Finite sample performance of sequential designs for model identification

Holger Dette
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum
Germany
email: holger.dette@ruhr-uni-bochum.de
FAX: +49 2 34 3214 559

Robert Kwiecien
RWTH Aachen
Institut für Medizinische Statistik
Pauwelstr. 30
52074 Aachen
Germany
email: rkwiecien@medfak.rwth-aachen.de

October 23, 2003

Abstract

Classical regression analysis is usually performed in two steps. In a first step an appropriate model is identified to describe the data generating process and in a second step statistical inference is performed in the identified model. An intuitively appealing approach to the design of experiment for these different purposes are sequential strategies, which use parts of the sample for model identification and adapt the design according to the outcome of the identification steps. In this paper we investigate the finite sample properties of two sequential design strategies, which were recently proposed in the literature. A detailed comparison of sequential designs for model discrimination in several regression models is given by means of a simulation study. Some non-sequential designs are also included in the study.

Keywords and Phrases: optimal design, robust design, discrimination design, sequential design, F-test

1 Introduction

In many applications precise knowledge of a statistical model describing the data generating process is not available and the analysis of the data is usually performed in two steps. In a first step the data is used to identify an appropriate model from a given class of competing models and the second step consists in the statistical analysis in the identified model (parameter estimation, prediction etc.). An experimental design for one task may be exceptionally inefficient for the other. Consider for example the regression model $Y = f(X) + \varepsilon$, where ε has a centered normal distribution. The real valued function f is assumed to belong to a given class of linear nested models, say $\mathcal{F} = \{g_1, \dots, g_k\}$, where

$$g_j(x) = \sum_{i=1}^{\ell_j} \beta_{ji} f_i(x), \qquad j = 1, \dots, k, \tag{1.1}$$

are competing nested models, $1 \le \ell_1 < \ell_2 < \dots < \ell_k$ and $f_1, \dots, f_{\ell_k}$ are given and known regression functions.

Typically, the discrimination between these models is performed by a sequence of step-wise F-tests for the hypotheses

$$H_{0j} : f = g_{j-1} \quad \text{versus} \quad H_{1j} : f = g_j,$$

[see Anderson (1962)]. More precisely, starting with the model g_k the hypotheses

$$H_{0j} := \begin{pmatrix} \beta_{j,\ell_{j-1}+1} \\ \vdots \\ \beta_{j,\ell_j} \end{pmatrix} = 0 \qquad j = k, \dots, 2 \tag{1.2}$$

are subsequently tested at a specified level, say $\alpha_j \in (0,1)$, and we select the model g_{j_0} for which the first test rejects the corresponding hypothesis. The test in the jth step is the classical F-test, which is based on the statistic

$$F_j = \frac{RSS_{j-1} - RSS_j}{RSS_j} \cdot \frac{N - \ell_j}{\ell_j - \ell_{j-1}}, \tag{1.3}$$

where RSS_j denotes the residual sum of squares from the total sample in the jth model. The inference in the identified model is usually done by classical methods for linear regression [see e.g. Seber (1977)].

An intuitive approach to the design of experiment for these different purposes are sequential strategies, which use parts of the sample for model identification and adapt the design according to the outcome of the experiments sequentially. In the literature there are essentially two proposals for sequential designs, which are useful for the rather different objects of discrimination and inference in the identified model. Montepiedra and Yeh (1997, 2002) proposed a two stage sequential procedure. One part of the sample is used for the identification of the appropriate model, while the design in the second stage is chosen for a most efficient inference in the identified model (see Section 2.2 for more details). In general, the analysis based on this sequential design is very difficult, because this procedure usually yields dependencies in the data and classical estimation and distribution theory is not directly applicable.

Recently, Biswas and Chaudhuri (2002) proposed a new sequential strategy, where the object of the design is to select the "correct" model from the family of nested models as well as to efficiently estimate the parameters associated with that model. They showed that the corresponding tests used in the discrimination step keep their preassigned levels and that the sequential design is able to identify the "correct" model with a probability converging to one if the sample size tends to infinity. Moreover, the sequential design converges to the optimal design corresponding to the "true" model.

The purpose of the present paper is to obtain a better understanding of the performance of sequential designs for model discrimination. We present a detailed finite sample size comparison of the two sequential design strategies for model identification with three main objectives. First, we investigate how the size of the samples in the different stages of the sequential designs affects the performance of the procedures. It is demonstrated that performance of the sequential designs depends sensitively on the choice of the sample sizes in the individual steps of the sampling procedure. Secondly, we compare the sequential designs of Montepiedra and Yeh (1997, 2002) with the designs proposed by Biswas and Chaudhuri (2002). Third, we compare these designs with some non-sequential designs proposed in the literature [see Lau and Studden (1985), Dette (1995), Dette and Röder (1997)].
In Section 2 we briefly review the two different concepts of sequential design for model discrimination. Section 3 contains a detailed investigation of the properties of sequential designs by means of a simulations study. We consider univariate, multivariate polynomials and Fourier regression models and investigate how the choice of the sampling proportions affects the performance of the different sequential procedures. For each sequential procedure we identify a "best configuration for the sample sizes in the different steps. Finally, in Section 4 we present a comparison of the best sequential procedures, which also includes some non-sequential designs. It is demonstrated that non-sequential discrimination designs and the procedure of Montepiedra and Yeh (1997, 2002) usually yield substantially smaller rates of misspecification compared to the sequential procedure of Biswas and Chaudhuri (2002). Moreover, the efficiencies between the different designs in the identified model are fairly comparable. Finally, some conclusions and recommendations for the design of experiments for model identification and parameter estimation in the identified model are given in Section 5.

2 Two sequential strategies revisited

2.1 The sequential design of Biswas and Chaudhuri (2002)

The method of Biswas and Chaudhuri (2002) starts with a convex combination of the D-optimal designs for the individual models and this design is updated in several steps. To be precise, let ξ_j denote the D-optimal design for the regression model g_j $(j = 1, \ldots, k)$ and assume that $N = m_0 + m_1 + \ldots + m_s$ experiments are permitted. Let $\alpha_i^{(0)} = \frac{1}{k}$ $(i = 1, \ldots, k)$ and consider as the design for the first stage $\xi^{(0)} := \sum_{i=1}^{k} \alpha_i^{(0)} \xi_i$ the uniform mixture of the D-optimal designs for the models $g_1, \ldots, g_k$.
The first m_0 observations are chosen at experimental conditions sampling randomly from the design $\xi^{(0)}$, which defines the initial design. This design is sequentially updated by $s \geq 0$ steps in the following way. For $r = 1, \ldots, s$ the sample of $m_0 + \ldots + m_{r-1}$ observations is used to test subsequently the hypotheses (1.2). The hypothesis H_{0j} will be rejected with some specified

level α_r^j if $T_r^j > c_r^j$, where the statistic T_r^j is defined by

$$T_r^j = \frac{\sum_{i=0}^{r-1} \left\{ RSS_{j-1}^{(i)} - RSS_j^{(i)} \right\}}{\sum_{i=0}^{r-1} RSS_j^{(i)}} \cdot \frac{\sum_{i=0}^{r-1} (m_i - \ell_j)}{r(\ell_j - \ell_{j-1})}, \tag{2.1}$$

and $RSS_j^{(i)}$ denotes the residual sum of squares based on a least squares fit in the model g_j from the m_i observations in the ith step ($i = 0, \ldots, r-1$). Note that in contrast to the classical F-statistic defined in (1.3) these sums are calculated separately for each sample of m_i observations ($i = 0, \ldots, r-1$). It was shown by Biswas and Chaudhuri (2002) that the critical values c_r^j can be obtained as the α_r^j quantiles of an F-distribution with $r(\ell_j - \ell_{j-1})$ and $\sum_{i=0}^{r} (m_i - \ell_j)$ degrees of freedom.
The design for the next m_r observations is then defined as follows. Let j_0 denote the first index for which the null-hypotheses $H_{0k}, \ldots H_{0j_0+1}$ are accepted and the null-hypothesis H_{0j_0} is rejected (if all tests accept the corresponding null-hypothesis we put $j_0 = 1$), then the model g_{j_0} is selected and the design is updated by $\xi^{(r)} = \sum_{i=1}^{k} \alpha_i^{(r)} \xi_i$, where the new weights $\alpha_i^{(r)}$ are defined by

$$\alpha_i^{(r)} = \begin{cases} \left(\alpha_i^{(r-1)} + \frac{1}{(k+r-1)} \right) / \sum & \text{if } i = j_0 \\ \alpha_i^{(r-1)} / \sum & \text{if } i \neq j_0 \end{cases}$$

and $\sum$ is a normalizing constant defined by the condition $\sum_{i=1}^{k} \alpha_i^{(r)} = 1$. The next m_r design points are then generated from the design $\xi^{(r)}$.
This procedure is repeated to obtain $m_0 + \ldots + m_s = N$ observations, where the observations at the rth stage are taken by sampling randomly from the design $\xi^{(r)}$. Finally the sequence of tests for the hypotheses $H_{0k}, H_{0k-1}, \ldots, H_{02}$ based on the statistics $T_{s+1}^k, \ldots, T_{s+1}^2$ is performed for the total sample and the model g_{j_0} is chosen for which the corresponding test rejects the null-hypothesis for the first time (in other words: $T_{s+1}^i \leq c_{s+1}^i; i = k, \ldots, j_0 + 1; T_{s+1}^{j_0} > c_{s+1}^{j_0}$). It was shown by Biswas and Chaudhuri (2002) that under appropriate (asymptotic) assumptions on $N, s, m_0, \ldots, m_s$ and α_r^j $(r = 0, \ldots, s, j = 1, \ldots, k)$ this procedure identifies the "correct" model with probability converging to one. Moreover, the information matrix of the design $\xi^{(s)}$ in the identified model , say g_{j_0}, converges to the information matrix of the D-optimal design for this model, say M_{j_0}, and the corresponding parameter estimate in the identified model has an asymptotic normal distribution with asymptotic mean $\beta_{j_0} \in \mathbb{R}^{\ell_{j_0}}$ and covariance matrix $\frac{1}{N} M_{j_0}^{-1}$ [see Biswas and Chaudhuri (2002) for more details].

2.2 The two-stage procedure of Montepiedra and Yeh (1997, 2002)

Recently, Montepiedra and Yeh (1997, 2002) proposed a two-stage procedure for the construction of efficient designs for model discrimination and estimation of the parameters in the identified model. The total sample of N observations is splitted into two parts, say $n_0 + n_1 = N$. In a first step n_0 observations are taken according to a non-sequential discrimination design. The data obtained from this design is used to select the appropriate model among $g_1, \ldots, g_k$, say

g_{j_0}. For the second step let $X_{j_0}(n_0) \in \mathbb{R}^{n_0 \times \ell_{j_0}}$ denote the design matrix of the design from the first step in the model g_{j_0}, then the next n_1 observations are taken such that the determinant

$$|f \cdot X_{j_0}^T(n_0)X_{j_0}(n_0) + (1-f) \cdot X_{j_0}^T(n_1)X_{j_0}(n_1)| \tag{2.2}$$

becomes maximal, where $X_{j_0}(n_1)$ is the design matrix in the model g_{j_0} obtained from these new observations and $f \in (0,1)$ is a weight to be determined by the experimenter. After all $N = n_0 + n_1$ observations have been collected from the two-stage design, the model selection step will be repeated using the total sample and the parameters in the identified model are estimated. Although this procedure is intuitively appealing, it has not been shown so far that this will yield a consistent procedure in a strong statistical sense. For example, the (F)-tests in the final step will usually not keep the preassigned level and it is not obvious that for an increasing sample size (in both steps) the "correct" model is identified with a probability converging to 1.

3 How many observations should be taken at each stage?

Usually efficiencies are used to compare different designs, but due to the randomness of the sequential procedures these efficiencies cannot be calculated directly. In this section we use a simulation to evaluate the performance of the different designs. We simulated 5.000 times a specific scenario for the designs described in Section 2. In each simulation we investigate the performance of the different discrimination strategies, where the level of the corresponding F-tests was always taken as 5%. These simulations are used for two purposes:

• The calculation of the rate of a "correct" identification of the underlying model (we simply count how often the "correct" model is chosen in the 5.000 trials).

• The determination of the D-efficiency

$$|X_{j_0}^T X_{j_0}|^{1/\ell_j}$$

in the identified model. For this purpose we define an "averaged" sequential design and this design is used for the efficiency calculations (for more details see Section 3.1 and 3.2).

We considered three cases of regression models, the univariate cubic polynomial model, a quadratic model with two factors and the Fourier-regression model of degree 2, for which the D-optimal designs in the nested submodels [which are needed in the sequential procedure of Biswas and Chaudhuri (2002)] are known from the literature. For the sake of simplicity we restrict ourselves to the approximate design case [see Kiefer (1974)] and in our simulation study the experimental conditions are chosen randomly from the corresponding approximate designs.

3.1 Sample sizes in the procedure of Biswas and Chaudhuri

In this subsection we investigate the performance of the sequential procedure of Biswas and Chaudhuri (2002) with respect to the choice of the number of stages and with respect to the sample sizes in the different steps. In Example 3.1 and 3.2 we consider a polynomial with one and two factors, respectively, while Example 3.3 deals with a Fourier regression model. We give a rather complete description of Example 3.1 and briefly summarize the essential features for the other two examples.

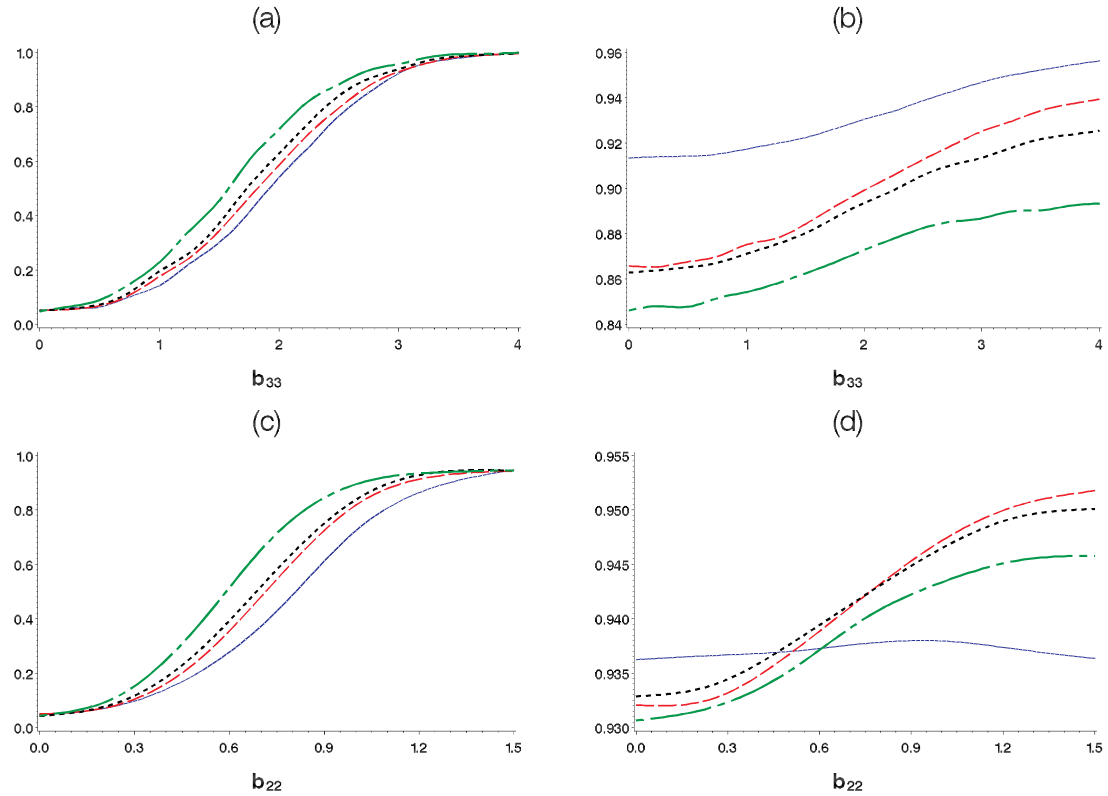

Figure 3.1: *Simulated probabilities of "correct" specification (a), (c) and efficiencies (b), (d) in the identified cubic and quadratic regression model for various sequential designs of Biswas and Chaudhuri (2002) with $N = 100$ observations. Solid lines: design (A); dashed lines: design (B); dotted lines: design (C); nearly solid lines: design (D). The designs are defined in (3.2).*

Example 3.1 (Cubic regression). Consider the cubic regression model

$$g_3(x) = \beta_{30} + \beta_{31}x + \beta_{32}x + \beta_{32}x^2 + \beta_{33}x^3; \tag{3.1}$$

on the interval $[-1, 1]$. For the sake of brevity we only report the results for the sample size $N = 100$ with four different choices for the stages; simulations for other cases showed a very similar picture. We considered the following partitions in the sequential scheme of Biswas and

Chaudhuri (2002)

$$
(3.2) \qquad \begin{array}{lll}
\frac{m_0}{N} = \ldots = \frac{m_9}{N} = \frac{1}{10} & & (A) \\
\frac{m_0}{N} = \ldots = \frac{m_4}{N} = \frac{1}{5} & & (B) \\
\frac{m_0}{N} = \frac{m_2}{N} = \frac{m_3}{N} = \frac{3}{10}, \frac{m_4}{N} = \frac{1}{10} & & (C) \\
\frac{m_0}{N} = \frac{m_1}{N} = \frac{1}{2} & & (D)
\end{array}
$$

where N denote the total sample size, here $N = 100$. Figures 3.1(a) and (c) show the rate of "correct" specification if the cubic or quadratic model are "correct", respectively. The model under consideration was $g_3(x) = 1 + x + x^2 + b_{33}x^3 \quad (b_{33} \neq 0)$ in the cubic case and $g_2(x) = 1 + x + b_{22}x^2$ for the quadratic case. We also performed a simulation study in the linear case $g_1(x) = 1 + b_{11}x$. If this model is the "correct" one, all designs in (3.2) behaved similary and for this reason the results are not depicted. From Figures 3.1(a) and (c) we observe that the rate of "correct" specification of the sequential procedure of Biswas and Chaudhuri (2002) is improved, if large sample sizes are chosen in the first steps of the sequential procedure. The allocation scheme (D) yields the uniformly best rates of "correct" model specification, independently of the "correct" model. This performance can even be improved by choosing all 100 observations in the first step which does not yield a sequential procedure.
The results for the corresponding efficiencies of the "averaged" designs are presented in Figures 3.1 (b) and (d). From Figure 3.1 (b) we observe the opposite behaviour for the efficiencies in the cubic case. Here the sequential design (A) yields the best efficiency and the performance is usually improved by using finer partitions. On the other hand, if the quadratic model is "correct", the sequential design with the finest partition has only a better performance if the parameter β_{22} is relatively small, for moderate values of β_{22}, the design (C) is preferrable, while for large values of the parameter β_{22} the design (B) yields the best efficiencies [see Figure 3.1 (d)]. However, it should be remarked that the differences between the efficiencies of the different designs are relatively small compared to the differences between the probabilities of "correct" specification. The D-efficiencies of the four designs in the cubic, quadratic and linear models differ by at most 7%, 2%, 6%, respectively. On the other hand, we observe more substantial differences between the probabilities of "correct" specification in the cubic and quadratic model (provided that the parameter β_{33} of the highest coefficient is not too large or too small). For example, in the cubic model the probabilities of "correct" specification between the best and worst sequential design can differ between 70% and 35% if the parameter β_{33} varies in the interval $[1, 2]$ [see Figure 3.1 (a)]. Similarly, in the quadratic case we observe a difference between 33% and 81% if the parameter β_{22} varies in the interval $[0.3, 1]$ [see Figure 3.1(c)].

Example 3.2 (Quadratic regression with two factors). In order to study the question if the properties of the sequential designs observed in the previous example depend on the number of factors in the model we investigate in our second example a quadratic regression model with two factors, that is

$$
(3.3) \qquad g_5(x) = 1 + x_1 + x_2 + x_1x_2 + x_1^2 + b_{55}x_2^2; \qquad x_1, x_2 \in [-1, 1].
$$

The nested models are given by

$$
(3.4) \qquad g_4(x) = 1 + x_1 + x_2 + x_1x_2 + b_{44}x_1^2
$$

$$
\begin{aligned}
(3.5) \qquad g_3(x) &= 1 + x_1 + x_2 + b_{33}x_1x_2 \\
(3.6) \qquad g_2(x) &= 1 + x_1 + b_{22}x_2 \\
(3.7) \qquad g_1(x) &= 1 + b_{11}x_1.
\end{aligned}
$$

The D-optimal designs used for the sequential procedure of Biswas and Chaudhuri (2002) can be found by standard methods and are not stated explicitly for the sake of brevity; see also Kono (1962) and Lim and Studden (1988) for some particular cases.

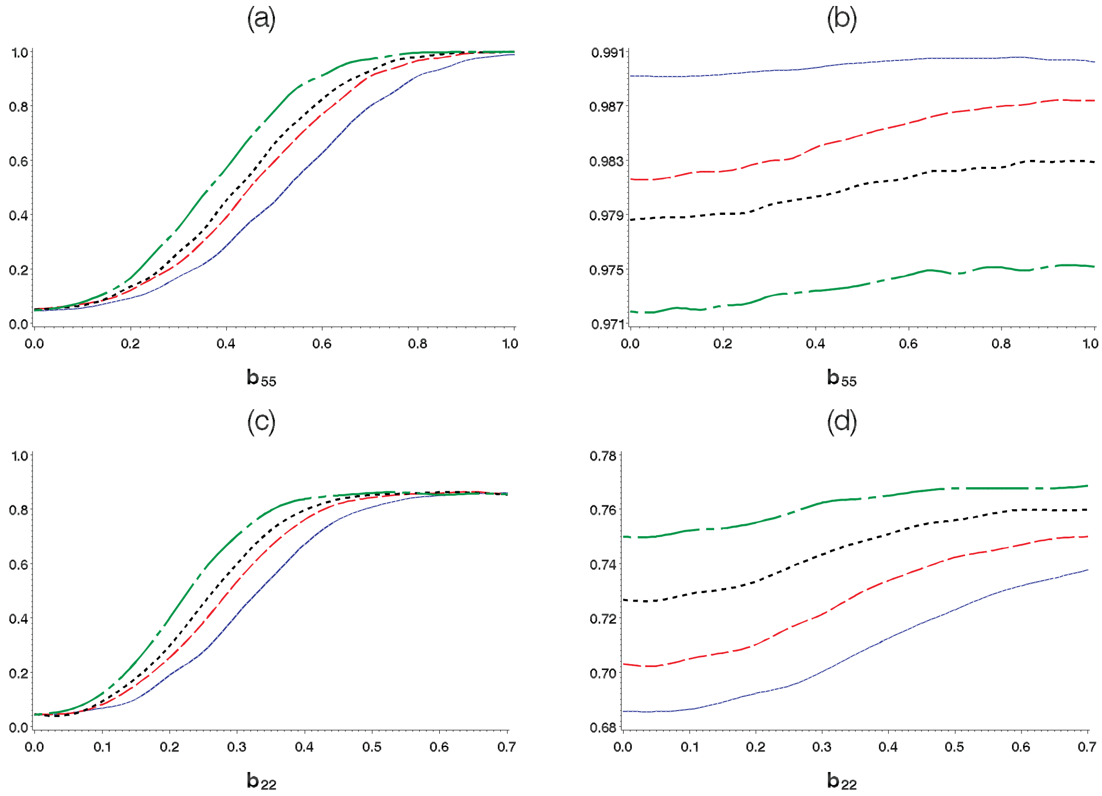

Figure 3.2: *Simulated probabilities of "correct" specification (a), (c) and efficiencies (b), (d) in the sub-models g_5 and g_2 of the quadratic regression model with two factors for various sequential designs of Biswas and Chaudhuri (2002) with $N = 200$ observations. Solid lines: design (A); dashed lines: design (B); dotted lines: design (C); nearly solid lines: design (D). The designs are defined in (3.2).*

The corresponding rejection probabilities and efficiencies are depicted in Figure 3.2(a), (b), (c) and (d) for the model g_5 and g_2, respectively. We observe the same phenomena as in the cubic case. Finer partititions in the sequential procedure of Biswas and Chaudhuri (2002) yield substantially smaller rates of "correct" specifiation but slightly larger efficiencies for the estimation of the parameters in the "true" model g_5. In the model g_2 the design (D) from (3.2) is the best with respect to the criterion of a "correct" specification and the efficiency criterion and all designs are ordered in the same way with respect to both criteria. In general the differences between the efficiencies are relatively small [between 0.97 and 0.99 in the model g_5 and between 0.68 and 0.75 in the model g_2] while more substantial differences can be observed for the probabilities of "correct" specification [see Figures 3.2 (a) and (c)]. We finally note that these differences are larger as in the case of the univariate cubic regression model considered in Example 3.1.

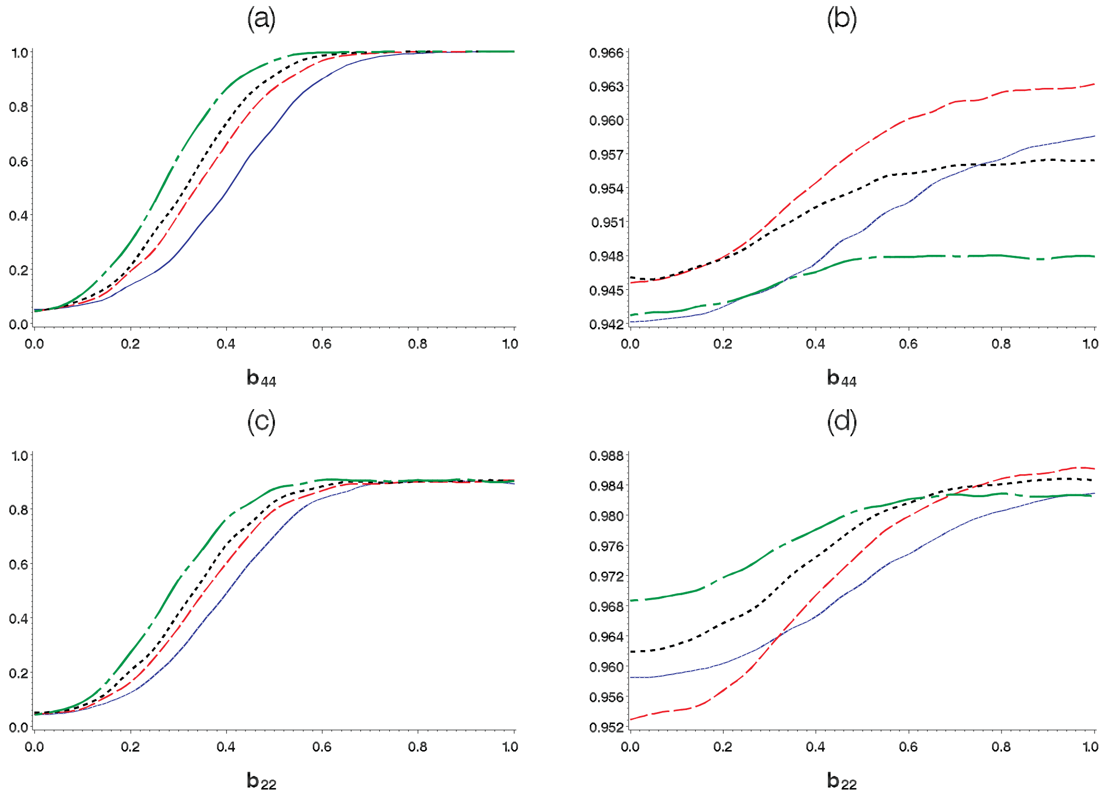

Figure 3.3: *Simulated probabilities of "correct" specification (a), (c) and efficiencies (b), (d) in the sub-models g_4 and g_2 of the Fourier regression model for various sequential designs of Biswas and Chaudhuri (2002) with $N = 200$ observations. Solid lines: design (A); dashed lines: design (B); dotted lines: design (C); nearly solid lines: design (D). The designs are defined in (3.2).*

Example 3.3 (Fourier regression). In the third example of our study we investigate the performance of the sequential designs of Biswas and Chaudhuri (2002) for the quadratic trigonometric regression model

$$g_4(x) = 1 + \sin x + \cos x + \sin(2x) + b_{44}\cos(2x);\;\; x \in [-\pi, \pi]\,. \tag{3.8}$$

Here the nested models are given by

$$g_3(x) \;=\; 1 + \sin x + \cos x + b_{33}\sin(2x) \tag{3.9}$$

$$g_2(x) = 1 + \sin x + b_{22} \cos x \tag{3.10}$$

$$g_1(x) = 1 + b_{11} \sin x. \tag{3.11}$$

The results of the corresponding rejection probabilities and efficiencies in the identified model are depicted in Figure 3.3(a), (b) and (c), (d) for the model g_4 and g_2 based on a sample with $N = 200$ observations, respectively. With respect to the criterion of "correct" specification we observe exactly the same picture as in the previous examples. Finer partitions of the total sample yield smaller probabilities of "correct" specification and the differences between the four designs in (3.2) are substantial. On the other hand the picture is not so clear with respect to the efficiency criterion. The differences between the four sequential designs are very small [at most 5% in the model g_4 and at most 3% in the model g_2; see Figure 3.3(b) and (d), respectively], but there exists no uniformly best design. In model g_4 the design (B) seems to be the best while in model g_2 the design (D) is most efficient for a broad range of the parameter b_{22}. Finally, we note that the design (A) with the finest partition does not yield the best efficiencies in both cases.

We have demonstrated in Example 3.1 - 3.3 that in general the probabilities of "correct" specification are substantially more sensitive with respect to the number of steps and the sample sizes in the different stages of the procedure of Biswas and Chaudhuri (2002) than the efficiencies. In all examples considered in our study designs with a finer partition yield lower rates of "correct" specification and the design (D) has a substantially better performance than the other designs in (3.2). In many cases this design is also the best with respect to the efficiency criterion in the identified model. In the other cases the differences with respect to the D-efficiency are very small. Further simulations, which are not given here for the sake of brevity show exactly the same picture. The sequential design proposed by Biswas and Chaudhuri (2002) should be constructed with only a few stages such that the sample sizes m_0, m_1 for the initial steps are not too small, because the loss of efficiency in the identified model is negligable, but the probability of "correct" specification is improved substantially. Based on our numerical experience we recomend to use only two stages if experiments are performed according to the sequential design of Biswas and Chandhuri (2002).

3.2 Sample sizes in the procedure of Montepiedra and Yeh

We now perform a similar study for the two-stage designs proposed by Montepiedra and Yeh (1997, 2002). For the sake of comparison the same examples as in Section 3.1 are considered. Again we give a rather detailed description for the (univariate) cubic regression model and summarize the essential features for the two other examples.

Example 3.4 (Cubic regression model). As design for the first stage Montepiedra and Yeh (1997, 2002) proposed the nonsequential discrimination design of Dette (1994), that is

$$\xi_{\text{Disc}} = \begin{pmatrix} -1 & -0.40825 & 0.40825 & 1 \\ 0.2 & 0.3 & 0.3 & 0.2 \end{pmatrix}. \tag{3.12}$$

We considered three possible allocations of the $N = 100 = n_0 + n_1$ observations, namely

$$
(3.13) \qquad \begin{array}{ll}
\frac{n_0}{N} = \frac{n_1}{N} = \frac{1}{2} & (E) \\
\frac{n_0}{N} = \frac{3}{4}, \ \frac{n_1}{N} = \frac{1}{4} & (F) \\
\frac{n_0}{N} = \frac{1}{4}, \ \frac{n_1}{N} = \frac{3}{4} & (G)
\end{array}
$$

where N denotes the total sample size and the design for the second stage is obtained by maximizing the function defined by (2.2) in the corresponding approximate setup with weight $f = n_0/(n_0 + n_1)$ [see Montediedra and Yeh (1997, 2002) for more details]. For example in the

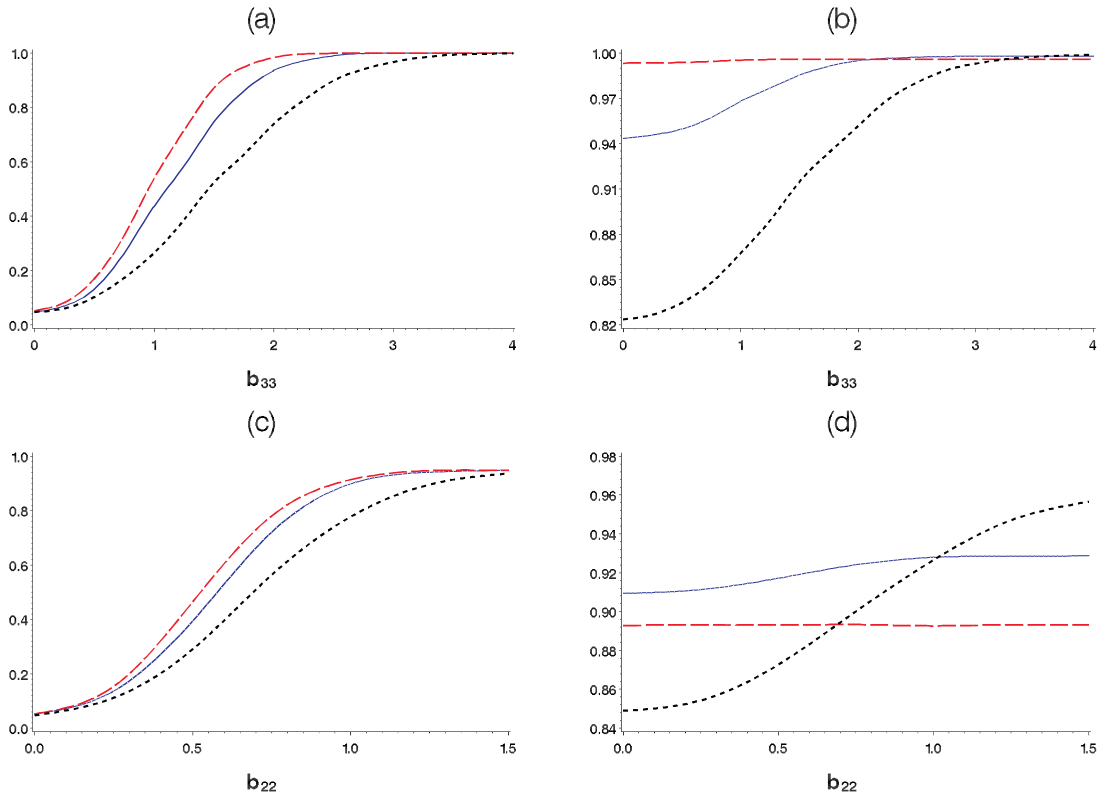

Figure 3.4: *Simulated probabilities of "correct" specification (a), (c) and efficiencies (b), (d) in the identified cubic and quadratic regression model for various two-stage designs of Montepiedra and Yeh (1997, 2002) with $N = 100$ observations. Solid lines: design (E); dashed lines: design (F); dotted lines: design (G). The designs are defined in (3.13).*

situation (E) we obtain the designs

$$
\xi_{E,3} = \begin{pmatrix} -1 & -0.4645 & 0.4645 & 1 \\ 0.298 & 0.202 & 0.202 & 0.298 \end{pmatrix}, \ \xi_{E,2} = \begin{pmatrix} -1 & 0 & 1 \\ 0.395 & 0.210 & 0.395 \end{pmatrix}, \ \xi_{E,1} = \begin{pmatrix} -1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}
$$

for the second stage if the cubic, quadratic or linear model have been chosen in the first stage, respectively. Similarly, in the case (F) and (G) these designs are given by

$$\xi_{F,3} = \begin{pmatrix} -1 & -0.4784 & 0.4784 & 1 \\ 0.3925 & 0.1075 & 0.1075 & 0.3925 \end{pmatrix} , \quad \xi_{F,2} = \xi_{F,1} = \begin{pmatrix} -1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

and

$$\xi_{G,3} = \begin{pmatrix} -1 & -0.4544 & 0.4544 & 1 \\ 0.266 & 0.234 & 0.234 & 0.266 \end{pmatrix}, \ \xi_{G,2} = \begin{pmatrix} -1 & 0 & 1 \\ 0.354 & 0.292 & 0.354 \end{pmatrix}, \ \xi_{F,1} = \begin{pmatrix} -1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

respectively. Figure 3.4 (a),(c) show the rates of "correct" specification of the underlying model and the corresponding efficiencies in the identified model for the three two-stage designs, if the cubic or quadratic model are "correct". The corresponding efficiencies in the "correct" model are depicted in Figure (b) and (d). Results for the linear case are not presented because the differences between the two-stage designs are negligible. We observe that the probabilities of "correct" specification for the quadratic model are increasing with the sample size for the first step [see Figure 3.4(a) and (c)]. In the cubic model the design (F) which allocates a larger sample size to the first case is the best while the design (G) is the worst. A comparison of the efficiencies also yields substantial differences between the two-stage designs in the cubic model [see Figure 3.4(c)]. The efficiencies differ by more than 10% if $\beta_{33} \leq 1$. In the quadratic model we observe only differences of approximately 5% [see Figure 3.4(d)]. Note that the design (F) yields the best efficiencies in the cubic model.

Example 3.5 (Quadratic regression with two factors). In this example we investigate the performance of the different two-stage designs in the quadratic model with two factors defined by (3.3). The probabilities of "correct" specification and the corresponding efficiencies of the designs in the "true" model are depicted in Figure 3.5 for the model g_5 and g_2. In the other models we observe a similar picture. It is interesting to note that in the model g_5 the designs have a similar performance with respect to the criterion of "correct" specification, where the design (E) with equal sample sizes for both stages is the best [see Figure 3.5(a)]. However, the differences with respect to the efficiency criterion are more remarkable and the design (F) is the best if g_5 is the "correct" model [see Figure 3.5(b)]. On the other hand in model g_2 the designs (E) and (F) behave very similar (with slight advantages for the design (F)) and the two-stage design (G) cannot be recommended [see Figure 3.5(c) and (d)].

Example 3.6 (Fourier regression). The final example of this section considers again the quadratic Fourier regression model (3.8) discussed in Example 3.3. For the design in the first stage of the procedure of Montepiedra and Yeh (1997, 2002) an optimal discrimination design proposed by Dette and Haller (1998) was used. This design maximizes a product of D_1-efficiencies corresponding to the different models and is given by

(3.14)
$$\xi_{\text{Disc}} = \begin{pmatrix} -\pi & -\frac{3}{5}\pi & -\frac{1}{5}\pi & \frac{1}{5}\pi & \frac{3}{5}\pi & \pi \\ \frac{1}{10} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{10} \end{pmatrix} .$$

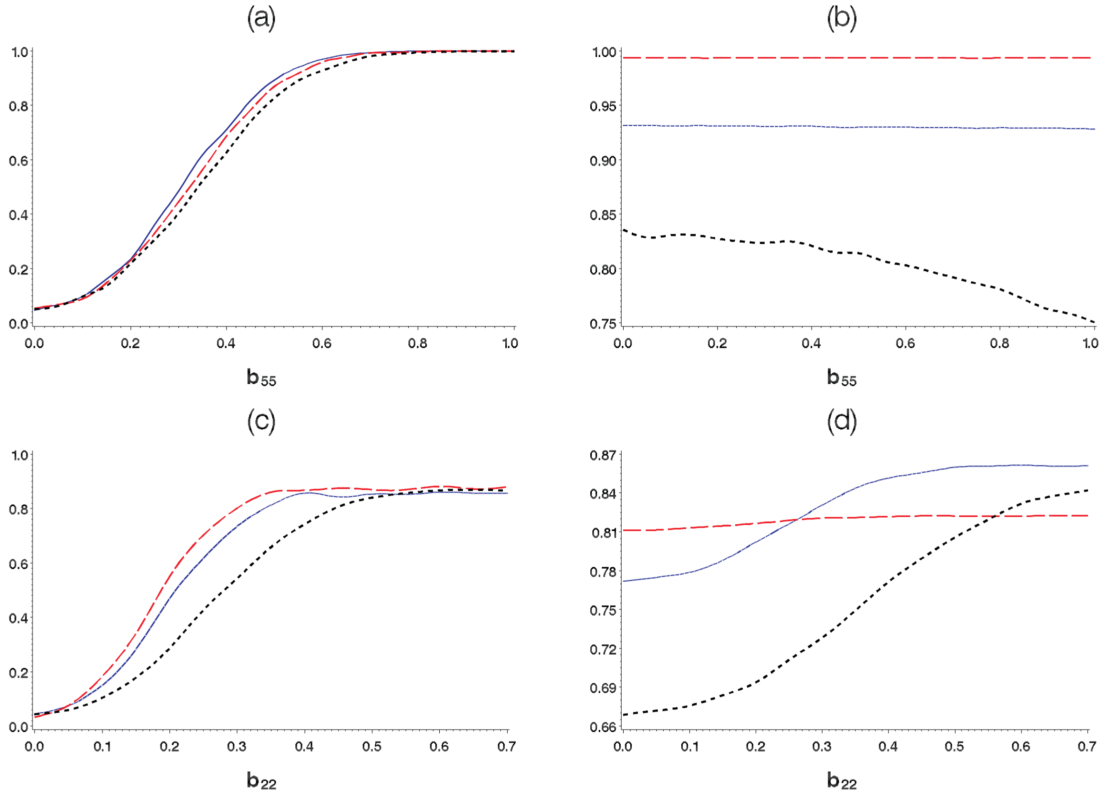

Figure 3.5: *Simulated probabilities of "correct" specification (a), (c) and efficiencies (b), (d) in the identified sub-model g_5 and g_2 of the quadratic regression model with two factors for various two-stage designs of Montepiedra and Yeh (1997, 2002) with* $N = 200$ *observations. Solid lines: design (E); dashed lines: design (F); dotted lines: design (G). The designs are defined in (3.13).*

The probabilities of "correct" specification and the corresponding efficiencies in the "correct" model are depicted in Figure 3.6 (a), (b) and (c), (d) for the model g_4 and the model g_2, respectively. In the other models we observed a similar picture. From Figure 3.6 we observe a substantial superiority of the design (F) (taking 75% of the observations in the first stage) with respect to the efficiency criterion and the criterion of "correct" specification in both models g_4 and g_2. Again the design (G) has the worst performance.

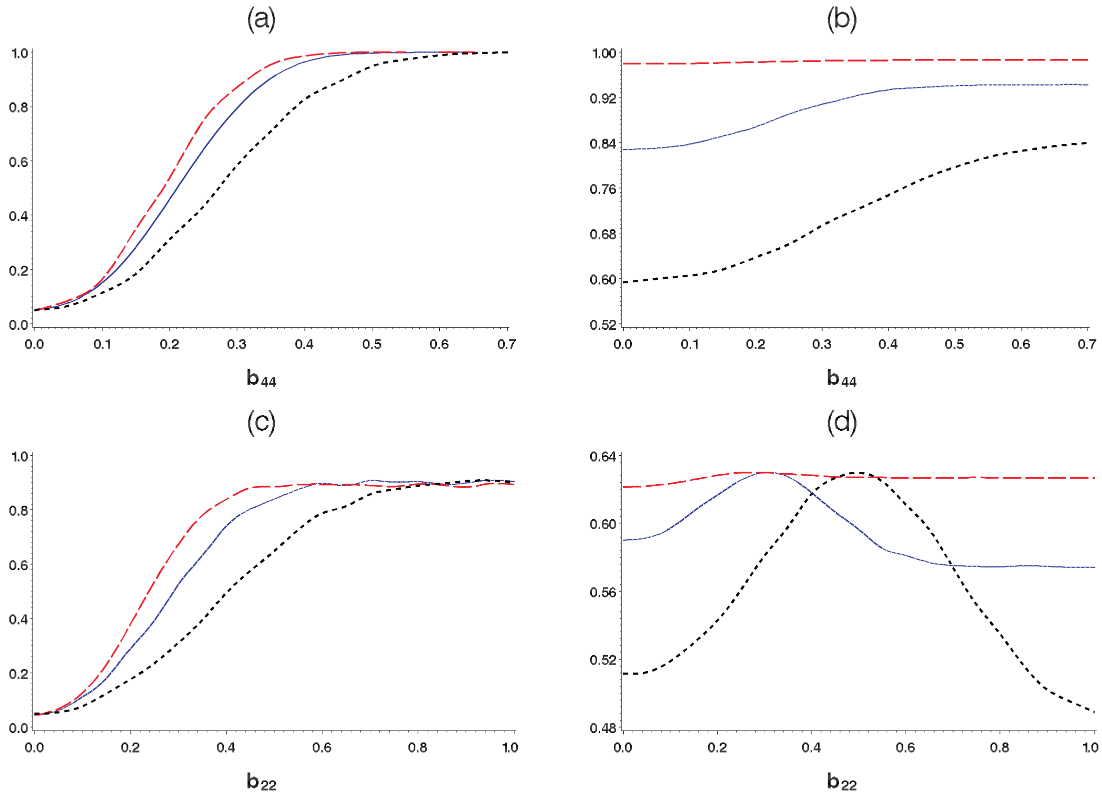

Figure 3.6: *Simulated probabilities of "correct" specification (a), (c) and efficiencies (b), (d) in the identified sub-model g_4 and g_2 of the Fourier regression model for various two-stage designs of Montepiedra and Yeh (1997, 2002) with $N = 200$ observations. Solid lines: design (E); dashed lines: design (F); dotted lines: design (G). The designs are defined in (3.13).*

The results of Example 3.4 - 3.6 can be summarized to obtain a recommendation for the choice of the sample sizes in the two-stage design proposed by Montepiedra and Yeh (1997, 2002). In most cases the design (F) is the best with respect to the criterion of "correct" specification of the underlying model. There is one situation, where the design (E) is better, but in this case the differences between all two-stage designs are negligible. In contrast to the sequential designs of Biswas and Chaudhuri (2002) the differences between the two-stage designs with respect to the efficiency criterion are remarkable. In all cases we observe a substantial improvement of efficiency by the two-stage design (F). Additional simulations for other models, which are not shown here for the sake of brevity, showed a similar picture. Thus taking 75% of the total observation in the first stage and only 25% of the observations in the second stage seems to be a good strategy for the two-stage design of Montepiedra and Yeh (1997, 2002). Finally, we remark that the corresponding F-tests do not keep the preassigned level, if the designs are constructed by the two-stage procedure of Montepiedra and Yeh (1997, 2002). For example, in the univariate cubic and quadratic model the simulated level varies between 7% and 9%, while the preassigned level is 5% [see Example 3.4].

4 A comparison of sequential designs

In this section we compare the two sequential designs proposed in Section 2. For the sake of brevity and transparency we investigate from each sequential approach only one design, which is in our opinion the best. Additionally we include in our comparison two non-sequential designs in each example. These designs were introduced by various authors to obtain good designs for discrimination or robust designs with respect to the assumption on the regression model [see Spruill (1990) or Läuter (1974) among many others].

For the sequential design proposed by Biswas and Chaudhuri (2002) we used the design (D) from (3.2) with only two stages, because the impact of these designs on the probability of "correct" specification in finite samples is more severe than on the efficiencies with respect to estimation in the identified model. For the two-stage design proposed by Montepiedra and Yeh (1997, 2002) we used the design (F) in (3.13) which uses 75 % of the observations in the first step. Moreover, in order to obtain a fair comparison the nominal level of the F-test was adjusted such that the effective level of the resulting tests is precisely 5% (see the discussion at the end of Section 3). The results are presented in the same way as in Section 3. We show the simulated rejection probabilities and D-efficiencies under the assumption that a specific model is "correct". Again we consider the univariate polynomial model, a quadratic model with two factors and the Fourier-regression model of degree 2.

Example 4.1 (Cubic regression). The sequential designs of Biswas and Chaudhuri (2002) and Montepiedra and Yeh (1997, 2002) have been described in Example 3.1 and 3.4 respectively. As model robust design we chosed the design

$$\xi_{\text{Robust}} = \begin{pmatrix} -1 & -0.4011 & 0.4011 & 1 \\ 0.3194 & 0.1806 & 0.1806 & 0.3194 \end{pmatrix}, \tag{3.15}$$

which was proposed by Dette (1990). The discrimination design is the design ξ_{Disc} defined in (3.12).

If the linear model is "correct" the differences between the four designs are negligible and therefore not depicted. Figure 4.1 (a) and (c) show the probabilities of "correct" specification in the cubic and quadratic model, respectively. We observe that the discrimination, model robust and the two-stage design of Montepiedra and Yeh (1997, 2002) have a substantially better performance than the sequential procedure of Biswas and Chaudhuri (2002) with respect to this criterion. This superiority is more visible in the cubic model. Note that we chosed the best design (D) in (3.2) with respect to the criterion of "correct" specification [see Figure 3.1] and all other designs in (3.2) have a worse performance. Comparing the remaining three designs we observe that the discrimination design ξ_{Disc} ist the best, while the two-stage design of Montepiedra and Yeh (1997, 2002) is better than the model robust design ξ_{Robust}. The picture in the quadratic model is similar. There is only a minor difference between the design ξ_{Disc} and the two-stage design of Montepiedra and Yeh (1997, 2002). The model robust design ξ_{Robust}

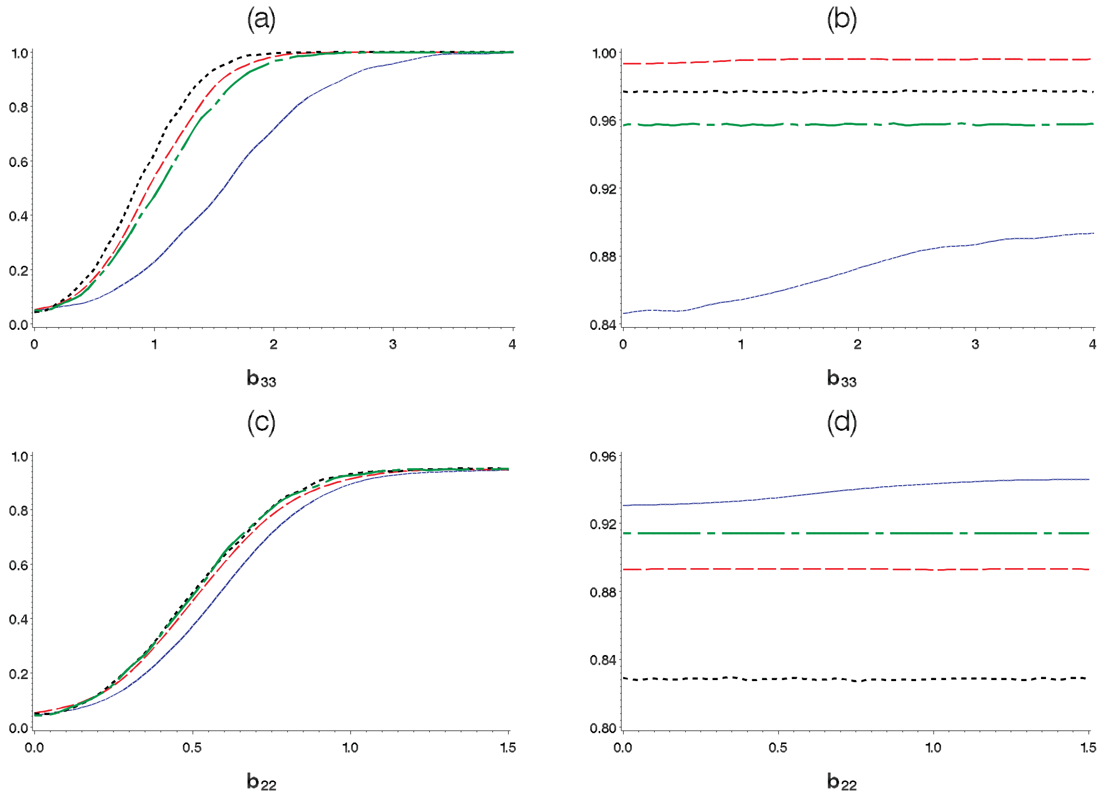

Figure 4.1: *Simulated probabilities of "correct" specification (a), (c) and efficiencies (b), (d) in the cubic and quadratic regression model for various sequential and non-sequential designs with* $N = 100$ *observations. Solid lines: the sequential design (D) of Biswas and Chaudhuri (2002); dashed lines: the two-stage design (F) of Montepiedra and Yeh (1997, 2002); dotted lines: non-sequential discrimination design; nearly solid lines: non-sequential model-robust design.*

has a slightly worse performance but it has a substantial better rate of "correct" specification then the sequential design of Biswas and Chaudhuri (2002) [see Figure 4.1 (c)]. We mention again that the nominal level for the two-stage design of Montepiedra and Yeh (1997, 2002) was chosen substantially smaller than 5% such that the effective (simulated) level is precisely 5%. A comparison of the efficiencies of the four designs can be found in Figure 4.1(b) and (d) for the cubic and quadratic model, respectively (in the linear model the differences are negligible). In the cubic polynomial the design of Biswas and Chaudhuri (2002) has about 10 % less efficiency than the discrimination, model robust and two-stage design, while the efficiencies of the lastnamed designs differ only between 2% and 4%. In the quadratic model [see Figure 4.1(d)] the sequential design of Biswas and Chaudhuri (2002) yields the best efficiencies, while the model robust and the two-stage design of Montepiedra and Yeh (1997, 2002)) are 2% and 4% less efficient. The discrimination design ξ_{Disc} has the worst performance with a loss of efficiency between 10% – 15% compared to the best case.

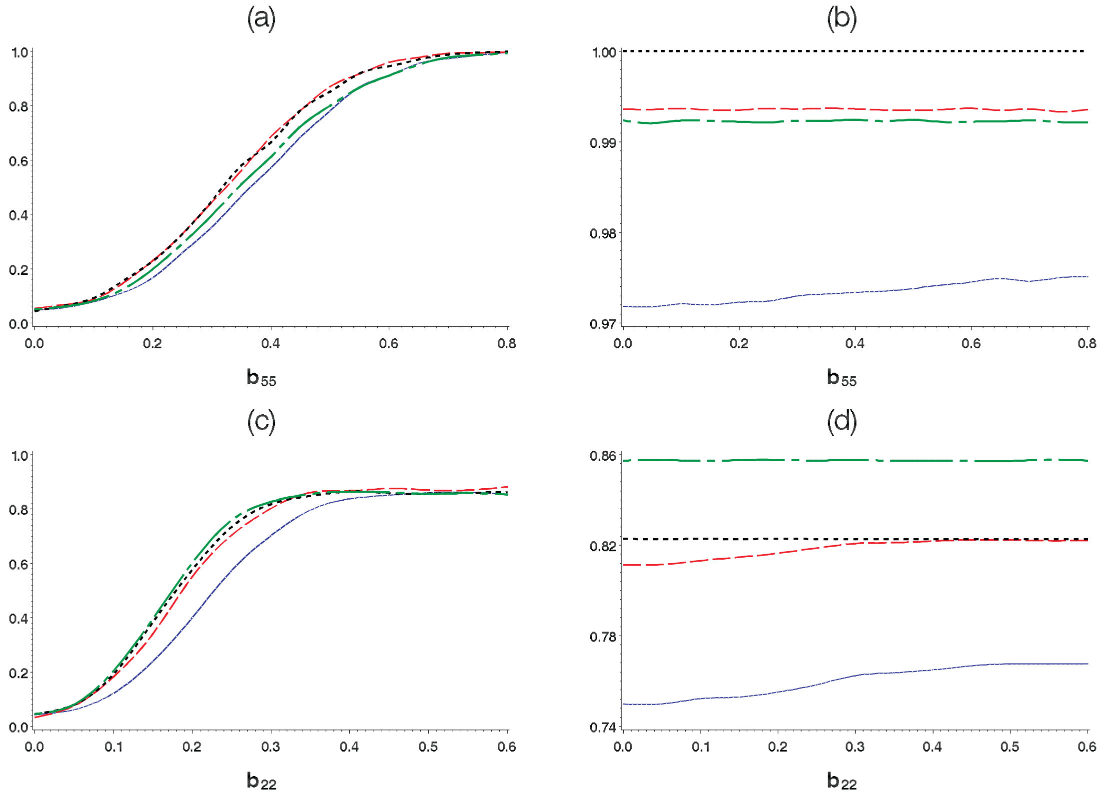

Figure 4.2: *Simulated probabilities of "correct" specification (a), (c) and efficiencies (b), (d) in the the sub-models g_5 and g_2 of the quadratic regression model with two factors for various sequential and non-sequential designs with $N = 200$ observations. Solid lines: the sequential design (D) of Biswas and Chaudhuri (2002); dashed lines: the two-stage design (F) of Montepiedra and Yeh (2002); dotted lines: non-sequential discrimination design; nearly solid lines: non-sequential model-robust design.*

Example 4.2 (Quadratic regression with two factors). Model-robust designs and optimal discrimination designs for the class $\{g_1, \ldots, g_5\}$ defined by the quadratic model with two factors in (3.13) have been determined by Dette and Röder (1997). The discrimination design ξ_{Disc} with respect to a uniform prior coincides with the D-optimal design for the full model (Kono, 1962), while the model robust design ξ_{Robust} has mass 0.16 at the four points $(\mp 1, \mp 1)$, mass 0.08 at $(\mp 1, 0)$ and $(0, \mp 1)$, and mass 0.04 at the origin. The total sample size is $N = 200$ and for the sequential method of Biswas and Chaudhuri (2002) we considered the case of $m_0 = m_1 = 100$.

Figure 4.2 (a) and (c) show the corresponding probabilities of "correct" specification if the model g_5 or g_2 is the underlying model. Figures corresponding to the cases g_4 and g_3 are very similar and are therefore not presented, while we did not find any substantial differences between the four designs in the model g_1. Again we observe substantially higher probabilities of "correct" identification for the non-sequential designs and the two-stage design of Montepiedra and Yeh (1997, 2002). In the model g_5 the non-sequential discrimination design and the two-stage procedure yield the highest rates of "correct" specification, while the model robust and the design of Biswas and Chaudhuri (2002) have a slightly worse performance [see Figure 4.2 (a)]. In the model g_2 all designs except the sequential design of Biswas and Chaudhuri (2002) have a similar performance [see Figure 4.2 (c)]. A comparison of the efficiencies for the four designs is presented in Figure 4.2 (b) and (d) corresponding to the cases where the model g_5 or g_2 is the "correct" regression model, respectively. Here we observe much smaller differences between the different designs. For example, in the model g_5 the design ξ_{Disc} is the best, because it is in fact also D-optimal for the quadratic regression, but the other designs are at most 2% less efficient. Similarly, if g_2 is the "true" model the model robust design is the best. Thus for models with two factors the differences between the efficiencies are either negligible or non-sequential designs or two-stage designs are more efficient than the design of Biswas and Chaudhuri (2002).

Example 4.3 (Fourier regression). A model robust design for the trigonometric regression model has been proposed by Lau and Studden (1985)

$$(3.16) \qquad \xi_{\text{Robust}} = \begin{pmatrix} -2.4294 & -1.2530 & 0 & 1.2530 & 2.4294 \\ 0.1965 & 0.2252 & 0.1566 & 0.2252 & 0.1965 \end{pmatrix}.$$

Figure 4.3 (a) and (c) show the corresponding probabilities of "correct" specification if the model g_4, or g_2 is the underlying model. In the case where the model g_1 is "correct" all designs produced similar results, while the situation for model g_3 is similar as presented here for the cases g_2 and g_4. We observe substantially larger probabilities of "correct" identification for the non-sequential designs and the design of Montepiedra and Yeh (1997, 2002). For example, if the model g_4 is "correct", the sequential design of Biswas and Chaudhuri (2002) is the worst with respect to the criterion of "correct" specification and the two stage design of Montepiedra and Yeh (1997, 2002) is the best. On the other hand, in the model g_2 the non-sequential discrimination design yields the highest rates of "correct" identification, while the sequential design of Biswas and Chaudhuri (2002) is the worst. In all considered cases the non-sequential designs and the two-stage designs of Montepiedra and Yeh (1997, 2002) performed substantially more reliable than the sequential procedure of Biswas and Chaudhuri (2002) with respect to the criterion of "correct" model identification. A comparison of the efficiencies for the four designs is presented in Figure 4.3 (b) and (d) corresponding to the cases where the model g_4 and g_2 is the "correct" regression model. Here we observe much smaller differences between the sequential and non-sequential designs. For example in the model g_4 the discrimination design ξ_{Disc} defined in (3.14) is the best [because it is in fact also D-optimal for the quadratic trigonometric regression], but the sequential designs are at most 6% less efficient [see Figure 4.3(b)]. Similarly, in the model g_2 the discrimination design ξ_{Disc} is again D-optimal, but the model robust and the design of Biswas and Chaudhuri (2002) are at most 4% less efficient [see Figure 4.3(d)]. However, in this case the sequential procedure of Montepiedra and Yeh (2002) yields a loss of efficiency of more than 30% (although it has the highest rates of "correct" specification.

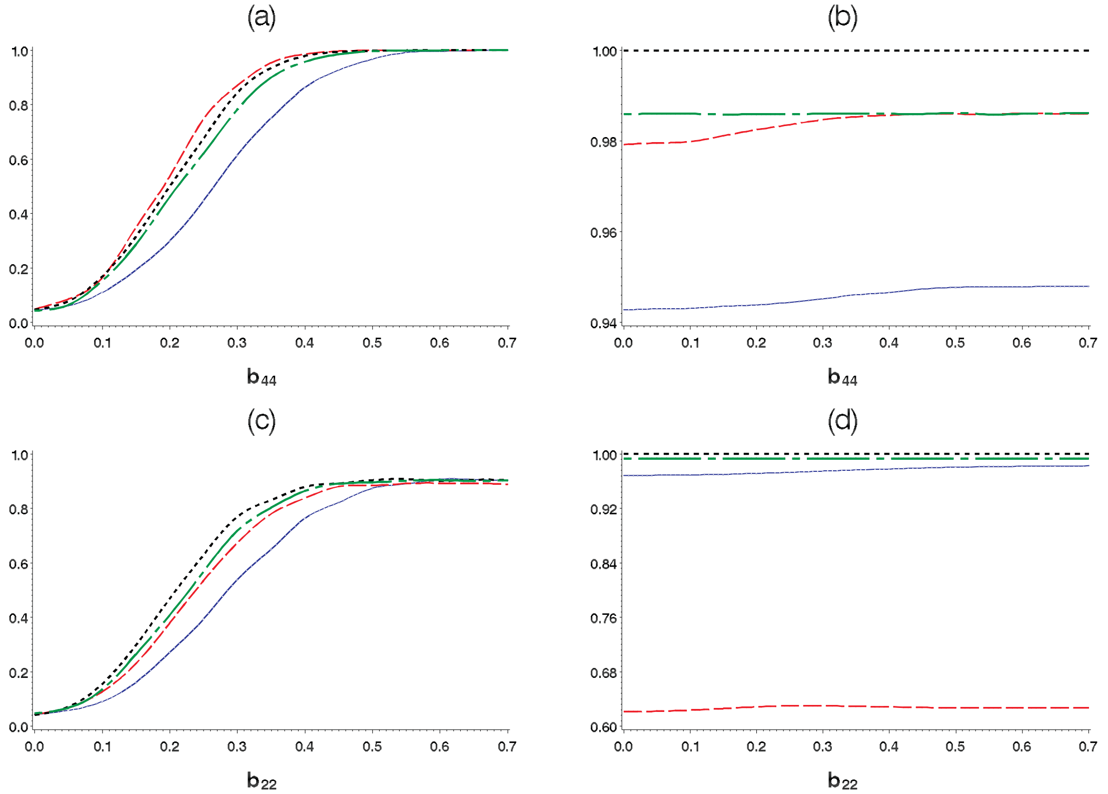

Figure 4.3: *Simulated probabilities of "correct" specification (a), (c) and efficiencies (b), (d) in the the sub-models g_4 and g_2 of the Fourier regression model for various sequential and non-sequential designs with $N = 200$ observations. Solid lines: the sequential design (D) of Biswas and Chaudhuri (2002); dashed lines: the two-stage design (F) of Montepiedra and Yeh (2002); dotted lines: non-sequential discrimination design; nearly solid lines: non-sequential model-robust design.*

5 Conclusions

In this paper we presented a numerical comparison of different design strategies for two different goals in regression models

- identification of an appropriate model
- efficient estimation of the parameters in the identified model

Two sequential and non-sequential design strategies are investigated in a univariate polynomial, a quadratic model with two factors and a trigonometric regression model. For the sequential designs of Biswas and Chaudhuri (2002) and the two-stage designs of Montepiedra and Yeh (1997, 2002) it is shown that a relative large proportion of the total sample should be taken at the initial stage in order to obtain good rates of "correct" specification of the underlying model. The influence of the size (and number) of the stages in a sequential procedure on the efficiency in the identified model is much smaller compared to the effect on the probability of "correct" specification. The two-stage designs proposed by Montepiedra and Yeh (2002) provide good probabilities of "correct" identification of the underlying model and reasonable efficiencies for the estimation of the parameters in the identified model. However, due to the dependencies of the observations obtained from sequential sampling, the corresponding F-tests for the identification of the model do not keep the preassigned level and the statistical properties of the estimator in the underlying model are not clear in general. Additionally, the determination of the design for the second step seems to be difficult, because this problem is equivalent to an optimal design problem in a Bayesian linear regression model.
The sequential strategy proposed by Biswas and Chaudhuri (2002) yields to a sequence of non-standard F-tests which keep the preassigned level. Moreover, for large sample sizes the "correct" model is identified with high probability, the sequential design is close to the D-optimal design for the "correct" model and the estimates obtained from this design are consistent and asymptotically normal. Although these properties are interesting from a theoretical (asymptotic) point of view, our numerical results show that these designs can only be recommended for very large sample sizes, because of their low rate of "correct" identification of the underlying model. In the situations considered in our study the two-stage designs of Montepiedra and Yeh (2002) and the non-sequential discrimination and model robust designs of Dette (1990, 1994) yield substantially higher rates of "correct" identification of the underlying model and (at least) comparable efficiencies for the estimation of the parameters in the underlying "correct" model. In many cases the efficiencies of these designs are even higher. The poor performance of the sequential designs proposed by Biswas and Chaudhuri (2002) can be explained by the fact that a modified version of the classical F-test is used in the identification steps. This modification is required to keep the preassigned level of the corresponding tests from the dependent data, and usually causes a substantial loss of power. The two-stage designs introduced by Montepiedra and Yeh (1997, 2002) do not yield exact tests, because inference is based on the standard F-test with dependent data. The non-sequential designs yield a sequence of F-tests for the identification step, which keep the preassigned levels and the calculation is usually simpler because it can be performed by standard algorithms [see e.g. Läuter (1974)].

Acknowledgements. The work of H. Dette was supported by the Deutsche Forschungsgemeinschaft (SFB 475, Komplexitätsreduktion in multivariaten Datenstrukturen). The authors would like to thank Isolde Gottschlich who typed numerous versions of the paper with considerable technical expertise and A. Biswas and G. Montepiedra for sending us their work before publication.

References

T.W. Anderson (1962). The choice of the degree of a polynomial regression as a multiple decision problem. Ann. Math. Statist. 33, 255-265.

A. Biswas, P. Chaudhuri (2002). An efficient design for model discrimination and parameter estimation in linear models. Biometrika, 89, 709-718.

H. Dette (1990). A generalization of D- and D_1-optimal designs in polynomial regression. Ann. Statist. 18, 1784-1801.

H. Dette (1994). Discrimination designs for polynomial regression on a compact interval. Ann. Statist. 22, 890-904.

H. Dette (1995). Optimal designs for identifying the degree of a polynomial regression. Ann. Statist. 23, 1248-1267.

H. Dette, I. Röder (1997). Optimal discrimination designs for multi-factor models. Ann. Statist. 25, 1161-1175.

H. Dette, G. Haller (1998). Optimal discriminating designs for Fourier regression. Ann. Statist. 26, 1496-1521.

J. Kiefer (1974). General equivalence theory for optimum designs (approximate theorey). Ann. Statist. 2, 849-879.

Kono, M. (1962). Optimal designs for quadratic regression on the k-cube. Mem. Fac. Sci. Kyushu Univ. Ser. A16, 114-22.

E. Läuter (1974). Experimental design for a class of models. Math. Oper. Forsch. Statist. 5, 379-398.

T.S. Lau, W.J. Studden (1985). Optimal designs for trigonometric and polynomial regression using canonical moments. Ann. Statist. 13, 383-394.

Lim, Y.B., Studden, W.J. (1988). Efficient D_s-optimal designs for multivariate polynomial regression on the q-cube. Ann. Statist. 16, 1225-40.

G. Montepiedra, A.B. Yeh (1997). A two stage strategy for the construction of D-optimal experimental designs. Commun. in Statistics, Simul. and Comput., 27, 277-402.

G. Montepiedra, A.B. Yeh (2002). Two stage designs for estimation and identification of polynomial models. Technical Report, Department of Applied Statistics and Operations Research. Bowling Green State University.

F. Pukelsheim (1993). Optimal Design of Experiments. Wiley, N.Y.

G.A.F. Seber (1977). Linear Regression Analysis. Wiley, N.Y.

S.D. Silvey (1980). Optimum Design. Chapman and Hall, London.

M.G. Spruill (1990). Good designs for testing the degree of a polynomial mean. Sankhyā, Ser. B, 52, 67-74.