

Optimal Sampling for Repeated Binary Measurements

Fernando A. Quintana* Peter Müller†

October 31, 2003

Key words and phrases: Bayesian decision problem; non-parametric Bayes; binary sequence data; optimal sampling.

MSC 2000: primary 62C10; secondary 62F15.

Abstract: We consider optimal design of sampling schedules for binary sequence data. The motivating example is a clinical trial where the measured responses are repeated measurements on a binary outcome. For example, in a study of smoking cessation the outcome could be an indicator for smoking. The decision is related to the trade-off between obtaining more information by more frequent sampling versus the incurred sampling cost.

We propose an approach which allows to incorporate a variety of goals in the utility function. We include deterministic sampling cost, a term related to prediction, and if relevant, a term related to learning about a treatment effect. To avoid dependence on a specific parametric form we use a non-parametric probability model, relying on minimal assumptions only. Quintana and Newton (1998) define partial ex-

*Departamento de Estadística, Pontificia Universidad Católica de Chile, Casilla 306, Santiago 22, CHILE. e-mail: quintana@mat.puc.cl. Partially supported by grant FONDECYT 1990430.

†Institute of Statistical Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251, USA. e-mail: pm@stat.duke.edu

changeability for a binary sequence. Assuming partial exchangeability the sampling distribution can be written as a mixture of order k homogeneous Markov chains. We use an implementation from Quintana and Müller (2003) that assumes a Dirichlet process prior for the mixture.

1 INTRODUCTION

We consider optimal sampling design for binary repeated measurements data. We approach the sampling design as a Bayesian decision problem. The motivating application is the choice of sampling schedules in a clinical trial with binary outcomes. The optimal choice is a trade-off of sampling cost for frequent observations versus the information lost if sampling is scheduled too infrequently.

Bayesian and decision theoretic approaches to optimal design in clinical trials are reviewed, for example, in Berry (1993), or Spiegelhalter et al. (1994). Important applications include optimal design for dose-finding trials (Thall and Russell, 1998; Whitehead and Brunier, 1995; Whitehead and Williamson, 1998), multi-arm clinical trials (Thall et al., 1995), choice of optimal sampling times (Stroud et al., 2001), and dose individualization (Wakefield, 1994). Another area of related design problems are network design problems in spatial contexts. See, for example Clayton et al. (1999). Optimal sampling for repeated measurements shares some features with spatial network design problems, albeit the univariate action space greatly simplifies the problem. We are not aware of any existing approaches specifically for repeated binary measurements.

We take a decision theoretic perspective and frame the optimal sampling design problem as expected utility maximization. The general setup of Bayesian decision theoretic designs is laid out, for example, in DeGroot (1970) or Berger (1985). Chaloner and Verdinelli (1995) and Verdinelli (1992) review applications of Bayesian design. The main elements of a decision problem are a set of possible actions $d \in \mathcal{D}$, a probability model $p_d(y|\theta)$ for future data y , a prior distribution $p(\theta)$ for the parameters and a utility function $u(d, \theta, y)$ that specifies the worth of decision d for assumed

values of parameters θ and data y . The sampling model $p_d(y|\theta)$ typically depends on the chosen action d , but the prior probability model usually does not. Therefore we include no d subindex in $p(\theta)$, although little changes if it does. Often decisions are made conditional on some pilot or historical data y_o , changing the relevant probability models to $p_d(y|\theta, y_o)$ and $p(\theta|y_o)$. It can be argued (DeGroot, 1970) that a rational decision maker should choose the action that maximizes expected utility $U(d) = \int u(d, \theta, y) dp_d(\theta, y|y_o)$. The approach we propose to optimal sampling for repeated binary measurements follows this paradigm.

The choice of sampling times is intrinsically linked with the nature of the serial dependence in the observed binary sequence. It is therefore important that the underlying probability model make no overly strict parametric assumptions unless they reflect genuine prior information. We will use a class of flexible non-parametric models for binary sequences introduced in Quintana and Müller (2003), who build on Quintana and Newton (1998) to define models for partially exchangeable sequences of random order k . The notion of order k partial exchangeability introduced in Quintana and Newton (1998) is a generalization of the traditional notion of exchangeability to dependent binary sequences. A representation result similar to de Finetti's representation theorem for infinite exchangeable sequences allows to consider probability models for repeated binary sampling that are based on invariance under a slightly less restrictive class of permutations than that implied by the symmetry of exchangeability assumptions.

In Section 2 we formally state the decision problem, still without reference to a particular probability model. In Section 3 we develop a semi-parametric probability model for repeated binary measurements. The model includes a regression on covariates, for example a treatment effect in a clinical trial application. In Section 3.2 we discuss implementation issues, including Monte Carlo and Markov chain Monte Carlo (MCMC) simulation. We conclude with an application example in Section 4.

2 THE DECISION PROBLEM

2.1 The Design Criterion

We start the formal description of the decision problem with a discussion of the design criterion, i.e., the utility function. The proposed criterion includes a deterministic sampling cost, a term related to predicting the response for a future patient, and a term related to learning about treatment effects. The criterion does not depend on specific features of the underlying probability model. In the following discussion we only need to assume some minimal structure for this model. Actual implementation of course requires to adopt a model. Later, in Section 3 we will introduce the model used in our implementation.

To simplify notation and terminology we describe our design approach assuming an application to a clinical trial design where each patient gives rise to a sequence of binary observations. Let $y_{ij} \in \{0, 1\}$ denote the responses for patient i at time j , $i = 1, \dots, N$ and $j = 1, \dots, n_i$. Write $y_i = (y_{ij}, j = 1, \dots, n_i)$ for $i = 1, \dots, N$. We assume a mixed effects model with patient specific random effects α_i , fixed effects β , occasion specific covariates w_{ij} , patient specific covariates x_i and possibly hyperparameters ϕ :

$$y_i \sim p(y_i | w_{ij}, x_i, \alpha_i, \beta), \quad \alpha_i \sim p(\alpha_i | x_i, \phi), \quad p(\beta, \phi), \quad (1)$$

For example, the longitudinal data model $p(y_i | w_{ij}, x_i, \alpha_i, \beta)$ could be a Markov chain with transition probabilities parametrized by (α_i, β) . The random effects distribution could be a regression on treatment indicators x_i , completed with a hyperprior on β and ϕ . The probability model introduced later, in Section 3, has a similar structure, with specific choices for the Markov chain transition probabilities and the random effects model $p(\alpha_i | x_i, \phi)$. We allow regression on patient specific covariates x_i at the level of both, the random effects distribution $p(\alpha_i | \dots)$ and the sampling model $p(y_{ij} | \dots)$.

We do not include a probability model for x_i and w_{ij} . As usual in regression models we assume that the mechanism of assigning covariates is independent of the hyperparameters (β, ϕ) , random effects α_i and responses y_{ij} and is therefore irrelevant

for posterior inference. However, in the design problem we need to consider the prior predictive distribution of future data. In this context we will assume that there is a known mechanism to generate covariates. In the example we use resampling from the covariates of patients in a previous trial. Alternatively, one could assume some deterministic process to assign covariates to future patients.

Some data might already be available at the time of decision making, for example an earlier study, or preliminary pilot data. Assume patients $i = 1, \dots, N_o$, $N_o < N$ are already observed. If patients can not be considered exchangeable across studies, a study-specific random effect can be included in α_i . Use $N_o = 0$ for no historical data.

After observing the first N_o patients we wish to decide on a sampling schedule for the future patients $i = N_o + 1, \dots, N$. A sampling schedule could be, for example, to record $y_{ij}, j = 1, 3, 5, \dots, n_i$. We use d to generically denote the sampling design and partition the data vector into (y_o, y_d, y_c) , where $y_o = (y_{ij}, i = 1, \dots, N_o)$ is the observed data from the first study, y_d are the responses of future patients that will be observed under design d , and y_c are the responses of future patients that will not be observed under design d . Also, we will use $y = (y_c, y_d)$, $\theta = (\alpha, \beta, \phi)$, and $\theta_o = (\alpha_i, i = 1, \dots, N_o)$ to denote data and random effects specific to the new and old study, respectively, where $\alpha = (\alpha_i, i = N_o + 1, \dots, N)$. Note that fixed effects β and hyperparameters ϕ are included in θ . We might parametrize the sampling design, for example, by letting d denote the sampling frequency, with $d = 1$ indicating that all responses are observed, $d = 2$ indicating that every 2nd observation is observed, etc. Then $y_d = (y_{ij}, i = N_o + 1, \dots, N, j = 1, 1 + d, 1 + 2d, \dots, n_j)$. But the sampling design d might be more complicated, including, for example, more frequent sampling in the first few periods and longer periods between later observations. The only constraint of the proposed approach is that the number of considered sampling designs be moderately small, say less than 100. The constraint is imposed by the need for simulation based evaluation of expected utilities for each possible design.

Choice of an optimal sampling design requires the definition of a utility function

that quantifies what alternative designs are worth. We start by defining a utility function $u(d, y, \theta)$ for an observed experiment (y, θ) , i.e., the value of the design d for hypothetical future data y and parameters θ . The first term of the utility function is a deterministic sampling cost $C_1(d)$, for example $c_1 \cdot |d|$, where $|d|$ is the number of observations under design d . The second term introduces a penalty $R_{ij}(y_{ij}, y_d)$ for residuals in predicting the responses y_{ij} for sampling times j that are not observed under design d . For example, this could be $|y_{ij} - E(y_{ij}|y_d)|$. The third term is related to estimating treatment effects. It is only included if one of the patient specific covariates x_i is a treatment indicator. Without loss of generality assume $x_i = 0$ for placebo and $x_i = 1$ for patients administered an experimental treatment under consideration. We use a statistic T_{ij} that quantifies learning about the treatment effect. For each patient i , we assume that the study includes a matching patient i' with identical covariates, but opposite treatment, $x_{i'} = 1 - x_i$. If this is not already true we can always augment the future data y to include a patient i' with such covariates. Of course, this hypothetical patient should then not be included in the first two terms of the loss function. To simplify notation we write y'_{ij} for $y_{i',j}$ and y'_i for $y_{i'}$. The measurement y'_{ij} is the response that we could have observed for patient i if the treatment assignment was reversed. We define

$$T_{ij}(y_{ij}, y'_{ij}, y_d) = \left| [E(y_{ij}|y_d) - E(y'_{ij}|y_d)] - [y_{ij} - y'_{ij}] \right|.$$

In words, T_{ij} is the absolute error in estimating the treatment effect for patient i , time j , with treatment effect defined as the difference between mean response under the treatment versus no treatment for patient i .

In summary, we use

$$u(d, y, \theta) = -C_1(d) - c_2 \frac{1}{|I_1|} \sum_{(ij) \in I_1} R_{ij}(y_{ij}, y_d) - c_3 \frac{1}{|I_2|} \sum_{(ij) \in I_2} T_{ij}(y_{ij}, y'_{ij}, y_d) \quad (2)$$

with the three terms corresponding to sampling cost, prediction and learning about treatment effects. Here $|I_j|$ denotes the cardinality of the set I_j . I_1 is some subset of responses which are not observed under design d . For example, $I_1 = \{(i, j), i =$

$N, y_{ij} \in y_c\}$, i.e., we might use prediction for the missing responses of the last patient as design criterion. Since patients are a priori exchangeable this is equivalent to considering prediction for all responses missing under design d , i.e., $I_1 = \{(i, j), y_{ij} \in y_c\}$. Similarly, I_2 indexes some subset of the responses in the new study. For example, we might use $I_2 = \{N, n_N\}$ to design for the treatment effect in the last period for the last patient. The optimal design d^* is formally defined by maximizing expected utility $U(d) = E[u(d, y, \theta)]$. The expectation is with respect to the joint distribution $p(\theta, y) = p(\theta)p(y_d, y_c|\theta)$ on parameters and data (including latent data y_c). If the decision is made conditional on some already observed data y_o , the appropriate distribution is the joint posterior (predictive) distribution $p(\theta, y_d, y_c | y_o)$, and the conditional expectations $E(y_{ij} | y_d)$ in the definition of R_{ij} and T_{ij} are replaced by $E(y_{ij} | y_d, y_o)$. Exploiting conditional independence in (1)

$$U(d) = \int u(d, y, \theta) p(y|\theta) p(\theta|y_o) d\theta dy. \quad (3)$$

Although $u(\cdot)$ as defined in (2) involves only y and no θ , it is technically convenient to augment the expected utility integral to an integral with respect to $p(y, \theta|y_o)$ instead of $p(y|y_o)$. The augmented integral is easier to evaluate because posterior predictive simulation from $p(y|y_o)$ is typically implemented in a two step procedure: simulating $\theta \sim p(\theta|y_o)$ and then $y \sim p(y|\theta)$. Also, the augmentation with θ greatly simplifies the evaluation of the conditional expectations in $u(d, y, \theta)$, as we will show below.

2.2 Evaluating Expected Utility

Substituting (2) into (3) the expected utility becomes

$$U(d) = -C_1(d) - \int \left\{ \frac{c_2}{|I_1|} \sum_{I_1} R_{ij}(y_{ij}, y_d) + \frac{c_3}{|I_2|} \sum_{I_2} T_{ij}(y_{ij}, y'_{ij}, y_d) \right\} dp(y, \theta|y_o). \quad (4)$$

We evaluate (4) by Monte Carlo simulation for the integral with respect to $p(y, \theta | y_o)$, together with nested MCMC simulation to evaluate $E(y_{ij} | y_d, y_o)$ for use in T_{ij} and R_{ij} . We argue below how the setup of (4) greatly simplifies the implementation of this

MCMC. We generate simulated experiments $(\theta^{(h)}, y^{(h)}) \sim p(\theta, y|y_o)$, $h = 1, \dots, H$, from the joint probability model on parameters θ and data y , evaluate the integrand

$$v(\theta^{(h)}, y^{(h)}) = \frac{c_2}{|I_1|} \sum_{I_1} R_{ij}(y_{ij}^{(h)}, y_d^{(h)}) + \frac{c_3}{|I_2|} \sum_{I_2} T_{ij}(y_{ij}^{(h)}, y_{ij}'^{(h)}, y_d^{(h)}),$$

and use the sample averages of these $v(\theta^{(h)}, y^{(h)})$ values to estimate expected utility. In Section 3 we discuss simulation from $p(\theta, y|y_o)$ in a specific probability model. The evaluation of $v(\cdot)$ requires $E(y_{ij} | y_d, y_o)$ to substitute into the definition of R_{ij} and T_{ij} . We evaluate these posterior predictive means by a nested MCMC simulation from $p(\theta, y_c | y_d, y_o)$. The nested MCMC simulation greatly simplifies by the following procedure. First rewrite

$$E(y_{ij} | y_d, y_o) = \int E(y_{ij} | \theta, y_{-ij}) p(\theta, y_c | y_d = y_d^{(h)}, y_o) d\theta dy_c, \quad (5)$$

where $y_{-ij} = y_i \setminus y_{ij}$ includes elements from y_d and y_c . We evaluate (5) by running MCMC simulation to generate a Monte Carlo sample from the posterior predictive $p(\theta, y_c | y_d^{(h)}, y_o)$. The MCMC is initialized with the values $(\theta^{(h)}, y_c^{(h)})$ that are generated to evaluate (4). But the pair $(\theta^{(h)}, y_c^{(h)})$ can be considered a draw from $p(\theta, y_c | y_d^{(h)}, y_o)$, since $p(\theta^{(h)}, y^{(h)} | y_o) = p(y_d^{(h)} | y_o) p(\theta^{(h)}, y_c^{(h)} | y_d^{(h)}, y_o)$, i.e., the MCMC simulation is initialized with an (exact) draw from the desired stationary distribution, and thus no burn-in is required. By construction the chain is in equilibrium from the start.

We use Monte Carlo integration to evaluate expected utility for each design d . If only a moderate number of possible designs are considered we can evaluate $U(d)$ for all of them and find the optimal design by inspection. In the application presented in Section 4, for example, only 9 possible designs are considered. This is fairly typical for sampling designs in binary sequences. However, if the design space is more complex, the Monte Carlo evaluation of expected utilities needs to be combined with an appropriate maximization routine to find the optimal design.

A practically important concern is the dependence of the optimal design on the trade-off parameters (c_2, c_3) . This is a generic problem in any design problem which combines multiple goals and no simple solutions exist. As a practical approach in the

absence of other information we suggest to proceed as follows. Consider the three terms in the expected utility function (4), related to sampling cost, prediction and inference loss, respectively. The additive nature of $U(d)$ allows to evaluate each term separately, writing $U(d) = -C_1(d) - c_2 R(d) - c_3 T(d)$, with the definition for $R(d)$ and $T(d)$ implied by the corresponding terms in (4). We propose to consider the range of $C_1(d)$, $R(d)$, and $T(d)$ over all possible designs and adjust the values of c_2 and c_3 to achieve approximately matching ranges for each term. Also, separate evaluation of C_1 , R and T allows computationally simple exploration of sensitivity of the optimal design with respect to c_2 and c_3 . See the discussion in Section 4 for an example.

3 THE PROBABILITY MODEL

3.1 *Partially Exchangeable Binary Sequences*

The discussion of the design criterion was without reference to a specific probability model, highlighting the model independent nature of the proposed approach to choosing a sampling design. We now propose a model which is suitable for the desired design problem. In choosing a probability model we face the following important considerations. The model should not impose strict parametric assumptions. In particular, the model should be flexible with respect to serial dependence in the binary sequence and ideally include the order of serial dependence as an explicit parameter. The model needs to include structure to enable joint inference for y_o and future data $y = (y_d, y_c)$, possibly allowing study-specific random effects. Within these constraints any model could be used to implement the discussed optimal design approach. In the following discussion we propose a Bayesian non-parametric model that we suggest as a default. It minimizes the model dependence of the final sampling design. However the use of a non-parametric model is in no way critical to the proposed sampling design choice. In fact, for the example reported in Section 4 we find that analogous parametric models would lead to very similar expected utilities and to the same sampling design. See the discussion at the end of Section 4.

Based on the above considerations we use an extension of the semi-parametric Bayesian model proposed in Quintana and Müller (2003). The model assumes that the binary sequence $y_i = (y_{ij}, j = 1, \dots, n_i)$ is partially exchangeable. Specifically, we assume that, conditional on k , the joint distribution of y_i is invariant under any permutation of $(y_{ij}, j = 1, \dots, n_i)$ that leaves the order k transition counts unchanged. For example, when $k = 1$, this means invariance under permutations that leave the number of transitions $0 \rightarrow 0$, $0 \rightarrow 1$, $1 \rightarrow 0$ and $1 \rightarrow 1$ unaltered. Under some additional technical conditions, it can be shown (Quintana and Newton, 1998) that this implies that the sampling distribution is a mixture of homogeneous Markov chains, with the mixture being with respect to the transition probabilities and the order of dependence in the Markov chain. Consider now the transition matrix for the i -th subject assuming an order K Markov chain. Instead of introducing parameters for all 2^K transition probabilities, we follow the more parsimonious approach described in Quintana and Müller (2003) and consider a simplified reparametrization in terms of a log-linear model:

$$\text{logit } P(y_{ij} = 1 | y_{i,j-1}, \dots, y_{i,j-K}, \alpha_i, \beta, K) = \alpha_{i0} + \sum_{1 \leq \ell \leq K} \alpha_{i\ell} y_{i,j-\ell} + \beta_x x_i + \beta_w w_{ij}. \quad (6)$$

Here $y_{i,j-\ell}$ are the lagged responses, x_i are patient specific covariates and w_{ij} are occasion specific covariates. The random effects vector $\alpha_i = (\alpha_{i0}, \dots, \alpha_{iK})$ collects all logistic regression parameters $\alpha_{i\ell}$. Additional fixed effects $\beta = (\beta_x, \beta_w)$ are common across subjects. Order k dependence, $k < K$, is expressed by vanishing higher order logistic regression parameters. If desired, (6) could be extended to include higher order interactions of the type $\alpha_{i\ell h} y_{i,j-\ell} y_{i,j-h}$, etc. Alternatively, one could define parameter vectors $\alpha(k) = (\alpha_\ell(k), \ell = 1, \dots, K)$ for each k and define α_i as the combined vector of all $\alpha(k)$. However, joint inference on $\alpha(k)$ and $\alpha(k')$ for any two different $k \neq k'$ is meaningless. And for marginal inference on $\alpha(k)$, the two representations are equivalent.

To translate this structure into the desired partially exchangeable probability model for y_i we still need to parametrize a mixing measure for (α_i, k) with respect to

which we mix the order k homogeneous Markov chains. As a compromise between ease of implementation and generality we use a Dirichlet process (DP) prior for α_i and a multinomial distribution on $k \in \{1, \dots, K\}$. DP priors are defined in Ferguson (1973). See, for example, MacEachern and Müller (2000) for a recent review. Denoting the right-hand side of (6) by $\text{logit}\gamma_{ij}(\alpha_i, \beta, k)$ our model can be summarized as

$$y_i \sim p(y_i | \alpha_i, \beta, k), \quad \alpha_i \sim G, \quad G \sim DP(M, G^o), \quad k \sim p_k(k), \quad \beta \sim p_\beta(\beta), \quad (7)$$

where $i = 1, \dots, N$ ranges over all subjects (from the old and new studies) and

$$p(y_i | \alpha_i, \beta, k) = \prod_{j=K+1}^{n_i} \gamma_{ij}(\alpha_i, \beta, k)^{y_{ij}} [1 - \gamma_{ij}(\alpha_i, \beta, k)]^{1-y_{ij}}. \quad (8)$$

Here M and G^o are the total mass parameter and base measure for the DP.

3.2 Posterior Predictive Simulation

Evaluation of the expected utility integrals in (4) requires posterior (predictive) simulation from $p(\theta, y | y_o)$, where $\theta = (\alpha_i, i = N_o + 1, \dots, N, \beta, k)$, including fixed effects and hyperparameters (β, k) . Exploiting conditional independence, posterior predictive sampling in model (7) is easily achieved by considering

$$\begin{aligned} p(y, \theta | y_o) &= \int p(y | \theta) p(\theta | \theta_o) p(\theta_o | y_o) d\theta_o \\ &= \int p(y | \theta) p(\alpha_i, i = N_o + 1, \dots, N | \theta_o) p(\beta, k, \theta_o | y_o) d\theta_o. \end{aligned} \quad (9)$$

We simulate from (9) by drawing in turn from each of the three distributions. Generating from $p(\beta, k, \theta_o | y_o)$ is posterior simulation conditional on the historical data. Generating from $p(\alpha_i, i = N_o + 1, \dots, N | \theta_o)$ is prior simulation for the future patient specific random effects α_i . It is implemented as Polya urn sampling. Finally, simulating from $p(y | \theta)$ requires generation of order k Markov chains. Details of all three steps are described in the Appendix.

A common issue in decision problems is that the model used to simulate future data, i.e., the probability model with respect to which expected utilities are defined,

needs more structure than the analysis model which is used to eventually derive inference conditional on the data. For example, in the regulatory environment of clinical trial design it is common to use informative priors for the design, but conservative, skeptical priors for the analysis. See, for example, Vlachos and Gelfand (1998) for a discussion. We refer to the two, possibly different, probability models as the design model and the inference model. A related issue arises in designing sampling strategies for binary sequences in the context of model (7). Posterior predictive simulation of a new trial requires to generate n_i , x_i and starting values y_{ij} , $j = 1, \dots, K$ for the Markov chain $p(y_i | \alpha_i, k, \beta)$ defined in (7). The analysis model (7) does not include any modeling for these variables. Instead of formal statistical modeling we propose to use the empirical distribution of observed sequence lengths n_i , covariates x_i and starting sequences $(y_{ij}, j = 1, \dots, K)$ from the earlier study, i.e., resample the values from $i = 1, \dots, N_o$.

4 EXAMPLE

Davis and Wei (1988) describe a bladder cancer study with $N_o = 82$ patients, and with up to $\max n_i = 12$ observations taken every third month for each patient. Each observation records an indicator for recurrence of bladder cancer tumors $y_{ij} \in \{0, 1\}$ for patient i in period j . The data set does not report y_{ij} for all patients and periods. For the purpose of this example we imputed the missing historical data points. Patients are grouped into controls, $x_i = 0$, and treatment, $x_i = 1$. In the context of this study we consider the problem of designing a future study. We assume the following prior probability model. For the DP base measure G^o we assume independent normal $N(0, \sigma^2)$ distributions for the logistic regression coefficients α_ℓ , with standard deviation $\sigma = 2$. For the treatment effect we use $\beta \sim N(0, \sigma^2)$ and for the order of dependence we assume a uniform prior $p(k) = 1/(K + 1)$ for $k = 0, \dots, K$. The total mass parameter is fixed at $M = 1$. While it would be straightforward to extend the model to include a prior on M , we found that the expected utility for the sampling designs is not sensitive to the choice of M .

We take $K = 2$, and y_{i1} and y_{i2} , $i = N_o + 1, \dots, N$, are chosen according to the empirical distribution of the corresponding responses in the observed (first) study. We design the sampling schedule d for a future study of $(N - N_o)$ patients, considering possible designs $d \in \{1, 2, 3, 4, \dots, 9\}$ and with $n_i = 13$ three-monthly responses for all future patients, $N_o + 1 \leq i \leq N$. Here $d = 1$ indicates that responses will be recorded at each possible occasion, $j = 1, \dots, n_i$, and for $d = 2, \dots, 8$ responses will be observed at $j \in \{3, 3 + d, 3 + 2d, \dots\}$. For example, for design $d = 2$ the observed data will be $\{y_{ij}, j = 3, 5, 7, 9, 11, 13\}$ while for design $d = 6$ recorded responses will be $\{y_{ij}, j = 3, 9\}$. For comparison we include as an additional design choice the option of collecting no data. We label this design as $d = 9$ and report the corresponding prior expected utilities.

We set up posterior predictive simulation to evaluate expected utilities in (4) with $p(y, \theta | y_o)$ conditional on the first N_o patients from the earlier study. We ran a chain of length 30,000. We evaluated several diagnostics implemented in the BOA software (Smith, 2000) and found no evidence for practical convergence problems. The multivariate potential scale reduction factor (Gelman and Rubin, 1992) was evaluated as 1.004, including k , β , and the logistic regression coefficients α_i for 4 randomly chosen patients. All parameters passed the stationarity tests proposed in Geweke (1992) and Heidelberger and Welch (1983). Conditional on the historical data y_o we find the following posterior means and posterior standard deviations for key parameters. The marginal posterior means for k , β and N^* are 1.4, -0.55 and 5.0, respectively. The corresponding marginal posterior standard deviations are 0.54, 0.60, and 1.59, respectively. The posterior from the old study becomes the prior for the design study. Since posterior inference in the old study is found to be robust with respect to the prior choices for the old study (data not shown) we conclude that the entire design approach is robust with respect to the initial hyperparameter choices.

We simulated $H = 250$ posterior predictive draws $(\theta^{(h)}, y^{(h)})$. For each simulated $y^{(h)}$ we set up a nested MCMC run to evaluate the conditional means $E(y_{ij} | y_d, y_o)$. For this nested MCMC we used 250 iterations, starting with the (true) simulated

$y_c^{(h)}, \theta^{(h)}$. Since the chain starts in equilibrium we can accumulate ergodic averages without any burn-in.

Table 1 reports the estimated expected utilities across alternative designs. Numerical uncertainties for the reported expected utilities are all below 0.06. The first columns report the expectations for the three terms in the utility function (2) corresponding to sampling cost, prediction and treatment effect. We use $C_1(d) = -|d|$. Figure 1 plots the values against $|d|$, the number of repeated measurements per patient. The terms $R(d)$ and $T(d)$ rise slowly from three-monthly sampling, $d = 1$, to sampling every six months, $d = 2$. Under $d = 1$, by definition $R = 0$ since $y_c = \{\}$. In both terms a critical change occurs when moving from 3 to only 2 measurements (designs $d = 5$ to $d = 6$). The last column in Table 1 combines the three terms using weights $c_2 = -10$ and $c_3 = -5$. The optimal design is found for $d = 2$. Little is lost when moving to more frequent sampling, $d = 1$, slightly more is lost when moving to $d = 3$. Dropping below 3 repeated measurements per patient is almost equivalent to prior inference, without any data.

The proposed optimal design approach is independent of the specific model. We used a non-parametric Bayesian model as a default choice to minimize model dependence of the final sampling recommendation. But the use of this non-parametric Bayesian model is by no means critical for the proposed design approach. To explore the use of alternative models and also to investigate sensitivity of the design choice with respect to the underlying probability model we considered two parametric models. To facilitate a comparison the models were chosen to be similar to the non-parametric model. In particular, we defined a parametric version of model (7) by replacing the unknown random effects distribution G with the base measure G^o . The resulting model takes the form of longitudinal data models as in Zeger and Karim (1991), including the lagged response $y_{i,j-\ell}$ as occasion specific covariates. The second parametric model assumes Markov transition probabilities parametrized as in (6), using a common α for all patients, with $\alpha \sim G^o$. The model is a mixture of order k Markov chains with prior G^o for the transition probabilities. We refer to the two

Table 1: Expected utilities under alternative designs. Let $R(d)$ and $T(d)$ denote the terms in (2) related to prediction and inference on the treatment effect, respectively. Each row reports one design, sampling cost, R and T . Numerical uncertainties are all below 0.06.

design d	$C_1(d)$	$E(R d)$	$E(T d)$	$U(d)$
1	11	0.00	0.00	-11.0
2	6	0.15	0.29	-9.0
3	4	0.34	0.61	-10.5
4	3	0.54	1.02	-13.5
5	3	0.50	0.99	-13.0
6	2	0.99	1.87	-21.3
7	2	1.07	1.84	-21.9
8	2	0.97	1.73	-20.3
9	0	1.26	2.22	-23.7

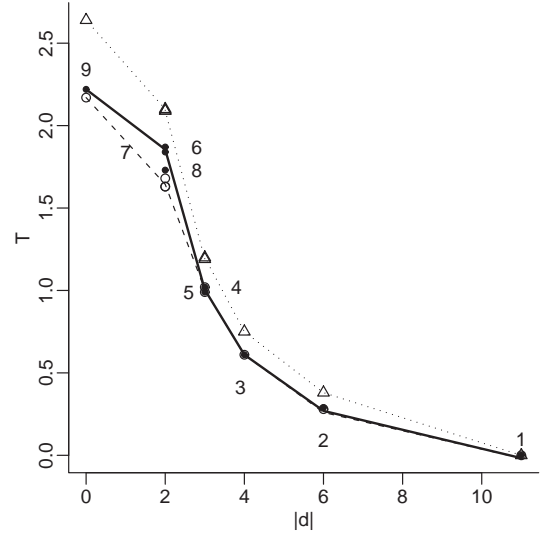
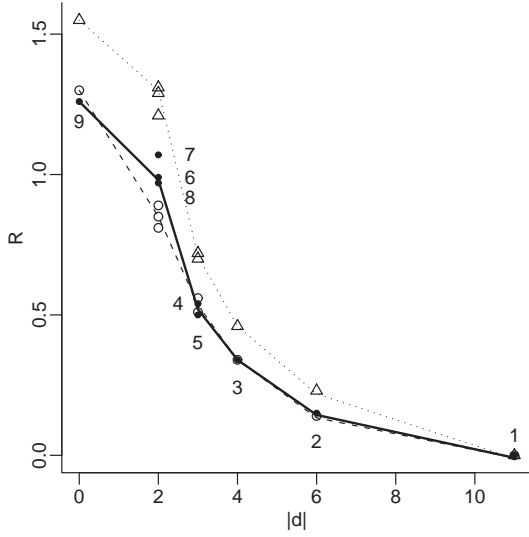
parametric models as models I and II. Both parametric models can be interpreted as special cases of (7), with $M \rightarrow 0$ (model II) and $M \rightarrow \infty$ (model I), respectively. Expected utilities under the two models are shown in Figure 1. While expected utilities differ slightly under the alternative models, the finally recommended sampling design remains unchanged.

Inference about the optimal sampling design typically involves trading off competing goals related to prediction, sampling cost and inference loss. This is formalized in the utility function proposed in (2). As in many biomedical decision problems specification of the trade-off weights c_2 and c_3 is a challenging problem. We have earlier proposed a pragmatic default choice for c_2 and c_3 . The additive nature of (2) allows an easy implementation of informal sensitivity analysis. Separately computing the expectations $R(d)$ and $T(d)$ corresponding to the second and third term in (2) facilitates a computationally efficient evaluation of expected utility $U(d)$ for alternative choices of c_2 and c_3 . Figure 2 plots expected utility using (c_2, c_3) equal $(5, 5)$, $(5, 10)$ and $(10, 0)$.

5 DISCUSSION

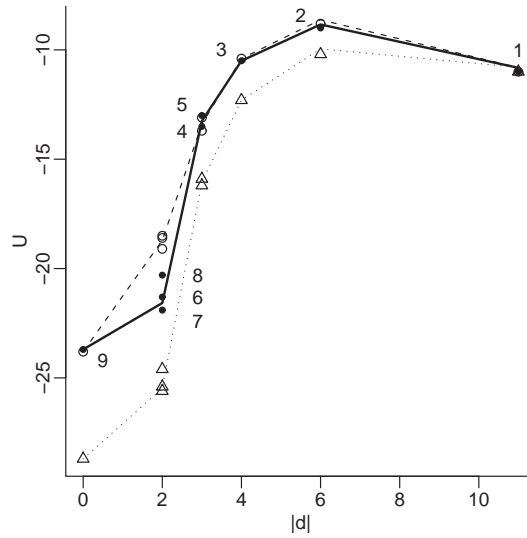
We have proposed an approach to optimal design for repeated binary measurements. The strengths of the proposed method are the flexibility of the underlying semi-parametric probability model, and the generality of the design criterion. The main limitations of the proposed approach are the assumption of equally spaced data, the numerical evaluation of the expected utility function, and the need to define trade-off parameters for the competing goals of minimizing sampling cost, predicting missing observations and optimizing inference about treatment effects, or other relevant summaries of the probability model. The constraint to equally spaced data stems from the underlying probability model. The Markov chain model would be inappropriate for unequally spaced data.

The numerical evaluation of the expected utility function is a limitation and an opportunity at the same time. On one hand, MCMC simulation is required to com-



(a) Prediction term $R(d)$

(b) Treatment term $T(d)$



(c) Total utility $U(d)$

Figure 1: Expected utilities. Panels (a) and (b) show the individual terms of the expected utility function. Panel (c) shows the expected utility using weights $c_2 = -10$ and $c_3 = -5$. In all three plots the points are labeled with the design number given in Table 1. The x-axis is the number of observations per patient. The solid lines show expected utilities under the non-parametric model. The dashed lines (with open circles for the points) show inference under parametric model I. The dotted lines (with triangles for the points) show expected utilities under parametric model II.

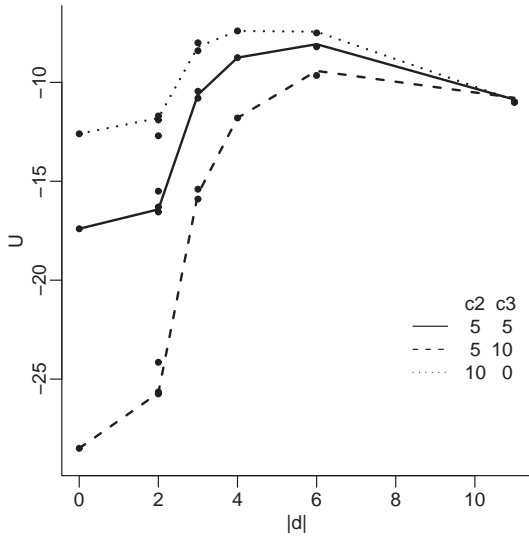


Figure 2: Expected utilities under alternative weights c_2 and c_3 . Separately computing the terms $C_1(d)$, $T(d)$ and $R(d)$ of the expected utility function facilitates easy exploration of alternative weights to trade off the competing goals.

pute expected utilities. On the other hand, the simulation based approach allows to substitute essentially arbitrary utility functions $u(d, \theta, y)$ and thus provides great flexibility in specifying the design criterion. For instance, the inclusion of more than two treatment levels can be easily accommodated within the general framework discussed. To do so, we can simply add as many fixed effect coefficients as needed, and modify $T(d)$ in (2) to any desired specification that reflects learning for the new multi-treatment scenario.

APPENDIX: POSTERIOR PREDICTIVE SIMULATION

We describe details of the posterior predictive simulation (9). Simulating from $p(y | \theta)$ is a straightforward generation of order k Markov chains with transition probabilities defined in (6). Recall that $\theta_o = (\alpha_i, i = 1, \dots, N_o)$ are the random effects specific to the first N_o patients. Generating samples from $p(\alpha_i, i = N_o + 1, \dots, N | \theta_o)$ is accomplished by exploiting the representation of the marginal prior on $\alpha_i, i =$

$1, \dots, N$, as a Polya urn scheme (Blackwell and MacQueen, 1973). Given θ_o , α_{N_o+1} is either equal to one of the components of θ_o with probability proportional to the corresponding cluster size, or a draw from G^o with probability proportional to M . The remaining $(\alpha_i, i = N_0 + 2, \dots, N)$ are drawn similarly, after upgrading the configurations and cluster sizes according to the result of the previous draw. See Blackwell and MacQueen (1973) for a general description of this type of urn schemes.

Finally, generating from $p(\theta_o, k, \beta | y_o)$ is posterior simulation for the earlier study y_o . It is implemented by standard Markov chain Monte Carlo posterior simulation for DP mixture models. See, for example MacEachern and Müller (2000) for details. Only one move is different, namely changing k . In our implementation we proceed as follows. The discrete nature of the random measure G implies the possibility of ties among the imputed α_i . Let $\{\alpha_1^*, \dots, \alpha_{N^*}^*\}$ denote the $N^* \leq N_o$ distinct values among the $\alpha_i, i = 1, \dots, N_o$. We introduce configuration indicators s_i with $s_i = j$ if $\alpha_i = \alpha_j^*$. The use of such configuration indicators is a standard tool in posterior simulation for DP mixture models. Consider now a Metropolis-Hastings (MH) proposal for a move from k to a proposed new value \tilde{k} drawn from a uniform distribution over $\{k - 1, k, k + 1\} \cap \{1, \dots, K\}$. Two cases arise. When $\tilde{k} = k$ the proposal will have exactly the same number of nonzero auto-logistic regression coefficients in (6). The proposal is then generated as a series of draws from all the full conditionals, including the configuration indicators $s = (s_1, \dots, s_n)$. In practice, this is the same as one iteration of the usual Gibbs sampler for DP mixture models, as discussed in, e.g. MacEachern and Müller (2000). This also implies great simplifications when computing the MH acceptance ratio. When $\tilde{k} \neq k$ the set of nonzero coefficients in (6) is different in the proposal and in the currently imputed θ . We define a proposal for the next state of the Markov chain as follows. We leave the configurations s unchanged, and define proposals $\tilde{\alpha}_j^*$ by considering the conditional posterior

$$p(\alpha_j^* | s, y_o, \beta, k) \propto G^o(\alpha_j^*) \prod p(y_i | \alpha_j^*, \beta, k), \quad (10)$$

where $p(y_i | \alpha_j^*, \beta, k)$ is given in (8). The product goes over the set of indices $\{i : i \leq N_o \text{ and } s_i = j\}$. We generate $\tilde{\alpha}_j^*$ from a multivariate normal approximation of (10).

Similarly, $\tilde{\beta}$ is generated from a normal approximation to the conditional posterior for β given $(\tilde{\alpha}_1^*, \dots, \tilde{\alpha}_{N^*}^*)$. The joint proposal $(\tilde{k}, \tilde{\beta}, \tilde{\alpha}_1, \dots, \tilde{\alpha}_{N^*})$ is then accepted with the appropriate MH acceptance probability. See additional details in Quintana and Müller (2003).

REFERENCES

- Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag.
- Bernardo, J. M., et al., eds. (1992), *Bayesian Statistics 4*, Oxford: Oxford University Press.
- Berry, D. (1993), “A case for Bayesianism in clinical trials (with discussion),” *Statistics in Medicine*, 12, 1377–1404.
- Blackwell, D. and MacQueen, J. B. (1973), “Ferguson distributions via Pólya urn schemes,” *The Annals of Statistics*, 1, 353–355.
- Chaloner, K. and Verdinelli, I. (1995), “Bayesian experimental design: a review,” *Statistical Science*, 10, 273–304.
- Clayton, M. K., et al. (1999), “Bayesian Sequential Design of a Network of Sensors,” Technical Report 1013, Department of Statistics, University of Wisconsin–Madison.
- Davis, C. S. and Wei, L. J. (1988), “Nonparametric methods for analyzing incomplete nondecreasing repeated measurements,” *Biometrics*, 44, 1005–1018.
- DeGroot, M. (1970), *Optimal Statistical Decisions*, New York: McGraw Hill.
- Ferguson, T. S. (1973), “A Bayesian analysis of some nonparametric problems,” *The Annals of Statistics*, 1, 209–230.
- Gelman, A. and Rubin, D. (1992), “Inference from iterative simulation using multiple sequences,” *Statistical Science*, 7, 457–511.

- Geweke, J. (1992), “Evaluating the accuracy of sampling-based approaches to calculating posterior moments,” in Bernardo et al. (1992).
- Heidelberger, P. and Welch, P. (1983), “Simulation run length control in the presence of an initial transient,” *Operations Research*, 31, 1109–1144.
- MacEachern, S. N. and Müller, P. (2000), “Efficient MCMC Schemes for Robust Model Extensions using Encompassing Dirichlet Process Mixture Models,” in *Robust Bayesian Analysis*, eds. F. Ruggeri and D. Ríos-Insua, 295–316, New York: Springer-Verlag.
- Quintana, F. A. and Müller, P. (2003), “Nonparametric Bayesian Assessment of the Order of Dependence for Binary Sequences,” *Journal of Computational and Graphical Statistics*, in press.
- Quintana, F. A. and Newton, M. A. (1998), “Assessing the Order of Dependence for Partially Exchangeable Binary Data,” *Journal of the American Statistical Association*, 93, 194–202.
- Smith, B. J. (2000), “Bayesian Output Analysis Program (BOA), Version 0.5.0 for S-PLUS and R.” Available at <http://www.public-health.uiowa.edu/BOA>.
- Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1994), “Bayesian Approaches to Randomized Trials,” *Journal of the Royal Statistical Society, Series A, General*, 157, 357–387.
- Stroud, J., Müller, P., and Rosner, G. (2001), “Optimal Sampling Times for Population Pharmacokinetic Studies,” *Applied Statistics*, 15, 345–359.
- Thall, P. and Russell, K. (1998), “A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials,” *Biometrics*, 54, 251–264.

- Thall, P., Simon, R., and Estey, E. (1995), “Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes,” *Statistics in Medicine*, 14, 357–379.
- Verdinelli, I. (1992), “Advances in Bayesian Experimental Designs,” in Bernardo et al. (1992), 467–482.
- Vlachos, P. and Gelfand, A. (1998), “Nonparametric Bayesian group sequential design,” in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Müller, and D. Sinha, 115–132, New York: Springer-Verlag.
- Wakefield, J. (1994), “An expected loss approach to the design of dosage regimes via sapling-based methods,” *The Statistician*, 43, 13–29.
- Whitehead, J. and Brunier, H. (1995), “Bayesian decision procedures for dose determining experiments,” *Statistics in Medicine*, 14, 885–893.
- Whitehead, J. and Williamson, D. (1998), “Bayesian decision procedures based on logistic regression models for dose-finding studies determining experiments,” *Journal of Biopharmaceutical Statistics*, 8, 445–467.
- Zeger, S. and Karim, M. (1991), “Generalized linear models with random effects,” *Journal of the American Statistical Association*, 86, 79–86.