Statistical Science
2017, Vol. 32, No. 3, 385–404
DOI: 10.1214/16-STS604

Principles of Experimental Design for Big Data Analysis

Christopher C. Drovandi, Christopher C. Holmes, James M. McGree, Kerrie Mengersen, Sylvia Richardson and Elizabeth G. Ryan

Abstract. Big Datasets are endemic, but are often notoriously difficult to analyse because of their size, heterogeneity and quality. The purpose of this paper is to open a discourse on the potential for modern decision theoretic optimal experimental design methods, which by their very nature have traditionally been applied prospectively, to improve the analysis of Big Data through retrospective designed sampling in order to answer particular questions of interest. By appealing to a range of examples, it is suggested that this perspective on Big Data modelling and analysis has the potential for wide generality and advantageous inferential and computational properties. We highlight current hurdles and open research questions surrounding efficient computational optimisation in using retrospective designs, and in part this paper is a call to the optimisation and experimental design communities to work together in the field of Big Data analysis.

Key words and phrases: Active learning, Big Data, dimension reduction, experimental design, sub-sampling.

1. INTRODUCTION

In this "Big Data" age, massive volumes of data are collected from a variety of sources at an accelerating pace. Traditional measurements and observations are now complemented by a wide range of digital data obtained from images, audio recordings and other sensors, and electronic data that are often available as real-time data streams. These are further informed by domain-specific data sources such as multi-source time series in finance, spatio-temporal monitors in the neurosciences and geosciences, internet and social media in marketing and human systems and "omic" information in biological studies.

Many of these data sets have the potential to provide solutions to important problems in health, science, sociology, engineering, business, information technology and government. However, the size, complexity and quality of these data sets often makes them difficult to process and analyse using standard statistical methods or equipment. It is computationally prohibitive to store and manipulate these large data sets on a single desktop computer and one may instead require parallel or distributed computing techniques that involve the use of hundreds or thousands of processors. Similarly, the analysis of these data often exceeds the capacity of standard computational and statistical software platforms, demanding new technological or methodological solutions. This motivates the development of tailored statistical methods that not only address the inferential question of interest, but also account for the inherent characteristics of the data, address potential

Christopher C. Drovandi is Senior Lecturer, James M. McGree is Associate Professor, Kerrie Mengersen is Professor, School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia, 4000 (e-mail: c.drovandi@qut.edu.au, james.mcgree@qut.edu.au, k.mengersen@qut.edu.au). Christopher C. Holmes is Professor of Biostatistics, Department of Statistics, and Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom, OX1 3TG (e-mail: cholmes@stats.ox.ac.uk). Sylvia Richardson is Professor, MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, United Kingdom, CB2 0SR (e-mail: sylvia.richardson@mrc-bsu.cam.ac.uk). Elizabeth G. Ryan is Statistician, Biostatistics & Health Informatics Department, King's College London, United Kingdom, SE5 8AF (e-mail: elizabeth.ryan@kcl.ac.uk).

biases and data gaps and appropriately adjust for the methods used to deal with the storage and analysis of the data.

A number of break-through approaches have emerged to address these challenges in managing, modelling and analysing Big Data. With respect to data management, the most popular current approaches employ a form of "divide-and-conquer" or "divide-and-recombine" (e.g., Xi et al., 2010, Guhaa et al., 2012) in which subsets of the data are analysed in parallel by different processors and the results are then combined. Similar approaches have also been promoted, such as "consensus Monte Carlo" (Scott, Blocker and Bonassi, 2013) and "bag of little bootstraps" (Kleiner et al., 2014), while others have studied the properties of Markov chain Monte Carlo (MCMC) subsampling algorithms (Bardenet, Doucet and Holmes, 2014, 2015).

With respect to modelling, the focus has turned from traditional statistical models to more scalable techniques that can more successfully accommodate the large sample sizes and high dimensionality. Some popular classes of scalable methods are based on dimension reduction such as principal components analysis (PCA) and its variants (Kettaneha, Berglund and Wold, 2005, Elgamal and Hefeeda, 2015), clustering (Bouveyron and Brunet-Saumard, 2014), variable selection via independence screening (Fan and Lv, 2008, Fan, Feng and Rui Song, 2011) and least angle regression (Efron et al., 2004). Other methods have been developed for specific types of data, such as sequential updating for streaming data (Schifano et al., 2016) or sketching (Liberty, 2013). Many popular statistical software packages such as R are also starting to include libraries of models for Big Data (Wang et al., 2015). The development of these methods represents an active point of intersection in both the statistical and machine learning communities (Leskovec, Rajaraman and Ullman, 2014) under the umbrella of Data Science.

Finally, the library of computational algorithms for the analysis of Big Data has also been multi-focused. Because of the size of the data, traditional estimation methods have been overshadowed by optimisation algorithms such as gradient descent and stochastic approximations (Liang et al., 2013, Toulis, Airoldi and Renni, 2014) and a wide variety of extensions and alternatives (Fan, Han and Liu, 2014, Cichosz, 2015, Suykens, Signoretto and Argyriou, 2015). Many algorithms also exploit sparsity in high-dimensional data to improve speed, efficiency and scalability of algorithms; see, for example Hastie, Tibshirani and Friedman (2009).

Summaries of these technological, methodological and computational approaches can be found in a number of excellent reviews (e.g., Fan, Han and Liu, 2014, Wang et al., 2015). Reviews of discipline-specific methods for analysing Big Data are also emerging (e.g., Yoo, Ramirez and Juan Liuzzi, 2014, Gandomi and Haider, 2015, Oswald and Putka, 2015). Despite the highlighted advantages, almost all of these authors concur that substantial challenges still remain. For example, Fan, Han and Liu (2014) identify three ongoing challenges: dealing adequately with accumulation of errors (noise) and spurious patterns in high-dimensional data; continuing to improve computational and algorithmic efficiency and stability; and accommodating heterogeneity, experimental variations and statistical biases associated with combining data from different sources using different technologies. Indeed, given the acceleration of size and diversity of data, it could be argued that these will remain as stumbling blocks for the foreseeable future.

In this paper, we explore an alternative approach that has the potential to circumvent or overcome many of these issues. Our approach is targeted toward applications of regression models with large N number of observations and small to moderate p predictors, so-called "tall data" situations (see also Bardenet, Doucet and Holmes, 2015 and Xi et al., 2010). We suggest that, depending on the aim of the analysis, one could adopt an optimal experimental design perspective whereby instead of (or as well as) analysing all of the data, a retrospective sample set is drawn in accordance with a sampling plan or experimental design, based on an identified statistical question and corresponding utility function. The analyses and inferences are then based on this designed sample. This allows the analyst to consider an ideal experiment or sample to answer the question of interest and then "lay" that experiment over the data. Thus, the Big Data management challenge becomes one of being able to extract the required design points; the modelling problem reduces to a designed analysis with reduced noise and less potential for spurious correlations and patterns relative to a randomly selected sub-sample of the same size.

There are several Big Data inferential goals for which this approach might be applicable. Goals for which design principles and corresponding utility functions are well established include estimation and testing of parameters and distributions, prediction, identification of relationships between variables and variable selection. Other aims include identification of subgroups and their characteristics, dimension reduction and model testing.

The suggested approach can also be considered as a targeted way of undertaking sampling in divide-and-conquer algorithms or for "sequential learning" in which a given design is applied to incoming data or new data sets until the question of interest is answered with sufficient precision or a pre-determined criterion is reached. It can also be used for evaluating the quality of the data, including potential biases and data gaps, since these will become apparent if the required optimal or near-optimal design points cannot be extracted from the data.

Finally, it is worth emphasising that this approach is a first exploration into the potential for retrospective experimental design for improved Big Data analysis. Many open research questions and challenges exist, not least of which is the need for new computational optimisation methods coupled to design criteria that can deliver a targeted sample set in a time compatible with that of a randomised sampling strategy.

2. BRIEF OVERVIEW OF EXPERIMENTAL DESIGN

In this section, we provide a brief introduction to the principles of optimal experimental design that are relevant to our approach, referring the interested reader to Appendix A for a more extensive overview.

The design of experiments is an example of decision analysis where the decision is to select the optimal experimental settings, $\mathbf{d}$, under the control of the investigator in some design space of options, $\mathbf{d} \in \mathsf{D}$. This is to maximise the expected return as quantified through a known utility function, $U(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y})$, that depends on some, possibly unknown, state of the world $\boldsymbol{\theta} \in \Theta$ and on a potential future dataset $\mathbf{y} \in \mathsf{Y}$ that may be observed when design $\mathbf{d}$ is applied. For example, in a regression analysis with continuous response Y, measurement covariates X, and where the study objective is to learn about the parameters $\boldsymbol{\theta}$ of a mean regression function, $E[Y] = f(X; \boldsymbol{\theta})$; then the design space might be points in X with $\mathbf{d} \in \mathsf{D} \subseteq \mathsf{X}$, and the utility function might be based on the variance of an unbiased estimator $\hat{\boldsymbol{\theta}} = S(Y, X)$ that targets the true unknown $\boldsymbol{\theta}$.

Following the Savage axioms (Savage, 1972), the coherent way to proceed is to select the design that maximises the expected utility,

$$\begin{aligned}\mathbf{d}^* &= \arg\max_{\mathbf{d}\in\mathsf{D}} E_{\mathsf{Y},\Theta}\{U(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y})\} \\ &= \arg\max_{\mathbf{d}\in\mathsf{D}} \int_{\mathsf{Y}} \int_{\Theta} U(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y}) p(\mathbf{y}|\mathbf{d}, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\mathbf{y}.\end{aligned} \tag{1}$$

In classical experimental design, the utility is often a scalar function of the Fisher information matrix, which already considers the expectation with respect to the future data $\mathbf{y}$, and in this case we can write the utility as $U(\mathbf{d}, \boldsymbol{\theta})$ and the integral over $\mathbf{y}$ is no longer required. Further, if the model parameter $\boldsymbol{\theta}$ is assumed known then the problem reduces to an optimization task over the design space. When $\boldsymbol{\theta}$ is unknown the expected utility can be considered with respect to the distribution of $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$, which is a probability measure that quantifies the decision maker's current state of uncertainty on the unknown value of $\boldsymbol{\theta}$. This is often referred to as a pseudo-Bayesian design, as the prior information $p(\boldsymbol{\theta})$ is discarded upon the collection of the actual data.

In a fully Bayesian experimental design, the utility function is often some functional of the posterior distribution, $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d})$. For example, a common parameter estimation utility is $U(\mathbf{d}, \mathbf{y}, \boldsymbol{\theta}) = \log p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d}) - \log p(\boldsymbol{\theta})$, which is the Shannon information gain. Integrating with respect to $\boldsymbol{\theta}$ produces the Kullback–Leibler divergence (KLD) between the prior and the posterior, $U(\mathbf{d}, \mathbf{y}) = \mathrm{KLD}(p(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d}))$. If the KLD can be computed or approximated directly, the integral over $\boldsymbol{\theta}$ is not required. In this case, the expected utility is formed by integrating over the prior predictive distribution, $p(\mathbf{y}|\mathbf{d})$. Integrals are typically approximated by Monte Carlo methods (see, e.g., Drovandi and Tran, 2016).

Of relevance to what follows, in some experimental design situations one may not be able to sample at specific design points or regions, so that D is restricted, in which case "design windows" or "sampling windows" may be required. These consist of a range of near optimal designs and represent regions of planned sub-optimality. Examples of the use of sampling windows include the design of population pharmacokinetic studies (e.g., Ogungbenro and Aarons, 2007, Duffull et al., 2012), which consisted of specific sampling time intervals.

3. EXPERIMENTAL DESIGN IN THE CONTEXT OF BIG DATA

As motivation, we consider a general regression set up where the response data $\mathbf{Y} \in \mathsf{Y}^N$ consists of N observations and the ith response $\mathbf{Y}_i \in \mathsf{Y} \subseteq \mathbb{R}^m$ is the realisation of an m dimensional random variable. Covariate or predictor information is provided in the matrix $\mathbf{X} \in \mathsf{X}^N$ where the ith row is $\mathbf{X}_i \in \mathsf{X} \subseteq \mathbb{R}^p$ where p is the number of predictors. We assume that N is very large and that p is small relative to N. Our objective is to avoid the analysis of the Big Data of size N by selecting a subset of the data of size n_d using the principles of optimal experimental design where the goal

of the analysis is pre-defined. Below we outline a sequential design approach that can achieve this in a sub-optimal but computationally feasible manner and then point to some possible extensions.

3.1 The Algorithm

At a high level, the experimental design principles described in Section 2 can be applied directly to a Big Dataset in order to obtain a sub-sample. Here, we consider a generic procedure inspired by sequential experimental design in order to obtain a close-to-optimal sub-sample of the data with respect to a pre-defined goal of the analysis. This is shown in Algorithm 1. There are two main motivations for our sequential approach: (1) iteratively gain information so that in subsequent iterations more informative data can be extracted, and (2) an optimal design problem only needs to be solved for a single observation at each iteration. In the algorithm, $\mathbf{d} \in \mathsf{D} \subseteq \mathsf{X}$ represents some or all values of the covariates for a hypothetical single observation. We denote as $\mathbf{x} \in \mathsf{X}$ values for the covariates for a single observation that is actually present in the dataset. Let $\mathbf{x}_s \in \mathsf{D}$ be the covariate values for a single observation in the dataset that correspond to the same covariates in $\mathbf{d}$. We denote the observed response corresponding to $\mathbf{x}$ as $\mathbf{y}$, re-defining the notation $\mathbf{y}$ used in Section 2.

Algorithm 1 Proposed algorithm to subset Big Data using experimental design methodology

1: Use a training sample of size n_t to obtain $\hat{\boldsymbol{\theta}}$ or to form a prior distribution $p(\boldsymbol{\theta})$. Set $n_c = n_t$.
2: **while** $n_c \leq n_d$ (where n_c is the current sample size, and n_d is the desired sample size) or when the goal of the analysis is not met **do**
3: Solve the optimisation problem $\mathbf{d}^* = \arg\max_{\mathbf{d}\in\mathbf{D}} E\{U(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y}_\mathbf{d})\}$. Note that $U(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y}_\mathbf{d})$ may not depend on $\boldsymbol{\theta}$ and/or $\mathbf{y}_\mathbf{d}$ depending on the utility function selected.
4: Find $\mathbf{x}$ in the remaining dataset that has not already been sampled such that $\|\mathbf{x}_s - \mathbf{d}^*\|$ is minimised. Take the corresponding observation $\mathbf{y}$. This step may be performed multiple times to sub-sample a batch of data of size m, $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^m$. Increase the size of the sub-sample, $n_c = n_c + m$.
5: Add $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^m$ into the data subset and re-estimate $\hat{\boldsymbol{\theta}}$ or update the prior $p(\boldsymbol{\theta})$ using all available data in the subset. Remove the data $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^m$ from the original dataset.
6: **end while**

We now denote the potential future observation collected at design $\mathbf{d}$ as $\mathbf{y}_\mathbf{d}$.

The objective is to first solve a design optimisation problem with the utility function incorporating the goal of the analysis (e.g., parameter estimation), which produces an optimal $\mathbf{d}^*$. It is important to note that this design optimisation problem is informed by the data currently in the sub-sample in the form of a point estimate $\hat{\boldsymbol{\theta}}$ or a "prior" distribution $p(\boldsymbol{\theta})$, which is a posterior conditional on the data sampled thus far. Given that $\mathbf{d}^*$ is unlikely to be exactly present in the data, as a pragmatic approach we propose to find the $\mathbf{x}$ in the remaining dataset that has not been sampled that minimises the distance $\|\mathbf{x}_s - \mathbf{d}^*\|$ between the relevant covariate values of each observation and the optimal design $\mathbf{d}^*$. Finally, take the corresponding $\mathbf{y}$ and update the information we have about the parameter $\boldsymbol{\theta}$.

It is interesting to note that when selecting a sub-sample of size n_d from the Big Data of size N, the optimal search would involve a comparison across all of the $\binom{N}{n_d}$ potential designs, which is computationally prohibitive. Hence, we propose to solve an approximate, but computable design problem, by first searching over all designs $\mathbf{d}$ in D, and then subsequently searching in the Big Dataset for the best matching collection of samples $\mathbf{x}$ minimising the distance to the approximating design solution $\mathbf{d}^*$. If at each step of our sequential design process the utility function $U(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y})$ and model $p(\boldsymbol{\theta}, \mathbf{y}, \mathbf{x})$ are "smooth" in the design space $\mathbf{d}$, meaning that for a small change in the design we can expect a small change in the expected utility, then for Big Data we can expect to lose little information from using this computable approximation.

3.2 Algorithm Discussion

For classical analysis problems, the training sample is an important component of the algorithm, since it affects the reliability of the parameter estimates. The training sample size n_t is likely to depend on the quality of the data available and the complexity of the data analysis that is to be performed. In the context of a Bayesian analysis, the training sample is used to form a prior distribution. The more data used in the training sample, the more precisely parameter estimates (classical) or parameters (Bayesian) can be determined, which helps to facilitate more optimal choices of data to take from the original dataset during subsequent iterations. However, the training sample is not optimally extracted from the data and, therefore, one may want to limit its size. We suggest that the training data can be

selected on the basis of a design with generally "good" properties, for example, balance, orthogonality, etc.

Line 3 of Algorithm 1 is the most challenging. If the number of design variables (covariates) is small enough, then a simple discrete grid search might suffice to obtain a near-optimal design. For more complex design spaces, it may be necessary to perform some numerical optimisation procedure. Some approaches that have been used in the design literature are the exchange algorithm (e.g., Fedorov, 1972), numerical quadrature (e.g., Long et al., 2013), MCMC simulation (e.g., Müller, 1999), or sequential Monte Carlo methods (e.g., Kück, de Freitas and Doucet, 2006, Amzal et al., 2006). This step of the algorithm may be computationally intensive and is currently the largest stumbling block for the general applicability of our approach. However, we demonstrate in several case studies in Section 5 that our approach is applicable in a number of nontrivial settings. Nonetheless, there is interest in developing new approaches to accelerate this step, which is an on-going research direction in the experimental design literature.

To reduce the number of design optimisations that need to be performed, we may extract from the Big Data a cluster of m data points where the $\mathbf{x}_s$ is closest to $\mathbf{d}^*$ (Line 4 in Algorithm 1). The optimal value of m is a trade-off between computational cost and information loss, which we do not explore here. In other applications using standard design criteria, such as D-optimality, means that the optimal design may be simple to determine (e.g., Pukelsheim, 1993, Tan and Berger, 1999, Ryan, Drovandi and Pettitt, 2015).

In the examples we consider later, we find that the Euclidean distance for the norm $\|\mathbf{x}_s - \mathbf{d}^*\|$ on standardised covariates works reasonably well. It should be noted that the user is free to choose an appropriate norm for their data.

To reduce the computational burden to implement Line 4 in Algorithm 1, the data set may need to be split up amongst multiple CPUs using a framework such as Hadoop. The minimisation problem (Line 4) can be performed on each of the CPUs, and then a minimisation can be performed over the results of all of the CPUs. This is similar to the "split-and-conquer" approach (e.g., Xi et al., 2010). Rather than finding an optimal design that consists of fixed points, as in Line 3 of Algorithm 1, we could instead find sampling windows, since the optimal design points $\mathbf{d}^*$ may not be present in the data set, and so we may require regions of near optimal designs. Moreover, in Line 2, one could instead run the algorithm until the utility function reached a certain pre-specified value (e.g., a certain level of precision).

A similar design algorithm is considered in follow-up studies, where only a small proportion of subjects are measured on the second occasion to reduce costs (Karvanen, Kulathinal and Gasbarra, 2009, Reinikainen, Karvanen and Tolonen, 2016). In these studies, the objective is to determine the best n out of N individuals to consider for the next follow-up. This difficult computational problem is solved by Karvanen, Kulathinal and Gasbarra (2009) and Reinikainen, Karvanen and Tolonen (2016) in a greedy manner by sequentially adding participants for the next follow-up that lead to the largest improvement in expected information gain until n subjects are selected. Our approach is different to this. First, it is not feasible in our context to scan through the entire Big Data to find the next observation that leads to the largest improvement in expected or observed utility. Instead, we solve an optimal design problem first, and then we simply need to find the design in the Big Data that is close to this optimal design. Second, our design approach uses the information from each selected data point to make better decisions about which design (and observation) to include next. In contrast, the applications of Karvanen, Kulathinal and Gasbarra (2009) and Reinikainen, Karvanen and Tolonen (2016) are static design problems (parameter values are not updated) that are solved in an approximate sequential manner. Third, our approach allows for the detection of potential holes in the data. Finally, our framework is more general as it is inclusive of both classical and Bayesian frameworks, whereas Karvanen, Kulathinal and Gasbarra (2009) and Reinikainen, Karvanen and Tolonen (2016) only consider classical designs.

3.3 Computational Overheads

The key challenge in the practical application of our approach is being able to implement algorithms, such as Algorithm 1, in a computational time such that the extra effort of obtaining design points does not outweigh the information benefits. That is, if n_d is the maximum sample size available through the designed approach given the constraints in compute infrastructure and runtime, and n_s is the corresponding sample size from using random subset selection. For our approach to be worthwhile, we require that the expected utility of the designed approach learned from n_d samples to be higher than the expected utility using n_s random samples, where typically $n_s > n_d$. Clearly, this

will be study dependent but, given the potential benefits shown below, it also motivates the need for new computational optimisation strategies targeted to general design criteria for Big Data analysis.

4. SIMULATION STUDY

Here, we apply our methods to data that are simulated from a logistic regression model that contains two covariates, x_1 and x_2. The following logistic model is used to describe the binary response variable $Y_i \sim \text{Binary}(\pi_i)$ where $\text{logit}(\pi_i) = \theta_0 + \theta_1 x_{1,i} + \theta_2 x_{2,i}$. We assume that the true parameter values are $(\theta_0, \theta_1, \theta_2) = (-1, 0.3, 0.1)$ and that the sample size for the full data set is $N = 10{,}000$. Although N is small for Big Data, it will serve to illustrate and motivate the essential features of our approach. We simulate the covariate values of x_1 and x_2 for each observation from a multivariate normal distribution with a mean vector of zeros and three different options for the covariance matrix:

1. $\left[\begin{smallmatrix} 3 & 0 \\ 0 & 3 \end{smallmatrix}\right]$, no dependence between covariates;
2. $\left[\begin{smallmatrix} 3 & 1.5 \\ 1.5 & 3 \end{smallmatrix}\right]$, positive correlation between covariates;
3. $\left[\begin{smallmatrix} 3 & -1.5 \\ -1.5 & 3 \end{smallmatrix}\right]$, negative correlation between covariates.

Here, we are interested in finding the "best" $n_d = 1000$ observations from the full data set to most precisely estimate the model parameters $(\theta_0, \theta_1, \theta_2)$. We demonstrate the use of both classical and Bayesian sequential design methods to subset the data.

The design variable for Line 3 of Algorithm 1 is given by potential values for the two covariates for a single observation, $\mathbf{d} = (x_1, x_2)$.

4.1 Classical Approach

For Line 1 of Algorithm 1, we select $n_t = 20$ training samples randomly from the full data and determine the MLE of the parameter, $\hat{\boldsymbol{\theta}}$. We denote the data present in the subset currently as $\mathcal{D}_s$, which consists of response and covariate values. The utility function we use for Line 3 is given by

$$U(\mathbf{d}) = \left|\mathcal{O}(\mathcal{D}_s, \hat{\boldsymbol{\theta}}) + \mathcal{I}(\mathbf{d}, \hat{\boldsymbol{\theta}})\right|,$$

where $\mathcal{O}(\mathcal{D}_s, \hat{\boldsymbol{\theta}})$ is the observed information matrix based on data collected so far and $\mathcal{I}(\mathbf{d}, \hat{\boldsymbol{\theta}})$ is the expected information matrix if we apply the design $\mathbf{d}$ for the next observation. For the optimal design procedure in Line 3 of Algorithm 1, we use a grid search over the design region $[-5, 5] \times [-5, 5]$. The grid consists of evenly-spaced points that are separated by an increment of 0.1.

Once the optimal design $\mathbf{d}^*$ is estimated from Line 3, we use the Euclidean distance in Line 4 to determine the next observation to take from the remaining Big Data. Following this, the parameter estimate $\hat{\boldsymbol{\theta}}$ is updated using maximum likelihood in Line 5. Then the process is repeated. Conditional on the training sample, the overall subsetting procedure is deterministic so we only perform the procedure once and obtain a single subset of size 1000. We denote the subsetted data generated from our design procedure as $\mathcal{D}_s^d$.

For comparison purposes, we generated 10,000 subsets of size 1000 randomly, with each subset denoted by $\mathcal{D}_s^{r_i}$ for $r = 1, \ldots, 10{,}000$. The final estimates for $(\theta_0, \theta_1, \theta_2)$ based on $\mathcal{D}_s^d$ are given in Table 1. Table 1 also displays $|\mathcal{O}(\mathcal{D}_s^d, \hat{\boldsymbol{\theta}})|$ where for notational simplicity the MLE $\hat{\boldsymbol{\theta}}$ is always based on the dataset present as the first argument of the observed information matrix. The largest observed information obtained out of the 10,000 randomly sampled data subsets is displayed in the final column of Table 1. Each row of Table 1 corresponds to the different correlation structures that were investigated for the simulated covariate data. Table 2 contains the estimates for $(\theta_0, \theta_1, \theta_2)$ (and their associated variance–covariance matrix) based on the full data under each of the covariance structures for the covariate data.

TABLE 1

Estimated $\boldsymbol{\theta}$ values [where $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)$] and the observed information value for the sub-sample of size $n_d = 1000$ obtained using the principled design approach ($|\mathcal{O}(\mathcal{D}_s^d, \hat{\boldsymbol{\theta}})|$). The last two columns contain the median (IQR) and maximum utility function values that were obtained from 10,000 randomly drawn sub-samples of data, each of size $n_d = 1000$

Covariance structure	$\hat{\boldsymbol{\theta}}$ based on $\mathcal{D}_s^d$	$\lvert\mathcal{O}(\mathcal{D}_s^d, \hat{\boldsymbol{\theta}})\rvert$	median ($\lvert\mathcal{O}(\mathcal{D}_s^{r_i}, \hat{\boldsymbol{\theta}})\rvert_{i=1}^{10,000}$) (IQR)	max($\lvert\mathcal{O}(\mathcal{D}_s^{r_i}, \hat{\boldsymbol{\theta}})\rvert_{i=1}^{10,000}$)
No correlation	$(-1.03, 0.34, 0.11)$	2.8×10^8	$5.3(4.9, 5.8) \times 10^7$	9.0×10^7
Positive correlation	$(-1.01, 0.32, 0.08)$	1.4×10^8	$3.9(3.6, 4.2) \times 10^7$	6.0×10^7
Negative correlation	$(-0.94, 0.41, 0.17)$	8.0×10^7	$4.1(3.8, 4.5) \times 10^7$	6.6×10^7

TABLE 2
Estimated $\boldsymbol{\theta}$ values [where $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)$] and their covariance using the full data sets that were simulated under different covariance structures of **X**

Covariance structure of X	$\hat{\boldsymbol{\theta}}$	Estimated covariance of $\hat{\boldsymbol{\theta}}$
No correlation	$(-0.98, 0.28, 0.08)$	$\begin{bmatrix} 5.4\times10^{-4} & -6.5\times10^{-5} & -2.1\times10^{-5} \\ -6.5\times10^{-5} & 1.9\times10^{-4} & 8.5\times10^{-6} \\ -2.1\times10^{-5} & 8.5\times10^{-6} & 1.8\times10^{-4} \end{bmatrix}$
Positive correlation	$(-1.02, 0.30, 0.08)$	$\begin{bmatrix} 5.6\times10^{-4} & -7.3\times10^{-5} & -1.9\times10^{-5} \\ -7.3\times10^{-5} & 2.6\times10^{-4} & -1.2\times10^{-4} \\ -1.9\times10^{-5} & -1.2\times10^{-4} & 2.4\times10^{-4} \end{bmatrix}$
Negative correlation	$(-1.00, 0.29, 0.08)$	$\begin{bmatrix} 5.4\times10^{-4} & -6.4\times10^{-5} & -1.9\times10^{-5} \\ -6.4\times10^{-5} & 2.4\times10^{-4} & 1.1\times10^{-4} \\ -1.9\times10^{-5} & 1.1\times10^{-4} & 2.3\times10^{-4} \end{bmatrix}$

From Table 1, it can be seen that the estimates of $(\theta_0, \theta_1, \theta_2)$ that were based on the subsets of data that were obtained via the principled design approach are quite close to the true parameter values, as well as the values that were obtained using the full data set (displayed in Table 2). Only a small amount of precision for the parameter estimates was lost by using the subset of data rather than the full dataset (Tables 1 and 2). This indicates that our method is fairly accurate in this example for subsetting the data so that our model parameters can be estimated precisely. One run of the optimal design process had a similar computational time to running 10,000 random subsets (approximately 40 seconds). However, the determinants of the observed information from the subsets of data that were obtained via our design approach were higher than the determinant of the observed information obtained from 10,000 randomly selected data subsets of the same sample size (Table 1). This highlights the potential of our designed approach.

However, the extra time used to determine the designed subset could be used to analyse a larger random sample. We investigated different sample sizes for the subsets that were selected randomly from the full data set and ran 10,000 replicates for each sample size. The results are displayed in Figure 1. For the simulation studies where the data were generated using a covariate structure with no correlation, or with positive correlation, the randomly selected data subset size had to be roughly doubled to obtain a higher utility (overall) than for the designed approach. For the simulation study where the data were generated using negative correlation between the covariates, the subset size of 1500 showed higher utility than the designed approach. We provide more discussion on the negative correlation case later.

Figure 2 shows the **x** values that minimise the Euclidean distance to the optimal designs at each iteration/time point based on the observed information thus far (and were thus extracted into the subset) against the optimal designs at that iteration. Ideally, these points should be equal and would lie along the 45-degree line in Figure 2. From Figure 2, it appears that there are two support points for both x_1 and x_2, one at either end of the design region. Since the covariates were drawn from a normal distribution, there are not many design values in the data that occur on the boundaries of the design region, and so less than optimal values will have to be chosen for the data subset and the data will be less informative than if the full dataset contained more values on the boundaries of the design region. When the covariates are correlated with one another, there are even less design values on the "corners" or boundaries of the design region and so the full dataset will be generally less informative. We discuss this further in the Bayesian section below.

4.2 Bayesian Approach

For the Bayesian approach, we use an SMC algorithm, similar to that used by Drovandi, McGree and Pettitt (2013), to sequentially generate samples from the posterior as more data are added into the subsample. We place independent normal priors on the parameters each with a mean of 0 and a standard deviation of 5; we do not use any training data in Line 1 of Algorithm 1. For the utility function required in Line

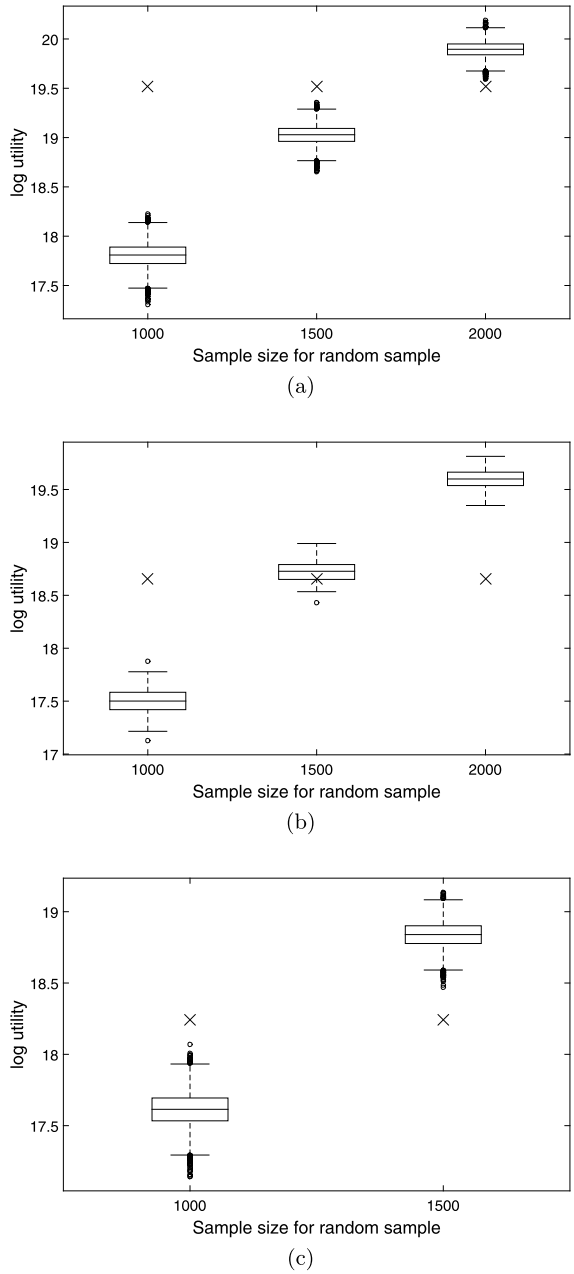


FIG. 1. *Boxplots of the log utility (determinant of the observed information matrix) for* 10,000 *randomly drawn data subsets of various sample sizes (x-axis) to compare against the utility function value of the designed approach (for a data subset size of* 1000; *displayed as a cross in each boxplot), for each of the correlation structures of the covariates*: (a) *no correlation between* x_1 *and* x_2, (b) *positive correlation between* x_1 *and* x_2, (c) *negative correlation between* x_1 *and* x_2.

3, we use

$$U(\mathbf{d}, y) = -\log\det\big(\mathrm{cov}(\boldsymbol{\theta}|\mathbf{d}, \mathcal{D}_s, y)\big),$$

where $\mathcal{D}_s$ represents the data currently in the subset and $y \in \{0, 1\}$ is a possible outcome for the next observation. The expected utility is given by

$$U(\mathbf{d}) = \sum_{y\in\{0,1\}} U(\mathbf{d}, y)\, p(y|\mathbf{d}, \mathcal{D}_s).$$

The quantities inside the summation are estimated using the posterior samples maintained through SMC and additional importance sampling to accommodate the possible outcomes for the next observation (see Drovandi, McGree and Pettitt, 2013 for more details). The optimisation procedure used in Line 3 is the same as that used for the classical approach above. For Line 4, we again use the Euclidean distance to obtain the next observation to add to the subset in Line 5. This process is repeated until 1000 observations are obtained. The final posterior mean estimates from one run of our sequential design approach for $(\theta_0, \theta_1, \theta_2)$ are given in Table 3, along with the estimates based on the full data set.

For one run of our algorithm, the optimal designs and the corresponding covariate values actually extracted from the data are shown in Figure 3 for the three different correlation structures. Most of the optimal design values appear in the "corners" of the design search space. When there is no correlation between the predictors, it is easier to find covariate values that are close to the optimal design values (top row of Figure 3). When there is correlation (middle and bottom rows of Figure 3), the corners of the design space are not as well covered by the data. We found that this was a particular issue when there was negative correlation. It can be seen from the bottom row of Figure 3 that the optimal design requested by the algorithm was often in the top right corner but there was no data there to satisfy this request. Thus, there is a chance that the actual data selected may not have a relatively high utility value. In this respect, below we demonstrate that the optimal design approach can perform worse than a simple random sample. We plan to develop methods to address this issue in future research. In the least, plots as in Figure 3 can be used as an exploratory tool to determine how close the data sub-sample is to the ideal design for the chosen research objectives.

The subset obtained from our method is affected by the Monte Carlo variability of the SMC approximation to the posteriors and also the Monte Carlo variability from the importance sampling procedure to determine the optimal designs. Thus, we repeated our process 1000 times independently. To determine how well our data subsets perform, we compared it to randomly selected datasets of size 1000 from the original dataset.

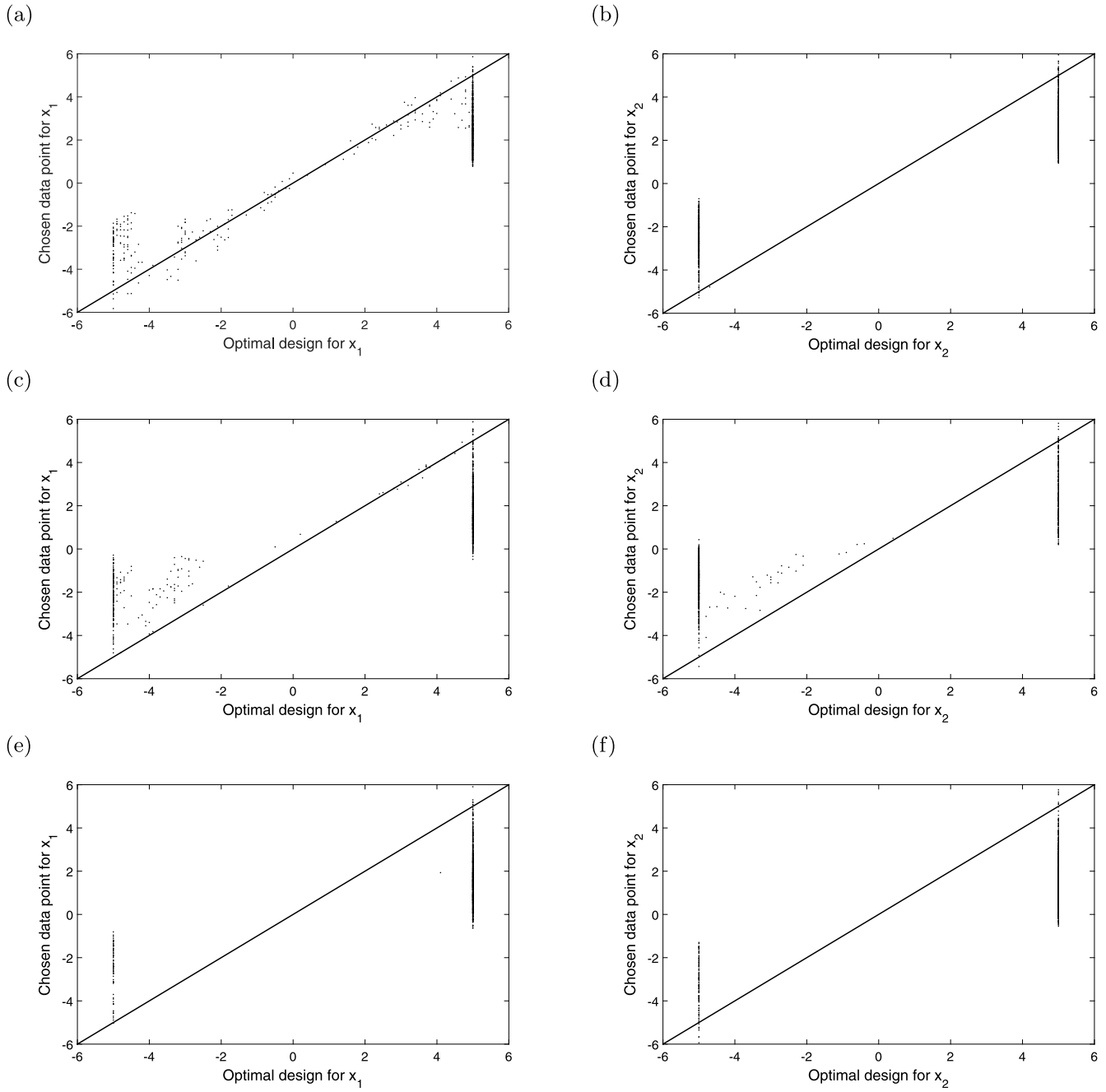


FIG. 2. *Chosen designs for x_1 and x_2 vs optimal designs for x_1 and x_2 where the correlation structures for the covariates are*: (a) *and* (b) *no correlation between x_1 and x_2*; (c) *and* (d) *positive correlation between x_1 and x_2*; (e) *and* (f) *negative correlation between x_1 and x_2. The* 45*-degree line indicates where the selected data points for the subset are equal to the optimal design.*

We obtained 1000 such datasets. For each of these randomly chosen datasets, we estimated the posterior distribution via SMC, which was then used to estimate the utility function. The distribution of the utility values obtained from our designed subsets is compared with that of the random datasets in Figure 4. It can be seen that our data subset outperforms the randomly selected designs as it generally produces a higher utility value than the randomly chosen datasets, except for the negative correlation structure (see earlier discussion). One run of the optimal design process took approximately 22 seconds, whereas investigating 1000 random subsets took approximately 25 minutes. Therefore, we were generally able to obtain a data subset of a par-

TABLE 3
Posterior estimates of $\boldsymbol{\theta}$ [where $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)$] and the observed utility using the data subset obtained via the designed approach and using the full data sets that were simulated under different covariance structures of **X**

Full or subset data	Covariance structure of X	$\hat{\theta}$	Observed utility
Subset	No correlation	$(-1.11, 0.33, 0.11)$	18.9
Full	No correlation	$(-1.02, 0.31, 0.10)$	24.7
Subset	Positive correlation	$(-0.91, 0.27, 0.13)$	19.3
Full	Positive correlation	$(-1.00, 0.31, 0.10)$	24.4
Subset	Negative correlation	$(-1.04, 0.31, 0.15)$	17.3
Full	Negative correlation	$(-1.03, 0.32, 0.12)$	24.6

ticular size that produced higher observed utility in a shorter amount of time.

To determine the sample size savings for the designed approach (compared to random subsets of data), we varied the size of the subsets that were selected randomly from the full data set and repeated this process 100 times. The results are displayed in Figure 5. It was found that the random sample data subset would have to be increased to a size of 1500–2000 to obtain an overall higher utility than for the designed approach, except for the negative correlation structure.

Overall, in this simulation study, analysing a larger random subset would be more efficient than analysing the smaller designed subset. However, this example still demonstrates the potential of the designed approach.

5. CASE STUDIES

The two case studies described here showcase our principled design approach applied to real data. Algorithm 1 is used together with a number of computational algorithms. For the purposes of cohesion and comparison, the first study employs a logistic regression model to predict risk of mortgage default, whereas the second study employs a more challenging mixed effects model. The cases differ with respect to the study aims; variable selection and precise regression parameter estimation. Comparisons are also made with results obtained from analysing the full (Big) data.

To further illustrate our design approach to subsetting Big Data, two additional case studies are provided in Appendices B and C, respectively. The case study in Appendix B is similar in spirit to case study 1 and involves the estimation of regression coefficients of covariates that might influence on-time flight arrivals. The case study in Appendix C highlights that experimental design principles may be useful in some applications for subsetting Big Data without needing to resort to optimal design methods such as those presented in Algorithm 1. This case study involves applying static experimental design principles for performing an ANOVA on a dataset of colorectal cancer patients in Queensland, Australia.

5.1 Case Study 1—Mortgage Default

In this case study, we consider the simulated mortgage defaults data set found here:

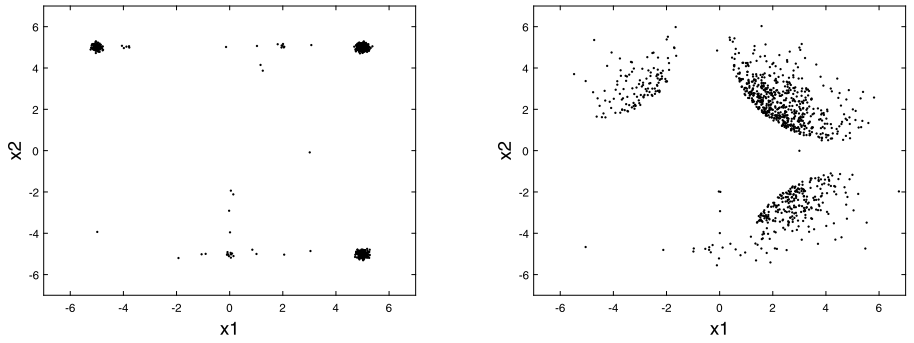
http://packages.revolutionanalytics.com/datasets/.

The scenario is that data have been collected every year for 10 years on mortgage holders, and contains the following variables:

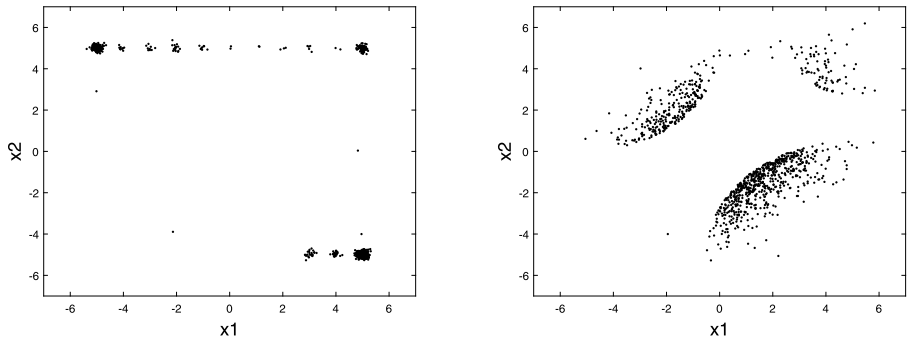
- default: a 0/1 binary variable indicating whether or not the mortgage holder defaulted on the loan (response variable);
- creditScore: a credit rating (x_1);
- yearsEmploy: the number of years the mortgage holder has been employed at their current job (x_2);
- ccDebt: the amount of credit card debt (x_3);
- houseAge: the age (in years) of the house (x_4); and
- year: the year the data were collected.

The proposed model for the binary outcome is the logistic regression model, with the above covariates as main effects (credit rating, years employed, credit card debt and house age) potentially significantly influencing the probability of defaulting. To determine which covariates are useful for prediction, we focus on the default data for the year 2000 which contains 1,000,000 records. We initially allowed all covariates to appear in the model, and obtained prior information about the parameters by extracting a random selection of $n_t = 5000$ data/design points from the full dataset in an initial learning phase.

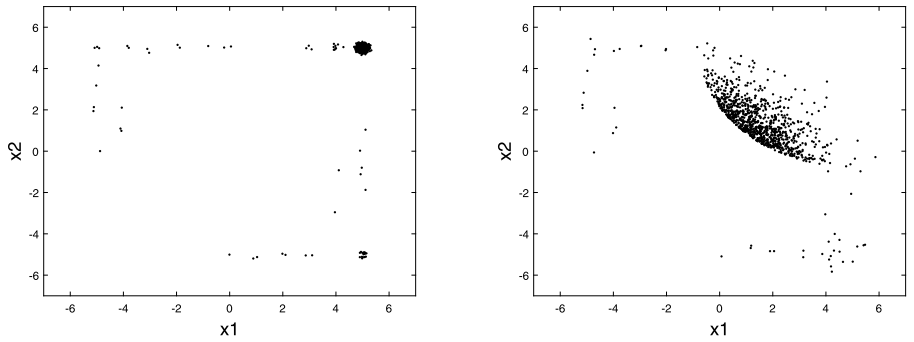
From this initial learning phase, it is useful to develop prior distributions about the model(s) appropriate for data analysis and the corresponding parameter values based on the extracted data. The primary motivation for this is the avoidance of the computational burden associated with continually considering a potentially large dataset within a (full) Bayesian analysis. To facilitate this, maximum likelihood estimates (MLEs) of parameters (and standard errors) were found for all potential models. Prior information



(a) no corr, optimal (b) no corr, selected



(c) +ve corr, optimal (d) +ve corr, selected



(e) −ve corr, optimal (f) −ve corr, selected

FIG. 3. *Left hand column shows the optimal designs* (*with slight jittering*) *selected for one run of Algorithm* 1 *for the Bayesian logistic regression example. The right-hand column shows the corresponding covariate values that were actually selected from the data. Results are shown for different covariance structures of* $\mathbf{X}$ *in the original* (*full*) *data*: (*top row*) *no correlation*, (*middle row*) *positive correlation and* (*bottom row*) *negative correlation*.

about the parameters was then constructed by assuming all parameters follow a normal distribution with the mean being the MLE and the standard deviation being the standard error of the MLE.

The next step was to "value add" to the information gained from the initial learning phase through our sequential design process. To do this, we implemented the SMC algorithm of Drovandi, McGree and Pettitt (2013) to approximate the sequence of target distributions which will be observed as data are extracted from the full data set (see Section 4.2). For Line 3 of Algorithm 1, we used a similar estimation utility to the simulation study in Section 4.2 to select designs which should yield precise estimates of the model parameters, and this utility was approximated via importance sampling. The optimisation procedure we apply in Line 3 is again a simple grid search by considering potential design points based on all combinations of the follow-

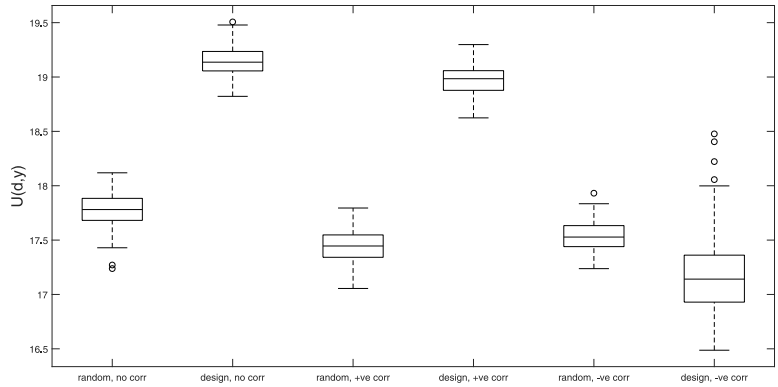


FIG. 4. *Boxplots of the observed utility values obtained for* 1000 *runs of a random design and the optimal sequential design for the Bayesian logistic regression example for the different covariance structures of* **X** *in the original* (*full*) *data.*

ing covariate levels (formed by inspecting the full data set—see Table 4).

This results in the consideration of 2205 potential design points at each iteration of the sequential design process. For Line 4, we find the data point in the remaining Big Data with a design closest to the optimal design in terms of Euclidean distance.

As we are interested in determining which variables are useful for prediction, each time the prior information was updated to reflect the information gained from a new data point, a 95% credible interval was formed for all regression coefficients in the model. If any credible interval was contained within $(-tol, tol)$, then this parameter/variable was dropped from the model. The

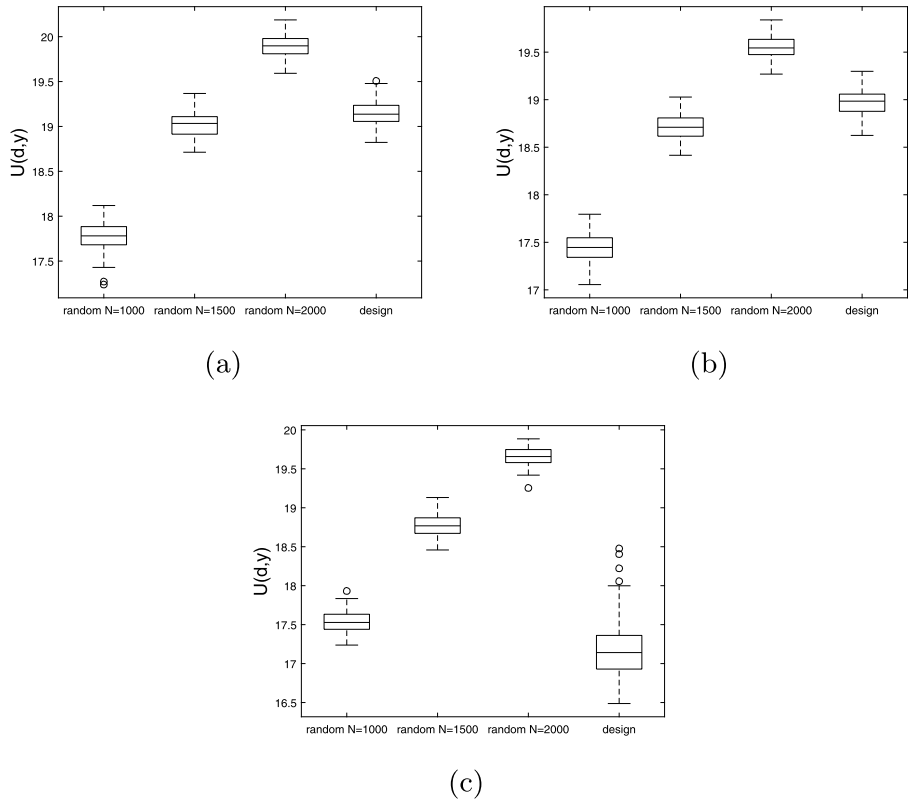


FIG. 5. *Boxplots of the observed utility values obtained for* 100 *runs of randomly drawn data subsets of various size* (*x-axis*) *and the optimal sequential design for the simulation study for the different covariance structures of* **X** *in the original* (*full*) *data*: (a) *no correlation*, (b) *positive correlation*, *and* (c) *negative correlation*.

TABLE 4
Scaled values of covariates in the mortgage case study (credit score, years employed, credit card debit and house age) considered when searching for Bayesian optimal designs

Covariate	Scaled levels
creditscore	$-4, -3, -2, -1, 0, 1, 2, 3, 4$
yearsemploy	$-2, -1, 0, 1, 2, 3, 4$
ccDebt	$-2, -1, 0, 1, 2, 3, 4$
houseAge	$-2, -1, 0, 1, 2$

TABLE 6
Summary of the posterior distribution of the parameters for the full main effects model based on all mortgage default data for the year 2000

Parameter	Mean	SD	2.5th	Median	97.5th
β_0	−11.40	0.13	−11.67	−11.40	−11.16
β_1	−0.42	0.03	−0.48	−0.42	−0.36
β_2	−0.63	0.03	−0.68	−0.63	−0.56
β_3	3.03	0.05	2.94	3.03	3.13
β_4	0.20	0.03	0.12	0.20	0.26

reduced model was then re-fit using SMC based on the appropriate prior information on the parameters (from the initial learning phase) and all sequentially extracted data. This process iterated until 1000 data points had been extracted in this sequential design process. To investigate this methodology, we considered four different values for *tol* (0.25, 0.5, 0.75 and 1).

Table 5 shows the covariates that remained in the model after an additional 1000 data points had been observed for $tol = 0.25, 0.50, 0.75$ and 1.00. The results suggest that if we are only interested in large effect sizes (>1.0), then credit card debit appears to be the only useful covariate. In contrast, if effects larger than 0.25 are deemed important then all variables remain in the model. Most notably, these results agree with the results from fitting the full main effects model based on the full data set; see Table 6. This model was fitted with a prior distribution based on 5000 randomly drawn data points (as in the above described initial learning phase), and sequentially updating these priors using a single data point at a time until all data had been included. The covariate information found here should be useful to lenders, as it seems to indicate that individual information is more informative than property characteristics for determining if someone will default on their mortgage.

TABLE 5
The covariates in the mortgage case study which were deemed useful for prediction based on $tol = 0.25, 0.50, 0.75$ *and* 1.00

tol	Remaining covariates
0.25	x_1, x_2, x_3, x_4
0.50	x_2, x_3
0.75	x_3
1.00	x_3

At each iteration of the sequential design process, the Bayesian optimal design and the corresponding extracted design was recorded. A comparison of the two is shown in Figure 6, by covariate (for $tol = 0.25$). Ideally, one would like to observe a one-to-one relationship between the two designs. Unfortunately, this was not observed in this study. For example, from Figure 6(c) which corresponds to credit card debt, values of around \$11,000 and \$13,000 were found as optimal, but the values extracted from the Big Data set varied between \$5,000 and \$13,000. This suggests that there is potentially a lack of mortgage default data on those with large credit card debts. Intuitively, this might make sense as such individuals may generally not have their loan approved.

For each of the four different values of *tol*, it took approximately 40 minutes to run the learning phase and sequential design process. To explore the computational gains/losses involved in implementing this designed approach, a comparison with randomly selected subsets of the same size was undertaken. The comparison was conducted such that priors from the initial learning phase were formed in the same manner as the designed approach but, for the sequential design process, instead of searching for an optimal design, a design was randomly selected from the data set. The analysis of such subsetted data in general took approximately 2 minutes, which means that around 20 randomly selected data sets could be analysed in the time it took to implement a designed approach. When these 20 random designs were run, only x_1 was removed from any model (across all values of *tol*). Indeed, this only occurred when $tol = 0.75$ or 1. In these cases, x_1 was removed from the model $1/20$ and $20/20$ times, respectively. Such results show that no random design provided more information about the parameter values as that of the designed approach. Thus, despite the designed approach having relatively high computational requirements, the benefits are seen in analysing highly

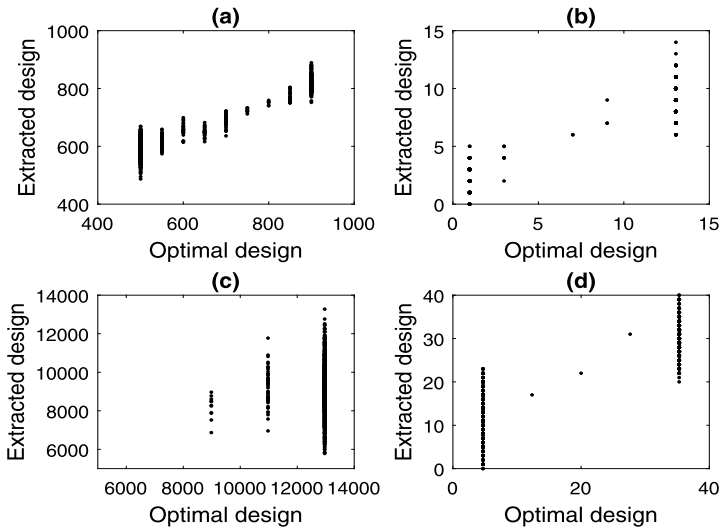


FIG. 6. *Extracted versus optimal design points for the mortgage default case study for tol* $= 0.25$ *for* (a) *credit score*, (b) *years employed*, (c) *credit card debit and* (d) *house age*.

informative data, and thus efficiently addressing analysis aims.

In summary, in this mortgage case study, through analysing a small fraction of the full data set, we were able to determine, with confidence, which covariates appear important for prediction, and also identify potential "holes" in the full data set in regards to our analysis aim.

5.2 Case Study 2

To illustrate the method with a more complex statistical model, we consider an analysis performed on accelerometer data (see, e.g., Trost et al., 2011). Here, 212 participants performed a series of 12 different activities at four different time points, approximately one year apart. The age range of the data was 5 to 18 years old. The purpose of the analysis was to assess the performance of 4 different so-called "cut-points", which are used to predict the type of activity performed based on the output of the accelerometer. The response variable is whether or not the cut-point correctly classifies the activity. Each individual at each time point performed all 12 activities and all 4 cut-points are applied (observations with missing classification responses were discarded). There are roughly 35,000 observations in the dataset. Although this sample size is not as large as in the previous case study, we show that it is sufficient to demonstrate our proposed approach.

A logistic regression mixed effects model was fitted to the data that included age as a continuous covariate (linear in the logit of the probability of correct classification), and the type of activity (12 levels) and the cut-point (4 levels) as factor variables. The model also included all two-way interactions, which resulted in a total of 63 fixed effects parameters. A normal random intercept was included for each participant. For completeness, we assume that the observed classification for the ith observation for subject t is $Y_{ti} \sim \text{Binary}(\pi_{ti})$ with

$$\begin{aligned}
&\text{logit}(\pi_{ti}) \\
&\quad = \beta_0 + b_t + \beta_{\text{age}}\text{age}_{ti} + \sum_{j=1}^{3} \beta^j_{\text{cut}}\text{cut}^j_{ti} \\
&\qquad + \sum_{j=1}^{11} \beta^j_{\text{trial}}\text{trial}^j_{ti} + \sum_{j=1}^{33} \beta^j_{\text{cut,trial}}\text{cut}^j_{ti} \times \text{trial}^j_{ti} \\
&\qquad + \sum_{j=1}^{3} \beta^j_{\text{age,cut}}\text{age}^j_{ti} \times \text{cut}^j_{ti} \\
&\qquad + \sum_{j=1}^{11} \beta^j_{\text{age,trial}}\text{age}^j_{ti} \times \text{trial}^j_{ti},
\end{aligned}$$

where $b_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \phi)$ for $t = 1, \ldots, 212$ (with ϕ being the between subject variance), $i = 1, \ldots, s_t$ where s_t is the number of observations taken on subject t, age_{ti} is the age in years, cut^j_{ti} is a dummy variable defining which cut-point is applied, trial^j_{ti} is a dummy variable defining which trial is applied and the β parameters are the fixed effects. The intercept parameter β_0 relates to an age of 0 years, cut-point 1 and trial 1.

Here, we assume that interest is in estimating the age effect on correct classification (both main and interaction effects, consisting of 15 fixed effects parameters). Thus, for our utility in Line 3 of Algorithm 1 we considered the negative log of the determinant of the posterior covariance matrix for these 15 parameters, and aimed to maximise this utility. For Line 1, we took a pilot dataset consisting of $n_t \approx 500$ observations where a full replicate was taken from different individuals at ages 6 to 18 years with an increment of two years. Then we performed our sequential design strategy to continually accrue data until at least 3000 observations were obtained. Thus, we attempted to obtain a close-to-optimal sub-sample of size $n_d \approx 3000$ to precisely estimate the age related parameters. The optimisation procedure for Line 3 is a simple grid search over the age covariate (between 6 and 18 with 2 year increments) to guide the next selection of data. For Line 4, we took all the data from the individual with the closest age (in terms of Euclidean distance) to the optimal design selected (48 observations when a full replicate is available). Note that we did not force data to be collected from different individuals than what has already been collected in the sub-sample. The optimal design was near the boundaries of the age range, so that naturally the sub-sampled data were usually taken from different individuals. However, a different design strategy could be adopted where data is taken from an individual who is not already present in the sub-sample with the closest age to the optimal design. Ryan, Drovandi and Pettitt (2015) considered Bayesian design for mixed effects models and found that it is not obvious whether to sample a few individuals heavily or sample many individuals sparsely, highlighting the importance of optimal Bayesian design in the context of mixed effects models. Furthermore, the amount of computation required to analyse the extracted data may depend not only on the size of the subset but on how many distinct individuals are sampled. We have not factored this in to our subsetting procedure, but it may be possible to do so.

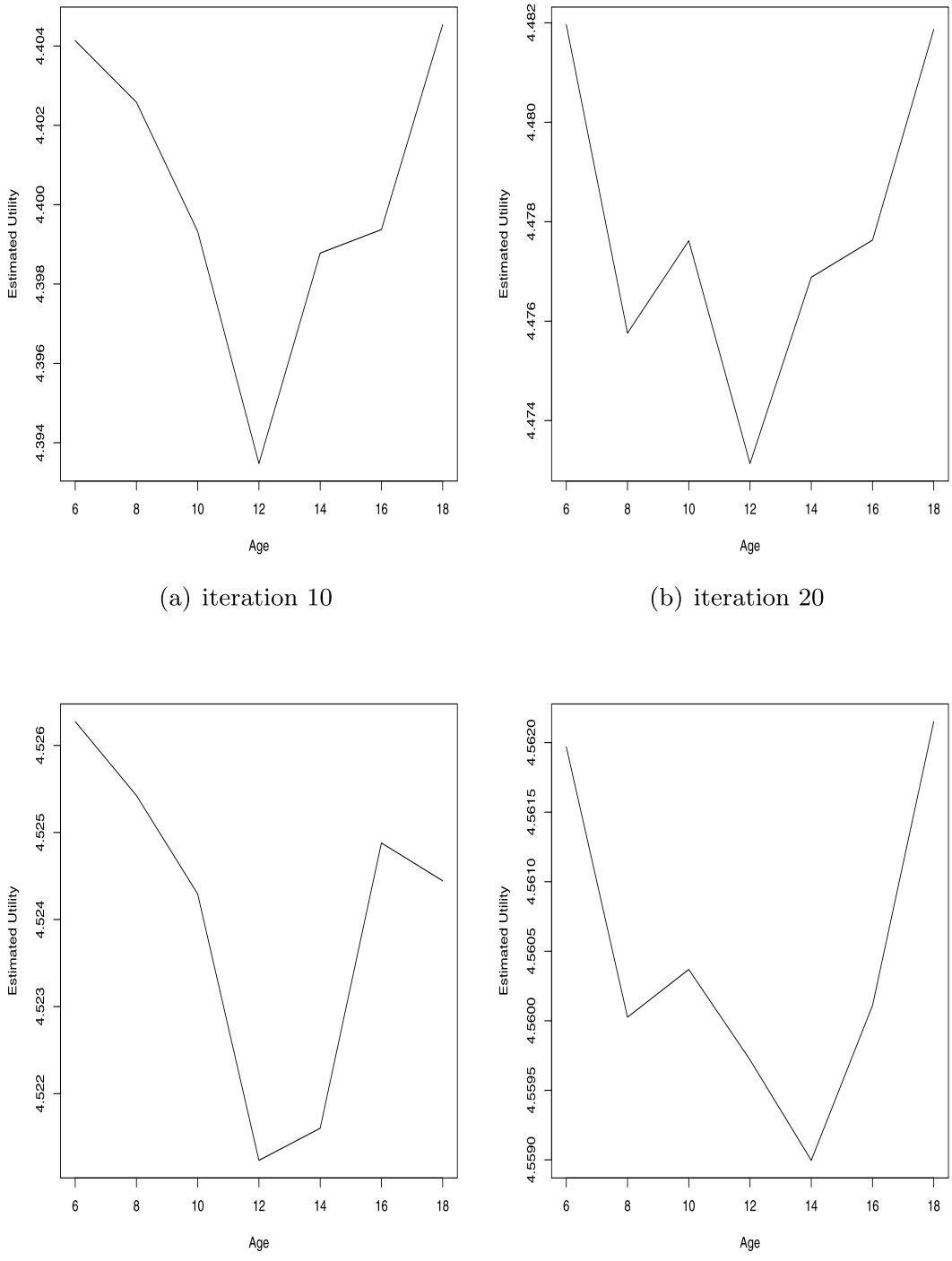
We required a fast method to approximate the posterior distribution, and hence the utility function. Here, we used the integrated nested Laplace approximation (INLA, Rue, Martino and Chopin, 2009) with the default priors in the R-INLA package (www.r-inla.org). Note that high accuracy of the posterior distribution is not required, it is only necessary that the method produces the appropriate ranking of potential designs. To estimate the expected utility at some proposed age, we took only a single sample from the current INLA posterior distribution and simulated a full replicate for a new individual at that age and estimated the new posterior distribution based on all the data in the sub-sample so far and the simulated data. This is performed for each proposed age, and the age that produced the highest utility was selected. We found that it was sufficient to use a single simulation to obtain a close-to-optimal design. A more precise determination of the optimal design could be obtained by considering more posterior predictions. The only stochastic part of the algorithm is the fact that we only draw a single posterior simulation. To investigate the variability in the observed utility of the subsetted data determined from the optimal design, we repeated our process 20 times.

Figure 7 shows the estimated utility for each proposed age at different stages of the algorithm for 1 of the 20 runs. It is evident that it is not difficult to estimate the optimal age to select, even with a single posterior simulation. We compared our sub-sample with two other more standard designs. The first (design 1) takes a completely random sample without replacement of the data with the same size as our optimally designed sub-sample. The second design (design 2) randomly samples without replacement from the unique combinations of individuals and age (with all the data taken at that combination) until a sample size not less than the size of the optimally designed sub-sample is taken. Designs 1 and 2 are repeated 1000 times. The boxplots of the estimated utilities for these two design schemes together with that obtained from the optimal design procedure are shown in Figure 8(a). It is evident that the optimally designed sub-sample approach leads to a much higher utility than those taken from designs 1 and 2. One run of the optimal design process took roughly 1 hour while investigating 1000 random subsets took roughly 3 hours.

The actual ages selected by the algorithm over the iterations for one of the runs is shown in Figure 8(b). It is evident that the optimal ages to sample are generally at the age boundaries.

6. DISCUSSION

This paper has explored the concept of a designed approach to analysing Big Data in order to answer specific aims. The proposed approach exploits established ideas in statistical decision theory and experimental design. The decision-theoretic framework facilitates formal articulation of the purpose of the analysis, desired decisions and associated utility functions. This forms the basis for designing an optimal or near-optimal sample of data that can be extracted from the Big Dataset

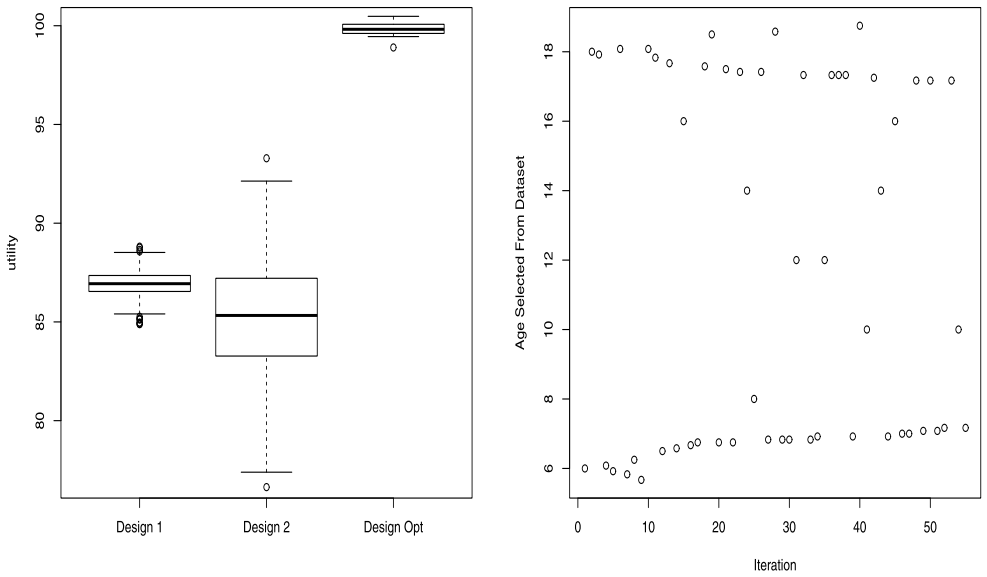


(a) iteration 10 (b) iteration 20

(c) iteration 30 (d) iteration 40

FIG. 7. *Estimated utility values for each proposed design* (*age*) *at iteration* 10 (a), 20 (b), 30 (c) *and* 40 (d) *of the optimal design sub-sampling algorithm for the cut-point dataset.*

in order to make the required decisions. Under this regime, there may be no need to analyse all of the Big Data. This has potential benefits with respect to data manipulation, modelling and computation. The extracted sample of data can be analysed according to the design, avoiding the need to accommodate complex features of the Big Data such as variable data quality, aggregated datasets with different collection methods



(a) distribution of observed utilities (b) selected ages

FIG. 8. *Results from the sub-sampling optimal design process for the cut-point dataset*. (a) *Comparison of observed utilities of sub-samples obtained from designs* 1 *and* 2 *with the observed utilities of the sub-samples obtained from the optimal design process*. (b) *Ages selected from the dataset during the optimal design sub-sampling process*.

and so on. The model can be extended in a more deliberate and structured manner to accommodate remaining biases such as nonrepresentativeness or measurement error, and the design can be replicated to facilitate critical evaluation of issues such as model robustness and "concept drift".

Consideration of the issue of model fit serves to illustrate the potential versatility of the designed approach to Big Data analysis. A natural by-product of the analysis of Big Data is very little statistical uncertainty for many models. However, this rarely reflects reality: in practice, we know that the model can be wrong in many ways. Through the designed approach, the aim of assessing model robustness can be incorporated into the design, in particular into the utility function, and a corresponding optimal sample can be extracted that will facilitate this investigation. For example, the experimental design can incorporate the intention to apply posterior predictive checks, or include a designed hold-out sample set drawn from the Big Data to evaluate goodness-of-fit via a posterior predictive check. Indeed, the utility function can be used as a vehicle to express a very wide range of statistical ambitions. There are implications for robustness in that the initial design of the training data set is predicated on the statistical model, and all models are wrong. Our suggestion here is to proceed in the spirit of Box (1980) in that the analysis should take place as an iterative process of criticism and estimation. The design subsetting procedure takes place as if the model were true, but following this model criticism should be used to help identify artefacts and systematic discrepancies of model fit with the aim to improve the robustness of inference. Such model criticism may of course lead to the consideration of more complex models in the design step. Unfortunately, optimal design has typically been restricted to low to moderate dimensional problems for linear models (Myers, Montgomery and Anderson-Cook, 2009), GLMs (Woods et al., 2006) and nonlinear mixed effects models (Mentré, Mallet and Baccar, 1997). Thus, in order for optimal design to adequately facilitate a wide range of analyses across Big Data sets which are invariably complicated, messy, heterogeneous, heteroscedastic with a large number of different types of variables and fraught with missing data, it seems there is a need for further developments in this area.

This designed approach may also be used not as a substitute for the Big Data analysis but as a complementary evaluation. Thus, the question of interest can be investigated in multiple ways and although the same data are being used for both analyses, the insights and inferences drawn from the two approaches can potentially provide a deeper understanding of the problem; see Appendix D for further discussion.

We have presented throughout the paper various advantages to the designed approach for subset selection that are not computational. An additional advantage is that the design generated could be re-used or harnessed for future/other datasets collected under similar conditions.

There are many ways in which the approach described above can be extended. In addition to expansions to accommodate more complex experimental aims, the designs and models can be extended to accommodate features of the obtained data. We discuss three examples: adjustment for inadequacies in the dataset from which the samples are extracted, extensions to allow for aggregation of information from different sources and the inclusion of replication.

The mismatch between the Big Data and the target population is widely acknowledged as a concern in many disciplines (Wang et al., 2015). Other widely acknowledged inadequacies include measurement error in variables of interest and missing data. If characteristics of these attributes are known in advance, they can be included in the design. There is a large classical literature on adjusting for non-coverage and selection bias in sampling design, for example, through the use of sampling weights (Kish and Hess, 1950, Lessler and Kalsbeek, 1992, Levy and Lemeshow, 1999), design-adjusted regression and its variants (Chambers, 1988) and propensity scores (Dagostino, 1998, Austin, 2011). Analogous weighting methods have been developed to account for missing data and measurement error (Brick and Montaquila, 2009). A growing literature is also available for Bayesian approaches to weighting (Si, Pillai and Gelman, 2015, Gelman, 2007, Oleson et al., 2007). Further, although the experimental design approach described here mitigates the endemic problem of data quality to some extent by extracting only those observations corresponding (at least approximately) to the design points and ignoring the remaining (possibly poorer quality) data, issues such as bias, nonrepresentativeness, missingness and so on may persist. In this case, a variety of methods can be adopted for adjusting the data (Chen et al., 2011), the likelihood (Wolpert and Mengersen, 2004a), the model (Espiro-Hernandez, Gustafson and Burstyn, 2011), the prior (Lehmann and Goodman, 2000) or the utility (Fouskakis, Ntzoufras and Draper, 2009), and the experimental design can be modified accordingly.

An alternative to adjusting the design is to augment the corresponding statistical model used to analyse the extracted data. In a Bayesian framework, this can be implemented through specification of informative priors in a Bayesian hierarchical or joint model (Wolpert and Mengersen, 2004a, Richardson and Gilks, 1993, Mason et al., 2012, Muff et al., 2015).

In a Big Data context, aggregation of data from different sources can be cumbersome due to the different characteristics of the datasets and the very large precisions of the obtained parameter estimates. A designed approach can provide at least partial solutions to these issues. For example, the experimental design can be augmented to sample efficiently from each data source, taking into account the characteristics associated with the source and the overall aim of the analysis. The corresponding statistical model can then be extended hierarchically to allow for the aggregation (McCarron et al., 2011). One could also conceive this problem as a meta-analysis, in which each data source is sampled and analysed according to an independently derived design and the results are combined via a random effects model or similar (Pitchforth and Mengersen, 2012, Schmid and Mengersen, 2013).

The designed approach can also be augmented to allow for potential deficiencies in the statistical model. For example, if the data are "big enough", then replicate samples can be extracted from the data using the same design strategy. The methodology for replication can be adapted in a straightforward manner from classical design principles (Nawarathna and Choudhary, 2015). These replicates can be employed for a variety of purposes, such as more accurate estimation and analysis of sources of variation or heterogeneity in the data, identification of potential unmodelled covariates or confounders, assessment of random effects, or evaluation of the robustness of the model itself. They can also be extracted according to a hyper-design to allow for evaluation of issues such as concept drift, whereby the response variable changes over time (or space) in ways that are not accounted for in the statistical model; see Gama et al. (2014) for a recent survey of this issue.

Finally, we stress once more that the benefits of the designed approach must be weighed up against the computational overheads and potentially reduced sample size in comparison to say a random sub-sampling strategy. We see this as motivation for the study of new computational optimisation methods that can exploit modern computer architectures to deliver designed samples for Big Data analysis.

ACKNOWLEDGEMENTS

CCD was supported by an Australian Research Council's Discovery Early Career Researcher Award

funding scheme (DE160100741). CH would like to gratefully acknowledge support from the Medical Research Council (UK), the Oxford-MAN Institute and the EPSRC UK through the *i-like* Statistics programme grant. CCD, JMM and KM would like to acknowledge support from the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). Funding from the Australian Research Council for author KM is gratefully acknowledged.

REFERENCES

AMZAL, B., BOIS, F. Y., PARENT, E. and ROBERT, C. P. (2006). Bayesian-optimal design via interacting particle systems. *J. Amer. Statist. Assoc.* **101** 773–785. MR2281248

AUSTIN, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* **46** 399–424.

BARDENET, R., DOUCET, A. and HOLMES, C. (2014). Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *Proceedings of the* 31*st International Conference on Machine Learning* (*ICML*-14) 405–413.

BARDENET, R., DOUCET, A. and HOLMES, C. (2015). On Markov chain Monte Carlo methods for tall data. Preprint. Available at arXiv:1505.02827 [stat.ME].

BOUVEYRON, C. and BRUNET-SAUMARD, C. (2014). Model-based clustering of high-dimensional data: A review. *Comput. Statist. Data Anal.* **71** 52–78. MR3131954

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. R. Stat. Soc., A* **143** 383–430.

BRICK, J. M. and MONTAQUILA, J. M. (2009). Nonresponse and weighting. In *Sample Surveys*: *Design*, *Methods and Applications*. *Handbook of Statist.* **29** 163–185. Elsevier, Amsterdam. MR2654638

CHAMBERS, R. (1988). Design-adjusted regression with selectivity bias. *Appl. Stat.* **37** 323–334.

CHEN, C., GRENNAN, K., BADNER, J., ZHANG, D., JIN, E. G. L. and LI, C. (2011). Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLoS ONE* **6** e17238.

CICHOSZ, P. (2015). *Data Mining Algorithms*: *Explained Using R*. Wiley, United Kingdom.

DAGOSTINO, R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* **17** 2265–2281.

DROVANDI, C. C., MCGREE, J. M. and PETTITT, A. N. (2013). Sequential Monte Carlo for Bayesian sequentially designed experiments for discrete data. *Comput. Statist. Data Anal.* **57** 320–335. MR2981091

DROVANDI, C. C. and TRAN, M.-N. (2016). Improving the efficiency of fully Bayesian optimal design of experiments using randomised quasi-Monte Carlo. Available at http://eprints.qut.edu.au/97889.

DUFFULL, S. B., GRAHAM, G., MENGERSEN, K. and ECCLESTON, J. (2012). Evaluation of the pre-posterior distribution of optimized sampling times for the design of pharmacokinetic studies. *J. Biopharm. Statist.* **22** 16–29. MR2872632

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499.

ELGAMAL, T. and HEFEEDA, M. (2015). Analysis of PCA algorithms in distributed environments. Preprint. Available at arXiv:1503.05214v2 [cs.DC].

ESPIRO-HERNANDEZ, G., GUSTAFSON, P. and BURSTYN, I. (2011). Bayesian adjustment for measurement error in continuous exposures in an individually matched case-control study. *BMC Med. Res. Methodol.* **11** 67–77.

FAN, J., FENG, Y. and RUI SONG, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557.

FAN, J., HAN, F. and LIU, H. (2014). Challenges of big data analysis. *Int. Reg. Sci. Rev.* **1** 293–314.

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911.

FEDOROV, V. V. (1972). *Theory of Optimal Experiments*. Academic Press, New York.

FOUSKAKIS, D., NTZOUFRAS, I. and DRAPER, D. (2009). Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care. *Ann. Appl. Stat.* **3** 663–690.

GAMA, J., ŽLIOBAITĖ, I., BIFET, A., PECHENIZKIY, M. and BOUCHACHIA, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys* (*CSUR*) **46** Article Number 44.

GANDOMI, A. and HAIDER, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *Internat. J. Inform. Management Sci.* **35** 137–144.

GELMAN, A. (2007). Struggles with survey weighting and regression modeling (with discussion). *Statist. Sci.* **22** 153–164.

GUHAA, S., HAFEN, R., ROUNDS, J., XIA, J., LI, J., XI, B. and CLEVELAND, W. S. (2012). Large complex data: Divide and recombine (D&R) with RHIPE. *Stat.* **1** 53–67.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

KARVANEN, J., KULATHINAL, S. and GASBARRA, D. (2009). Optimal designs to select individuals for genotyping conditional on observed binary or survival outcomes and non-genetic covariates. *Comput. Statist. Data Anal.* **53** 1782–1793. MR2649544

KETTANEHA, N., BERGLUND, A. and WOLD, S. (2005). PCA and PLS with very large data sets. *Comput. Statist. Data Anal.* **48** 68–85.

KISH, L. and HESS, I. (1950). On noncoverage of sample dwellings. *J. Amer. Statist. Assoc.* **53** 509–524.

KLEINER, A., TALWALKAR, A., SARKAR, P. and JORDAN, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 795–816.

KÜCK, H., DE FREITAS, N. and DOUCET, A. (2006). SMC samplers for Bayesian optimal nonlinear design. Technical Report, Univ. British Columbia, Vancouver, BC.

LEHMANN, H. P. and GOODMAN, S. N. (2000). Bayesian communication: A clinically significiant paradigm for electronic communication. *J. Am. Med. Inform. Assoc.* **7** 254–266.

LESKOVEC, J., RAJARAMAN, A. and ULLMAN, J. D. (2014). *Mining of Massive Datasets*. Cambridge Univ. Press, Cambridge.

LESSLER, J. T. and KALSBEEK, W. D. (1992). *Nonsampling Error in Surveys*. Wiley, New York. MR1193229

LEVY, P. S. and LEMESHOW, S. (1999). *Sampling of Populations*: *Methods and Applications*, 3rd ed. Wiley, New York.

LIANG, F., CHENG, Y., SONG, Q., PARK, J. and YANG, P. (2013). A resampling-based stochastic approximation method for analysis of large geostatistical data. *J. Amer. Statist. Assoc.* **108** 325–339. MR3174623

LIBERTY, E. (2013). Simple and deterministic matrix sketching. In *Proceedings of the* 19*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 581–588. ACM, New York.

LONG, Q., SCAVINO, M., TEMPONE, R. and WANG, S. (2013). Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Comput. Methods Appl. Mech. Engrg.* **259** 24–39.

MASON, A., BEST, N., PLEWIS, I. and RICHARDSON, S. (2012). Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. *J. Off. Stat.* **28** 279–302.

MCCARRON, C. E., PULLENAYEGUM, E. M., THABANE, L., GOEREE, R. and TARRIDE, J.-E. (2011). Bayesian hierarchical models combining different study types and adjusting for covariate imbalances: A simulation study to assess model performance. *PLoS ONE* **6** e25635.

MENTRÉ, F., MALLET, A. and BACCAR, D. (1997). Optimal design in random-effects regression models. *Biometrika* **84** 429–442.

MUFF, S., RIEBLER, A., HELD, L., RUE, H. and SANER, P. (2015). Bayesian analysis of measurement error models using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 231–252.

MÜLLER, P. (1999). Simulation-based optimal design. In *Bayesian Statistics*, 6 (*Alcoceber*, 1998) 459–474. Oxford Univ. Press, New York. MR1723509

MYERS, R. H., MONTGOMERY, D. C. and ANDERSON-COOK, C. M. (2009). *Response Surface Methodology*: *Process and Product Optimization Using Designed Experiments*, 3rd ed. Wiley, Hoboken, NJ. MR2464113

NAWARATHNA, L. S. and CHOUDHARY, P. K. (2015). A heteroscedastic measurement error model for method comparison data with replicate measurements. *Stat. Med.* **34** 1242–1258. MR3322747

OGUNGBENRO, K. and AARONS, L. (2007). Design of population pharmacokinetic experiments using prior information. *Xenobiotica* **37** 1311–1330.

OLESON, J. J., HE, C., SUN, D. and SHERIFF, S. (2007). Bayesian estimation in small areas when the sampling design strata differ from the study domains. *Surv. Methodol.* **33** 173–185.

OSWALD, F. L. and PUTKA, D. J. (2015). Statistical methods for big data. In *Big Data at Work*: *The Data Science Revolution and Organisational Psychology*. Routledge, New York.

PITCHFORTH, J. and MENGERSEN, K. (2012). Bayesian meta-analysis. In *Case Studies in Bayesian Statistics* 121–144. Wiley, New York.

PUKELSHEIM, F. (1993). *Optimal Design of Experiments*. Wiley, New York. MR1211416

REINIKAINEN, J., KARVANEN, J. and TOLONEN, H. (2016). Optimal selection of individuals for repeated covariate measurements in follow-up studies. *Stat. Methods Med. Res.* **25** 2420–2433. MR3572861

RICHARDSON, S. and GILKS, S. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *Am. J. Epidemiol.* **138** 430–442.

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392.

RYAN, E. G., DROVANDI, C. C. and PETTITT, A. N. (2015). Simulation-based fully Bayesian experimental design for mixed effects models. *Comput. Statist. Data Anal.* **92** 26–39. MR3384249

SAVAGE, L. J. (1972). *The Foundations of Statistics*, revised ed. Dover Publications, New York. MR0348870

SCHIFANO, E. D., WU, J., WANG, C., YAN, J. and CHEN, M.-H. (2016). Online updating of statistical inference in the big data setting. *Technometrics* **58** 393–403. MR3520668

SCHMID, C. H. and MENGERSEN, K. (2013). Handbook of Meta-analysis in Ecology and Evolution. 145–173 Bayesian meta-analysis. Princeton Univ. Press, Princeton.

SCOTT, S. L., BLOCKER, A. W. and BONASSI, F. V. (2013). Bayes and big data: The consensus Monte Carlo algorithm. In *Bayes* 250.

SI, Y., PILLAI, N. and GELMAN, A. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Anal.* **10** 605–625.

SUYKENS, J. A. K., SIGNORETTO, M. and ARGYRIOU, A. (2015). *Regularization*, *Optimization*, *Kernels*, *and Support Vector Machines*. Chapman and Hall/CRC, Boca Raton, FL.

TAN, F. E. S. and BERGER, M. P. F. (1999). Optimal allocation of time points for the random effects models. *Comm. Statist.* **28** 517–540.

TOULIS, P., AIROLDI, E. and RENNI, J. (2014). Statistical analysis of stochastic gradient methods for generalized linear models. In *Proceedings of the* 31*st International Conference on Machine Learning* 667–675.

TROST, S. G., LOPRINZI, P. D., MOORE, R. and PFEIFFER, K. A. (2011). Comparison of accelerometer cut-points for predicting activity intensity in Youth. *Med. Sci. Sports Exerc.* **43** 1360–1368.

WANG, C., CHEN, M. H., SCHIFANO, E., WU, J. and YAN, J. (2015). A survey of statistical methods and computing for big data. Preprint. Available at arXiv:1502.07989v1 [stat.CO].

WOLPERT, R. L. and MENGERSEN, K. L. (2004a). Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: Effects of environmental tobacco smoke. *Statist. Sci.* **3** 450–471.

WOLPERT, R. L. and MENGERSEN, K. L. (2004b). Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: Effects of environmental tobacco smoke. *Statist. Sci.* **19** 450–471. MR2185626

WOODS, D. C., LEWIS, S. M., ECCLESTON, J. A. and RUSSELL, K. G. (2006). Designs for generalized linear models with several variables and model uncertainty. *Technometrics* **48** 284–292.

XI, B., CHEN, H., CLEVELAND, W. S. and TELKAMP, T. (2010). Statistical analysis and modelling of Internet VoIP traffic for network engineering. *Electron. J. Stat.* **4** 58–116.

YOO, C., RAMIREZ, L. and JUAN LIUZZI, J. (2014). Big data analysis using modern statistical and machine learning methods in medicine. *International Neurology Journal* **18** 50–57.