Gibbs optimal design of experiments

Antony M. Overstall
School of Mathematical Sciences, University of Southampton,
Southampton SO17 1BJ, United Kingdom,
A.M.Overstall@soton.ac.uk
and
Jacinta Holloway-Brown
School of Computer and Mathematical Sciences, University of Adelaide,
Adelaide 5005, Australia,
jacinta.holloway-brown@adelaide.edu.au
and
James M. McGree
School of Mathematical Sciences, Queensland University of Technology,
Brisbane 4001, Australia,
james.mcgree@qut.edu.au

January 2, 2025

Abstract

Bayesian optimal design is a well-established approach to planning experiments. A distribution for the responses, i.e. a statistical model, is assumed which is dependent on unknown parameters. A utility function is then specified giving gain in information in estimating the true values of the parameters, using the Bayesian posterior distribution. A Bayesian optimal design is given by maximising expectation of the utility with respect to the distribution implied by statistical model and prior distribution for the true parameter values. The approach accounts for the experimental aim, via specification of the utility, and of assumed sources of uncertainty. However, it is predicated on the statistical model being correct. Recently, a new type of statistical inference, known as Gibbs inference, has been proposed. This is Bayesian-like, i.e. uncertainty for unknown quantities is represented by a posterior distribution, but does not necessarily require specification of a statistical model. The resulting inference is less sensitive to misspecification of the statistical model. This paper introduces Gibbs optimal design: a framework for optimal design of experiments under Gibbs inference. A computational approach to find designs in practice is outlined and the framework is demonstrated on exemplars including linear models, and experiments with count and time-to-event responses.

Keywords: Gibbs statistical inference; loss function, robust design of experiments; utility function

1 Introduction

Experiments are key to the scientific method and are a pillar of statistical science (e.g. Stigler, 2016, Chapter 6). They are used to systematically investigate the relationship between a series of controllable variables and a response variable. In this paper, an experiment consists of a fixed number of runs, where each run involves the specification of all controllable variables and the subsequent observation of a response. The experimental aim is then addressed, usually by the estimation of a statistical model, or models. The quality of this analysis can be highly dependent on the design of the experiment: the specification of the controllable variables for each run (see, e.g., Mead et al., 2012).

Decision-theoretic Bayesian optimal design of experiments (Chaloner and Verdinelli, 1995, termed here as *Bayesian optimal design* for brevity) provides a principled approach to planning experiments. The Bayesian optimal design approach can briefly be described as follows. First, a statistical model is assumed. This is a probability distribution for the responses, fully specified apart from a vector of unknown parameters. A prior distribution is also assumed representing prior knowledge about the true values of the parameters, i.e. those parameter values that make the statistical model coincide with the true probability distribution for the responses. A utility function is then specified, returning the gain in information, in estimating the true values of the parameters, from the responses obtained via a given design. The expected utility is given by the expectation of the utility with respect to the responses and true parameter values, under the joint probability distribution implied by the statistical model and prior distribution. Lastly, a Bayesian optimal design maximises the expected utility over the space of all designs.

The advantages of the Bayesian optimal design approach include the following. Firstly, the approach explicitly takes account of the aim of the experiment through specification of the utility function. For example, the approach can easily be extended for model selection and/or prediction. Second, by taking the expectation of the utility, it incorporates all known sources of uncertainty.

The patent disadvantage is that the objective function given by the expected utility is rarely available in closed form. Instead, an often computationally expensive numerical routine is used to approximate the expected utility which then needs to be maximised over a potentially high-dimensional design space. However, in the last decade, new computational methodology has been proposed to find Bayesian designs more efficiently than previously possible, or find designs for scenarios which were previously unattainable (see, e.g., Rainforth et al., 2023, and reference therein).

A more fundamental disadvantage is that the statistical model for the responses is specified before the responses are observed. The resulting design can be highly tailored to a statistical model which may actually be significantly misspecified, compromising the accuracy and precision of future statistical inference.

Recently, there has been progress in Bayesian-like statistical inference that does not require the specification of a probability distribution, i.e. a statistical model. Instead a loss function is specified identifying desirable parameter values for given responses (those that minimise the loss). The specification of the loss does not necessarily follow from specification of a probability distribution for the responses. A so-called (unnormalised) Gibbs posterior distribution for the parameters is then given by the product of the exponential of the negative loss and a prior distribution for target parameter values. The target parameter values are the parameters which minimise the expectation of the loss with respect the true (but unknown and unspecified) probability distribution of the responses (Bissiri et al., 2016). Gibbs inference (inference using the Gibbs posterior distribution) can be seen as a Bayesian-like analogue of classical M-estimation (e.g., Hayashi, 2000, Chapter 7). Bissiri et al. (2016) provide a thorough theoretical treatment of Gibbs inference. In particular, they show that the Gibbs posterior distribution provides coherent inference about the target parameter values.

This paper proposes a *decision-theoretic Gibbs optimal design of experiments* framework (referred to as *Gibbs optimal design* for brevity). Similar to Bayesian optimal design, a utility function is specified: a function

returning gain in information, in estimating the target parameter values, using responses obtained via a given design, where dependence on the responses is through the Gibbs posterior distribution. However the expected utility does not immediately follow. The absence of a statistical model, means there is no joint probability distribution for the responses and target parameter values. Instead, we propose taking expectation of the utility function with respect to a probability distribution for the responses which we term the designer distribution. This user-specified designer distribution should be flexible enough to be close to the true probability distribution for the responses, certainly closer than the statistical model. Specification of the designer distribution can be aided by the fact that it does not need to be "useful", i.e. it will not be fitted to the observed responses of the experiment and need not be capable of addressing the experimental aim.

To demonstrate the concept of the designer distribution, consider the following simple example. Suppose the aim of the experiment is to investigate the relationship between a single controllable variable, x, and a continuous response. It is assumed that the mean response is given by a polynomial of x where the coefficients are unknown parameters. A reasonable loss is sum of squares: the sum of the squared differences between the responses and their mean (given by the polynomial). A suitable designer distribution is the unique-treatment model (e.g. Gilmour and Trinca, 2012): a probability distribution where each unique x value in the experiment exhibits a unique mean. The unique-treatment model is not a useful model, i.e. it does not allow one to effectively learn the relationship between x and the response. However, it is a flexible model likely to be closer to capturing the true relationship between x and the response, than a polynomial.

The topic considered in this paper fits within the field of *robust design of experiments*. One of the first treatments of this topic was the work of Box and Draper (1959) who considered design of experiments for a statistical model with mean response a linear model with an intercept and first-order term, for a single controllable variable, when the true probability distribution has a mean response with an additional quadratic term. A relatively up-to-date review of robust design of experiments is given by Wiens (2015) who considers the topic under misspecification of both the mean and error structure of the statistical model, as well as from the point of view of different applications (including dose-response, clinical trials and computer experiments).

The remainder of the paper is organised as follows. Section 2 provides brief mathematical descriptions of Bayesian optimal design and Gibbs inference, before Section 3 introduces the concept of Gibbs optimal design. Section 4 considers Gibbs optimal design for the class of linear models. Finally, we demonstrate Gibbs optimal design on illustrative examples in Section 5.

2 Background

2.1 Design problem setup

Suppose the experiment aims to investigate the relationship between k controllable variables denoted $\mathbf{x} = (x_1, \ldots, x_k)^{\mathrm{T}} \in \mathbb{X}$, and a measurable response denoted by y. The experiment consists of n runs, where the ith run involves specifying $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^{\mathrm{T}} \in \mathbb{X}$ and observing response y_i, for $i = 1, \ldots, n$. Let $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ denote the $n \times 1$ vector of responses and X the $n \times k$ design matrix with ith row $\mathbf{x}_i^{\mathrm{T}}$, for $i = 1, \ldots, n$. It is assumed that $\mathbf{y}$ are realisations from a multivariate probability distribution denoted $\mathcal{T}(X)$, which is the true but unknown response-generating probability distribution. This paper addresses the specification of X to learn the most about the true response-generating probability distribution.

2.2 Bayesian optimal design

Bayesian optimal design relies on the specification of a statistical model and utility function. The statistical model, denoted $\mathcal{S}(\mathbf{t}; X)$, is a probability distribution for the responses $\mathbf{y}$ used to represent the true response-generating probability distribution, $\mathcal{T}(X)$. The statistical model is fully specified up to a $p \times 1$ vector of parameters $\mathbf{t} = (t_1, \ldots, t_p)^{\mathrm{T}} \in \Theta$ with parameter space $\Theta \subset \mathbb{R}^p$.

Suppose there exist values $\boldsymbol{\theta}_T$ of the parameters such that the statistical model and true response-generating probability distribution coincide, i.e. $\mathcal{T}(X) = \mathcal{S}(\boldsymbol{\theta}_T; X)$. If $\boldsymbol{\theta}_T$ is known, then the true response-generating distribution is completely known so, in some form, estimating $\boldsymbol{\theta}_T$ is the aim of the experiment. Bayesian inference for this aim is achieved by evaluating the Bayesian posterior distribution, denoted $\mathcal{B}(\mathbf{y}; X)$ and given by

$$\pi_{\mathcal{B}}(\mathbf{t}|\mathbf{y}; X) \propto \pi(\mathbf{y}|\mathbf{t}; X)\pi_T(\mathbf{t}). \tag{1}$$

In (1), $\pi(\mathbf{y}|\mathbf{t}; X)$ is the likelihood function: the probability density/mass function of $\mathcal{S}(\mathbf{t}; X)$, and $\pi_T(\cdot)$ is the probability density function (pdf) for the prior distribution of $\boldsymbol{\theta}_T$ (denoted $\mathcal{P}_T$). For simplicity, we assume that the prior distribution $\mathcal{P}_T$ does not depend on the design X. However, the technicalities of Bayesian optimal design change little if $\mathcal{P}_T$ depends on X.

The utility function is denoted $u_{\mathcal{B}}(\mathbf{t}, \mathbf{y}, X)$. It gives the gain in information of estimating parameter values $\mathbf{t}$ using the Bayesian posterior distribution conditional on responses $\mathbf{y}$ collected via design X. The Bayesian expected utility is

$$U_{\mathcal{B}}(X) = \mathrm{E}_{\mathcal{P}_T}\left\{\mathrm{E}_{\mathcal{S}(\boldsymbol{\theta}_T, X)}\left[u_{\mathcal{B}}(\boldsymbol{\theta}_T, \mathbf{y}, X)\right]\right\}. \tag{2}$$

In (2), the utility is evaluated at the true parameter values $\boldsymbol{\theta}_T$, indicating that the experimental aim is to estimate these values. The inner expectation is with respect to the responses $\mathbf{y}$ under $\mathcal{S}(\boldsymbol{\theta}_T; X)$, and the outer expectation with respect to the true parameter values $\boldsymbol{\theta}_T$, under the prior distribution, $\mathcal{P}_{\mathcal{T}}$.

A Bayesian optimal design is given by maximising $U_{\mathcal{B}}(X)$ over the design space $\mathbb{D} = \mathbb{X}^n$.

Two commonly-used utility functions are negative squared error (NSE; see, e.g., Robert 2007, Section 2.5.1) and Shannon information (SH; see, e.g., Chaloner and Verdinelli 1995). The NSE utility is

$$u_{\mathcal{B},NSE}(\mathbf{t}, \mathbf{y}, X) = -\left\|\mathbf{t} - \mathrm{E}_{\mathcal{B}(\mathbf{y}; X)}\left(\boldsymbol{\theta}_T\right)\right\|_{I_p},$$

where $\mathrm{E}_{\mathcal{B}(\mathbf{y}; X)}\left(\boldsymbol{\theta}_T\right)$ is the Bayesian posterior mean of $\boldsymbol{\theta}_T$ and $\|\mathbf{u}\|_V = \mathbf{u}^{\mathrm{T}} V \mathbf{u}$ for a vector $\mathbf{u}$ and a matrix V of appropriate dimensions. The SH utility is

$$u_{\mathcal{B},SH}(\mathbf{t}, \mathbf{y}, X) = \log \pi_{\mathcal{B}}(\mathbf{t}|\mathbf{y}, X).$$

The Bayesian expected NSE and SH utilities are equal to the negative expected trace of variance matrix and entropy, respectively, of the Bayesian posterior distribution of $\boldsymbol{\theta}_T$, where expectation is with respect to the marginal distribution of $\mathbf{y}$, under the statistical model and prior distribution $\mathcal{P}_T$. Thus maximising the Bayesian expected NSE and SH utilities is equivalent to minimising expected uncertainty in the posterior distribution of $\boldsymbol{\theta}_T$, as measured by variance and entropy, respectively.

Notwithstanding the conceptual advantages of Bayesian optimal design listed in Section 1, the disadvantage of finding Bayesian optimal designs in practice is clear. Firstly, the utility depends on $\mathbf{y}$ through the Bayesian posterior distribution. For most cases, this distribution is not available in closed form meaning the utility will not be available in closed form. Second, the integration required to evaluate the Bayesian expected utility given by (2) will also not be analytically tractable. Although it is straightforward to derive a numerical approximation to the Bayesian expected utility $U_{\mathcal{B}}(X)$, this approximation then needs to be maximised over the nk-dimensional design space $\mathbb{D}$. The dimensionality of $\mathbb{D}$ can be high if the number of runs n and/or number of controllable variables k are large. This limits the effectiveness of commonly

used numerical approximations such as, for example, Monte Carlo integration. However, in the last decade significant progress has been in the development of new computational methodology to find Bayesian optimal designs for a range of problems. See, for example, the review papers of Ryan et al. (2016) and Rainforth et al. (2023), and references therein.

A more fundamental disadvantage is that Bayesian optimal design assumes that the statistical model is correct, i.e. there exist $\boldsymbol{\theta}_T$ such that $\mathcal{T}(X) = \mathcal{S}(\boldsymbol{\theta}_T; X)$. This can be observed in the expected utility in (2), where the utility is evaluated at $\boldsymbol{\theta}_T$, and expectation is taken with respect to the prior distribution of $\boldsymbol{\theta}_T$. When the statistical model is incorrect, there does not exist $\boldsymbol{\theta}_T$ such that $\mathcal{T}(X) = \mathcal{S}(\boldsymbol{\theta}_T; X)$. This makes evaluating the utility at $\boldsymbol{\theta}_T$ in (2) nonsensical. However, it is well known (see, e.g. Berk, 1966; Walker, 2013; Bissiri et al., 2016) that when the model is incorrect, the Bayesian posterior provides coherent inference about the parameter values, $\boldsymbol{\theta}_{SI}$, that minimise the Kullback-Liebler divergence between $\mathcal{T}(X)$ and $\mathcal{S}(\cdot; X)$. This raises the possibility of, in (2), replacing evaluation of the utility at $\boldsymbol{\theta}_T$ by evaluation at $\boldsymbol{\theta}_{SI}$. The utility would then be representing gain in information on $\boldsymbol{\theta}_{SI}$, rather than the non-existent $\boldsymbol{\theta}_T$. In this paper, we go further and consider a more general framework for optimal design revolving around Gibbs inference (of which Bayesian inference is a special case).

2.3 Gibbs inference

Gibbs inference (also known as general Bayesian inference) is an approach to statistical inference which is Bayesian-like, i.e. information is summarised by a probability distribution for unknown quantities, but does not necessarily rely on specification of a statistical model. Instead a loss function, denoted $\ell(\mathbf{t}; \mathbf{y}, X)$ is specified. This function identifies desirable parameter values for observed responses $\mathbf{y}$. Notably, the corresponding classical M-estimators (see, e.g., Hayashi, 2000, Chapter 7) are defined

$$\hat{\boldsymbol{\theta}}_\ell(\mathbf{y}; X) = \arg\min_{\mathbf{t}\in\Theta} \ell(\mathbf{t}; \mathbf{y}, X).$$

The Gibbs posterior distribution, denoted $\mathcal{G}_\ell(\mathbf{y}; X)$, is given by

$$\pi_\ell\left(\mathbf{t}|\mathbf{y}, X\right) \propto \exp\left[-w\ell(\mathbf{t}; \mathbf{y}, X)\right] \pi_\ell(\mathbf{t}), \tag{3}$$

where $\pi_\ell(\cdot)$ is the pdf of the prior distribution (denoted $\mathcal{P}_\ell$) for the target parameter values (discussed below) and $w > 0$ is a calibration weight (also discussed below). In (3), the quantity $\exp\left[-w\ell(\mathbf{t}; \mathbf{y}, X)\right]$ is known as the generalised likelihood. Similar to Bayesian inference, we assume that the prior distribution $\mathcal{P}_\ell$ does not depend on the design X.

Bissiri et al. (2016) showed that the Gibbs posterior distribution provides coherent inference about the target parameter values defined as

$$\boldsymbol{\theta}_\ell = \arg\min_{\mathbf{t}\in\Theta} L_{\mathcal{T}(X)}(\mathbf{t}),$$

where $L_{\mathcal{T}(X)}(\mathbf{t}) = \mathrm{E}_{\mathcal{T}(X)}\left[w\ell(\mathbf{t}; \mathbf{y}, X)\right]$ is the expected loss with respect to the responses $\mathbf{y}$ under the true response-generating distribution, $\mathcal{T}(X)$.

Suppose the responses and design can be partitioned as $\mathbf{y}^{\mathrm{T}} = (\mathbf{y}_1^{\mathrm{T}}, \mathbf{y}_2^{\mathrm{T}})^{\mathrm{T}}$ and $X = (X_1^{\mathrm{T}}, X_2^{\mathrm{T}})^{\mathrm{T}}$. Coherent inference, in this case, means that the Gibbs posterior formed from the complete data, i.e. by using generalised likelihood $\exp\left[-w\ell(\mathbf{t}; \mathbf{y}, X)\right]$ with prior $\mathcal{P}_\ell$ is identical to the Gibbs posterior from first evaluating the Gibbs posterior $\mathcal{G}_\ell(\mathbf{y}_1; X_1)$ and then using this as a prior distribution with generalised likelihood $\exp\left[-w\ell(\mathbf{t}; \mathbf{y}_2, X_2)\right]$.

The calibration weight, w, controls the rate of learning from prior, $\mathcal{P}_\ell$, to Gibbs posterior, $\mathcal{G}_\ell(\mathbf{y}; X)$. To see this, consider the extreme cases. As $w \to 0$, then $\mathcal{G}_\ell(\mathbf{y}; X) \to \mathcal{P}_\ell$, and, as $w \to \infty$, then $\mathcal{G}_\ell(\mathbf{y}; X)$ converges to a point mass at $\hat{\boldsymbol{\theta}}_\ell(\mathbf{y}; X)$. Therefore, the specification of the calibration weight is crucial for Gibbs inference.

Developing methodology for the automatic specification of the calibration weight is an active area of research (see, e.g. Catoni, 2007; Grünwald, 2012; Bissiri et al., 2016; Holmes and Walker, 2017; Lyddon et al., 2019; Jewson and Rossell, 2022). For example, Bissiri et al. (2016) described various approaches to the specification of the calibration weight. These include (a) choosing w so that the prior distribution $\mathcal{P}_\ell$ contributes a fraction (e.g. $1/n$) of the information in the generalised likelihood (b) choosing w to maintain operational characteristics, i.e. so that properties of the Gibbs posterior match those of the classical M-estimators; or (c) allowing w to be unknown with a hierarchical prior distribution. This paper does not advocate for any singular approach for specifying the calibration weight. The Gibbs optimal design framework, as introduced in the next section, does not rely on any particular calibration method and can be used, in principle, with any method.

In summary, the prior distribution, $\mathcal{P}_\ell$, for the target parameter values, $\boldsymbol{\theta}_\ell$, is updated to the Gibbs posterior by using information in the generalised likelihood $\exp\left[-w\ell(\mathbf{t};\mathbf{y},X)\right]$. This is analogous to how the prior on the true parameter values is updated to the Bayesian posterior by using information in the likelihood. Note that the prior distribution for Gibbs inference can be non-informative reflecting a lack of prior knowledge about $\boldsymbol{\theta}_\ell$ in a similar way to how we can specify a non-informative prior on the true parameter values.

2.4 Bayesian inference as Gibbs inference

The Bayesian posterior distribution as a Gibbs posterior can be obtained under the self-information loss, i.e. the negative log-likelihood

$$\ell_{SI}\left(\mathbf{t};\mathbf{y},X\right) = -\log\pi\left(\mathbf{y}|\mathbf{t};X\right),$$

with $w = 1$, i.e. $\mathcal{G}_{SI}(\mathbf{y};X) = \mathcal{B}(\mathbf{y};X)$. In this case, the M-estimators, $\hat{\boldsymbol{\theta}}_{SI}(\mathbf{y};X)$, are the maximum likelihood estimators. The corresponding expected loss is

$$L_{SI,\mathcal{T}(X)}(\mathbf{t}) = \mathrm{E}_{\mathcal{T}(X)}\left[-\log\pi\left(\mathbf{y}|\mathbf{t};X\right)\right],$$

which, up to a constant independent of $\mathbf{t}$, is the Kullback-Leibler (KL) divergence between the true response-generating distribution and the statistical model. Thus the target parameter values, denoted $\boldsymbol{\theta}_{SI}$, minimise this divergence. Under a correctly specified statistical model, where there exist $\boldsymbol{\theta}_T$ such that $\mathcal{S}(\boldsymbol{\theta}_T;X) = \mathcal{T}(X)$, then $\boldsymbol{\theta}_{SI} = \boldsymbol{\theta}_T$ with a minimised Kullback-Leibler divergence of zero. Now consider the case of a misspecified statistical model, where there do not exist parameter values $\boldsymbol{\theta}_T$ such that $\mathcal{S}(\boldsymbol{\theta}_T;X) = \mathcal{T}(X)$. In this case, Bayesian inference still provides coherent inference about $\boldsymbol{\theta}_{SI}$.

3 Gibbs optimal design of experiments

3.1 Expected utility under the design distribution

The proposal is to extend the Bayesian optimal design framework to Gibbs statistical inference. Similar to Bayesian optimal design, the Gibbs optimal design framework will involve maximising an expected utility function over the design space $\mathbb{D}$. The challenge is that, with Bayesian optimal design, the expectation of the utility is taken with respect to the responses and true parameter values, $\boldsymbol{\theta}_T$, under a joint distribution given by the statistical model and prior distribution for $\boldsymbol{\theta}_T$. Neither, a statistical model nor true parameter values necessarily exist under a Gibbs inference analysis.

We propose that the utility function be evaluated at the target parameter values. As defined in Section 2.3, the target parameter values depend on the true response-generating distribution. However, this is unknown,

so instead we replace these by the target parameter values under a distribution we term the *designer distribution*. Furthermore, the expectation of the utility function is then taken under this designer distribution to give the objective function.

The purpose of the designer distribution is to represent the unknown true response-generating distribution $\mathcal{T}(X)$, in a similar way to the statistical model $S(\mathbf{t}, X)$. This prompts the question: why not use the designer distribution as the statistical model? However, note that, post-experiment, the designer distribution will not be used as a statistical model for the observed responses. This means that it does not need to be "useful", i.e. it does not need to be able to attain the aims of the experiment. Consider the simple example in Section 1. A reasonable designer distribution is the unique-treatment model. Every unique vector of controllable variables (i.e. treatment) is allowed a unique mean response with no relationship between the mean response at different controllable variables. This should provide a flexible representation of the true response-generating distribution but would be ineffective in learning the relationship between the controllable variables and response. We derive expected utilities under this designer distribution in Section 4.

Formally, let $\mathcal{D}(\mathbf{r}; X)$ denote the designer distribution, which depends on hyper-variables $\mathbf{r}$. The idea is that we may formulate more than one approximation to $\mathcal{T}(X)$, indexed by different hyper-variables. A hyper-variable distribution, denoted $\mathcal{C}$, is specified for $\mathbf{r}$. Similar to the prior distributions $\mathcal{P}_T$ and $\mathcal{P}_\ell$, the hyper-variable distribution is assumed not to depend on design X.

The target parameter values under the designer distribution are

$$\boldsymbol{\theta}_{\ell,\mathcal{D}}(\mathbf{r}; X) = \arg\min_{\mathbf{t}\in\Theta} \mathrm{E}_{\mathcal{D}(\mathbf{r};X)} \left[w\ell(\mathbf{t}; \mathbf{y}, X) \right]$$

and are a function of the hyper-variables $\mathbf{r}$ and design X.

The utility function is denoted $u_{\mathcal{G}}(\mathbf{t}, \mathbf{y}, X)$. This can have the same functional form as a utility used under Bayesian optimal design. The difference is dependence on responses $\mathbf{y}$ and design X is via the Gibbs posterior (instead of the Bayesian posterior).

The Gibbs expected utility objective function is then given by

$$U_{\mathcal{G}}(X) = \mathrm{E}_{\mathcal{C}} \left\{ \mathrm{E}_{\mathcal{D}(\mathbf{r};X)} \left[u(\boldsymbol{\theta}_{\ell,\mathcal{D}}(\mathbf{r}; X), \mathbf{y}, X) \right] \right\}. \tag{4}$$

The inner expectation is with respect to the responses, $\mathbf{y}$, under the designer distribution, $\mathcal{D}(\mathbf{r}; X)$ and the outer expectation is with respect to the hyper-variables, $\mathbf{r}$, under $\mathcal{C}$. Note that the utility is evaluated at the target parameter values $\boldsymbol{\theta}_{\ell,D}(\mathbf{r}; X)$ rather than the true parameter values $\boldsymbol{\theta}_T$. This reflects the fact that the experimental aim is to estimate the target parameter values.

Bayesian optimal design is a special case of Gibbs optimal design, under the self-information loss and the designer distribution being the statistical model. This leads to the target parameters and hyper-variables being the true parameter values. Finally $\mathcal{C}$ is chosen as the prior distribution for the true parameter values. However, this tells us that the use of Bayesian optimal design implicitly assumes that the statistical model is true, i.e. a very strong assumption.

In the literature, we are aware of two modifications of Bayesian optimal design which are also special cases of Gibbs optimal design. Firstly, Etzioni and Kadane (1993) considers the case where the outer expectation in the Bayesian expected utility (2) is taken under a different prior distribution to that used to form the Bayesian posterior. This can represent the scenario where two separate individuals are (a) designing the experiment and (b) analysing the observed responses, and who may have differing prior beliefs on the true parameter values. In this case, the loss is self-information, and the designer distribution coincides with the statistical model. However, $\mathcal{P}_\ell$ is the prior distribution representing the prior beliefs of the individual analysing the observed responses, and $\mathcal{C}$ is the prior distribution representing the prior beliefs of the individual designing the experiment.

Overstall and McGree (2022) considered a Bayesian optimal design framework whereby the inner expectation in the Bayesian expected utility (2), is taken under an alternative model for the responses. The motivation for doing so was to introduce robustness into the Bayesian optimal design process, for example, by making the alternative model more complex than the statistical model. The proposal in this paper is to extend this further to where the inference does not have to be Bayesian.

3.2 Computational approach

In general, it will not be possible to find closed form expressions for the Gibbs expected utility, given by (4). This is because the Gibbs posterior distribution and target parameter values will not be available in closed form. This matches how the Bayesian expected utility, given by (2), is also typically not available in closed form.

In this section, we outline an exemplar computational approach that can be used to approximately find Gibbs optimal designs. It essentially uses a combination of a normal approximation to approximate the Gibbs posterior and a Monte Carlo integration to approximate the Gibbs expected utility. The approximate Gibbs expected utility is then maximised using the approximate coordinate exchange (ACE; Overstall and Woods 2017) algorithm. In the literature, the combination of Monte Carlo and normal approximations has been used to find Bayesian designs (Long et al., 2013), including in partnership with ACE (Overstall et al., 2018). It is anticipated that many existing computational approaches used to find Bayesian optimal designs can be repurposed to find Gibbs optimal designs (see Section 6 for more discussion).

3.2.1 Approximating the expected utility

Let $\tilde{\boldsymbol{\theta}}_\ell(\mathbf{y}; X) = \arg\max_{\mathbf{t}\in\Theta} \left[-w\ell(\mathbf{t}; \mathbf{y}, X) + \log \pi_\ell(\mathbf{t})\right]$ denote the Gibbs posterior mode and let

$$\begin{aligned}\tilde{\Sigma}_\ell(\mathbf{y}; X) &= -\left[-w\frac{\partial \ell(\tilde{\boldsymbol{\theta}}_\ell(\mathbf{y}; X); \mathbf{y}, X)}{\partial \mathbf{t}\partial \mathbf{t}^{\mathrm{T}}}\right. \\ &\qquad \left. + \frac{\partial \log \pi_\ell(\tilde{\boldsymbol{\theta}}_\ell(\mathbf{y}; X))}{\partial \mathbf{t}\partial \mathbf{t}^{\mathrm{T}}}\right]^{-1},\end{aligned}$$

denote the negative inverse Hessian matrix of the log Gibbs posterior pdf evaluated at the Gibbs posterior mode. Then the Gibbs posterior distribution is approximated by a normal distribution with mean $\tilde{\boldsymbol{\theta}}_\ell(\mathbf{y}; X)$ and variance $\tilde{\Sigma}_\ell(\mathbf{y}; X)$. This is analogous to normal approximations used for Bayesian posterior distributions (see, e.g. O'Hagan and Forster, 2004, page 237). In the examples in Section 5, we use a quasi-Newton method to numerically evaluate the Gibbs posterior mode.

Now, due to the tractability of the normal distribution, approximations for many utility functions ensue. We denote such approximations by $\tilde{u}(\mathbf{t}, \mathbf{y}, X)$. For example, approximations for the NSE and SH utilities are

$$\begin{aligned}&\tilde{u}_{NSE}(\mathbf{t}, \mathbf{y}, X) = -\|\mathbf{t} - \tilde{\boldsymbol{\theta}}_\ell(\mathbf{y}; X)\|_2^2, \\ &\tilde{u}_{SH}(\mathbf{t}, \mathbf{y}, X) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\tilde{\Sigma}_\ell(\mathbf{y}; X)| \\ &-\frac{1}{2}\left(\mathbf{t} - \tilde{\boldsymbol{\theta}}_\ell(\mathbf{y}; X)\right)^{\mathrm{T}} \tilde{\Sigma}_\ell(\mathbf{y}; X)^{-1} \left(\mathbf{t} - \tilde{\boldsymbol{\theta}}_\ell(\mathbf{y}; X)\right),\end{aligned}$$

respectively.

A Monte Carlo approximation to the expected utility is then given by

$$\hat{U}_G(X) = \frac{1}{B}\sum_{b=1}^{B} \hat{u}(\boldsymbol{\theta}_{\ell,D}(\mathbf{r}_b, X), \mathbf{y}_b, X),$$

where $\{\mathbf{r}_b, \mathbf{y}_b\}_{b=1}^{B}$ is a sample from the joint distribution of responses $\mathbf{y}$ and hyper-variables $\mathbf{r}$ implied by designer distribution $\mathcal{D}(\mathbf{r}, X)$ and $\mathcal{C}$.

For clarity, the following algorithm can be used to form the Monte Carlo approximation to the expected Gibbs utility.

1. Inputs are design, X; loss function, $\ell(\mathbf{t}; \mathbf{y}, X)$; designer distribution, $\mathcal{D}(\mathbf{r}; X)$, hyper-variable distribution, $\mathcal{C}$; and Monte Carlo sample size B.
2. For $b = 1, \ldots, B$, complete the following steps.
   (a) Generate hyper-variables $\mathbf{r}_b \sim \mathcal{C}$.
   (b) Determine target parameter values under designer distribution $\mathcal{D}(\mathbf{r}_b; X)$, i.e.

$$\boldsymbol{\theta}_{\ell,\mathcal{D}}(\mathbf{r}_b; X) = \arg\min_{\mathbf{t}\in\Theta} \mathrm{E}_{\mathcal{D}(\mathbf{r}_b;X)} \left[\ell(\mathbf{t}; \mathbf{y}, X)\right].$$

   (c) Generate responses $\mathbf{y}_b \sim \mathcal{D}(\mathbf{r}_b; X)$.
   (d) Find Gibbs posterior mode

$$\tilde{\boldsymbol{\theta}}(\mathbf{y}_b; X) = \arg\max_{\mathbf{t}\in\Theta} \left[-w\ell(\mathbf{t}; \mathbf{y}_b, X) + \log \pi_\ell(\mathbf{t})\right]$$

   and

$$\tilde{\Sigma}_\ell(\mathbf{y}_b; X) = -\left[-w\frac{\partial \ell(\tilde{\boldsymbol{\theta}}_\ell(\mathbf{y}_b; X); \mathbf{y}_b, X)}{\partial \mathbf{t}\partial \mathbf{t}^{\mathrm{T}}} + \frac{\partial \log \pi_\ell(\tilde{\boldsymbol{\theta}}_\ell(\mathbf{y}_b; X))}{\partial \mathbf{t}\partial \mathbf{t}^{\mathrm{T}}}\right]^{-1}.$$

   (e) Using the normal distribution

$$\mathrm{N}\left[\tilde{\boldsymbol{\theta}}_\ell(\mathbf{y}_b; X), \tilde{\Sigma}_\ell(\mathbf{y}_b; X)\right]$$

   as an approximation to the Gibbs posterior, form the approximation to the utility

$$\tilde{u}_b = \tilde{u}_{\mathcal{G}}\left[\boldsymbol{\theta}_{\ell,\mathcal{D}}(\mathbf{r}_b; X), \mathbf{y}_b, X\right].$$

3. The Monte Carlo approximation to the Gibbs expected utility is

$$\tilde{U}_{\mathcal{G}} = \frac{1}{B}\sum_{b=1}^{B} \tilde{u}_b.$$

In Step 2(b), we are required to find the target parameter values $\boldsymbol{\theta}_{\ell,\mathcal{D}}(\mathbf{r}_b; X)$ under the designer distribution $\mathcal{D}(\mathbf{r}_b; X)$. This is because the utility function in the inner expectation of the expected Gibbs utility (4) is evaluated at the target parameter values. How this is accomplished will depend on the application but will typically require a numerical optimisation. We address this issue for each example in Section 5.

3.2.2 Maximising the approximate expected utility

Unfortunately, the Monte Carlo approximation to the Gibbs expected utility, given by (2) is not straightforward to maximise to yield the Gibbs optimal design. This is because it is computationally expensive to evaluate (one evaluation requires B numerical maximisations to find the Gibbs posterior mode, and possible B further numerical maximisations to find the target parameter values) and it is stochastic. To solve this problem, we use the ACE algorithm. Very briefly, ACE uses a cyclic descent algorithm (commonly called coordinate exchange in the design of experiments literature: see, e.g., Meyer and Nachtsheim 1995) to maximise a Gaussian process prediction of the expected utility, sequentially over each one-dimensional element of the design space. For more details on the ACE algorithm, see Overstall and Woods (2017).

While many methods could be used to maximise the approximate Gibbs expected utility, the ACE algorithm has been chosen due to its compatibility with any Monte Carlo approximation to the expected utility and an implementation being readily availability in software (Overstall et al., 2020).

4 Linear models

4.1 Introduction

In this section we consider Gibbs optimal design for the special case of linear models. The design of experiments for linear models is a significant topic in the literature. For textbook treatments, see, for example, Atkinson et al. (2007), Goos and Jones (2011) and Morris (2011). For specific treatments of robust design of experiments for linear models, see, for example, Cook and Nachtsheim (1982), Li and Nachtsheim (2000), Bingham and Chipman (2007), Tsai and Gilmour (2010) and Smucker and Drew (2015). One difference between the proposed framework and these papers, is that these papers attempt to find designs for a set of models. On the other hand, we acknowledge that the model we are fitting is incorrect and attempt to find a design to allow the fitted model to be as close to the truth as possible. In this way, our work is closer in spirit to, for example, Waite and Woods (2022) and Azriel (2023).

Suppose y is a continuous response and the true data-generating process $\mathcal{T}(X)$ has

$$y_i = \mu(\mathbf{x}_i) + e_i,$$

for $i = 1, \ldots, n$, where $\mathbf{x}_i \in \mathbb{X} = [-1, 1]^k$, and $e_1, \ldots, e_n$ are independent and identically distributed random errors with mean zero and variance $\sigma^2 < \infty$. The true mean response, $\mu(\mathbf{x})$, for controllable variables $\mathbf{x}$ is unknown and is approximated by a linear function $\mathbf{f}(\mathbf{x})^{\mathrm{T}}\mathbf{t}$ where $\mathbf{f} : \mathbb{X} \to \mathbb{R}^p$ is a specified regression function.

Bayesian optimal design proceeds by assuming there exist $\boldsymbol{\theta}_T$ such that $\mathbf{f}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\theta}_T = \mu(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{X}$. Under an improper uniform prior distribution, $\mathcal{P}_T$, for $\boldsymbol{\theta}_T$, the Bayesian expected NSE and SH utilities are proportional to the following two objective functions

$$\begin{aligned} U^*_{\mathcal{B},NSE}(X) &= -\mathrm{tr}\left[(F^{\mathrm{T}}F)^{-1}\right] h_1(d) \\ U^*_{\mathcal{B},SH}(X) &= \log|F^{\mathrm{T}}F|, \end{aligned} \tag{5}$$

respectively, where F is an $n \times p$ matrix with ith column given by $\mathbf{f}(\mathbf{x}_i)^{\mathrm{T}}$, for $i = 1, \ldots, n$. The Bayesian optimal designs found by maximising the objective functions in (5) are also known as A- and D-optimal, respectively. These are commonly considered in the classical design of experiments literature.

We now consider Gibbs optimal design for the linear model for two different loss functions: sum of squares (Section 4.2) and L^2 loss (Section 4.3). In both cases, we specify the designer distribution to be the same

as $\mathcal{T}(X)$ above. For the distribution of $e_1, \ldots, e_n$, we specify a scale mixture of normal distributions, i.e. $e_i|\kappa \sim \mathrm{N}(0, \kappa)$, where κ is a random variable with $\mathrm{E}_{\mathcal{C}}(\kappa) = \sigma^2$. When considering the marginal distribution of e_i, a scale mixture of normal distributions results in a wide range of symmetric, uni-modal distributions (e.g. West, 1987) thus the designer distribution exhibits a large amount of flexibility. The hyper-variables are $\mathbf{r} = \left(\mu(\cdot), \sigma^2, \kappa\right)$.

4.2 Sum of squares loss

Consider the sum of squares (SS) loss function

$$\ell_{SS}(\mathbf{t}; \mathbf{y}, X) = \sum_{i=1}^{n} \left[y_i - \mathbf{f}(\mathbf{x}_i)^{\mathrm{T}}\mathbf{t}\right]^2.$$

Under the SS loss, the M-estimators are the familiar least squares estimators $\hat{\boldsymbol{\theta}}_{SS}(\mathbf{y}; X) = \left(F^{\mathrm{T}}F\right)^{-1} F^{\mathrm{T}}\mathbf{y}$.

We consider two different specifications for the calibration weight, w. In both cases, we specify a uniform prior distribution for the target parameter values, $\boldsymbol{\theta}_{SS}$.

4.2.1 Fixed calibration weight based on maintaining characteristics

Under a fixed calibration weight, the Gibbs posterior distribution for $\boldsymbol{\theta}_{SS}$ is

$$\mathrm{N}\left[\hat{\boldsymbol{\theta}}_{SS}(\mathbf{y}; X), \frac{1}{2w}\left(F^{\mathrm{T}}F\right)^{-1}\right].$$

The calibration weight is chosen so that the Gibbs posterior variance is equal to the variance of $\hat{\boldsymbol{\theta}}_{SS}(\mathbf{y}; X)$. Therefore, $w = 1/2\tilde{\sigma}^2$, where $\tilde{\sigma}^2$ is an estimate of the error variance under the unique-treatment model (e.g. Gilmour and Trinca, 2012). Under this model, each unique treatment induces a unique mean response. Specifically, let $\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_q$ be the unique treatments, i.e. the unique values of $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $q \leq n$. If $q < n$, i.e. there is at least one repeated treatment, then an estimator of σ^2 is given by

$$\tilde{\sigma}^2 = \frac{1}{n-q}\mathbf{y}^{\mathrm{T}}\left(I_n - H_Z\right)\mathbf{y},$$

where $H_Z = Z\left(Z^{\mathrm{T}}Z\right)^{-1}Z^{\mathrm{T}}$ and Z is the $n \times q$ matrix with ijth element $Z_{ij} = 1$ if $\mathbf{x}_i = \bar{\mathbf{x}}_j$ and zero otherwise, for $i = 1, \ldots, n$ and $j = 1, \ldots, q$. This estimator is independent of the choice of regression function $\mathbf{f}(\mathbf{x})$.

Under the designer distribution described in Section 4.1, the target parameter values are

$$\boldsymbol{\theta}_{SS,\mathcal{D}}(\mathbf{r}; X) = \left(F^{\mathrm{T}}F\right)^{-1}F^{\mathrm{T}}\boldsymbol{\mu} \tag{6}$$

where $\boldsymbol{\mu} = (\mu(\mathbf{x}_1), \ldots, \mu(\mathbf{x}_n))^{\mathrm{T}}$. See Section A of the Appendix for a justification of (6).

The Gibbs expected NSE and SH utilities are proportional to the following two objective functions

$$\begin{aligned} U^*_{\mathcal{G},NSE}(X) &= -\mathrm{tr}\left[\left(F^{\mathrm{T}}F\right)^{-1}\right]h_1(d) \\ U^*_{\mathcal{G},SH}(X) &= -ph_2(d) + \log\left|F^{\mathrm{T}}F\right|, \end{aligned} \tag{7}$$

where $d = n - q$ is the pure error degrees of freedom (Gilmour and Trinca, 2012), $h_1(d) = 1$ if $d > 0$ and ∞ if $d = 0$, and $h_2(d) = \psi(d/2) - \log d + d/(d-2)$, if $d > 0$ and ∞ if $d = 0$, with $\psi(\cdot)$ the digamma function. Justification of the expressions in (7) is given in Section B of the Appendix.

The objective functions in (7) can be seen as modifications of the objective functions for A- and D-optimality given in (5), respectively. In each case, the modification incentivises increasing d. In $U^*_{\mathcal{G},NSE}(X)$, the incentive given by $h_1(d)$ ensures the estimate $\tilde{\sigma}^2$ (and therefore the Gibbs posterior) exists. A corollary is that if the A-optimal design has at least one repeated design point, then the Gibbs optimal design under the NSE utility will be the A-optimal design. Although replication is a core principle of design of experiments (e.g. Morris, 2011, Section 1.2.4), whether or not an optimal design replicates design points is solely based on the objective function, through the balance of exploitation (replication) versus exploration (unique treatments). In $U^*_{\mathcal{G},SH}(X)$, the incentive for increasing d is more significant.

We consider a numerical example in Section 4.2.3.

4.2.2 Random calibration weight

We now allow w to be random. The target parameter values are unchanged from $\boldsymbol{\theta}_{SS,\mathcal{D}}(\mathbf{r}; X)$ in Section 4.2.1. We specify a prior distribution for the calibration weight given by $\pi(w) \propto w^{n/2-1}$. This choice of prior distribution results in the marginal Gibbs posterior distribution for $\boldsymbol{\theta}_{SS,\mathcal{D}}(\mathbf{r}; X)$ being the same as the corresponding Bayesian posterior distribution based on normally distributed errors. However, the formation of the Gibbs posterior does not require the assumption of normal errors. Specifically, the marginal Gibbs posterior distribution is a multivariate t-distribution (e.g. Kotz and Nadarajah, 2004, page 1) with mean $\hat{\boldsymbol{\theta}}_{SS}(\mathbf{y}; X)$, scale matrix $\hat{\sigma}^2 \left(F^{\mathrm{T}}F\right)^{-1}$ and $n-p$ degrees of freedom, where $\hat{\sigma}^2 = \mathbf{y}^{\mathrm{T}} \left(I_n - H_F\right) \mathbf{y}/(n-p)$ and $H_F = F \left(F^{\mathrm{T}}F\right)^{-1} F^{\mathrm{T}}$. Note that $\hat{\sigma}^2$ is the familiar regression function-dependent estimate of the error variance.

The Gibbs expected NSE and SH utilities are proportional to the following two objective functions

$$\begin{aligned} U^*_{\mathcal{G},NSE}(X) &= -\mathrm{tr}\left[\left(F^{\mathrm{T}}F\right)^{-1}\right] h_1(d) \\ U^*_{\mathcal{G},SH}(X) &= \log\left|F^{\mathrm{T}}F\right| - \mathrm{E}_{\mathcal{C}}\left(\mathrm{E}_{\mathcal{D}(\mathbf{r},X)}\left\{u^*_{SH}(\mathbf{y}; \mathbf{r}, X)\right\}\right), \end{aligned} \tag{8}$$

respectively, where

$$u^*_{SH}(\mathbf{y}; \mathbf{r}, X) = p \log \hat{\sigma}^2 + n \log\left[1 + \frac{(n-p)\|\mathbf{y} - \boldsymbol{\mu}\|_{H_F}}{\hat{\sigma}^2}\right].$$

Note that the Gibbs optimal design under the NSE utility will be the A-optimal design. Similar to Section 4.2.1, the objective function under the SH utility is a modification of the objective function for D-optimality. The modification, $\mathrm{E}_{\mathcal{C}}\left(\mathrm{E}_{\mathcal{D}(\mathbf{r},X)}\left\{u^*_{SH}(\mathbf{y}; \mathbf{r}, X)\right\}\right)$, is not available in closed form.

Before we consider a numerical example, we reiterate our position from Section 2.3 that we do not advocate either of these two approaches for specifying the calibration weight. However, in Section C of the Appendix we present the results of a simulation study which suggests that the fixed calibration weight of Section 4.2.1 should be favoured based on the coverage and precision of probability intervals for the target parameter values.

4.2.3 Numerical example

To illustrate the concepts, we consider a numerical example from Gilmour and Trinca (2012), where $n = 16$, $k = 3$ and

$$\mathbf{f}(\mathbf{x}) = \left(1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3\right)^{\mathrm{T}}$$

implementing an intercept, main and quadratic effects, and two-way interactions resulting in $p = 10$.

Table 1: For the squared error loss function, values of the objective function (and efficiencies), under each design, and for each design.

	Design		
	Fixed w (Section 4.2.1)	Random w (Section 4.2.2)	D-optimal
Fixed w	11.95 (100%)	11.95 (100%)	$-\infty$ (0%)
Random w	15.73 (100%)	15.73 (100%)	12.05 (69%)
D-optimal	18.26 (85%)	18.26 (85%)	19.92 (100%)
Pure error degrees of freedom, d	6	6	0

For this setup, the A-optimal design has $d = 2$, i.e. two repeated design points, meaning the Gibbs optimal design, under the NSE utility and fixed calibration weight is the A-optimal design. Therefore, we focus on Gibbs optimal designs under the SH utility. To find the design under the random calibration weight, for which the Gibbs expected utility is not available in closed form, we use the computational approach described in Section 3.2 but where the normal approximation is not required since the Gibbs posterior is available in closed form as a multivariate t-distribution. For this design, we also need to specify a hyper-variable distribution $\mathcal{C}$ for $\mu(\cdot)$, σ^2 and κ. We let $\mu(\cdot)$ be a realisation of a zero-mean Gaussian process with Matérn correlation function. Let $\bar{\boldsymbol{\mu}} = (\mu(\bar{\mathbf{x}}_1), \ldots, \mu(\bar{\mathbf{x}}_q))^{\mathrm{T}}$ be the unique mean responses, with $\boldsymbol{\mu} = Z\bar{\boldsymbol{\mu}}$, then

$$\bar{\boldsymbol{\mu}} \sim \mathrm{N}\left(\mathbf{0}_q, \tau^2 A\right). \tag{9}$$

In (9), the ijth element of A is $A_{ij} = c_1(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\rho})$ with

$$\begin{aligned} c_1(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\rho}) &= \prod_{z=1}^{k} \left[(1 + \rho|\bar{x}_{iz} - \bar{x}_{jz}|)\right. \\ &\qquad \left. \times \exp\left(-\rho|\bar{x}_{iz} - \bar{x}_{jz}|\right)\right], \end{aligned}$$

being the Matérn correlation function. To complete, we let κ be exponentially distributed with mean $\sigma^2 = 1$ and set $\rho_z = 1$ (for $z = 1, \ldots, k$), and $\tau^2 = 1$.

Table 1 shows the values of the objective functions (and efficiencies), under each design, for each design. The Gibbs optimal designs have identical performance under the two Gibbs expected utilities, indicating that the approach is invariable to the two different specifications of the calibration weight. These designs also have relatively good performance (85% efficiency) compared to the D-optimal design. On the other hand, the D-optimal design has $d = 0$ leading to a Gibbs expected utility, under fixed w, of $-\infty$ because the Gibbs posterior does not exist. The performance of the D-optimal design when compared against the Gibbs optimal design, under random w, is reasonable (69% efficiency). Table 1 also shows the pure error degrees of freedom, d, for each design. Note that the Gibbs optimal designs have $d = 6$, leading to $q = p = 10$ unique design points. This is the minimum possible under a non-informative prior distribution for the target parameter values meaning that replication is as high as possible.

4.3 L^2 loss

4.3.1 Derivation

We now consider an alternative loss inspired by Tuo and Wu (2015). The loss is given by a quadratic approximation to $\int_{\mathbb{X}} \left[\bar{\mu}(\mathbf{x};\mathbf{y},X) - \mathbf{f}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\theta}\right]^2 \mathrm{d}\mathbf{x}$, where $\bar{\mu}(\mathbf{x};\mathbf{y},X)$ is a non-parametric prediction of $\mu(\mathbf{x})$ based on $\mathbf{y}$ and X. Specifically, the loss is given by

$$\ell_{L^2}(\mathbf{t};\mathbf{y},X) = \sum_{m=1}^{M} s_m \left[\bar{\mu}(\boldsymbol{\chi}_m;\mathbf{y},X) - \mathbf{f}(\boldsymbol{\chi}_m)^{\mathrm{T}}\mathbf{t}\right]^2 \tag{10}$$

where $\{s_m, \chi_n\}_{m=1}^{M}$ are the quadrature weights and nodes, respectively, with M being the number of quadrature nodes. If $\bar{\mu}(\mathbf{x};\mathbf{y},X)$ is an unbiased predictor and the quadrature is exact, the target parameter values minimise the L^2 norm of the difference between $\mu(\mathbf{x})$ and $\mathbf{f}(\mathbf{x})^{\mathrm{T}}\mathbf{t}$. For this reason, the loss given by (10) is referred to as the L^2 loss.

For $\bar{\mu}(\mathbf{x};\mathbf{y},X)$, we use the predictive mean from a non-zero nugget, zero-mean Gaussian process. We assume

$$\mathbf{y} \sim \mathrm{N}\left[\mathbf{0}_n, \phi^2\left(\nu I_n + B\right)\right],$$

where $\nu > 0$ is the nugget and the ijth element of B is

$$\begin{aligned} B_{ij} &= c_2(\mathbf{x}_i,\mathbf{x}_j;\boldsymbol{\alpha}) \\ &= \exp\left[-\sum_{z=1}^{k}\alpha_z\left(x_{iz}-x_{jz}\right)^2\right], \end{aligned}$$

for $i,j = 1,\ldots,n$. The parameters ϕ^2, ν and $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_k)^{\mathrm{T}}$ are estimated by leave-one-out cross validation, and these estimates are denoted $\bar{\phi}^2$, $\bar{\nu}$ and $\bar{\boldsymbol{\alpha}}$, respectively. The predictive mean is given by

$$\bar{\mu}(\mathbf{x};\mathbf{y},X) = \mathbf{b}(\mathbf{x};\bar{\alpha})^{\mathrm{T}}\bar{V}\mathbf{y},$$

where $\mathbf{b}(\mathbf{x};\bar{\alpha}) = [c_2(\mathbf{x}_1,\mathbf{x};\bar{\alpha}),\ldots,c_2(\mathbf{x}_n,\mathbf{x};\bar{\alpha})]^{\mathrm{T}}$, the ijth element of $\bar{B}$ is $c_2(\mathbf{x}_i,\mathbf{x}_j;\bar{\boldsymbol{\alpha}})$, and $\bar{V} = \left(\bar{\nu}I_n + \bar{B}\right)^{-1}$.

The M-estimators are given by minimising (10) with respect to $\mathbf{t}$ and are

$$\hat{\boldsymbol{\theta}}_{L^2}(\mathbf{y};X) = E_Q^{-1}F_Q^{\mathrm{T}}\bar{D}_Q\bar{V}\mathbf{y},$$

where $E_Q = \sum_{m=1}^{M} s_m\mathbf{f}(\boldsymbol{\chi}_m)\mathbf{f}(\boldsymbol{\chi}_m)^{\mathrm{T}}$, F_Q is an $M \times p$ matrix with mth row $s_m\mathbf{f}(\boldsymbol{\chi}_m)^{\mathrm{T}}$ and $\bar{D}_Q$ is an $M \times n$ matrix with mith element $c_2(\boldsymbol{\chi}_m,\mathbf{x}_i;\bar{\boldsymbol{\alpha}})$, for $m = 1,\ldots,M$ and $i = 1,\ldots,n$.

We consider fixing w to maintain operational characteristics. Under a fixed w, the Gibbs posterior distribution of the target parameter values is

$$\mathrm{N}\left[\hat{\boldsymbol{\theta}}_{L^2}(\mathbf{y};X), \frac{1}{2w}E_Q^{-1}\right]. \tag{11}$$

We specify w such that the trace of the Gibbs posterior variance is approximately equal to the trace of the variance of $\hat{\boldsymbol{\theta}}_{L^2}(\mathbf{y};X)$. Specifically,

$$w = \frac{\mathrm{tr}\left[E_Q^{-1}\right]}{2\bar{\sigma}^2\mathrm{tr}\left[E_Q^{-1}F_Q^{\mathrm{T}}\bar{D}_Q\bar{V}\bar{V}\bar{D}_Q^{\mathrm{T}}F_QE_Q^{-1}\right]},$$

where

$$\bar{\sigma}^2 = \frac{\left\| \left(I_n - \bar{B}\bar{V}\right) \mathbf{y} \right\|_{I_n}}{\operatorname{tr}\left\{ \left(I_n - \bar{V}\bar{B}^{\mathrm{T}}\right) \left(I_n - \bar{B}\bar{V}\right) \right\}},$$

is an estimate of the error variance under the Gaussian process prediction.

The target parameter values, $\boldsymbol{\theta}_{L^2,\mathcal{D}}(\mathbf{r}; X)$, are given by minimising

$$\mathrm{E}_{\mathcal{D}}(\mathbf{r}; X) \left[w\ell_{L^2}(\mathbf{t}; \mathbf{y}, X) \right],$$

with respect to $\mathbf{t}$. These values are not available in closed form, so we approximate, via the delta method, to obtain

$$\boldsymbol{\theta}_{L^2,\mathcal{D}}(\mathbf{r}; X) = E_Q^{-1} F_Q^{\mathrm{T}} \bar{D}_Q \bar{V} \boldsymbol{\mu}.$$

The Gibbs expected NSE and SH utilities give the following two objective functions

$$\begin{aligned} U^*_{\mathcal{G},NSE}(X) &= \mathrm{E}_{\mathcal{C}} \left\{ \mathrm{E}_{\mathcal{D}(\mathbf{r};X)} \left[u^*_{NSE}(\mathbf{y}; \mathbf{r}, X) \right] \right\} \\ U^*_{\mathcal{G},SH}(X) &= \log \pi_{\mathcal{G}} \left[\boldsymbol{\theta}_{L^2,\mathcal{D}}(\mathbf{r}; X) | \mathbf{y}, X \right] \end{aligned} \tag{12}$$

where $u^*_{NSE}(\mathbf{y}; \mathbf{r}, X) = \| \boldsymbol{\theta}_{L^2,\mathcal{D}}(\mathbf{r}; X) - \hat{\boldsymbol{\theta}}_{L^2}(\mathbf{y}; X) \|$ and $\pi_{\mathcal{G}}\left[\mathbf{t}|\mathbf{y}, X\right]$ is the pdf of (11). Neither of the two expressions in (12) are available in closed form.

4.3.2 Numerical example

We return to the numerical example in Section 4.2.3 using the same designer distribution. Note that the correlation functions under the designer distribution and the Gaussian process prediction, $\bar{\mu}(\mathbf{x}; \mathbf{y}, X)$, are different.

For the quadrature, we use a Gauss-Legendre scheme (Weiser, 2016) with $M = 1000$. Since the two objective functions given by (12) are not available in closed form we use the computational methodology of Section 3.2 but where the normal approximation is not required since the Gibbs posterior is available in (11).

Figure 1 shows two-dimensional projections of the Gibbs optimal designs under the NSE and SH utilities. They appear similar to space-filling designs, which motivates a comparison with a maximum projection Latin hypercube design (Joseph et al., 2015). Therefore, also shown in Figure 1, are two-dimensional projections of a maximum projection Latin hypercube design (LHD) with $n = 16$ and $k = 3$ (where the design points are scaled to $\mathbb{X}$). Inspection of Figure 1 shows that the Gibbs optimal designs typically have points closer to the extremes of $\mathbb{X}$ than the LHD. Table 2 shows the values (and efficiencies) of the objective function under each design. The LHDs peform reasionably well compared to the Gibbs optimal designs (efficiencies of 70-75%). The Gibbs optimal design under the SH utility performs well (88%) when compared against the design under the NSE utility. However, the converse is not true: the Gibbs optimal design under the NSE utility has only 45% efficiency under the expected SH utility function. Figure 1 shows that the NSE design has clusters of designs points and it appears that these result in the observed lack of SH-efficiency.

5 Examples

5.1 Count responses and overdispersion

In this example, the aim of the experiment is to learn the relationship between controllable variables and count responses. Design of experiments for count responses has previously been considered by, for example,

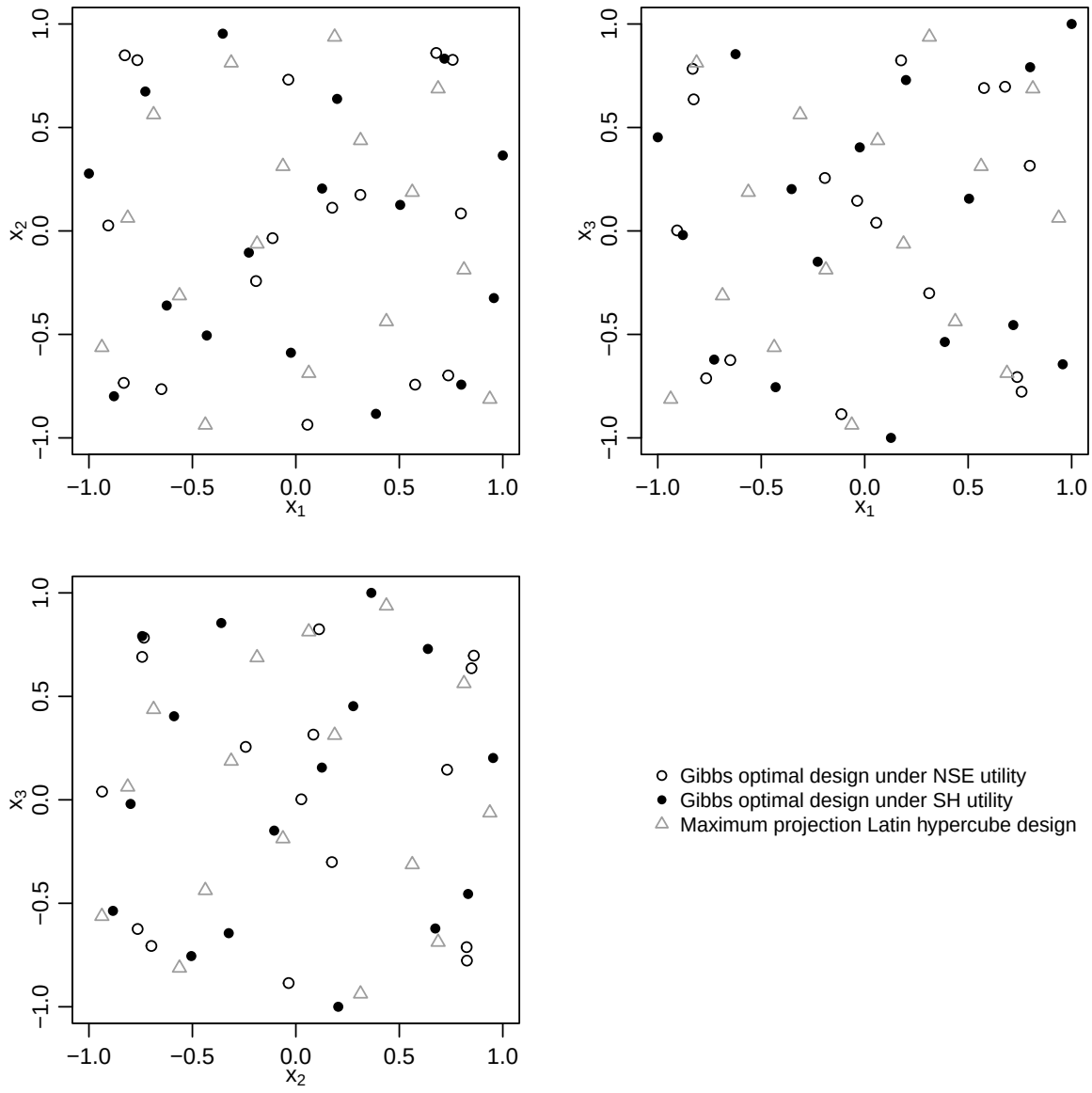

Figure 1: Plots of the Gibbs optimal designs, under the L^2 loss, and maximum projection Latin hypercube design.

Table 2: For the L^2 loss function, values of the objective function (and efficiencies), under each Gibbs expected utility, and for each design, including the maximum projection Latin hypercube design (LHD).

	Design		
	NSE utility	SH utility	LHD
NSE utility	-0.6950 (100%)	-0.7882 (88%)	-0.9191 (75%)
SH utility	-12.21 (45%)	-4.299 (100%)	-7.932 (70%)

Myers et al. (2010, Chapter 8), Russell (2019, Chapter 5) and McGree and Eccleston (2012). The typical approach is to assume the responses are realisations of a Poisson distribution. Accordingly, a Bayesian optimal design can then be found. Specifically, we may assume

$$\begin{aligned} y_i &\sim \text{Poisson}\left[\exp\left(\eta_i\right)\right] && (13)\\ \eta_i &= \mathbf{f}(\mathbf{x}_i)^{\mathrm{T}}\mathbf{t}, && (14) \end{aligned}$$

with $\mathbf{t} = \boldsymbol{\theta}_T$, along with a prior distribution, $\mathcal{P}_T$, for $\boldsymbol{\theta}_T$. In this example, we suppose that $\mathbf{f}(\mathbf{x}) = (1, x_1, x_2)^{\mathrm{T}}$, i.e. an intercept and main effects for $k = 2$ controllable variables, and $\mathbb{X} = [-1, 1]^2$

However, assuming a Poisson distribution is a strong assumption. For example, one property of the Poisson distribution is that its mean and variance are equal, $\text{var}(y_i) = \exp\left(\eta_i\right)$. In practice, this equality of mean and variance can often fail, a phenomenon known as over-dispersion (see, e.g., Pawitan, 2013, Section 4.5).

In the frequentist literature, a commonly-used approach is estimation via quasi-likelihood (see, e.g., Davison, 2003, pages 512-517). This does not require the assumption of a specified probability distribution for the response and allows the variance to be equal to the mean multiplied by a dispersion parameter ϕ.

The quasi log-likelihood is proportional to the log-likelihood under the statistical model given by (13), but multiplied by a factor of $1/\phi$, i.e.

$$l_{QL}(\mathbf{t}; \mathbf{y}, X) = \frac{1}{\phi}\sum_{i=1}^{n}\left[y_i\eta_i - \exp\left(\eta_i\right)\right],$$

where η_i is given by (14). This means that the maximum quasi-likelihood estimators are identical to the maximum likelihood estimators but their asymptotic variance is inflated by a factor of ϕ, to account for over-dispersion. The dispersion parameter is estimated from the responses (see, e.g. Davison, 2003, page 483).

We can use Gibbs inference using the negative quasi log-likelihood as the loss function (but removing the factor of $1/\phi$), i.e.

$$\ell_{QL}(\mathbf{t}; \mathbf{y}, X) = \sum_{i=1}^{n}\left[\exp\left(\eta_i\right) - y_i\eta_i\right],$$

where η_i is given by (14). The calibration weight, w, actually corresponds to $1/\phi$. We specify w to be the reciprocal of the estimator of ϕ, i.e.

$$w = \frac{n - p}{\sum_{i=1}^{n}\left[y_i - \exp(\hat{\eta}_i)\right]^2 / \exp(\hat{\eta}_i)},$$

where $\hat{\eta}_i = \mathbf{f}(\mathbf{x}_i)^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{QL}$ with

$$\hat{\boldsymbol{\theta}}_{QL} = \arg\min_{\mathbf{t}\in\Theta} \ell_{QL}(\mathbf{t}; \mathbf{y}, X) = \arg\max_{\mathbf{t}\in\Theta} l_{QL}(\mathbf{t}; \mathbf{y}, X),$$

being the maximum (quasi) likelihood estimators.

For the designer distribution, we modify the unique-treatment model from Section 4 to the count response scenario. We suppose

$$y_i \sim \text{neg} - \text{bin}\left(\mu_i, \alpha_i\right),$$

for $i = 1, \ldots, n$, where $\text{neg} - \text{bin}(\mu, \alpha)$ denotes a negative binomial distribution with mean μ and variance $\mu + \mu^2/\alpha$. The mean for the ith response is given by

$$\log \mu_i = \mathbf{f}(\mathbf{x}_i)^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\tau}, \quad (15)$$

where $\mathbf{z}_i = (Z_{i1}, \ldots, Z_{iq})^{\mathrm{T}}$ is the $q \times 1$ vector identifying which of the q unique treatments is subjected to the ith run, and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_q)^{\mathrm{T}}$ is the $q \times 1$ vector of unique-treatment effects. The mean is chosen as (15) so that the elements of $\boldsymbol{\tau}$ represent the discrepancy between the true mean and the assumed mean under the quasi-likelihood specification (on the log scale). This allows a type of parity when we compare Bayesian and Gibbs optimal designs later in this sub-section.

The scale parameter is $\alpha_i = \mu_i^2/(\kappa\mu_i - \mu_i)$, which results in $\text{var}(y_i) = \kappa\mu_i$. Thus $\kappa > 1$ controls the extent of over-dispersion.

Under the designer model, the hyper-variables are $\mathbf{r} = (\boldsymbol{\beta}, \boldsymbol{\tau}, \kappa)$. The target parameter values are

$$\boldsymbol{\theta}_{\ell_{QL},D}(X, \boldsymbol{\beta}, \kappa, \boldsymbol{\tau}) = \arg\min_{\mathbf{t}\in\Theta} \sum_{i=1}^{n} \left[\exp(\eta_i) - \mu_i\eta_i\right],$$

where η_i is given by (14) and μ_i is given by (15). Firstly, $\boldsymbol{\theta}_{\ell_{QL},D}(X, \boldsymbol{\beta}, \theta, \boldsymbol{\tau})$ are independent of κ. Second, the target parameter values, $\boldsymbol{\theta}_{\ell_{QL},D}(X, \boldsymbol{\beta}, \theta, \boldsymbol{\tau})$, under the designer distribution are the maximum likelihood likelihood estimators under a Poisson GLM. Therefore, in Step 2.(a) of the algorithm to approximate the Gibbs expected utility, in Section 3.2, we can readily use Fisher scoring to calculate the target parameter values.

We find Bayesian and Gibbs optimal designs, for $n = 10, 20, \ldots, 50$, under the NSE utility function. For the Bayesian optimal design, we assume the prior distribution, $\mathcal{P}_T$, is such that the elements of $\boldsymbol{\theta}_T$ are independent, each having a $\text{N}(0, 1)$ distribution. For the Gibbs optimal design, for the hyper-variable distribution, we assume $\boldsymbol{\beta}$, $\boldsymbol{\tau}$ and κ are independent. We assume the elements of $\boldsymbol{\tau}$ are independent, each having a $\text{U}(-2, 2)$ distribution, and suppose that the elements of $\boldsymbol{\beta}$ are independent, each having a $\text{N}(0, 1)$ distribution. We specify $\kappa \sim U(1, 5)$.

Figure 2(a) shows plots of the Gibbs expected negative squared error utility for both Gibbs (black circles) and Bayesian optimal designs (grey circles). It also shows the Bayesian expected negative squared error utility for both designs (triangles with Gibbs black and Bayesian grey). It can be seen that both designs are sub-optimal under the alternative expected utility. However, the difference on the scale of the expected Bayesian utility is larger than on the scale of the expected Gibbs utility. Informally, less is lost by not making strong assumptions and these assumptions being true, than by making these strong assumptions and these not being true.

Figures 2(b) and (c) shows plots of x_2 against x_1 for the Gibbs and Bayesian optimal designs, respectively. The size of the plotting character is proportional to the number of repeated designs points at each support point. Clearly, the Bayesian optimal design places points at the extremes of $\mathbb{X}$. Whereas, the Gibbs optimal design includes points in the interior.

5.2 Time-to-event responses and proportional hazards

In this example, the aim of the experiment it to learn the relationship between controllable variables and a time-to-event response (possible subjected to non-informative right censoring). Previous approaches for designing these type of experiments have been proposed by, for example, Chaloner and Larntz (1992), Konstantinou et al. (2015), and Konstantinou et al. (2017).

It is common for time-to-event responses to assume a proportional hazards model (see, e.g., Davison, 2003, Section 10.8.2). This is where the pdf for the response is

$$f(y_i; \mathbf{t}, \mathbf{x}_i) = h_0(y_i) \exp\left[\eta_i - e^{\eta_i} \int_0^{y_i} h_0(u)\mathrm{d}u\right], \tag{16}$$

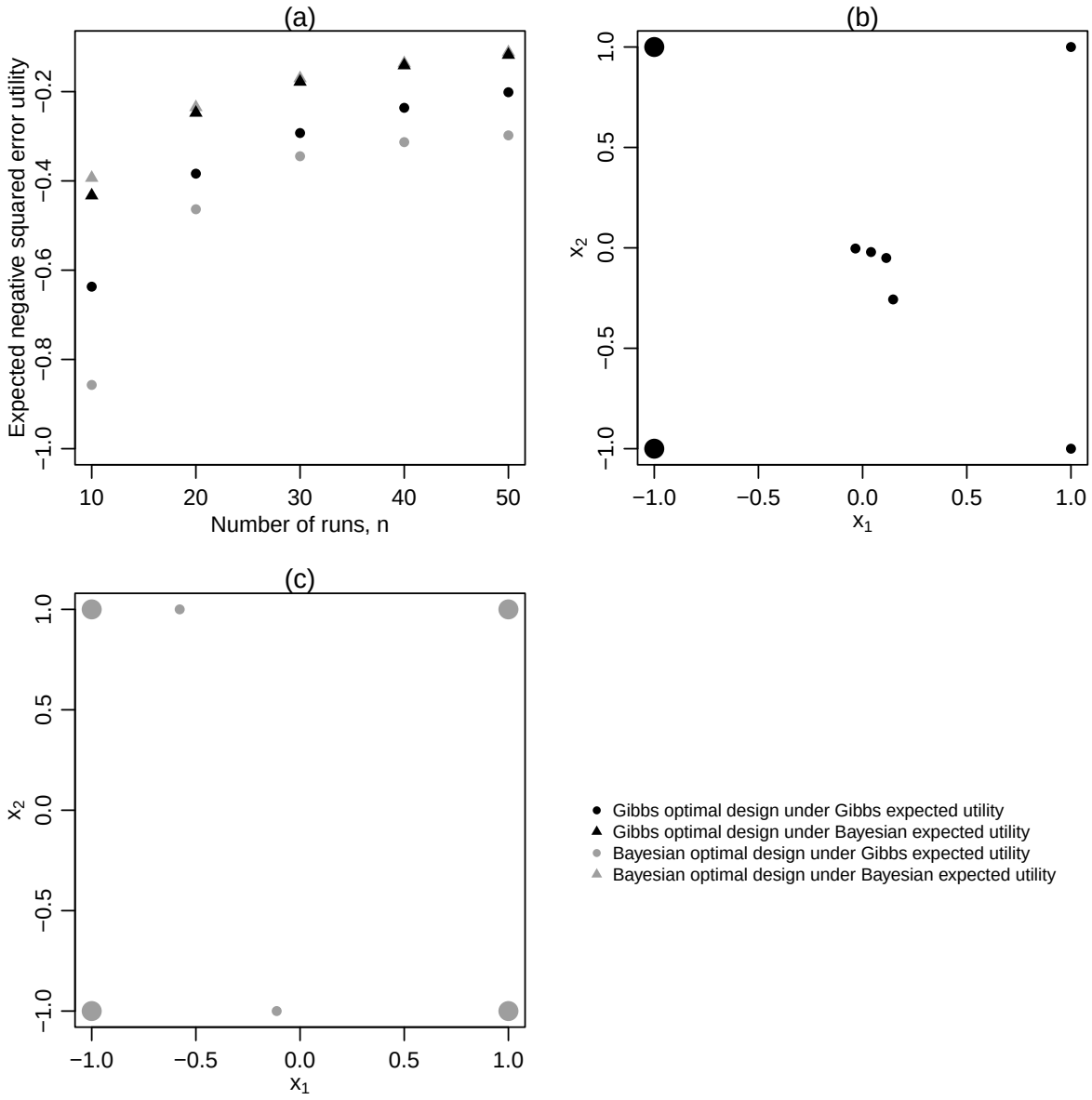

Figure 2: Plots for the count response experiment. Sub-panel (a) shows plots of the Gibbs and Bayesian expected negative squared error utility against number of runs, n, for the Gibbs and Bayesian optimal designs. Sub-panels (b) and (c) shows plots of x_2 against x_1 for the Gibbs and Bayesian optimal designs, respectively.

where $h_0(\cdot)$ is a baseline hazard and $\eta_i = \mathbf{f}(\mathbf{x}_i)^{\mathrm{T}}\mathbf{t}$. In this example, we suppose that $\mathbf{f}(\mathbf{x}) = (1, x_1, x_2, x_3)^{\mathrm{T}}$, i.e. an intercept and main effects for $k = 3$ controllable variables. Bayesian inference requires that a particular form for the baseline hazard is assumed (see, e.g., Ibrahim et al., 2001, Chapter 2). By contrast, under the frequentist approach, the Cox proportional hazards model does not requires a particular form for the baseline hazard to be assumed, with estimators given by maximising the partial likelihood.

5.2.1 Gibbs optimal designs

We consider Gibbs inference using the negative partial log-likelihood as a loss function. Bissiri et al. (2016) considered such an analysis for investigating the relationship of bio-markers and survival time for colon cancer. They set $w = 1$ which can be informally justified as ensuring the asymptotic Gibbs posterior variance is equal to the asymptotic variance of the maximum partial likelihood estimators.

If we assume that all responses are unique (i.e. there are no tied responses), then the loss function is

$$\ell_{PL}(\mathbf{y}, \mathbf{c}; \mathbf{t}) = \sum_{i=1}^{n} c_i \left[\log \sum_{j \in R_i} e^{\eta_j} - \eta_i \right],$$

where $\mathbf{c} = (c_1, \ldots, c_n)$, with $c_i = 0$ if the ith response is right censored at y_i, and $c_i = 1$ otherwise, and $R_i = \{j \in (1, \ldots, n) | y_j \geq y_i\}$ is the risk set, i.e. the subset of $\{1, \ldots, n\}$ of responses which have not been observed or right censored before y_i.

For the designer distribution, we assume y_i and c_i are independent. For the censoring indicators, it is assumed that $c_i \sim \text{Bernoulli}(\rho)$ for some $\rho \in [0, 1]$ controlling the probability of right censoring, with $\rho = 1$ corresponding to no censoring. If the designer distribution for y_i is a proportional hazards distribution, i.e. has pdf given by (16), with $\eta_i = \mathbf{f}(\mathbf{x}_i)^{\mathrm{T}}\boldsymbol{\beta}$, then the target parameter values are $\boldsymbol{\theta}_{PL,D}(\mathbf{r}, \mathbf{c}, X) = \boldsymbol{\beta}$. Moreover, the maximum partial likelihood estimators and Gibbs posterior are invariant to the actual choice of baseline hazard. Therefore, for simplicity, we use the exponential distribution with $h_0(u) = 1$. The hyper-variables are $\mathbf{r} = (\boldsymbol{\beta}, \rho)$. The hyper-variable distribution is such that $\boldsymbol{\beta}$ and ρ are independent. The elements of $\boldsymbol{\beta}$ are assumed independent with each element having a $\text{N}(0, 5)$ distribution. We fix ρ and investigate sensitivity to its specification.

We consider number of runs, $n = 10, 20, \ldots, 50$, censoring probability $\rho = 0.5, 0.75, 1.00$ and the NSE utility function. Figure 3(a) shows a plot of Gibbs expected NSE utility against number of runs, n, for the different values of ρ. It makes intuitive sense, that as ρ increases (probability of censoring decreases), the Gibbs expected utility increases since less information is lost. The effect of amount of the censoring on the actual design was investigated and it was found that the optimal design was only negligibly affected by probability of censoring. For example, the optimal design for $n = 10$ and $\rho = 0.5$ has a very similar value of expected Gibbs utility (with $\rho = 1$) as the optimal design found under $\rho = 1$.

We compare the Gibbs optimal designs to Bayesian optimal designs under a Weibull proportional hazards model. In this case, $h_0(u) = \psi u^{\psi - 1}$, where $\psi > 0$ is an unknown nuisance parameter. If $\psi > 1$ ($\psi < 1$), then the baseline hazard is an increasing (decreasing) function of u. The log-likelihood is given by

$$\begin{aligned} &\log \pi(\mathbf{y}, \mathbf{c} | \boldsymbol{\beta}, \psi; X) \\ &= \sum_{i=1}^{n} I(c_i = 1) \left[\eta_i + \log \psi + (\psi - 1) \log y_i \right] - e^{\eta_i} y_i^{\psi}. \end{aligned}$$

The elements of $\boldsymbol{\beta}$ and $\mathbf{c}$ are assumed to have the same distributions as under the Gibbs expected utility above. The parameter ψ is assumed to have a $\text{U}[0.5, 1.5]$ distribution.

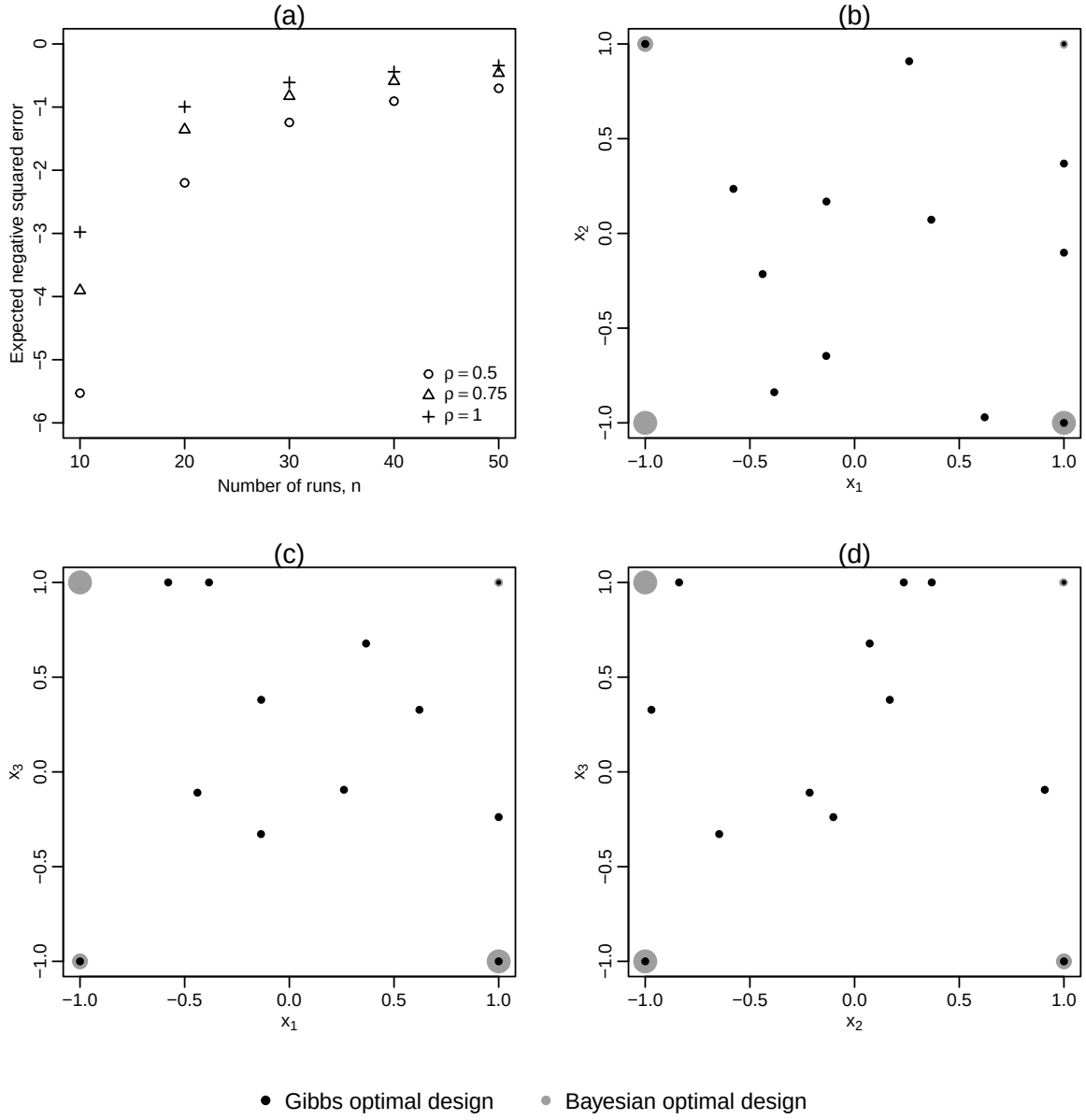

Figure 3: Panel (a): Plot of expected negative squared error utility against number of runs, n, for different levels of censoring (controlled by ρ). Panels (b)-(d): Plots of x_1, x_2 and x_3 for the Gibbs and Bayesian optimal designs. The size of the plotting character is proportional to the number of repetitions of the design points.

Number of runs, n	10	20	30	40	50
$\hat{U}_G(\bar{X}_B)/\hat{U}_G(\bar{X}_G)$	0.037	0.061	0.131	0.159	0.175
$\hat{U}_B(\bar{X}_G)/\hat{U}_B(\bar{X}_B)$	0.270	0.510	0.600	0.650	0.680

Table 3: Estimated expected utility efficiency of the Gibbs and Bayesian optimal designs under the alternative utility.

We present results for the case where the censoring probability is $\rho = 0.75$. The results for $\rho = 0.5$ and $\rho = 1$ were qualitatively the same. Table 3 shows the estimated efficiency of the Gibbs and Bayesian optimal designs under the alternative utility. Suppose for a given value of n and ρ, $\bar{X}_G$ and $\bar{X}_B$ denote the Gibbs and Bayesian optimal design, respectively. Then, for example, the efficiency of $\bar{X}_G$ under the expected Bayesian utility is $U_B(\bar{X}_G)/U_B(\bar{X}_B)$. The conclusion is similar to the count response experiment in Section 5.1, in that less information is lost by not making strong assumptions (i.e. assuming a Weibull distribution) and these assumptions being true than by making these strong assumptions and these not being true.

Figure 3(b)-(c) show projections of the design points. The size of the plotting symbol is proportional to the number of repeated design points. A similar phenomenon to the count response experiment in Section 5.1 occurs, in that the Gibbs optimal design has more points in the interior.

6 Discussion

This paper proposes a new design of experiments framework called Gibbs optimal design. This framework extends the decision-theoretic Bayesian optimal design of experiments approach to account for potentially misspecified statistical models. A computational approach is outlined for finding designs in practice. The framework is demonstrated on illustrative examples involving linear models, in addition to experiments involving count and time-to-event responses. From the examples, it was found that less information is lost by not making strong assumptions, and these being true, than by making strong assumptions and these assumptions not being true.

The key is the specification of the designer distribution. It needs to be flexible enough that it is plausible that the true response-generating distribution is a special case. In the illustrative examples, a unique-treatment model was employed as the designer distribution.

The framework should open up avenues for further research. One interesting area is that the target parameter values of Gibbs inference under a fixed design are dependent on the design. This can be viewed as an unattractive property, and future work will attempt to address this. One potential idea is that the utility function be evaluated at *desired target parameter values*, where these are the limit of the target parameter values as $n \to \infty$ (and under certain conditions) and have physical interpretation independent of the design. As an example, consider the linear model in Section 4. Under certain conditions, Wong et al. (2017) showed that the target parameter values under the sum of squares loss converged to the parameter values that minimise the L^2 norm of the difference between the designer distribution mean response and $\mathbf{f}(\mathbf{x})^{\mathrm{T}}\mathbf{t}$. These parameter values have desirable physical interpretation.

A further area of research is to investigate the sensitivity of the Gibbs optimal design to the method chosen to specify calibration weight. In the numerical example in Section 4.2, the Gibbs optimal designs were insensitive to the different specifications of the calibration weight. This will need further investigation in consort with current research on automatic specifications.

Lastly, in Section 3.2, a computational approach was outlined for finding Gibbs optimal designs by maximising the Gibbs expected utility. This was based on the simple premise of approximating the Gibbs posterior distribution by a normal distribution. This builds on existing approaches to find Bayesian optimal designs. We encourage researchers who develop computational methodology to find Bayesian designs to generalise to finding Gibbs optimal designs. There is an additional step in the latter case in that the target parameter values, $\boldsymbol{\theta}_{\ell,\mathcal{D}}(\mathbf{r}; X)$, need to be determined. This is exemplified by Step 2.(b) in the algorithm in Section 3.2. However, this step adds negligible computational complexity relative the ultimate task of approximating and maximising the expected utility over a design space of, potentially, high dimensionality.

A Justification of target parameter values given by equation (6)

The following results will prove useful.

Lemma A.1. *If a linear model has model matrix F, and the corresponding unique-treatment model has model matrix Z, with $H_Z = Z\left(Z^TZ\right)^{-1}Z$, then $F = H_ZF$.*

Proof. Without loss of generality, unique-treatment $\bar{\mathbf{x}}_j$ is repeated $n_j \geq 1$ times, for $j = 1, \ldots, q$. Then $n = \sum_{j=1}^{q} n_j$. Let $\bar{\mathbf{f}}_j = \mathbf{f}(\bar{\mathbf{x}}_j)$, for $j = 1, \ldots, q$. Then the model matrix F can be written

$$F = \begin{pmatrix} \bar{F}_1 \\ \vdots \\ \bar{F}_q \end{pmatrix}$$

where $\bar{F}_j$ is an $n_j \times p$ matrix with each row given by $\bar{\mathbf{f}}_j$, i.e. the rows of $\bar{F}_j$ are identical. The corresponding H_Z matrix can be written in the following block diagonal form

$$H_Z = \begin{pmatrix} \frac{1}{n_1}J_{n_1} & 0 & \\ 0 & \ddots & 0 \\ & 0 & \frac{1}{n_q}J_{n_q} \end{pmatrix},$$

where J_{n_j} is the $n_j \times n_j$ matrix of ones. Then

$$H_ZF = \begin{pmatrix} \frac{1}{n_1}J_{n_1}\bar{F}_1 \\ \vdots \\ \frac{1}{n_q}J_{n_q}\bar{F}_q \end{pmatrix},$$

where $J_{n_j}\bar{F}_j = n_j\bar{F}_j$. Therefore $H_ZF = F$ as required. □

Lemma A.2. *If the designer distribution, $\mathcal{D}(\mathbf{r}, X)$, is such that $\mathbf{y} \sim \mathrm{N}\left(Z\bar{\boldsymbol{\mu}}, \kappa I_n\right)$, then*

(i) $\|\mathbf{y}\|_{H_F}$ *and* $\|\mathbf{y}\|_{I_n - H_Z}$ *are independent;*

(ii) $F^T\mathbf{y}$ *and* $\|\mathbf{y}\|_{I_n - H_Z}$ *are independent;*

(iii) $\|\mathbf{y}\|_{I_n - H_Z} \sim \chi^2_d$*;*

(iv)

$$\begin{aligned} \mathrm{E}_{\mathcal{D}(\mathbf{r},X)}\left[\frac{1}{\|\mathbf{y}\|_{I_n - H_Z}}\right] &= \frac{1}{\kappa(d-2)} \\ \mathrm{E}_{\mathcal{D}(\mathbf{r},X)}\left[\log\|\mathbf{y}\|_{I_n - H_Z}\right] &= \log\kappa + \log 2 + \psi\left(\frac{d}{2}\right). \end{aligned}$$

Proof. Statements (i) to (iv) are proved as follows.

(i) From Lemma A.1, $(I_n - H_Z)H_F = H_F - H_ZH_F = H_F - H_ZF(F^TF)^{-1}F^T = H_F - F(F^TF)^{-1}F^T = H_F - H_F = 0_{n\times n}$. Statement (i) follows from Craig's theorem (see, for example, Hocking 2013, Theorem 10.2).

(ii) From Lemma A.1, $F^T(I_n - H_Z) = [(I_n - H_Z)F]^T = 0_{n\times n}$. Statement (ii) follows from, for example, Hocking 2013, Theorem 10.3.

(iii) Since $I_n - H_Z$ is idempotent and

$$\begin{aligned}\|Z\bar{\boldsymbol{\mu}}\|_{I_n-H_Z} &= \bar{\boldsymbol{\mu}}^T Z^T Z\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}^T Z^T H_Z Z\bar{\boldsymbol{\mu}} \\ &= \bar{\boldsymbol{\mu}}^T Z^T Z\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}^T Z^T Z\bar{\boldsymbol{\mu}} \\ &= 0,\end{aligned}$$

statement (iii) follows from, for example, Hocking 2013, Theorem 10.1.

(iv) The expectations in statement (iv) follow from properties of χ^2_d.

□

The target parameter values minimise

$$\begin{aligned}L_{SS,\mathcal{D}}(\mathbf{t};\mathbf{r},X) &= \mathrm{E}_{\mathcal{D}(\mathbf{r},X)}\left[\frac{\|\mathbf{y}-F\mathbf{t}\|_{I_n}}{2\tilde{\sigma}^2}\right] \\ &= \frac{n-q}{2}\mathrm{E}_{\mathcal{D}(\mathbf{r},X)}\left[\frac{\|\mathbf{y}-F\mathbf{t}\|_{I_n}}{\|\mathbf{y}\|_{I_n-H_Z}}\right]\end{aligned}$$

where $\mathcal{D}(\mathbf{r},X)$ has $\mathbf{y}\sim \mathrm{N}(Z\bar{\boldsymbol{\mu}},\kappa I_n)$ and $\mathbf{r} = \left(\mu(\cdot),\kappa,\sigma^2\right)$. Differentiating $L_{SS,\mathcal{D}}(\mathbf{t};\mathbf{r},X)$ with respect to $\mathbf{t}$ and setting equal to $\mathbf{0}_p$ gives the following target parameter values

$$\boldsymbol{\theta}_{SS,\mathcal{D}}(\mathbf{r},X) = \frac{1}{\mathrm{E}_{\mathcal{D}(\mathbf{r},X)}\left(\|\mathbf{y}\|^{-1}_{I_n-H_Z}\right)}\left(F^TF\right)^{-1}\mathrm{E}_{\mathcal{D}(\mathbf{r},X)}\left(\frac{F^T\mathbf{y}}{\|\mathbf{y}\|_{I_n-H_Z}}\right).$$

From Lemma A.2 (ii), $F^T\mathbf{y}$ and $\|\mathbf{y}\|_{I_n-H_Z}$ are independent, so

$$\begin{aligned}\boldsymbol{\theta}_{SS,\mathcal{D}}(\mathbf{r},X) &= \frac{1}{\mathrm{E}_{\mathcal{D}(\mathbf{r},X)}\left(\|\mathbf{y}\|^{-1}_{I_n-H_Z}\right)}\left(F^TF\right)^{-1}\mathrm{E}_{\mathcal{D}(\mathbf{r},X)}\left(F^T\mathbf{y}\right)\mathrm{E}_{\mathcal{D}(\mathbf{r},X)}\left(\|\mathbf{y}\|^{-1}_{I_n-H_Z}\right), \\ &= \left(F^TF\right)^{-1}F^T\boldsymbol{\mu},\end{aligned}$$

as required.

B Justification of Gibbs expected utilities given by equation (7)

B.1 NSE utility

The Gibbs posterior mean is given by

$$\mathrm{E}_{\mathcal{G}}\left[\boldsymbol{\theta}_{SS,\mathcal{D}}(\mathbf{r};X)\right] = \hat{\boldsymbol{\theta}}_{SS}(\mathbf{y};X).$$

Therefore, the inner expectation (i.e. with respect to $\mathbf{y}$ under $\mathcal{D}(\mathbf{r};X)$) in the Gibbs expected NSE utility is

$$\begin{aligned}\mathrm{E}_{\mathcal{D}(\mathbf{r};X)}\left[u_{\mathcal{G},NSE}\left(\boldsymbol{\theta}_{SS,\mathcal{D}}(\mathbf{r};X),\mathbf{y},X\right)\right] &= \mathrm{E}_{\mathcal{D}(\mathbf{r};X)}\left[\|\boldsymbol{\theta}_{SS,\mathcal{D}}(\mathbf{r};X) - \hat{\boldsymbol{\theta}}_{SS}(\mathbf{y};X)\|_2^2\right] \\ &= -\sigma^2\mathrm{tr}\left[\left(F^TF\right)^{-1}\right],\end{aligned}$$

since $\mathrm{E}_{\mathcal{D}(\mathbf{r};X)}\left[\hat{\boldsymbol{\theta}}_{SS}(\mathbf{y};X)\right] = \boldsymbol{\theta}_{SS,\mathcal{D}}(\mathbf{r};X)$. Now the Gibbs expected NSE utility is

$$\begin{aligned} U_{\mathcal{G},NSE}(X) &= \mathrm{E}_{\mathcal{C}}\left[-\sigma^2 \mathrm{tr}\left[\left(F^TF\right)^{-1}\right]\right], \\ &= \mathrm{E}_{\mathcal{C}}\left(-\sigma^2\right)\mathrm{tr}\left[\left(F^TF\right)^{-1}\right]. \end{aligned}$$

A constraint of $n > q$ ensures that the Gibbs posterior distribution exists.

B.2 SH utility

The Gibbs posterior distribution is $\mathrm{N}\left[\hat{\boldsymbol{\theta}}_{SS}(\mathbf{y};X), \hat{\sigma}^2\left(F^TF\right)^{-1}\right]$ with log pdf

$$\begin{aligned} \log \pi_{\mathcal{G}}\left(\mathbf{t}|\mathbf{y};X\right) &= -\frac{p}{2}\log\left(2\pi\right) - \frac{p}{2}\log\hat{\sigma}^2 + \frac{1}{2}\log|F^TF| - \frac{1}{2\hat{\sigma}^2}\left[\mathbf{t}-\hat{\boldsymbol{\theta}}_{SS}(\mathbf{y};X)\right]^T F^TF\left[\mathbf{t}-\hat{\boldsymbol{\theta}}_{SS}(\mathbf{y};X)\right] \\ &= -\frac{p}{2}\log\left(2\pi\right) + \frac{p}{2}\log d - \frac{p}{2}\log\left[\mathbf{y}^T\left(I_n - H_Z\right)\mathbf{y}\right] + \frac{1}{2}\log|F^TF| \\ &\quad -\frac{d\mathbf{t}^TF^TF\mathbf{t}}{2\mathbf{y}^T\left(I_n-H_Z\right)\mathbf{y}} + \frac{d\mathbf{t}^TF^T\mathbf{y}}{\mathbf{y}^T\left(I_n-H_Z\right)\mathbf{y}} - \frac{d\mathbf{y}^TH_F\mathbf{y}}{2\mathbf{y}^T\left(I_n-H_Z\right)\mathbf{y}}, \end{aligned} \tag{17}$$

where $H_F = F\left(F^TF\right)^{-1}F$. The line (17) follows from $\hat{\sigma}^2 = \mathbf{y}^T\left(I_n - H_Z\right)\mathbf{y}/(n-q)$ and $d = n-q$ is the pure error degrees of freedom.

From Lemma A.2 (i), (ii), and (iv), it follows that the inner expectation (i.e. with respect to $\mathbf{y}$ under $\mathcal{D}(\mathbf{r};X)$) in the Gibbs expected SH utility is

$$\begin{aligned} \mathrm{E}_{\mathcal{D}(\mathbf{r};X)}\left[u_{\mathcal{G},SH}\left(\boldsymbol{\theta}_{SS,\mathcal{D}}(\mathbf{r};X),\mathbf{y},X\right)\right] &= \mathrm{E}_{\mathcal{D}(\mathbf{r};X)}\left[\log\pi_{\mathcal{G}}\left(\boldsymbol{\theta}_{SS,\mathcal{D}}(\mathbf{r};X)|\mathbf{y};X\right)\right] \\ &= -\frac{p}{2}\left[\log(2\pi) + \log\sigma^2 + \log 2\right] - \frac{p}{2}h_2(d) + \frac{1}{2}\log|F^TF|, \end{aligned} \tag{18}$$

since $\boldsymbol{\theta}_{SS,\mathcal{D}}(\mathbf{r};X) = \left(F^TF\right)^{-1}F^TZ\bar{\boldsymbol{\mu}}$. Equation (7) follows from taking expectation of (18) with respect to the hyper-variable distribution $\mathcal{C}$.

C Simulation study comparing fixed and random specifications of the calibration weight

In this section, we present the results of a simulation study comparing the fixed and random specifications of the calibration weight, under the SS loss.

We let $k = 3$ and consider the same regression function from Sections 4.2.3 and 4.3.2, i.e.

$$\mathbf{f}(\mathbf{x}) = \left(1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3\right)^T.$$

We generate responses $\mathbf{y}$ via a uniformly distributed random design of n runs using the designer distribution described in Section 4.2.3. We evaluate the Gibbs posterior distributions under both the fixed and random specifications of the calibration weight. We repeat this $B = 1000$ times for $n = 20, 25, 30, \ldots, 200$.

The approaches to specify the calibration weights are compared using the median of the mean coverage and length (measuring precision) of the 95% probability intervals for $\boldsymbol{\theta}_{SS,\mathcal{D}}(\mathbf{r}, X)$. The mean is over the $B = 1000$ repetitions and the median is over the $p = 10$ elements of $\boldsymbol{\theta}_{SS,\mathcal{D}}(\mathbf{r}, X)$.

Figure 4 shows plots of the coverage and length against n. The random calibration weight approach appears to results in wider (less precise) 95% probability intervals that result in over-coverage. Moreover, the coverage of the 95% probability intervals does not appear to converge to 95% as n increases (in contrast to the fixed calibration weight approach). This is due to the random calibration weight approach using the estimate of the error variance which is based on the incorrect regression function.

References

Atkinson, A. C., Donev, A. and Tobias, R. D. (2007) *Optimum Experimental Designs, with SAS*. Oxford University Press.

Azriel, D. (2023) Optimal minimax random designs for weighted least squares estimators. *Biometrika* **110**, 273–280.

Berk, R. (1966) Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics* **37**, 51–58.

Bingham, D. and Chipman, H. (2007) Incorporating prior information in optimal design for model selection. *Technometrics* **49**, 155–163.

Bissiri, P., Holmes, C. and Walker, S. (2016) A general framework for updating belief distributions. *J. R. Stat. Soc. B* **78**, 1103–1130.

Box, G. and Draper, N. (1959) A basis for the selection of a response surface design. *Journal of the American Statistical Association* **54**, 622–654.

Catoni, O. (2007) PAC-bayesian supervised classification: the thermodynamics of statistical learning. *IMS Lecture Notes* **56**.

Chaloner, K. and Larntz, K. (1992) Bayesian design for accelerated life testing. *Journal of Statistical Planning and Inference* **33**, 245–259.

Chaloner, K. and Verdinelli, K. (1995) Bayesian experimental design: a review. *Stat. Sci.* **10**, 273–304.

Cook, R. and Nachtsheim, C. (1982) Model-robust, linear-optimal designs. *Technometrics* **24**, 49–54.

Davison, A. (2003) *Statistical Models*. Cambridge University Press.

Etzioni, R. and Kadane, J. (1993) Optimal experimental design for another's analysis. *Journal of the American Statistical Association* **88**, 1404–1411.

Gilmour, S. and Trinca, L. (2012) Optimum design of experiments for statistical inference (with discussion). *J. R. Stat. Soc. C* **61**, 345–401.

Goos, P. and Jones, B. (2011) *Optimal design of experiments: a case study approach*. John Wiley and Sons Inc.

Grünwald, P. (2012) The safe bayesian. In *International Conference on Algorithmic Learning* 169–183. Springer.

Hayashi, F. (2000) *Econometrics*. Princeton University Press.

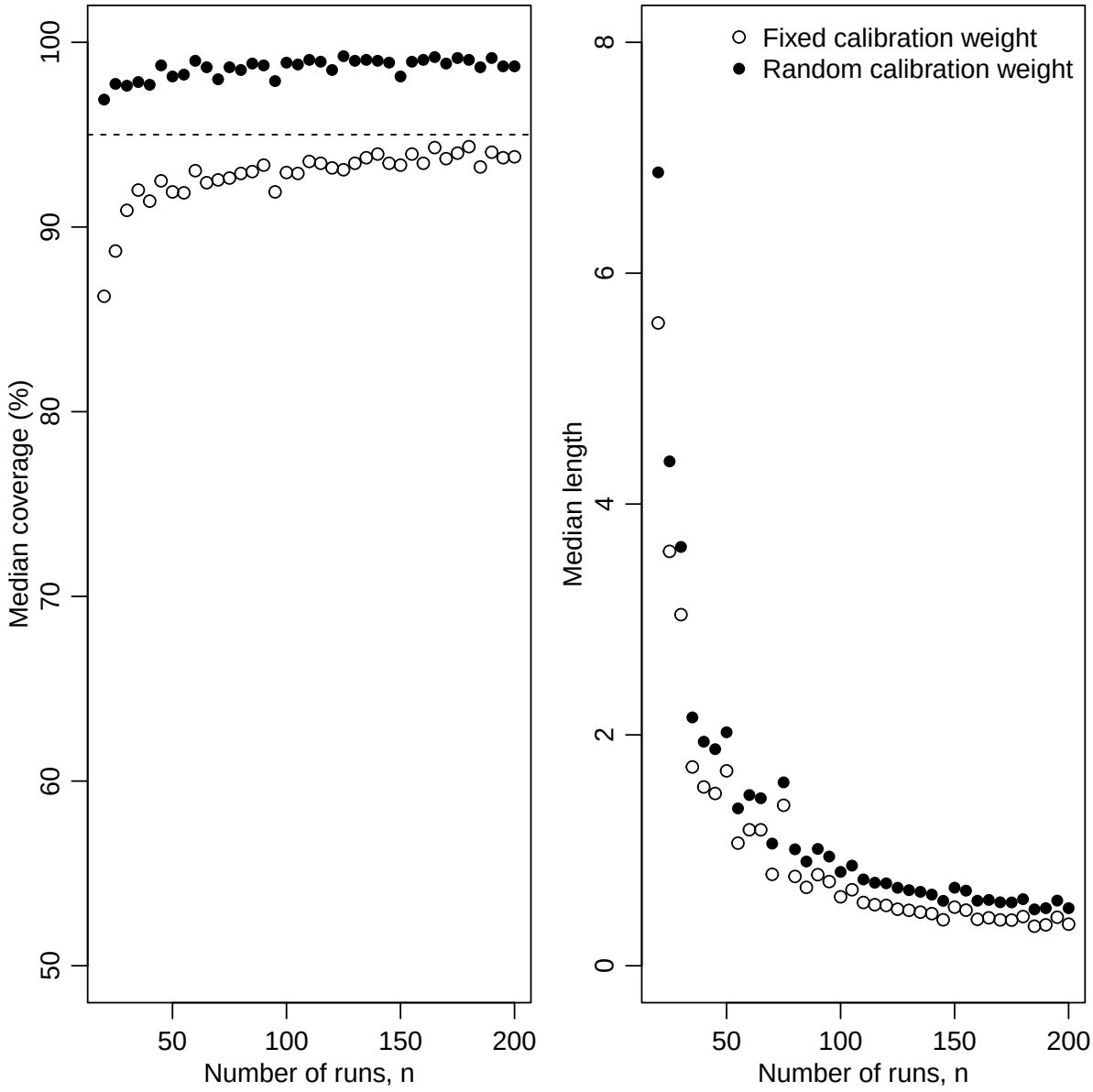

Figure 4: Plots coverage (left) and length (right) against n for the two different approaches to specifying the calibration weight.

Hocking, R. (2013) *Methods and Applications of Linear Models.* John Wiley & Sons 2nd edn.

Holmes, C. and Walker, S. (2017) Assigning a value to a power likelihood in a general bayesian model. *Biometrika* **104**, 497–503.

Ibrahim, J., Chen, M. and Sinha, D. (2001) *Bayesian survival analysis.* Springer.

Jewson, J. and Rossell, D. (2022) General bayesian loss function selection and the use of improper models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84**, 1640–1665.

Joseph, V., Gul, E. and Ba, S. (2015) Maximum projection designs for computer experiments. *Biometrika* **102**, 371–380.

Konstantinou, M., Biedermann, S. and Kimber, A. (2015) Optimal designs for full and partial likelihood information—with application to survival models. *Journal of Statistical Planning and Inference* **165**, 27–37.

— (2017) Model robust designs for survival trials. *Computational Statistics and Data Analysis* **113**, 239–250.

Kotz, S. and Nadarajah, S. (2004) *Multivariate t Distributions and their Applications.* Cambridge University Press, Cambridge.

Li, W. and Nachtsheim, C. (2000) Model-robust factorial designs. *Technometrics* **42**, 345–352.

Long, Q., Scavino, M., Tempone, R. and Wang, S. (2013) Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Computer Methods in Applied Mechanics and Engineering* **259**, 24–39.

Lyddon, S., C.C., H. and Walker, S. (2019) General Bayesian updating and the loss-likelihood bootstrap. *Biometrika* **106**, 465–478.

McGree, J. and Eccleston, J. (2012) Robust designs for Poisson regression models. *Technometrics* **54**, 64–72.

Mead, R., Gilmour, S. and Mead, A. (2012) *Statistical Principles for the Design of Experiments: Applications to Real Experiments.* Cambridge University Press.

Meyer, R. and Nachtsheim, C. (1995) The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometerics* **37**, 60–69.

Morris, M. (2011) *Design of experiments: an introduction based on linear models.* CRC Press.

Myers, R., Montgomery, D., Vining, G. and Robinson, T. (2010) *Generalized Linear Models with Applications in Engineering and the Sciences* vol. 2nd. Wiley Series in Probability and Statistics.

O'Hagan, A. and Forster, J. (2004) *Kendall's Advanced Theory of Statistics Volume 2B Bayesian Inference.* Wiley 2nd edn.

Overstall, A. and McGree, J. (2022) Bayesian decision-theoretic design of experiments under an alternative model. *Bayesian Analysis* **17**, 1021–1041.

Overstall, A., McGree, J. and Drovandi, C. (2018) An approach for finding fully Bayesian optimal designs using normal-based approximations to loss functions. *Statistics and Computing* **28**, 343–358.

Overstall, A., Woods, D. and Adamou, M. (2020) acebayes: An R package for Bayesian optimal design of experiments via approximate coordinate exchange. *Journal of Statistical Software* **95**, 1–33.

Overstall, A. M. and Woods, D. C. (2017) Bayesian design of experiments using approximate coordinate exchange. *Technometrics* **59**, 458–470.

Pawitan, Y. (2013) *In all likelihood*. Oxford University Press.

Rainforth, T., Foster, A., Ivanova, D. and Smith, F. (2023) Modern Bayesian experimental design. arXiv:2302.14545.

Robert, C. (2007) *The Bayesian Choice*. Springer 2nd edn.

Russell, K. G. (2019) *Design of experiments for generalized linear models*. Chapman and Hall/CRC Press.

Ryan, E., Drovandi, C., McGree, J. and Pettitt, A. (2016) A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review* **84**, 128–154.

Smucker, B. and Drew, N. (2015) Approximate model spaces for model-robust experiment design. *Technometrics* **57**, 54–63.

Stigler, S. (2016) *The Seven Pillars of Statistical Wisdom*. Harvard University Press.

Tsai, P.-W. and Gilmour, S. (2010) A general criterion for factorial designs under model uncertainty. *Technometrics* **52**, 231–242.

Tuo, R. and Wu, C. (2015) Efficient calibration for imperfect computer models. *Annals of Statistics* **43**, 2331–2352.

Waite, T. and Woods, D. (2022) Minimax efficient random experimental design strategies with application to model-robust design for prediction. *Journal of the American Statistical Association* **117**, 1452–1465.

Walker, S. (2013) Bayesian inference with misspecified models. *J. Statist. Plan. Inference* **143**, 1621–1633.

Weiser, C. (2016) *Methods for Multivariate Quadrature*. mvQuad R package version 1.0-6.

West, M. (1987) On scale mixture of normal distributions. *Biometrika* **74**, 646–648.

Wiens, D. (2015) Robustness of design. In *Handbook of Design and Analysis of Experiments* (eds. A. Dean, M. D. Morris, J. Stufken and D. Bingham) 719–754. Chapman & Hall.

Wong, R., Storlie, C. and Lee, T. (2017) A frequentist approach to computer model calibration. *J. R. Stat. Soc. B* **79**, 635–648.