

Contract no: IESBS20851A2/1/085

Experimental Design: A Bayesian Perspective

Merlise A. Clyde

Duke University, Durham, NC USA

This entry provides an overview of experimental design using a Bayesian decision-theoretic framework. Scientific experimentation requires decisions about how an experiment will be conducted and analyzed. Such decisions depend on the goals and purpose of the experiment, but certain choices may be restricted by available resources and ethical considerations. Prior information may be available from earlier experiments or from conjectures which motivate the investigation. The Bayesian approach provides a coherent framework where prior information and uncertainties regarding unknown quantities can be combined to find an experimental design that optimizes the goals of the experiment.

1 Introduction

Experimentation plays an integral part in the scientific method. Conjectures and hypotheses are put forth based on the current state of knowledge. Experimental data may be collected to address unknown aspects of the problem.

Finally, analysis of experimental results may lead to one hypothesis being favored over others or may lead to new questions and investigations, so that the process is repeated, with the accumulation of additional knowledge about the scientific process under investigation.

In some fields of scientific inquiry, physical models can be used to describe the outcome of an experiment given certain inputs with complete certainty. In the majority of applications, one cannot describe the scientific phenomena perfectly, leading to a distribution on possible outcomes, which can be described by probability models. For example, in comparing a new therapy to an existing treatment, individuals receive one of two treatments. The outcome or response is an indicator of “success” or “failure” of the given treatment, which can be modeled using Bernoulli distributions with unknown success probabilities. In psychological or educational testing, the outcome may be scores on a battery of tests. The outcomes or responses may depend on many other factors besides the assigned treatment, such as age, education, gender, or other explanatory variables. Hypotheses or quantities of interest are often phrased in terms of parameters θ of statistical models which relate the distribution of outcomes Y to levels of explanatory variables.

In conducting an experiment, there are many design issues to resolve, including deciding which treatments to study, which factors to control, and what aspects of an experiment to randomize. Other aspects of experimental design, such as how many experimental units are needed, how many observations should be

allocated to each treatment, or what levels of other input or control variables should be used, have traditionally fallen under the umbrella of statistical design (see *Experimental Design: Overview*). Because of costs, ethics, or other limitations on resources or time, sample sizes are usually restricted, and efficient use of available resources is critical. The purpose of optimal experimental design is to improve statistical inference regarding the quantities of interest by the optimal selection of values for design factors under the control of the investigator, within, of course, the constraints of available resources. Decision theory (see *Decision Theory, Bayesian*) provides a mathematical foundation for the selection of optimal designs. Prior information from earlier related experiments, observational studies, or subjective beliefs from personal observations, can be valuable in deciding how to allocate treatments efficiently, leading to more informative experiments. The Bayesian approach to experimental design provides a formal way to incorporate such prior information into the design process.

2 Bayesian Optimal Designs

The statistical aspects of an experiment e may be formally described by the sample space Ω (possible outcomes for the response Y), the parameter space Θ , and a probability model $p_e(y|\theta)$ that represents the distribution of observable random variables Y indexed by a parameter θ , an element of the parameter space Θ . Sample sizes, treatment levels, number of treatments, lev-

els of explanatory variables, or other aspects of the design to be selected are implicitly contained in $p_e(Y|\theta)$. The primary goal(s) or terminal decision(s) of an experiment may include, but are not limited to, estimating θ or other quantities of interest that are functions of θ , predicting future observations, selecting among competing models, or testing other hypotheses.

Lindley (1972) presented a two-part decision theoretic approach to experimental design, which provides a unifying theory for most work in Bayesian experimental design today. Lindley's approach involves specification of a suitable utility function reflecting the purpose and costs of the experiment; the best design is selected to maximize expected utility. In this framework, an experiment e is selected from the possible collection of experiments E (the first decision problem). After choosing an experiment e , outcomes Y are observed. Based on the observed data Y and experiment e , a terminal decision d is selected from possible decision rules D , which addresses the terminal goal(s) of the experiment. A utility function in the form $U(d, \theta, e, Y)$ encodes the costs and consequences of using experiment e and decision d with data Y and parameter θ . Assuming that the goals of an experiment and terminal decision can be formally expressed through a utility function, the Bayesian solution is to find the best design and best decision rule that maximize expected utility.

While the process of experimentation followed by inference/decision making proceeds in time order, it is easier to solve the optimal decision problem in reverse time order. The terminal stage decision problem involves finding the best

decision rule d given the observed data Y under experiment e that maximizes the *posterior* expected utility,

$$\max_d \int_{\Theta} U(d, \theta, e, Y) p(\theta|Y, e) d\theta = U(e, Y). \quad (1)$$

Here, the expectation or averaging over θ accounts for uncertainty regarding the unknown θ . The expectation is taken with respect to the posterior distribution of θ , which properly reflects uncertainty in θ at the terminal stage after Y has been observed under experiment e .

As the experiment e must be selected before data Y are observed, the second stage optimization problem involves finding the best experiment e that maximizes the *pre-posterior* expected utility. The pre-posterior expected utility is obtained by integrating the result in (1) over possible outcomes in the sample space Ω ,

$$U(e) = \int_{\Omega} U(e, Y) p(Y|e) dY = \int_{\Theta} \int_{\Omega} U(e, Y) p_e(Y|\theta) p(\theta) d\theta dY. \quad (2)$$

This integral is with respect to $p(Y|e)$, the marginal distribution of the data under experiment e , which is obtained by integrating $p_e(Y|\theta)$ over possible *prior* values for θ , described by prior distribution $p(\theta)$. The Bayesian solution to the experimental design problem is provided by the experiment e^* which maximizes $U(e)$:

$$U(e^*) = \max_e \int_{\Omega} \max_d \int_{\Theta} U(d, \theta, e, Y) p(\theta|y, e) p(y|e) d\theta dY. \quad (3)$$

This general formulation can be used to find optimal designs for a single exper-

iment, and can be extended to optimal selection of a sequence of experiments and sequential decision making (Lindley, 1972).

3 Choice of Utility Functions

It is important that utility functions be tailored to the goals of a given problem. Optimal designs for discriminating between two different models may be quite different than designs for prediction. In a one-way analysis of variance model, the best design for comparing k treatments to a control group, is not necessarily the optimal design for estimating the effects of $k + 1$ treatments, as these experiments have different goals. While taking equal number of observations in each of the $k + 1$ treatment groups is a possibility, other arrangements may provide more information, particularly when data from previous experiments are taken into account. For example, several previous studies may be available using an existing therapy, but limited information may be available on a new treatment. Differential costs of treatment also need to be considered, which may lead to other allocations of sample sizes or choice of experiments. Ethical considerations may also be incorporated that may constrain the assignment of treatments (Kadane, 1996).

Many of the papers in Bayesian design have based utility functions on Shannon information or quadratic loss. Motivation for these choices is discussed below. The reader is encouraged to refer to Chaloner and Verdinelli (1995) for additional details on these criteria, references, and relationships of other

Bayesian utility functions to standard optimality criteria.

3.1 Shannon Information

Shannon information is appropriate for inference problems regarding θ or functions in θ , without specification of particular hypotheses. The expected utility function is based on the expected change in Shannon information or equivalently the Kullback-Leibler divergence between the posterior and prior distributions. As the prior distribution does not depend on the design, this simplifies to the expected Shannon information of the posterior distribution,

$$U(e) = \int_{\Theta} \int_{\Omega} \log\{p(\theta|U, e)\} p_e(y|\theta) p(\theta) dY d\theta \quad (4)$$

thus the design goal is to find the design that maximizes the information provided by the experiment. In normal linear models with normal prior distributions, this leads to a criterion related to the well known D -optimality from classical design,

$$U(e) \propto \log |(X_e^T X_e + R)/\sigma^2| \quad (5)$$

where X_e is the design matrix for experiment e and R/σ^2 is the prior precision matrix (inverse of the prior covariance matrix).

If prediction of future observations is important, the expected gain in Shannon information for a future observation Y_{n+1} ,

$$U(e) = \int \log(p(Y_{n+1}|Y, e)p(Y_{n+1}|Y, e)p(Y|e)) dY dY_{n+1} \quad (6)$$

may be relevant. In normal linear models with prediction at a new point x_{n+1} this leads to

$$U(e) \propto \log[\{x_{n+1}^T (X_e^T X_e + R)^{-1} x_{n+1} + 1\}^{-1} / \sigma^2], \quad (7)$$

a Bayesian version of c -optimality.

3.2 Quadratic Loss

Point estimation based on quadratic loss leads to the expected utility function

$$U(e) = - \int_{\Theta} \int_{\Omega} (\theta - \hat{\theta})^T A ((\theta - \hat{\theta})) \quad (8)$$

where A is a symmetric non-negative definite matrix. The matrix A can be used to weight several different estimation problems where interest may be in estimating individual components of θ or linear combinations of θ . Under the normal linear model and normal prior distribution this results in

$$U(e) = -\sigma^2 \text{tr}\{A(X_e^T X_e + R)^{-1}\} \quad (9)$$

a Bayesian generalization of the A -optimality design criterion.

3.3 Other Utility and Loss Functions

Utility functions may be based on other loss functions besides quadratic loss. While a quadratic loss function may be appropriate in many cases, there are times when underestimation of a quantity incurs greater losses than overestimation. In such situations, asymmetric loss functions are more appropriate

for design and inference (Clyde et al., 1996). Discussion of other utility functions related to prediction, hypothesis testing and model discrimination and applications can be found in Chaloner and Verdinelli (1995).

3.4 *Multiple Objectives*

An experiment may often have several, possibly competing objectives which cannot be easily characterized by only one of the standard optimality criteria, and several design criteria may be appropriate. Weighted combinations of utility functions are still valid utility functions, so optimal designs for multiple objectives can be handled within the maximizing expected utility framework. For examples see Verdinelli (1992); Verdinelli and Kadane (1992); Clyde and Chaloner (1996). A difficulty with this approach is that utilities may be expressed in different scales which must be accounted for in the choice of weights. Equivalently, one can find the optimal experiment under one (primary) optimality criterion subject to constraints on the minimal *efficiency* of the experiment under criteria for other objectives. These approaches are a potential way to address robustness of experimental designs under multiple objectives, models, and prior distributions.

4 **Prior Distributions**

Prior elicitation is an important step in designing Bayesian experiments, as well as analysis. Clyde et al. (1996) use historical data from previous experi-

ments to construct a hierarchical prior distribution to use for design of future experiments. Kadane (1996) consider many of the practical issues in subjective elicitation for clinical trials. Tsai and Chaloner (2001) describe a design problem where prior distributions are elicited from over 50 clinical experts.

While some researchers may agree to using prior information to help design an experiment, they may want to have their final conclusions stand on their own, such as in a frequentist analysis. One of the goals of the experiment may be to convince a skeptic, who has different beliefs from the designer of the experiment, that a treatment is effective. In such cases the prior distribution used in constructing the posterior distribution $p(\theta|Ye)$ in the expected utility (1) for the terminal decision problem may be different than the prior distribution $p(\theta)$ used in finding the best design in (2). Etzioni and Kadane (1993) consider the problem of design when the designer and terminal decision maker have different prior beliefs corresponding to different decision rules.

5 Calculations

Calculation of expected utility, as in (1-2), requires evaluation of potentially high dimensional integrals, combined with difficult optimization problems, and in part, has limited the use of Bayesian optimal design in the practice. In normal linear models for many of the standard utility functions presented, the terminal decision rule can be solved in closed form and integrals can be computed analytically, leaving the optimization problem of finding the best

experiment e^* . Except in special cases, numerical optimization must often be used to find the optimal design.

One can often relax the problem in the following way. Finding optimal “exact” designs is often a difficult problem (similar to the traveling salesman problem). Designs can often be viewed in terms of a collection of support points (treatment levels) with weights that indicate the number of observations to be assigned at each support point. An exact design is one where the number of observations at each support point is an integer. A “continuous” design is obtained by allowing the solution for the weights to be any real number. Rather than finding the optimal exact design, in the relaxed problem, the class of experiments is enlarged to include continuous designs. Mathematically, finding the optimal design in the continuous problem is easier to solve, and methods for checking optimality are often feasible. If the solution corresponds to an exact design, one can then show that it is the globally optimal design. While continuous designs cannot be used in practice, rounding of continuous designs often provides exact solutions that are close to optimal.

For nonlinear models, generalized linear models and other “nonlinear” design problems (i.e. interest in nonlinear functions in an otherwise linear model), expected utility generally cannot be calculated in closed form and must be approximated. Many of the results for nonlinear design rely on asymptotic normal approximations in calculating expectations (Chaloner and Verdinelli, 1995). Integrals may also be approximated by numerical quadrature, Laplace

integration or Monte Carlo integration. With advances in computing resources, simulation-based optimal design is now an option, although methods often have to be designed for each specific application (see Müller (1999) for an overview of this area). As utility functions in many problems have been selected for their computational tractability, simulation-based design may soon open the way for greater use of scientific based utility functions that better match the goals of the experiment. Hierarchical models are becoming increasingly important in modelling latent and random effects and accounting for subject-to-subject variability, for example. Advances in Bayesian computation, such as Markov chain Monte Carlo, now mean that inference in such models can be carried out in real problems. It is becoming easier to accommodate hierarchical and other complex models in the design of experiments with simulation-based optimal design schemes. Clyde et al. (1996) compare analytic approximations and Monte Carlo based design schemes for a hierarchical binary regression model using an asymmetric loss function.

6 Applications

Pilz (1991) covers Bayesian design and estimation in linear models, although from a rather mathematical viewpoint. Atkinson (1996) reviews both classical and Bayesian optimal design for linear and nonlinear designs, and presents recent applications such as design for clinical trials. Chaloner and Verdinelli (1995) provide an thorough review of literature on Bayesian designs and ap-

plications, which includes experimental design for linear models such as regression and analysis of variance models, factorial and fractional factorial experiments, variance component models, mixtures of linear models, hierarchical models, nonlinear regression models, binary regression models, design for clinical trials and sequential experimentation. The article includes several worked out examples, including design for one-way analysis of variance. The article by Clyde et al. (1996) explores simulation-based design for a hierarchical logistic regression model, and illustrates how to construct prior distributions based on previous experiments. Other examples of simulation-based design appear in Müller (1999). For an example related to social sciences, see *Experimental design: large scale social experimentation*.

The area of Bayesian design and analysis for clinical trials in both sequential and non-sequential designs is an exciting and active area, with many practical developments. For a survey of literature on sequential design, see the entry *Sequential Statistical Methods*. The volume edited by Kadane (1996) describes a complete case study and discusses many of the critical issues in design and analysis of clinical trials, considering ethics, prior elicitation, randomization, treatment allocation, utilities, and decision making. Tsai and Chaloner (2001) describe the design of two large clinical trials, where prior distributions are based on eliciting prior opinions from over 50 clinicians. The design problem involves finding a sample size so that consensus of opinion is met with high probability, where consensus means that all clinicians would prescribe the

same treatment based on their posterior opinions after the trial. Other recent examples include Rosner and Berry (1995); Simon and Freedman (1997).

While the increase use of Bayesian design in applications is especially encouraging, specialized software is often required. There is a growing need for reliable, user-friendly software for Bayesian design so that the methods can be more accessible and see greater applicability.

7 Example

The following design problem is concerned with choosing a sample size for an experiment to confirm results from an earlier study. All individuals in the first study had been diagnosed with breast cancer. Each individual had a biopsy of tumor tissue analyzed for expression levels of a protein implicated in progression of the disease; measured protein expression levels were scored between 0 and 8. All 135 individuals received a chemotherapeutic agent, with the goal of achieving pre-operative tumor regression. The measured outcome for each subject was a binary indicator of tumor regression. The researchers expected that the probability of tumor regression would increase as protein expression levels increased, and were surprised that the estimated probabilities reached a maximum around a expression level of 5, and then declined with higher levels. A logistic regression model (see *Multivariate Analysis: Discrete Variables, Logistic Regression*) was used to relate expression levels to clinical outcome (indicator of tumor regression). The linear predictor in the logistic

regression model included a term with a parameter θ , such that for $\theta \geq 1$ the probability of tumor regression increases with expression level, while for $\theta < 1$ the probability of tumor regression increases up to a level of 5, then decreases. From a statistical perspective the evidence in favor of $\theta < 1$ was weak with a Bayes Factor of 1.65 (“not worth a bare mention” on Jeffreys’ scale of interpretation for Bayes Factors (Kass and Raftery, 1995)). The goal of the second study is to investigate whether this decline was a chance occurrence. If the decrease is real, then this may suggest that a higher dose is necessary at higher expression levels, leading to other experiments.

Spezzaferri (1988) presented a utility function for model discrimination which can be used in designs for testing $H_0 : \theta \geq 1$ versus $H_1 : \theta < 1$. Spezzaferri’s criterion can also incorporate parameter estimation, but for simplicity only model discrimination is considered. The expected utility function for model discrimination can be represented as

$$U(e) = \int_{\Omega} Pr(\theta < 1|Y, e)p(Y|e, H_1)dY$$

where $p(Y|e, H_1)$ is the marginal distribution of the new data under $H_1 : \theta < 1$.

The experimental design involves selecting the total sample size N_e , as individuals with specific protein expression levels cannot be selected in advance due to costs of obtaining the results. The expected utility versus negative costs ($-\$2000 \times N$) is illustrated in Figure 1, and was constructed using smoothing of utilities generated using Monte Carlo experiments (Müller, 1999). While

analytic calculations are intractable, calculation of utilities for this model are straightforward via simulation. The marginal distribution for Y given an experiment with sample size N must incorporate averaging over different configurations of expression levels among the N subjects and the unknown parameters in the logistic regression model. Configurations of size N are generated using a multinomial distribution with a probability vector π . The expression level distribution from the first study is used to construct a conjugate Dirichlet prior distribution for π . Data from the previous experiment is used to construct the prior distribution for θ and other parameters in the logistic model, and construct the predictive distribution of the new data given a vector of new expression levels. For each set of simulated data under the new study $Pr(\theta < 1|Y, e)$ is calculated, which incorporates both the old data (in the prior distribution) and the new data. For each sample size N , protein expression levels and outcomes are repeatedly generated from their predictive distributions, and the average of $Pr(\theta < 1|Y, e)$ over experiments e with the same sample size N_e provides an estimate of the expected utility for e . These averages were smoothed in order eliminate Monte Carlo variation in their estimates and are plotted in Figure 1.

For the above criterion, expected utility will increase with the total sample size, and as formulated the design problem has an infinite solution. Each point on the curve (Figure 1) is a combination of $U(e)$ and cost. Expected utility has a maximum at 1, which is obtained as N goes to infinity (and with infinite

costs). The experiment which minimizes costs is given by $N = 0$, where the posterior probability that $\theta < 1$ given the data from the previous study was 0.85. The researchers want to minimize total costs, but at the same time find a sample size such that the expected utility is above 0.95 (corresponding to a Bayes factor above 3, which would provide strong evidence in favor of H_1). Combining costs and model discrimination, the combined expected utility function can be written as

$$U(e) - \lambda(\$2000 * N_e) \quad (10)$$

where λ reflects the tradeoff between information for discriminating between the two hypotheses and costs of subjects. In general, it is difficult to specify these tradeoffs explicitly, as the different components are often measured in different units. Verdinelli and Kadane (1992) illustrate choosing λ when there are tradeoffs between two components in a utility function. If the expected utility for model discrimination is selected as 0.95, then the corresponding cost is \$421,327 (the point represented by the triangle in Figure 1), and λ corresponds to the negative of the slope of the curve at this point ($\lambda = 0.0001961/\$2000$). Figure 2 shows the combined utility given by equation (10) using this value of λ . The maximum (marked by the triangle) corresponds to an optimal sample size of 211. For utilities which combine more than two objectives, graphical solutions may not be feasible, but a constrained optimization approach can be used to determine the tradeoff parameters (Clyde and Chaloner, 1996).

8 Summary

Bayesian experimental design is a rapidly growing area of research, with many exciting recent developments in simulation-based design and a growing number of real applications, particularly in clinical trials. By incorporating prior information, the Bayesian approach can lead to more efficient use of resources with less-costly and more informative designs. Utility functions can be explicitly tailored to the given problem and can address multiple objectives. Models, prior distributions, and utility functions have often been chosen to permit tractable calculations. However, with increased computing power, simulation-based methods for finding optimal designs allow for more realistic model, prior and utility specifications. Optimal designs have often been criticized because the number of support points in the design equal the number of parameters (in nonlinear Bayesian design this is not always the case), which does not permit checking the assumed model form. Model uncertainty is almost always an issue in inference, and certainly at the design stage. Bayesian model averaging has been extremely successful in accounting for model uncertainty in inference problems, and has great potential for use in Bayesian experimental design for constructing more robust designs.

Bibliography

Atkinson, A. C., 1996. The usefulness of optimum experimental designs (disc: P95-111). *Journal of the Royal Statistical Society, Series B, Methodological*

- 58, 59–76.
- Chaloner, K., Verdinelli, I., 1995. Bayesian experimental design: A review. *Statistical Science* 10, 273–304.
- Clyde, M., Chaloner, K., 1996. The equivalence of constrained and weighted designs in multiple objective design problems. *Journal of the American Statistical Association* 91, 1236–1244.
- Clyde, M., Müller, P., Parmigiani, G., 1996. Inference and design strategies for a hierarchical logistic regression model. In: Berry, D., Stangl, D. (Eds.), *Bayesian Biostatistics*. Marcel Dekker (New York).
- Etzioni, R., Kadane, J. B., 1993. Optimal experimental design for another’s analysis. *Journal of the American Statistical Association* 88, 1404–1411.
- Kadane, J. B., 1996. *Bayesian Methods and Ethics in a Clinical Trial Design*. John Wiley & Sons (New York).
- Kass, R. E., Raftery, A. E., 1995. Bayes factors. *Journal of the American Statistical Association* 90, 773–795.
- Lindley, D. V., 1972. *Bayesian Statistics – A Review*. SIAM (Philadelphia).
- Müller, P., 1999. Simulation based optimal design (with discussion). In: Berger, J., Bernardo, J., Dawid, A., Smith, A. (Eds.), *Bayesian Statistics 6*. Clarendon Press (Oxford).
- Pilz, J., 1991. *Bayesian Estimation and Experimental Design in Linear Regression Models*. John Wiley & Sons (New York).
- Rosner, G. L., Berry, D. A., 1995. A Bayesian group sequential design for a multiple arm randomized clinical trial. *Statistics in Medicine* 14, 381–394.

- Simon, R., Freedman, L. S., 1997. Bayesian design and analysis of 2×2 factorial clinical trials. *Biometrics* 53, 456–464.
- Spezzaferri, F., 1988. Nonsequential designs for model discrimination and parameter estimation. In: Bernardo, J., DeGroot, M., Lindley, D., Smith, A. (Eds.), *Bayesian Statistics 3*. Clarendon Press (Oxford).
- Tsai, C.-P., Chaloner, K., 2001. Using prior opinions to examine sample size in two clinical trials. In: Gatsonis, C., et al. (Eds.), *Case Studies in Bayesian Statistics, Volume V*. Springer-Verlag (New York), to appear.
- Verdinelli, I., 1992. Advances in Bayesian experimental design (with discussion). In: Bernardo, J., Berger, J., Dawid, A., Smith, A. (Eds.), *Bayesian Statistics 4*. Clarendon Press (Oxford).
- Verdinelli, I., Kadane, J. B., 1992. Bayesian designs for maximizing information and outcome. *Journal of the American Statistical Association* 87, 510–515.

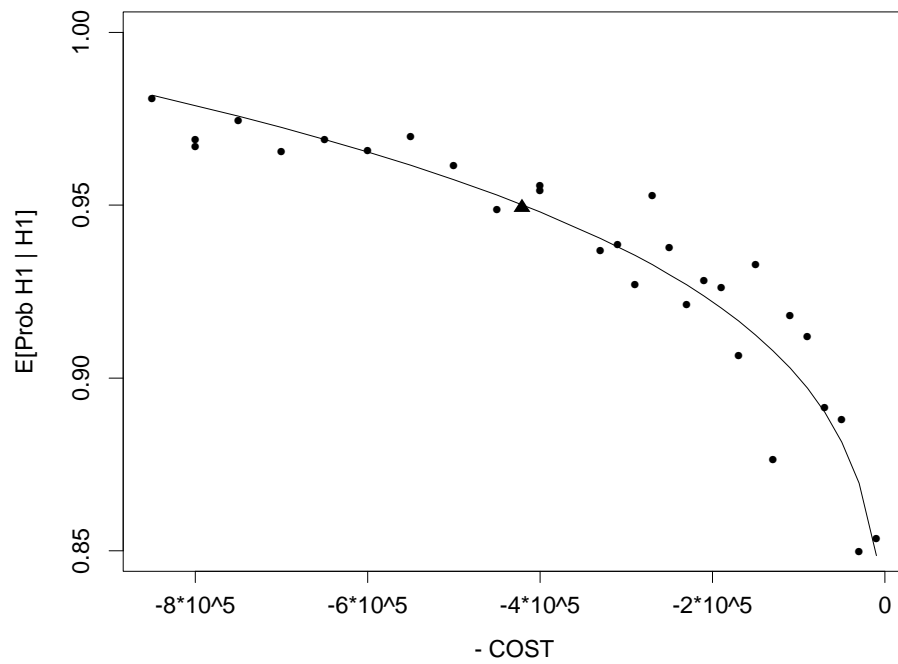


Figure 1

Smoothed expected utility for model discrimination versus costs; the points indicate simulated expected utilities which exhibit Monte Carlo variation

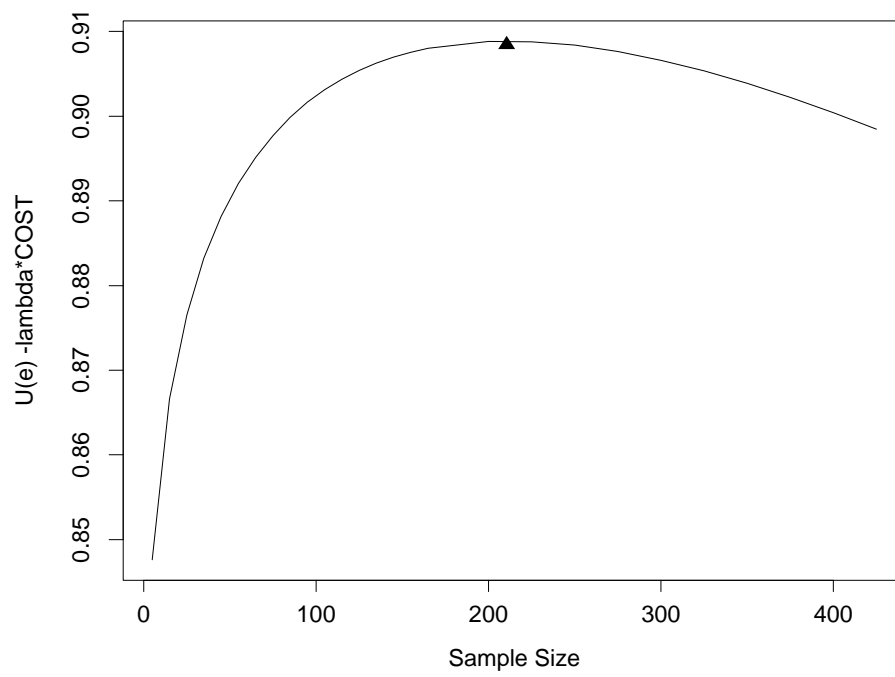


Figure 2

Combined utility for model discrimination with a penalty for costs; the triangle indicates the combined utility at the optimal design