Werner Müller and Antonio C. Ponce de Leon

Discrimination Between Two Binary Data Models. Sequentially Designed Experiments.

Working Paper

Discrimination Between Two Binary Data Models: Sequentially Designed Experiments

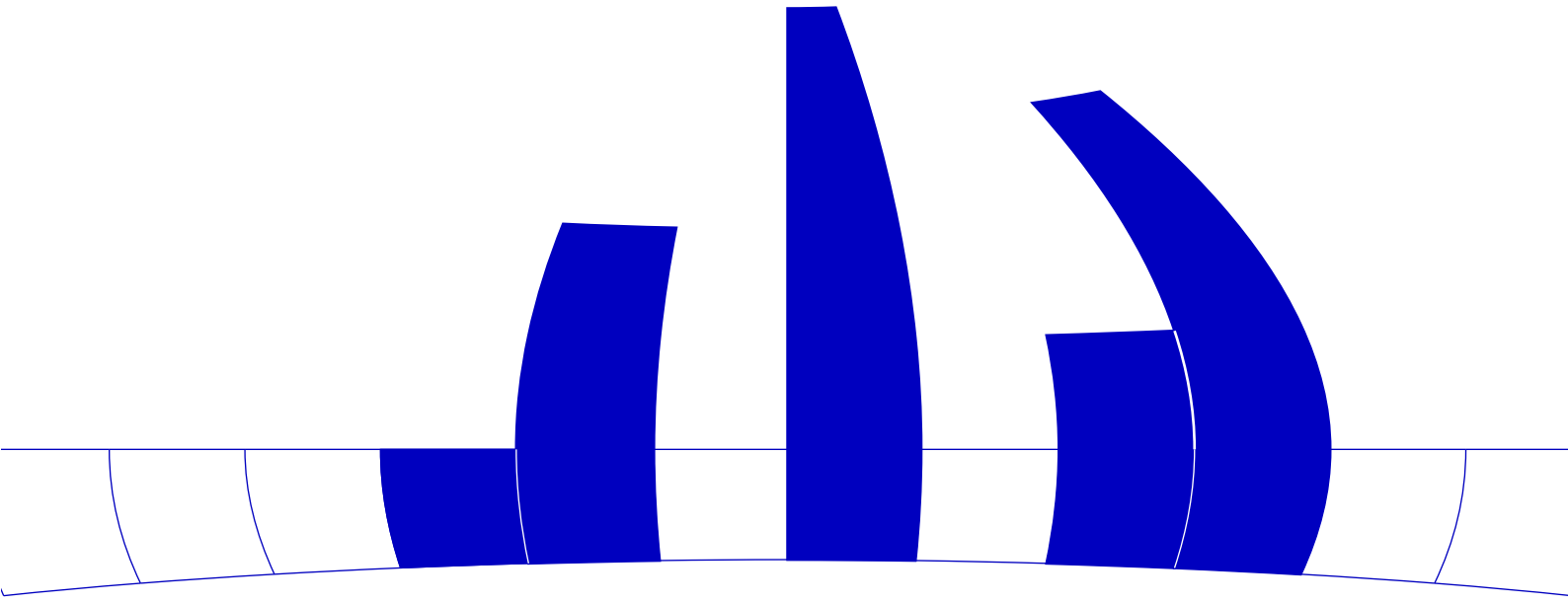
Werner Müller, Antonio C. Ponce de Leon

Institut für Statistik
Wirtschaftsuniversität Wien

Forschungsberichte

Bericht 37
July 1993

http://statmath.wu-wien.ac.at/



Discrimination Between Two Binary Data Models: Sequentially Designed Experiments

Werner G. Müller and Antonio C. Ponce de Leon *

Department of Statistics, University of Economics and Business Administration, Vienna
and
Department of Public Health Sciences, St. George's Hospital Medical School, London

Abstract

In this paper we describe and illustrate a sequential approach for the problem of designing experiments to discriminate between binary data models as efficiently as possible. The models considered ought to belong to the class of binary response models which is conveniently regarded as a subclass of Generalized Linear Models. Further they may be either nested or non-nested, and their linear predictor structures must be given.

We examine this problem for the case of two rival models. Optimal design theory is utilized so as to derive important properties of and relevant results with respect to optimal experimental designs. Two sequential methods of designing efficient experiments for discrimination between two models are suggested. Furthermore, simulation techniques are used not only to illustrate how the methods can be implemented but also to enable comparisons between sequentially designed experiments and local optimal ones, that are designs which are optimal with respect to specific parameter values and model assumptions.

1 Introduction

One of the main reasons for the increasing interest in optimum experimental designs under the framework of linear or nonlinear regression is that seldom only one specific model can be satisfactorily fitted to data resulting from unplanned experiments. In some cases a number of models might fit the data reasonably well, therefore, requiring that procedures of model choice and/or regression diagnostics are applied with the purpose of making a decision on which is the "best"

*On leave of absence from Universidad do Estado do Rio de Janeiro.

fit. Several of these methods are available in the literature. However, it is possible that even the most efficient of them is unable to make the correct model choice. In fact, a conflicting situation may arise if we consider that different models may score better under different evaluation procedures.

For binary response models, these conflicting diagnoses are even more likely to occur since some of the classical models such as those specified by the link functions logit, probit and complementary log-log provide expected responses of the dependent variable which coincide very closely over a large range of the explanatory variables. However, a better designed experiment, i.e. one for which the levels of the explanatory variables contain greater amount of information on discriminating between the expected values under the rival models, is more likely to supply data leading to an adequate fit for the most appropriate model. The notion "a greater amount of information", referred to above, must be clarified and formally defined if comparisons between different designs are to be made.

2 Framework

For two rival binary data models suppose that π_j denote the probabilities of success (an observation $y_i = 1$) related to the structures of the rival models. Under the framework of Generalized Linear Models, such structures are described by two components, the link function $g(.)$ and the linear predictor $\xi\beta$, namely such that

$$\pi_j = g_j(\xi_{nj}\beta_j), \qquad j = 1, 2. \tag{2.1}$$

The former function must be monotonic and differentiable whereas for the latter the columns of the design matrix ξ_{nj} display the values of the covariates $\{X_{j1}, ..., X_{jm_j}\}$, at which the responses $\{y_1, ..., y_n\}$ will be taken and β_j is the unknown set of parameters. By this notation the dependence of the probabilities of success on the linear predictor, and therefore on the design and the unknown parameters, is emphasized. An excellent reference to the presented class of models is Agresti (1990).

3 Problem Formulation

Consider a typical design to be a sequence of points belonging to the design space. Then it is reasonable to say that some of such sequences may be more or less informative to discriminate the rival models than others. Therefore, we can state the problem of model discrimination as the search for the most informative design according to a specific criterion, which ideally would give a measure of the information contained in the points of the design space with respect to the expected value of both models.

For a special case of binary data models, namely for those given in (9.1) and (9.2) a treatment of this problem was attempted by Chambers & Cox (1967), extended by two similar papers by Yanagisawa (1988) and Yanagisawa (1990). Unfortunately for the discrimination of binary data models the optimal design depends upon which is the true model and its true parameter values. Therefore those approaches must be considered as local ones, in a sense that it was

assumed that the true model and its parameters are known, an assumption that is hardly to hold in practice. Such designs can be understood to lie within the framework of so-called T-optimal designs introduced in a remarkable paper by Atkinson & Fedorov (1975).

A possible solution to this dependence problem is offered by the Bayesian optimum designs proposed in Ponce deLeon & Atkinson (1991) and Ponce deLeon (1993). In practical situations however, especially when observations are costly and quick responses are available, a natural alternative may be to design experiments sequentially. The main advantages usually associated with such experiments are feasibility and asymptotic optimality. The former is restricted, however, by the specific characteristics of each problem, whilst for the latter a detailed study should be carried out to assess important features such as the rate of convergence and the oscillation of a sequence of designs. For a strong motivation to sequential designs refer to Wu (1985)b, who uses them for the estimation of the probabilities π in binary data models.

4 Optimization Criterion

The problem of discriminating between two binary response models can be viewed as a search problem where the aim is to optimize a certain choice based on a specific criterion. In the context of optimal design theory it is assumed that the choice of points in the design space is arbitrary. Whatever the situation in which this assumption can be made an optimization process can be carried out in order to attain the specific aims of the experiment. For this purpose, another essential component of the general problem is the criterion of optimality which provides the means for comparison among different designs. This criterion of optimality is the main indicator for the choice of design that will be made.

A natural candidate for such a criterion depends upon any suitable measure of goodness-of-fit under a specific binary data model. For this purpose two strong alternative criteria are available. We may use either the Pearson chi-squared statistic or the log likelihood ratio statistic to test between the fits of candidate models. It is well known (cf. Agresti (1990)) that both of them are asymptotically χ^2_{k-m0} under the H_0 hypothesis and $\chi^2_{k-m0,\lambda}$ under the alternative H_A with noncentrality parameters:

$$\lambda = n\left(\frac{(\pi_A - \hat{\pi}_0)^2}{\hat{\pi}_0} + \frac{(\hat{\pi}_0 - \pi_A)^2}{1-\hat{\pi}_0}\right) \tag{4.1}$$

and

$$\lambda = 2n\left(\pi_A \log\left(\frac{\pi_A}{\hat{\pi}_0}\right) + (1-\pi_A)\log\left(\frac{1-\pi_A}{1-\hat{\pi}_0}\right)\right) \tag{4.2}$$

respectively. Here subscripts 0 and A indicate the respective hypothesis, the hatted terms denote the limit values of the maximum likelihood estimates and k refers to the number of points at which the n observations are taken.

Yanagisawa (1988) pointed out that from a table that gives the probabilities $P[\chi^2_{\nu,\lambda} > \chi^2_\nu(\alpha)]$ (see for instance Haynam et al. (1970)) it is seen, that for any significance level less than 0.2 the power for the alternative hypothesis decreases with increasing degrees of freedom. This leads us to consider only designs with

(asymptotically) $k = m_0 + 1$ distinct observation points, since the maximum value of λ is constant irrespective of the number of points k. Having k fixed we now observe a strict relation between the power of the test and the value of the noncentrality parameter λ.

Therefore we may like Yanagisawa (1988) use (4.1) as the design criterion. But because of the asymptotic equivalence to (4.2) we chose the latter due to its nice properties in model selection (see Agresti (1990)).

5 Sequential Procedure

For the discrimination of two nonlinear regression models Atkinson & Fedorov (1975) presented a sequential design optimization procedure that maximized the difference of the response under each model in each step. This technique would asymptotically reach the T-optimum design, that is the one maximizing the lack of fit of the false model based on the residual sum of squares.

Hence, for the binary data framework we suggest combining this method with the analogy that can be drawn between residual sum of squares in regression theory and the log likelihood ratio in Generalized Linear Models. This leads to a general iterative procedure to find an optimum experimental design to discriminate between two binary data models. This modified procedure now consists of the following steps:

Step 1 $\rightarrow$ Choose an initial design $\xi_{n_0} = \{x_1, ..., x_{n_0}\}^T$, take observations $\{y_1, ..., y_{n_0}\}^T$ and find the maximum likelihood estimates $\hat{\beta}_{00}$ and $\hat{\beta}_{A0}$ for the sets of parameters β_0 and β_A respectively.
Step 2 $\rightarrow$ Let ξ_i be the design at the $(i-n_0)$th iteration of the method. Let $x\hat{\beta}_{ji}$ be the linear predictor estimator based on the parameter estimates obtained in the $(i-n_0)$th iteration. Take the next observation y_{i+1} at the design point

$$x_{i+1} = \operatorname{Arg}\max_x \left\{ g_A(x\hat{\beta}_{Ai}) \log\left(\frac{g_A(x\hat{\beta}_{Ai})}{g_0(x\hat{\beta}_{0i})}\right) + (1 - g_A(x\hat{\beta}_{Ai})) \log\left(\frac{1 - g_A(x\hat{\beta}_{Ai})}{1 - g_0(x\hat{\beta}_{0i})}\right) \right\} \tag{5.1}$$

where $g_0(.)$ and $g_A(.)$ denote the respective link functions. This is the value that provides the largest increase in the noncentrality parameter estimate (4.2).
Step 3 $\rightarrow$ Find new maximum likelihood estimates $\hat{\beta}_{j(i+1)}$ for the parameter sets β_j based on the set of observations $\{y_1, ..., y_{n_0}, ..., y_{i+1}\}^T$ taken at the points of the updated design ξ_{i+1}.
Step 4 $\rightarrow$ Proceed with Step 2 and 3 until there is no change in the structure of the design.

Note that the procedure still gives a local solution, in the sense that the hypotheses are treated unsymmetrically. To avoid that, we suggest to replace the maximizing function in (5.1) by the weighted average

$$\hat{p}\left\{ g_0(x\hat{\beta}_{0i}) \log\left(\frac{g_0(x\hat{\beta}_{0i})}{g_A(x\hat{\beta}_{Ai})}\right) + (1 - g_0(x\hat{\beta}_{0i})) \log\left(\frac{1 - g_0(x\hat{\beta}_{0i})}{1 - g_A(x\hat{\beta}_{Ai})}\right) \right\}$$

$$+(1-\hat{p})\left\{g_A(x\hat{\beta}_{Ai})\log\left(\frac{g_A(x\hat{\beta}_{Ai})}{g_0(x\hat{\beta}_{0i})}\right)+(1-g_A(x\hat{\beta}_{Ai}))\log\left(\frac{1-g_A(x\hat{\beta}_{Ai})}{1-g_0(x\hat{\beta}_{0i})}\right)\right\}, \tag{5.2}$$

where $\hat{p}$ has to be a suitable estimator of the probability of the H_0 model, with the obvious properties $\text{plim}_{H_0,n\to\infty}\hat{p}=1$ and $\text{plim}_{H_A,n\to\infty}\hat{p}=0$.

6 Convergence Considerations

It is quite natural to assume that the design generated by (5.1) or by its modification via (5.2) will eventually reach the T-optimum design evaluated at the true model parameters. However, it is an unfortunate fact that convergence results for optimal sequential designs in nonstandard situations can only be sparsely found in the literature. Often the assumptions under which a rigorous mathematical proof can be found are quite restrictive, for an example refer to Müller & Pötscher (1992) and in a more general setting Wu (1985)a. However, there is a line of argumentation that nicely links numerical algorithms for the calculation of nonsequential designs, like those proposed by Fedorov (1986), to the type of procedures given above, providing a heuristic basis for the simulation experiment carried out in Section 9.

For the considered case (5.1) this means that the weak convergence

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\{g_A(\xi_i\hat{\beta}_{Ai})\log\left(\frac{g_A(\xi_i\hat{\beta}_{Ai})}{g_0(\xi_i\hat{\beta}_{0i})}\right)+(1-g_A(\xi_i\hat{\beta}_{Ai}))\log\left(\frac{1-g_A(\xi_i\hat{\beta}_{Ai})}{1-g_0(\xi_i\hat{\beta}_{0i})}\right)\right\}=$$

$$\sup_{\xi}\inf_{\pi_0}\int_{-\infty}^{\infty}\left\{\pi_A\log\left(\frac{\pi_A}{\pi_0}\right)+(1-\pi_A)\log\left(\frac{1-\pi_A}{1-\pi_0}\right)\right\}\xi dx$$

may follow from the convergence of the iterative algorithms of Fedorov (1986), if for the true model (under H_A) and for the design ξ_n the parameter estimators are consistent. Then we may pursue the argument that after any stage s of the procedure that is large enough to guarantee that the estimates are close to the true parameters, the future design will evolve like under full information and the dominating rest, the $n-s$ observation design, is known to be a good approximator for the optimal design measure. However, the proofs of consistency of parameters in a binary data model require independent observations and a fixed number of supporting points (besides other regularity conditions, cf. Agresti (1990)). Since for ξ_n both requirements only hold in the limit, it may be possible, although unlikely, to observe a counter-case in some situations. That this is not the case in the framework presented here is additionally supported by the simulation study in Section 9. Similar considerations can be undertaken for the modification by (5.2).

Another, more moderate danger to the consistency of the sequential design is the possibility that the optimal design may be singular for one of the candidate models. In these cases Fedorov (1986) suggests a regularisation by prescribing a small fraction of observations to a design that provides existence and uniqueness of $\hat{\beta}_{ji}$. A discussion of convergence of sequential D-optimal designs, providing guidelines for the present situation, is given by Chaudhuri & Mykland (1993).

7 Initial Design

The same problem occurs with the construction of the initial design ξ_0. For the case of models like those used in the example in Section (9) Wu (1985)b points out, that the existence and uniqueness of $\hat{\beta}_{j0}$ is not guaranteed if the intervals $[\min_{x_i \in \xi_{n_0}}\{x_i : y_i = 0\}, \max_{x_i \in \xi_{n_0}}\{x_i : y_i = 0\}]$ and $[\min_{x_i \in \xi_{n_0}}\{x_i : y_i = 1\}, \max_{x_i \in \xi_{n_0}}\{x_i : y_i = 1\}]$ do not overlap for more than one point. Therefore we suggest a sequential procedure also for the construction of the initial design: taking one observation at a point where the condition is likely to be fulfilled, checking the condition, taking the next observation and so forth until maximum likelihood estimators are available.

8 Nested Models

When the alternative models are such that they have the same link function and $\beta_0 = \{\beta_0', 0\}$ and $\beta_A = \{\beta_A', \beta_A''\}$, i.e. the models are nested, under certain regularity conditions the inverse of the covariance matrix of the maximum likelihood estimator $\hat{\beta}_A$ may be estimated by

$$\hat{M} = n\xi_n^T \left(\left(\frac{\partial \pi_A}{\partial \xi_n \hat{\beta}_A} \right) \operatorname{diag}\left(\hat{\pi}_A(1-\hat{\pi}_A)\right)^{-1} \left(\frac{\partial \pi_A}{\partial \xi_n \hat{\beta}_A} \right)^T \right) \xi_n \qquad (8.1)$$

with its limiting value not equaling the Fisher information matrix because of dependence, but being of similar form (cf. Silvey (1980)).

It can be partitioned according to the partioning of the parameters as

$$\hat{M} = \begin{pmatrix} \hat{M}_{A'A'} & \hat{M}_{A'A''} \\ \hat{M}_{A'A''}^T & \hat{M}_{A''A''} \end{pmatrix}.$$

The inverse of the determinant of the estimated covariance matrix of $\beta_{A''}$, the parameters that belong exclusively to the larger model, can be expressed as

$$|\hat{M}_{A''A''} - \hat{M}_{A'A''}^T \hat{M}_{A'A'}^{-1} \hat{M}_{A'A''}|, \qquad (8.2)$$

which is a scalar function of x, the last component of ξ_n and can be considered as a sort of generalized information measure of the estimator. Therefore we apply it to replace the rather clumsy expression in Step 2 of the proposed sequential procedure, which has the advantage that no separate estimates $\hat{\beta}_{0i}$ are required at each step. Alternatively we may apply the techniques of Harper & Neuhardt (1985) or multistage versions of the procedures of Minkin (1987) and Khan & Yazdi (1988), all developed for the parameter estimation case. (8.2) is usually referred to as the D_s-optimality criterion, cf. Atkinson & Donev (1992).

The relative performance of the technique proposed in Section 5 to the modified procedure of this section needs further investigation and we hope to discuss this problem in another paper.

9 Example

Within the framework (2.1) the two most widely used techniques are the logit and the probit analysis. In logit analysis the probabilities are related to the linear predictor in the form:

$$\pi_1 = \frac{\exp(\xi\beta_1)}{1+\exp(\xi\beta_1)} \tag{9.1}$$

whereas in probit analysis, the link function is supposed to be the inverse cumulative probability distribution for the standard normal random variable $g(.) = \Phi^{-1}(.)$, thus

$$\pi_2 = \Phi(\xi\beta_2). \tag{9.2}$$

For $m1 = m2$ their discrimination turns out to be quite difficult, not to say unfeasible, which is intuitive regarding the fact that they correspond quite closely over a wide region of the design space. As our reference example we use the univariate versions of (9.1) and (9.2), i.e. $m1 = m2 = 2$, with intercept $(X_{11} = X_{21} = 1)$.

The same example was used by Chambers & Cox (1967) and Yanagisawa (1988) to demonstrate the virtues of their local designs. Chambers & Cox (1967) restricted the number of different design points to $k = 3$ and assessed the minimum number of observations necessary to achieve a power of 50% at a significance level of 5%. Yanagisawa (1988) used the Pearson goodness-of-fit statistic (4.1) as design criterion, whilst we apply the log likelihood ratio (4.2) in all the calculations. $N(.5)$ the minimum number of observations for a power of 50% at a significance level of 5% of a discrimination test is asymptotically related to the noncentrality parameter as $N(.5) = 3.841/\lambda$.

Of course the values for $N(.5)$ reported by Yanagisawa (1988) underestimate the necessary effort in real life situations. Since it is hardly justifiable that one knows in advance, which is the true model (what would then discrimination be good for?) or what its true parameters are, one has to use estimates of the noncentrality parameter stemming from sequential designs for power calculations. The additional effort needed then solely depends on the quality of the initial design and the convergence rate of the parameter estimators.

The magnitude of these additional observations has to be evaluated by a simulation study. In the simulation process the true model and its parameter values are supposed to be known so that responses can be generated artificially according to local and general assumptions. Based on a randomized starting design and responses the sequential procedures of Section 5 are activated in order to find the next design point for which a new response will be generated by simulation. Prior to the next iteration the design, the sets of estimates for the parameters under each model, and all relevant variables must be updated. We continue this process until the desired power of 50% is achieved.

As data generating model both (9.1) and (9.2) were chosen, each with the parameter values $\beta_A = \{2, 2\}$. 10 runs were performed for each sequential design strategy considered and the minimum and maximum value of $N(.5)$ reported.

Concomitantly, based on the local assumptions concerning the rival models a local T-optimal design can be easily determined by means of optimal design theory, by using any algorithm from Fedorov (1986), for examples cf.

	β	$N(.5)$	T-optimum	(5.1)	(5.2)
H_0: probit	(1.091,1.091)	minimum	1036	1074	1127
H_A: logit	(2,2)	maximum	1036	1160	1316
H_0: logit	(3.618,3.618)	minimum	1323	1351	1422
H_A: probit	(2,2)	maximum	1323	1587	1661

Table 1: Power evaluation with various design techniques.

Ponce deLeon (1993). Such a design can be viewed as the target for the sequential process. In order to assess how fast convergence is achieved the design criterion may be plotted against the number of observations. This is done for the case H_0:probit, H_A:logit and procedure (5.2) in Figure 1, where the dashed line indicates the target value.

Insert Figure 1 here.
Figure1: $\hat{\lambda}$ vs. n for $\pi_A = \pi_1$ and $\pi_0 = \pi_2$.

The results of the simulation experiments are given in Table 1. It reports $N(.5)$ for sequential designs using (5.1) and (5.2) and the T-optimum design with logit (9.1) and probit (9.2) as the true model. From this table it can be seen that by using $N(.5)$ from the T-optimum design one considerably underestimates the number of observations necessary to achieve a power of 50%, namely by at least 8.1% and 7.0% for (9.1) or (9.2) true respectively.

Another interesting feature is not only that it is easier to discriminate between the models, when logit generates the data. This is a result that is in accordance with previous findings, cf. Yanagisawa (1988). Moreover, the number of additional observations attributed to allowing for symmetry in the hypotheses compared to those associated with the ignorance about parameter values is relatively much larger for logit than for probit generated data.

10 Conclusions

Although the applicability of the considered example is rather restricted it nevertheless provides a good reference point to problems of related kind and gives an indication of how different design strategies might perform in binary data model discrimination. It points out that one has to be especially careful in assessing power values based on local optimal designs, which inevitably leads to overestimating for fixed n. However the sequential techniques presented are valid in the whole class (2.1) and therefore have a wide range of application possibilities,

not only in the medical or biometrical field as Müller & Ponce deLeon (1993) demonstrate with an example from economics.

Acknowledgements

The authors are grateful to their phd-supervisors A.C. Atkinson and V.V. Fedorov for their authorship of a paper which drew attention to the subject.

References

A. Agresti. *Categorical Data Analysis.* Wiley, New York, 1990.

A.C. Atkinson and A.N. Donev. *Optimum Experimental Designs.* Oxford Statistical Science Series No.8, Oxford University Press, 1992.

A.C. Atkinson and V.V. Fedorov. The design of experiments for discriminating between two rival models. *Biometrika*, 62:57–70, 1975.

E.A. Chambers and D.R. Cox. Discrimination between alternative binary response models. *Biometrika*, 54:573–578, 1967.

P. Chaudhuri and P. Mykland. Nonlinear experiments: optimal design and inference based on likelihood. *Journal of the American Statistical Association*, 88(422):538–546, 1993.

V.V. Fedorov. *Model Choice: Design of T-Optimal Experiments.* Technical Report WP-86-000, IIASA, Laxenburg, 1986.

W.V. Harper and J.B. Neuhardt. Sequential D-optimal designs for binary response models. In *Proceedings of the Biopharmaceutical Section of the American Statistical Association*, pages 125–130, ASA, 1985.

G.E. Haynam, Z. Govindarajulu, and F.C. Leone. Tables of the cumulative non-central chi-square distribution. In H.L. Harter and D.B. Owen, editors, *Selected Tables in Mathematical Statistics*, Markham, Chicago, 1970.

M.K. Khan and A.A. Yazdi. On D-optimal designs for binary data. *Journal of Statistical Planning and Inference*, 18:83–91, 1988.

S. Minkin. Optimal design for binary data. *Journal of the American Statistical Association*, 82(400):1098–1103, 1987.

W.G. Müller and A.C.M. Ponce de Leon. *Optimal design of an experiment in economics.* Technical Report, submitted for publication, 1993.

W.G. Müller and B.M. Pötscher. Batch sequential design for a nonlinear estimation problem. In V.V. Fedorov, W.G. Müller, and I. Vuchkov, editors, *Model Oriented Data Analysis*, Physica, 1992.

A.C.M. Ponce de Leon and A.C. Atkinson. Optimal experimental design for discriminating between two rival binary data models in the presence of prior information. *Biometrika*, 78, 3:601–608, 1991.

A.C.M. Ponce de Leon. *Optimum experimental design for model discrimination and generalized linear models.* Technical Report, unpublished doctoral

thesis at the London School of Economics and Political Sciences, Department of Statistical and Mathematical Sciences, London, 1993.

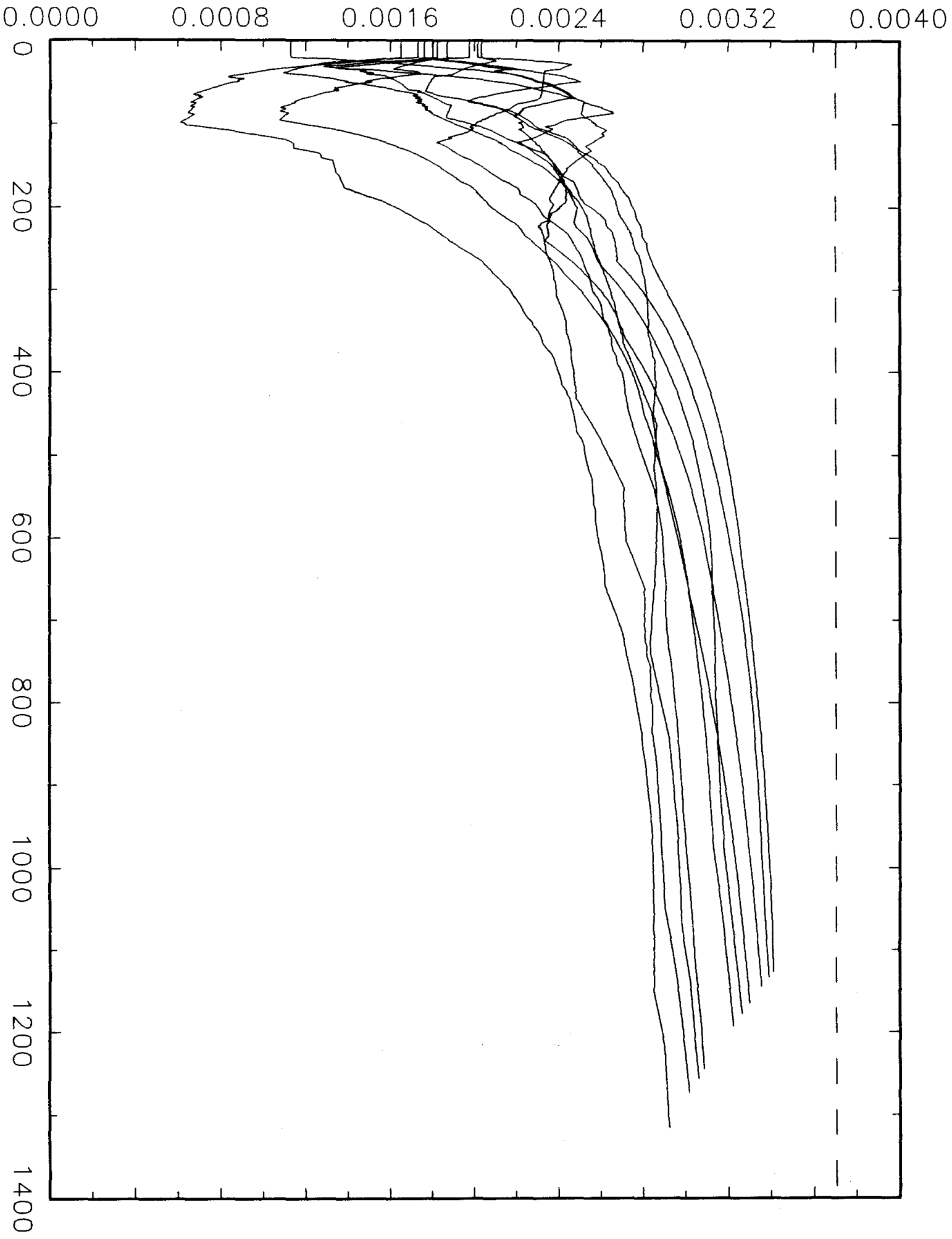
S.D. Silvey. *Optimal Design.* Chapman and Hall, London, 1980.

C.F.J. Wu. Asymptotic inference from sequential design in a nonlinear situtation. *Biometrika*, 72:553–558, 1985.

C.F.J. Wu. Efficient sequential designs with binary data. *Journal of the American Statistical Association*, 85:974–984, 1985.

Y. Yanagisawa. Designs for discrimination between binary response models. *Journal of Statistical Planning and Inference*, 19:31–41, 1988.

Y. Yanagisawa. Designs for discrimination between bivariate binary response models. *Biometrical Journal*, 32:25–34, 1990.


0.0000
0.0008
0.0016
0.0024
0.0032
0.0040
0
200
400
600
800
1000
1200
1400