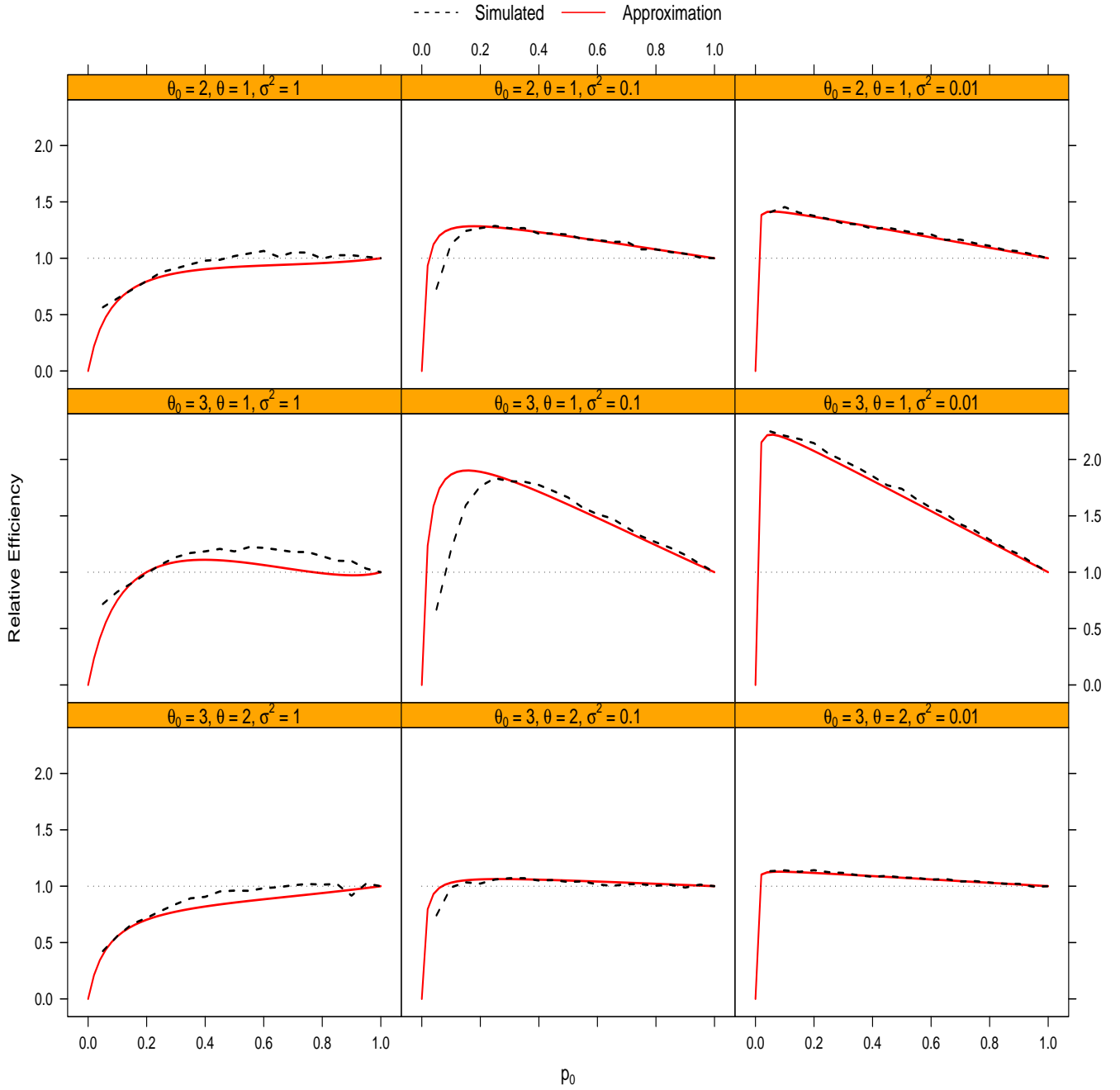


Figure 1: Plot of the approximation $r(\xi_A, \xi_B)$ defined in (4.2) for the MSE ratio (solid line) and corresponding simulation results (dashed line) as function of p_0 under the exponential model (4.1) for $N = 100$ and different configurations of σ, θ_0 and θ . From left to right: $\sigma^2 = 1, 0.1, 0.01$; from top to bottom: $(\theta_0, \theta) = (2, 1), (3, 1), (3, 2)$.

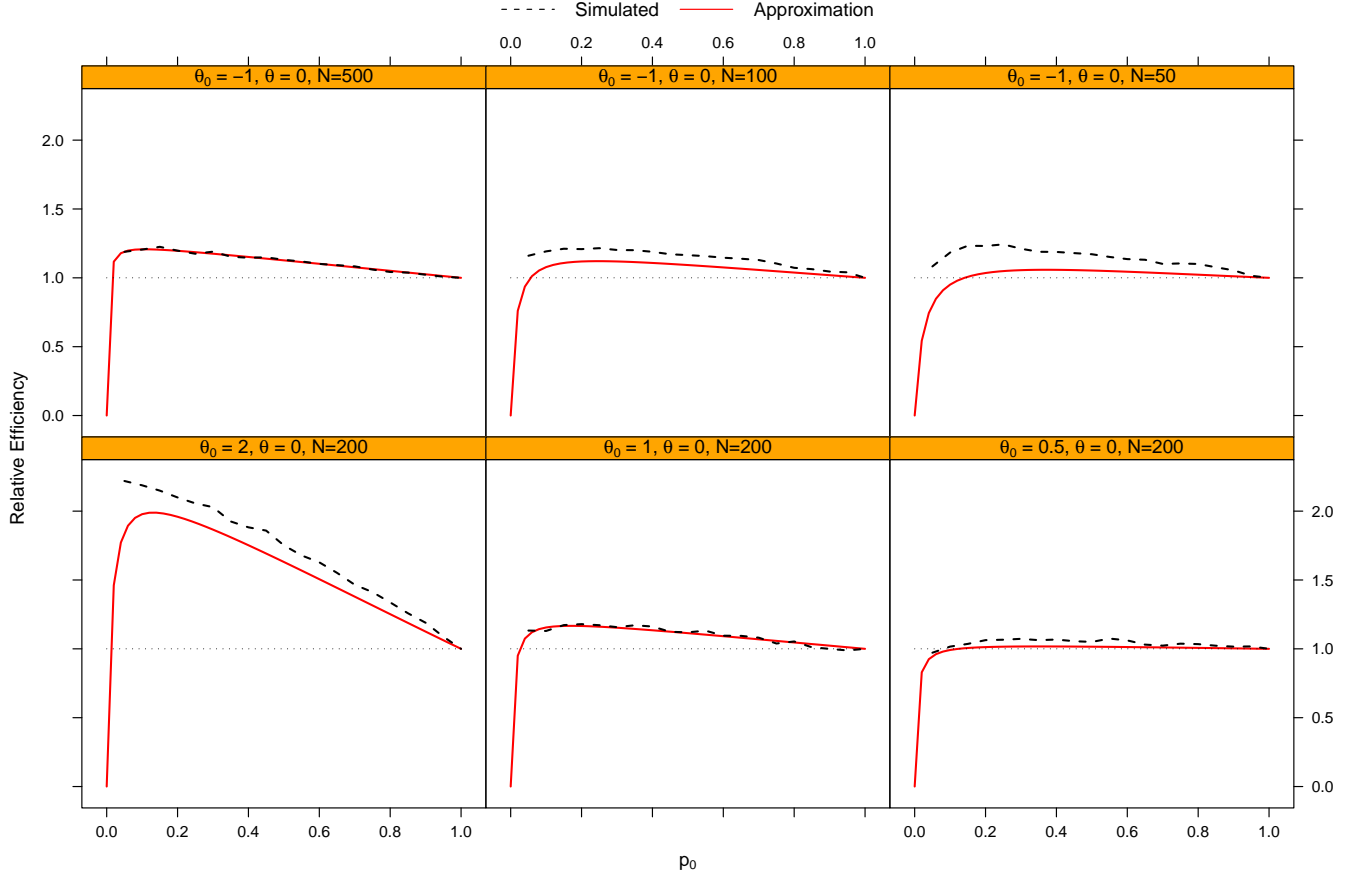


Figure 2: Plot of the approximation $r(\xi_A, \xi_B)$ defined by (4.3) for the MSE ratio (solid line) and corresponding simulation results (dashed line) as a function of p_0 under the logistic model for different configurations of θ, θ_0 , and N .

Standard calculation shows that the design concentrating its mass at the point $x(\theta) = \theta$ is locally optimal. Therefore we obtain

$$I(\theta, \tau) = \frac{e^{\tau-\theta}}{(1 + e^{\tau-\theta})^2},$$

which implies

$$I(\theta, \theta) = \frac{1}{4}, \quad g(\theta) = -\frac{1}{8}, \quad \text{and} \quad H(\theta, \theta_0) = p_0 \frac{e^{\theta_0-\theta}}{(1 + e^{\theta_0-\theta})^2} + p_1 \frac{1}{4}.$$

Consequently, it follows from (3.2) that

$$\text{eff}(\xi_A, \xi_B) \approx r(\xi_A, \xi_B) = \left\{ \left(p_0 + \frac{p_1(1 + e^\gamma)^2}{4e^\gamma} \right)^{-1} + p_1 \frac{(20p_0e^\gamma + p_1(1 + e^\gamma)^2)(1 + e^\gamma)^4}{N_0(4p_0e^\gamma + p_1(1 + e^\gamma)^2)^3} \right\}^{-1}, \quad (4.3)$$

where $\gamma = \theta_0 - \theta$ denotes the degree of initial misspecification of θ through θ_0 . In Figure 2 we plot the approximation $r(\xi_A, \xi_B)$ defined in (4.3) together with the corresponding simulation results as function

of p_0 for different values of θ_0, θ and N . Again the approximation obtained by the asymptotic theory is very accurate.

We observe that in most situations the adaptive design shows a better performance, although the improvement remains small, except for large values of $|\gamma|$ and N . Only for very small sample sizes the non-adaptive design ξ_A performs better than the adaptive design ξ_B . These results can be explained by the fact that the variance in the logistic regression model is relatively small. For example, if $\theta_0 - \theta = -1$ the variance of individual observations in the first stage is $p(\theta_0, \theta)(1 - p(\theta_0, \theta)) = 0.197$. As a consequence, the parameter estimate $\hat{\theta}_1$ obtained from the first stage is rather accurate and the corresponding design $\xi_{\hat{\theta}_1}$ is already close to the locally optimal design. If $\theta_0 - \theta = -2$ the variance of the observations from the first stage is even smaller (roughly 0.105), which explains the superiority of the adaptive design in this case.

4.3 Poisson regression model

In our final example we consider the Poisson regression model

$$P(Y = k|x) = \frac{(e^{\theta x})^k}{k!} e^{-e^{\theta x}}$$

for $x \in \mathbb{R}$. A straightforward calculation shows that the Fisher information at the point x is $x^2 e^{\theta x}$. A locally optimal design based on the initial first guess θ_0 advises the experimenter to take all observations at the point $x(\theta_0) = -2/\theta_0$. Consequently,

$$g(\theta) = -\frac{8}{\theta^4 e^2}, \quad I(\theta, \theta_0) = \frac{4}{\theta_0^2} e^{-2\theta/\theta_0}$$

and it follows from (3.2) that for $\gamma = \theta/\theta_0$

$$\text{eff}(\xi_A, \xi_B) \approx r(\xi_A, \xi_B) = \left\{ \frac{e^\gamma}{p_0 e^\gamma + p_1 \gamma^{-2}} + \frac{p_1 e^2 (5p_0 e^\gamma \gamma^2 + p_1)}{4\theta^4 N_0 (p_0 e^\gamma \gamma^2 + p_1)^3} \right\}^{-1}. \quad (4.4)$$

In Figure 3 we plot the MSE ratio as function of p_0 for two parameter specifications: $\theta_0 = 1.5, \theta = 1$ and $\theta_0 = -0.1, \theta = 0.1$. In the first case rather large sample sizes were chosen, to avoid situations where the interim MLE $\hat{\theta}_1 = 0$ and the optimal design point for the second stage cannot be calculated. The plots for the different scenarios show that the superiority or inferiority of the adaptive design depends sensitively on the parameter constellation. In the first case ($\theta_0 = 1.5, \theta = 1$) the adaptive design performs nearly uniformly better over the full range of p_0 , although the advantages are small (efficiency gain less than 10%). Only for very small p_0 values the non-adaptive designs yields a smaller mean squared error. On the other hand, it follows from the second row in Figure 3 that the non-adaptive design yields a substantially smaller mean squared error if the sign of the unknown parameter is misspecified. In this case there exists no situation, where the adaptive design is more efficient than the non-adaptive design and in many cases the difference is substantial. In this example, $\theta_0 = -0.1,$

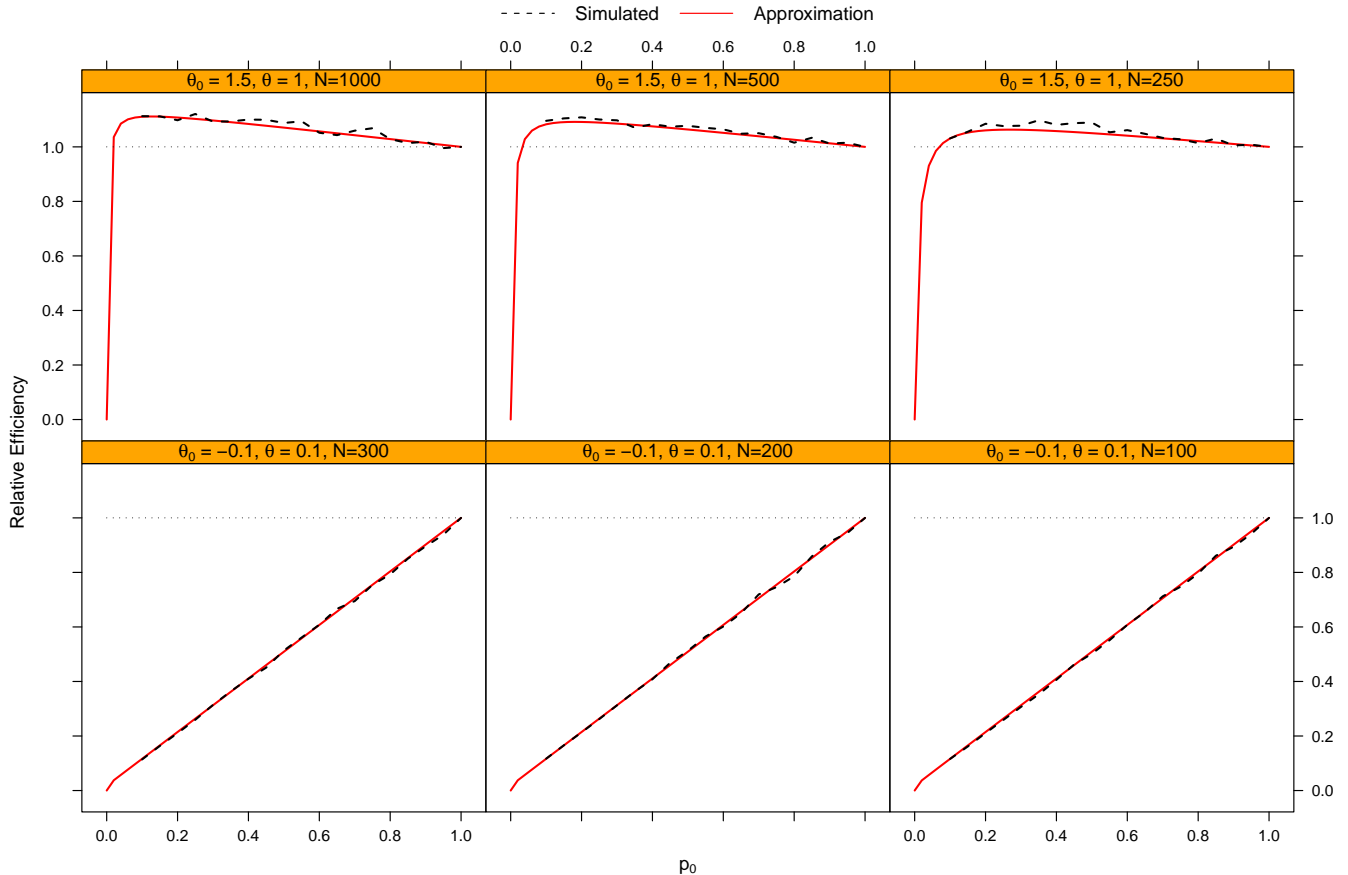


Figure 3: Plot of the asymptotic MSE ratio (solid line) and corresponding simulation results (dashed line) as a function of p_0 under the Poisson model for different configurations of θ , θ_0 , and N .

$\theta = 0.1$ and the MSE of the estimate from the non-adaptive design is roughly $100p_0\%$ of the MSE of the estimate from the adaptive design. Again the simulated MSE ratio is approximated well by the asymptotic theory.

5 Conclusions

A major motivation for this work was the observation from simulation studies that the benefit of adaptive designs in terms of estimation efficiency is sometimes less in magnitude than intuitively expected, and crucially depends on the underlying models and assumptions (Dragalin et al., 2011). This paper provides a theoretical confirmation of these empirical results in a well controllable situation, taking aside possible additional influence factors. We derive analytic expansions for the mean squared error of the MLE based on an adaptive design, which enables the analytical comparison of adaptive with non-adaptive designs.

One main result of this paper is that one can theoretically expect a benefit of adaptive designs for

sufficiently large sample sizes for a broad class of nonlinear regression models. When the sample size is small, however, the remainder in (2.20) is non-negligible. This can lead to situations, where the non-adaptive design outperforms the adaptive design, as illustrated with three practical examples. In some applications, further general conclusions can be derived. For example, the efficiency ratio (4.2) reveals that adaptive designs are always more efficient than non-adaptive designs for sufficiently small variances.

In practice more complex models than considered in this paper are often used. The methodology presented in this paper remains applicable when using computer algebra systems. Moreover, the results enable us to understand the relationship of key factors impacting the relative efficacy of adaptive designs compared to non-adaptive designs. For example, in the logistic regression example from Section 4.2 the relative efficiency depends only on three factors: the unknown degree of misspecification γ and the two design parameters p_0 and N_0 . Using analytical methods, closed form expressions can be derived for the relationship of these factors, giving insight into their impact on efficiency performance. By contrast, simulation studies, even if performed comprehensively, do not provide theoretical explanations and are mainly used to provide empirical evidence.

6 Appendix

6.1 Assumptions

We consider independent observations $(x_1, Y_1), \dots, (x_N, Y_N)$ at experimental conditions x_1, \dots, x_n and define by $f(Y_i, x_i, \theta)$ the density of Y_i at experimental condition x_i ($i = 1, \dots, N$). Let d denote the number of unknown model parameters. We assume that the approximate locally optimal design ξ_τ has at least $k \geq d$ support points, such that the matrix $M(\xi_\tau, \theta)$ is positive definite for all $\theta, \tau \in \Theta$. We assume further that the density f is three times continuously differentiable with respect to the parameter θ (for all x, y), continuously differentiable with respect to the variable x and that all derivatives of the integral

$$\int f(y, x, \theta) dy$$

can be obtained by differentiating under the integral sign. Suppose that for each $x \in \mathcal{X}$ and $\theta = (\vartheta_1, \dots, \vartheta_d)^T \in \Theta$ there exists an open neighbourhood \mathcal{U}_θ , such that for all $i, j \in \{1, \dots, d\}$ there exist functions $M_2(y, x, \theta), M_3(y, x, \theta)$, such that for all $\tau \in \mathcal{U}_\theta$

$$\left| \frac{\partial^2}{\partial \vartheta_i \partial \vartheta_j} \log f(y, x, \theta) \right| \leq M_2(y, x, \theta)$$

$$\left| \frac{\partial^3}{\partial \vartheta_i \partial \vartheta_j \partial \vartheta_k} \log f(y, x, \theta) \right|_{\theta=\tau} \leq M_3(y, x, \theta)$$

and

$$\int M_j(y, x, \theta) f(y, x, \theta) dy < \infty; \quad j = 2, 3.$$

Additionally, we assume that for $i = 1, 2, 3$ the random variables $\frac{\partial^i}{\partial \theta} \log f(Y_i, x, \theta)$ have bounded second and third absolute moments uniformly with respect to $x \in \mathcal{X}$.

6.2 Proof of Theorem 2.1

A) Asymptotic distribution of $\sqrt{N_0}(\hat{\theta}_1 - \theta)$: The following calculations are standard but included here, because a good understanding of the “classical” case turns out to be helpful for the situation of an adaptive design. The MLE $\hat{\theta}_1$ based on the observations $(x_1, Y_1), \dots, (x_{N_0}, Y_{N_0})$ is a solution of the equation

$$0 = \sum_{i=1}^{N_0} \frac{\partial}{\partial \theta} \log f(Y_i, x_i, \hat{\theta}_1). \quad (6.1)$$

Assume that the observations are taken according to a design ξ with a $k \geq d$ different experimental conditions, say t_1, \dots, t_k , with positive weights w_1, \dots, w_k . Because $N_0 w_i \rightarrow \infty$ ($i = 1, \dots, k$) it follows by the strong law of large numbers that

$$\frac{1}{N_0} \sum_{i=1}^{N_0} \frac{\partial^2}{\partial^2 \theta} \log f(Y_i, x_i, \theta) \longrightarrow -M(\xi, \theta) := \int \int \frac{\partial^2}{\partial^2 \theta} \log f(y, x, \theta) f(y, x, \theta) dy d\xi(x),$$

and a standard argument shows that

$$M(\xi, \theta) = \int \int \left(\frac{\partial}{\partial \theta} \log f(y, x, \theta) \right)^T \left(\frac{\partial}{\partial \theta} \log f(y, x, \theta) \right) dy d\xi(x).$$

Therefore a Taylor expansion yields for (6.1)

$$0 = \sqrt{N_0}(\hat{\theta}_1 - \theta) \frac{1}{N_0} \sum_{i=1}^{N_0} \frac{\partial^2}{\partial^2 \theta} \log f(Y_i, x_i, \theta) + \frac{1}{\sqrt{N_0}} \sum_{i=1}^{N_0} \frac{\partial}{\partial \theta} \log f(Y_i, x_i, \theta) + O_p\left(\frac{1}{\sqrt{N_0}}\right).$$

This gives for any design ξ with positive masses at $k \geq d$ points

$$\sqrt{N_0}(\hat{\theta}_1 - \theta) = M^{-1}(\xi, \theta) \frac{1}{\sqrt{N_0}} \sum_{i=1}^{N_0} \frac{\partial}{\partial \theta} \log f(Y_i, x_i, \theta) + O_p\left(\frac{1}{\sqrt{N_0}}\right).$$

Now the sum on the right hand side is a sum of independent random variables, and the central limit theorem shows that the random variable

$$\frac{1}{\sqrt{N_0}} \sum_{i=1}^{N_0} \frac{\partial}{\partial \theta} \log f(Y_i, x_i, \theta)$$

has an asymptotic (d -dimensional) normal distribution with mean 0 and covariance matrix $M(\xi, \theta)$. Therefore the first assertion of Theorem 2.1 follows using the locally optimal design ξ_{θ_0} and observing the definition of the matrix I in (2.2).

B) Asymptotic representation of $\sqrt{N}(\hat{\theta}_B - \theta)$: The MLE $\hat{\theta}_B$ from the total sample satisfies

$$0 = \sum_{i=1}^{N_0} \frac{\partial}{\partial \theta} \log f(Y_i, x_i(\theta_0), \hat{\theta}_B) + \sum_{i=n_0+1}^{N_0+N_1} \frac{\partial}{\partial \theta} \log f(Y_i, x_i(\hat{\theta}_1), \hat{\theta}_B)$$

where $x_{N_0+1}(\hat{\theta}_1), \dots, x_{N_0+N_1}(\hat{\theta}_1)$ are the design points from the second step (which depend on the parameter estimate $\hat{\theta}_1 = \tilde{\theta}_1 + O_p(\frac{1}{N_0})$ obtained in the first stage). The same argument as in the first part of the proof yields

$$\begin{aligned} 0 &= \frac{1}{N} \sum_{i=1}^{N_0} \frac{\partial^2}{\partial^2 \theta} \log f(Y_i, x_i(\theta_0), \theta) \sqrt{N}(\hat{\theta}_B - \theta) + \frac{1}{\sqrt{N}} \sum_{i=1}^{N_0} \frac{\partial}{\partial \theta} \log f(Y_i, x_i(\theta_0), \theta) + O_p\left(\frac{1}{\sqrt{N}}\right) \\ &+ \frac{1}{N} \sum_{i=N_0+1}^{N_0+N_1} \frac{\partial^2}{\partial^2 \theta} \log f(Y_i, x_i(\hat{\theta}_1), \theta) \sqrt{N}(\hat{\theta}_B - \theta) + \frac{1}{\sqrt{N}} \sum_{i=N_0+1}^{N_0+N_1} \frac{\partial}{\partial \theta} \log f(Y_i, x_i(\hat{\theta}_1), \theta) + O_p\left(\frac{1}{\sqrt{N}}\right) \\ &= -(p_0 I(\theta, \theta_0) + p_1 I(\theta, \tilde{\theta}_1)) \sqrt{N}(\hat{\theta}_B - \theta) + \sqrt{p_0} I^{1/2}(\theta, \theta_0) Z_{0, N_0} + \sqrt{p_1} I^{1/2}(\theta, \tilde{\theta}_1) Z_{1, N_1} + O_p\left(\frac{1}{\sqrt{N}}\right), \end{aligned}$$

where Z_{0, N_0} and Z_{1, N_1} are defined in (2.7) and (2.8), and we have used the fact that the design points $x_i(\theta)$ and the density f of the locally optimal design are continuously differentiable with respect to θ and x , respectively. This gives

$$\sqrt{N}(\hat{\theta}_B - \theta) = (p_0 I(\theta, \theta_0) + p_1 I(\theta, \tilde{\theta}_1))^{-1} (\sqrt{p_0} I^{1/2}(\theta, \theta_0) Z_{0, N_0} + \sqrt{p_1} I^{1/2}(\theta, \tilde{\theta}_1) Z_{1, N_1}) + O_p\left(\frac{1}{\sqrt{N}}\right). \quad (6.2)$$

and proves the second assertion of Theorem 2.1.

6.3 Proof of Corollary 2.1

Recall the definition of $\tilde{\theta}_1$ in (2.6). We consider a Taylor expansion of the function $I(\theta, \tilde{\theta}_1)$ defined in (2.2), which gives for the element in the position (i, j) of the matrix $I(\theta, \theta + \tau)$

$$(I(\theta, \theta + \tau))_{ij} = (I(\theta, \theta))_{ij} + \nabla (I(\theta, \theta))_{ij} \tau + \frac{1}{2} \tau^T \nabla^2 (I(\theta, \theta))_{ij} \tau + o(\tau^2), \quad (6.3)$$

where the derivatives are taken with respect to the second argument of the matrix $I(\theta, \tau)$ and evaluated at $\tau = \theta$. Writing the expansion (6.3) in matrix form and using the notation $\tau = \frac{1}{\sqrt{N_0}} I^{-1/2}(\theta, \theta_0) Z_{0, N_0}$ [see Theorem 2.1] yields

$$I(\theta, \tilde{\theta}_1) = I(\theta, \theta) + \frac{1}{\sqrt{N_0}} D_1(\theta, Z_{0, N_0}) + \frac{1}{2N_0} D_2(\theta, Z_{0, N_0}) + o_p\left(\frac{1}{N_0}\right) = I(\theta, \theta) + R(\theta, Z_{0, N_0}) + o_p\left(\frac{1}{N_0}\right),$$

where the matrices $R(\theta, Z_{0, N_0})$, $D_1(\theta, Z_{0, N_0})$ and $D_2(\theta, Z_{0, N_0})$ are defined in (2.15) - (2.17), respectively. Assuming $A, B \in \mathbb{R}^{d \times d}$ with $\det A \neq 0$ and letting $\varepsilon \rightarrow 0$, we use the expansion

$$(A + \varepsilon B)^{-1} = (I_d + \varepsilon A^{-1} B)^{-1} A^{-1} = A^{-1} (I_d - \varepsilon B A^{-1} + \varepsilon^2 B A^{-1} B A^{-1}) + o(\varepsilon^2)$$

and obtain from (2.10) and (2.11) the representations (2.12) and (2.13). \square

6.4 Proof of Theorem 3.1

Note that the mapping defined in (2.2) can be rewritten as

$$I(\theta, \tau) = \int \left(\frac{\partial}{\partial \theta} \log f(y, x(\tau), \theta) \right)^2 f(y, x(\tau), \theta) dy, \quad (6.4)$$

where we have used the fact that ξ_τ is a one-point design supported at the point $x(\tau)$. This yields for the derivative of the first order in (6.3)

$$\nabla I(\theta, \tau) \Big|_{\tau=\theta} = \frac{\partial}{\partial x} \int f(y, x, \theta) \left(\frac{\partial}{\partial \theta} \log f(y, x, \theta) \right)^2 dy \Big|_{x=x(\theta)} \cdot \frac{\partial}{\partial \tau} x(\tau) \Big|_{\tau=\theta} = 0,$$

where the last identity follows from (3.1). Similarly, we obtain for the second derivative

$$g(\theta) := \nabla^2 I(\theta, \tau) \Big|_{\tau=\theta} = \frac{\partial^2}{\partial^2 x} \int f(y, x, \theta) \left(\frac{\partial}{\partial \theta} \log f(y, x, \theta) \right)^2 dy \Big|_{x=x(\theta)} \cdot \left(\frac{\partial}{\partial \tau} x(\tau) \Big|_{\tau=\theta} \right)^2,$$

where $g(\theta) < 0$, because $x(\theta)$ maximizes the function in (6.4). Consequently, we have $D_1(\theta, Z_{0, N_0}) = 0$, $S(\theta, Z_{0, N_0}) = 0$ and obtain for the matrix $R(\theta, Z_{0, N_0})$ defined by (2.15)

$$R(\theta, Z_0) = \frac{1}{2N_0} \frac{g(\theta) Z_{0, N_0}^2}{I(\theta, \theta_0)},$$

which, together with (2.12) and (2.13), yields as approximation for the variance of

$$\begin{aligned} \text{Var}(\hat{\gamma}_B) &= \text{Var}(E[\hat{\gamma}_B | Y_1, \dots, Y_{N_0}]) + E[\text{Var}(\hat{\gamma}_B | Y_1, \dots, Y_{N_0})] \\ &\approx E \left[\frac{p_0 Z_{0, N_0}^2 I(\theta, \theta_0)}{H^2(\theta, \theta_0)} \left(1 - \frac{p_1 g(\theta) Z_{0, N_0}^2}{2N_0 I(\theta, \theta_0) H(\theta, \theta_0)} \right)^2 \right] \\ &\quad + p_1 E \left[\frac{1}{H^2(\theta, \theta_0)} \left(1 - \frac{p_1 g(\theta) Z_{0, N_0}^2}{2N_0 I(\theta, \theta_0) H(\theta, \theta_0)} \right)^2 \left(I(\theta, \theta) + \frac{Z_{0, N_0}^2 g(\theta)}{2N_0 I(\theta, \theta_0)} \right) \right] \\ &= \frac{p_0 I(\theta, \theta_0)}{H^2(\theta, \theta_0)} \left(1 - \frac{3p_1 g(\theta)}{N_0 I(\theta, \theta_0) H(\theta, \theta_0)} \right) \\ &\quad + \frac{p_1 I(\theta, \theta)}{H^2(\theta, \theta_0)} \left\{ 1 + \frac{g(\theta)(p_0 I(\theta, \theta_0) - p_1 I(\theta, \theta))}{2N_0 I(\theta, \theta_0) H(\theta, \theta_0) I(\theta, \theta)} \right\} \\ &= \frac{1}{H(\theta, \theta_0)} - \frac{g(\theta) p_1 (5p_0 I(\theta, \theta_0) + p_1 I(\theta, \theta))}{2N_0 H^3(\theta, \theta_0) I(\theta, \theta_0)}. \end{aligned}$$

This proves the assertion. \square

Acknowledgements The authors thank Martina Stein, who typed parts of this manuscript with considerable technical expertise. This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823) of the German Research Foundation (DFG).

References

- Atkinson, A. C., Donev, A., and Tobias, R. D. (2007). *Optimum Experimental Designs, with SAS*. Oxford University Press, Oxford.
- Bornkamp, B., Bretz, F., Dette, H., and Pinheiro, J. (2010). Response-adaptive dose-finding under model uncertainty. (submitted) <http://www.ruhr-uni-bochum.de/mathematik3/research/index.html>.
- Bornkamp, B., Bretz, F., Dmitrienko, A., Enas, G., Gaydos, B., Hsu, C. H., Koenig, F., Krams, M., Liu, Q., Neuenschwander, B., Parke, T., Pinheiro, J., Roy, A., Sax, R., and Shen, F. (2007). Innovative approaches for designing and analyzing adaptive dose-ranging trials (with discussion). *Journal of Biopharmaceutical Statistics*, 17:965–995.
- Box, M. J. (1971). Bias in nonlinear estimation. *Journal of the Royal Statistical Society, Ser. B*, 33:171–201.
- Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: a review. *Statistical Science*, 10(3):273–304.
- Chang, H. and Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *Annals of Statistics*, 37:1466–1488.
- Chernoff, H. (1953). Locally optimal designs for estimating parameters. *Annals of Mathematical Statistics*, 24:586–602.
- Dette, H. (1997). Designing experiments with respect to “standardized” optimality criteria. *Journal of the Royal Statistical Society, Ser. B*, 59:97–110.
- Dette, H., Bretz, F., Pepelyshev, A., and Pinheiro, J. (2008). Optimal designs for dose finding studies. *Journal of the American Statistical Association*, 103(483):1225–1237.
- Dette, H., Kiss, C., Bevanda, M., and Bretz, F. (2010). Optimal designs for the EMAX, log-linear and exponential models. *Biometrika*, 97(2):513–518.
- Dette, H. and Neugebauer, H. M. (1996). Bayesian optimal one point designs for one parameter nonlinear models. *Journal of Statistical Planning and Inference*, 52:17–31.
- Dragalin, V., Bornkamp, B., Bretz, F., Miller, F., Padmanabhan, S. K., Patel, N., Perevozskaya, I., Pinheiro, J., and Smith, J. R. (2011). A simulation study to compare new adaptive dose-ranging designs. *Statistics in Biopharmaceutical Research*, in press.
- Dragalin, V., Hsuan, F., and Padmanabhan, S. K. (2007). Adaptive designs for dose-finding studies based on sigmoid e_{\max} model. *Journal of Biopharmaceutical Statistics*, 17(6):1051–1070.

- Fang, X. and Hedayat, A. S. (2008). Locally D-optimal designs based on a class of composed models resulted from blending Emax and one-compartment models. *Annals of Statistics*, 36:428–444.
- Ford, I. and Silvey, S. D. (1980). A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika*, 67(2):381–388.
- Ford, I., Torsney, B., and Wu, C. F. J. (1992). The use of canonical form in the construction of locally optimum designs for nonlinear problems. *Journal of the Royal Statistical Society, Ser. B*, 54:569–583.
- Haines, L. M., Perevozskaya, I., and Rosenberger, W. F. (2003). Bayesian optimal design for Phase I clinical trials. *Biometrics*, 59:591–600.
- He, Z., Studden, W. J., and Sun, D. (1996). Optimal designs for rational models. *Annals of Statistics*, 24:2128–2142.
- Miller, F., Guilbaud, O., and Dette, H. (2007). Optimal designs for estimating the interesting part of a dose-effect curve. *Journal of Biopharmaceutical Statistics*, 17:1097–1115.
- Montepiedra, G. and Yeh, A. B. (2004). Two-stage designs for identification and estimation of polynomial models. *Computational Statistics and Data Analysis*, 46:531–546.
- Müller, C. H. and Pázman, A. (1998). Applications of necessary and sufficient conditions for maximum efficient designs. *Metrika*, 48:1–19.
- Pukelsheim, F. (2006). *Optimal Design of Experiments*. SIAM, Philadelphia.
- Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. University of Chicago Press, Chicago.
- Roy, A., Ghosal, S., and Rosenberger, W. (2008). Convergence properties of sequential Bayesian D-optimal designs. *Journal of Statistical Planning and Inference*, 139:425–440.
- Wiens, D. P. (2009). Robust discrimination designs, with Matlab code. *Journal of the Royal Statistical Society, Ser. B*, 71:805–829.
- Yang, M. (2010). On the de la Garza Phenomenon. *Annals of Statistics*, 38(4):2499–2524.
- Zhu, W. and Wong, W. K. (1998). Multiple-objective designs in a dose-response experiment. *IMS Lecture Notes*, 34:73–82.