A model robust subsampling approach for Generalised Linear Models in big data settings

Amalan Mahendran¹,² · Helen Thompson¹,² · James M. McGree¹,²

Received: 22 December 2022 / Revised: 11 March 2023 / Published online: 27 May 2023
© The Author(s) 2023

Abstract

In today’s modern era of big data, computationally efficient and scalable methods are needed to support timely insights and informed decision making. One such method is subsampling, where a subset of the big data is analysed and used as the basis for inference rather than considering the whole data set. A key question when applying subsampling approaches is how to select an informative subset based on the questions being asked of the data. A recent approach for this has been proposed based on determining subsampling probabilities for each data point, but a limitation of this approach is that the appropriate subsampling probabilities rely on an assumed model for the big data. In this article, to overcome this limitation, we propose a model robust approach where a set of models is considered, and the subsampling probabilities are evaluated based on the weighted average of probabilities that would be obtained if each model was considered singularly. Theoretical results are derived to inform such an approach. Our model robust subsampling approach is applied in a simulation study and in two real-world applications where performance is compared to current subsampling practices. The results show that our model robust approach outperforms alternative methods.

Keywords  Large data set · Model averaging · Optimal subsampling · Experimental design

1 Introduction

Analysis of big data has gained interest in recent years through providing new insights and unlocking hidden knowledge in different fields of study (Karmakar
and Mukhopadhyay 2020) including medicine (Rehman et al. 2021), fraud detection (Vaughan 2020), the oil and gas industry (Nguyen et al. 2020) and astronomy (Zhang and Zhao 2015). However, the analysis of big data can be challenging for traditional statistical methods and standard computing environments (Wang et al. 2016). Martinez-Mosquera et al. (2020) discuss storage and modelling solutions when handling such a large amount of data. In general, modelling solutions can be grouped into three broad categories: (1) subsampling methods, where the analysis is performed on an informative subsample obtained from the big data (Kleiner et al. 2014; Ma et al. 2015; Ma and Sun 2015; Drovandi et al. 2017; Wang et al. 2018, 2019; Yao and Wang 2019; Ai et al. 2021a, b; Cheng et al. 2020; Lee et al. 2021; Yao and Wang 2021); (2) divide and recombine methods, where the big data set is divided into smaller blocks, and then the intended statistical analysis is performed on each block and subsequently recombined for inference (Lin and Xi 2011; Guha et al. 2012; Cleveland and Hafen 2014; Chang et al. 2017; Li et al. 2020); (3) online updating of streamed data, where statistical inference is updated as new data arrive sequentially (Schifano et al. 2016; Xue et al. 2020). In recent years, compared to divide and recombine methods, subsampling has been applied to a variety of regression problems, while online updating is typically only used for streaming data. In addition, in cases where a large data set is not needed to answer a specific question with sufficient confidence, subsampling seems preferable as the analysis of the data can often be undertaken with standard methods. Moreover, the computational efficiency of subsampling over analysing large data sets has been observed for parameter estimation in linear (Wang et al. 2019) and logistic (Wang et al. 2018) regression models. For these reasons, we focus on subsampling methods in this article.

The key challenge for subsampling methods is how to obtain an informative subsample that can be used to efficiently answer specific analysis questions and provide results that align with the analysis of the whole big data set. Two approaches for this exist in the literature: (1) randomly sample from the big data with subsampling probabilities that are found based on a specific statistical model and objective (e.g., prediction and/or parameter estimation) (Wang et al. 2018; Yao and Wang 2019; Ai et al. 2021a, b; Lee et al. 2021; Yao and Wang 2021); (2) select subsamples based on an experimental design (Drovandi et al. 2017; Deldossi and Tommasi 2022). Randomly sampling with certain probabilities (based upon the definitions of $A$- or $L$-optimality criteria, see Atkinson et al. 2007) is the focus of this article, and has been applied for parameter estimation in a wide range of regression problems including softmax (Yao and Wang 2019) and quantile regression (Ai et al. 2021a), and Generalised Linear Models (GLMs; Wang et al. 2018; Ai et al. 2021b; Yao and Wang 2021). In contrast, the approach based on an experimental design has only been applied for: (1) parameter estimation in logistic fixed and mixed effects regression models (Drovandi et al. 2017); (2) parameter estimation and prediction accuracy in linear and logistic regression models (Deldossi and Tommasi 2022).

A key feature of both of the current subsampling approaches is that they rely on a statistical model that is assumed to appropriately describe the big data. Given this is a potentially limiting assumption, Yu and Wang (2022) proposed to select the best candidate model from a pool of models based on the Bayesian Information Criterion (BIC; Schwarz 1978). This was applied to linear models, and resulted in subsampling
probabilities that were more appropriate than those based on considering a single model. Similarly, for GLMs, Shi and Tang (2021), Meng et al. (2021) and Yi and Zhou (2023) explored using space-filling or orthogonal Latin hypercube designs so that a wide range of models could be considered, however, such approaches have notable limitations particularly as the design dimension increases. In this paper, we propose that, instead of selecting a single best candidate model for the big data, a set of models is considered and a model averaging approach is used to determine the subsampling probabilities. Through adopting such an approach, it is thought that the analysis goal (e.g., efficient parameter estimation) should be achieved regardless of the preferred model for the data. To implement this model robust approach, we consider subsampling based on $A$- and $L$-optimality within the Generalised Linear Modelling framework, and adopt a model averaged approach based on each of these criteria. Given we consider GLMs, our approach should be generally applicable across many areas of science and technology where a variety of data types are observed. This is demonstrated through applying our proposed methods within a simulation study, and for the analysis of two real-world big data problems.

The remainder of the article is structured as follows. Section 2 introduces GLMs and the existing probability-based subsampling approach of Ai et al. (2021b) and Yao and Wang (2021). Our proposed model robust subsampling approach is introduced in Section 3, which is embedded within a GLM framework. A simulation study is then used to assess the performance of our model robust approach in Sect. 4, and two real-world applications are presented. Section 5 concludes the article with a discussion of the results and some suggestions for future research.

2 Background

There are a variety of ways big data can be subsampled. In this section, we focus on the approach where subsampling probabilities are determined for each data point, and the big data is subsampled (at random) based on these probabilities. Such an approach was first proposed by Wang et al. (2018) for logistic regression problems in big data settings, and has been extended to a wide range of regression problems (e.g., Yao and Wang 2019, 2021; Ai et al. 2021a, b). In this section, we describe such a subsampling approach as applied to GLMs based on the work of Ai et al. (2021b).

2.1 Generalised Linear Models

Let a big data set be denoted as $F_N = (X_0, y)$, where $X_0 = (x_{01}, \ldots, x_{0N})^T \in \mathbb{R}^{N \times p}$ represents a data matrix based on the big data set with $p$ covariates, $y = (y_1, \ldots, y_N)^T$ represents the response vector and $N$ is the total number of data points. To fit a GLM, consider the model matrix $X = h(X_0) \in \mathbb{R}^{N \times (p+1)}$ where $h(.)$ is some function of $X_0$ which creates an additional $t$ columns representing, for example, an intercept and/or higher-order terms. A GLM can then be defined via three components: (1) distribution of response $y$, which is from the exponential family (e.g., Normal, Binomial or Poisson); (2) linear predictor $\eta = X \theta$, where $\theta = (\theta_1, \ldots, \theta_{p+t})^T$ is
the parameter vector; and (3) link function \( g(.) \), which links the mean of the response to the linear predictor (Nelder and Wedderburn 1972). Throughout this article, the inverse link function \( g^{-1}(.) \) is denoted by \( u(.) \).

A common exponential form for the probability density or mass function of \( y \) can be written as:

\[
f(y; \omega, \gamma) = \exp\left(\frac{y\omega - \psi(\omega)}{a(\gamma)} + b(y, \gamma)\right),
\]

where \( \psi(.) \), \( a(.) \) and \( b(.) \) are some functions, \( \omega \) is known as the natural parameter and \( \gamma \) the dispersion parameter. Based on Eq. (1) the link function \( g(.) \) can then be defined as \( g(\mu) = \eta \), where \( \mu = E[y|X, \theta] = d\psi(\omega)/d\omega \). A general linear model, or linear regression model, is a special case of a GLM where \( y \sim N(\mu, \Sigma) \) and \( g(.) \) is the identity link function \( g(\mu) = \mu \), such that \( \mu = X\theta \). For logistic regression, \( y \sim \text{Bin}(n, \pi) \) and \( g(.) \) is the logit link function \( g(\pi) = \log(\pi/(1 - \pi)) \) such that \( g(\pi) = X\theta \). Similarly, for Poisson regression, \( y \sim \text{Poisson}(\lambda) \) and \( g(.) \) is the log link function \( g(\lambda) = \log(\lambda) \) such that \( g(\lambda) = X\theta \). Note that the dispersion parameter \( \gamma = 1 \) for the logistic and Poisson regression models.

### 2.2 A general subsampling algorithm for GLMs

As described by Ai et al. (2021b), consider a general subsampling approach to estimate parameters \( \theta \) through a weighted log-likelihood function for GLMs (where the weights are the inverse of the subsampling probabilities). A weighted likelihood function is considered, since an unweighted likelihood leads to biased estimates of model parameters, see Wang (2019). Define \( \phi_i \) as the probability that, for a single draw, row \( i \) of \( F_N \) is randomly selected, for \( i = 1, \ldots, N \), where \( \sum_{i=1}^N \phi_i = 1 \) and \( \phi_i \in (0, 1) \). A subsample \( S \) of size \( r \) is then drawn with replacement from \( F_N \) based on \( \phi = (\phi_1, \ldots, \phi_N) \). The selected responses, covariates and subsampling probabilities are then used to estimate the model parameters. Pseudo-code for this general subsampling approach is provided in Algorithm 1.

**Algorithm 1 General subsampling algorithm (Ai et al. 2021b)**

1: Sampling: Assign \( \phi_i, i = 1, \ldots, N \), for \( F_N \). Based on \( \phi \), draw an \( r \) size subsample with replacement from \( F_N \) to yield \( S = \{x^*_i, y^*_i, \phi^*_i\}_{i=1}^r = (X^*, y^*, \phi^*) \).

2: Estimation: Based on \( S \), using \( L(\theta|X^*, y^*, \phi^*) \) the likelihood function, find:

\[
\hat{\theta} = \arg\max_{\theta} \log L(\theta|X^*, y^*, \phi^*)
\]

\[
\equiv \arg\max_{\theta} \frac{1}{r} \sum_{i=1}^r \left( \frac{y^*_i u(\theta^T x^*_i) - \psi(u(\theta^T x^*_i))}{\phi^*_i} \right).
\]

3: Output: \( \hat{\theta} \) and \( S \).

From Algorithm 1, the first step is to assign subsampling probabilities \( \phi_i \) to the rows of \( F_N \). The simplest approach is to assign each data point an equal probability of
being selected. These probabilities could also depend on the composition of \( y \), e.g., for binary data, one could sample proportional to the inverse of the number of successes and failures. Based on these probabilities, a subsample \( S \) of size \( r \) is then drawn at random (with replacement) from \( F_N \) to yield \( \{x_i^*, y_i^*, \phi_i^* \}_{i=1}^r \). Based on this subsample, model parameters \( \tilde{\theta} \) are estimated via maximising a weighted log-likelihood function. The estimates \( \tilde{\theta} \) can then be considered as estimates of what would be obtained if the whole big data set were analysed.

The asymptotic properties of \( \tilde{\theta} \) based on the general subsampling approach given in Algorithm 1 have been derived by Ai et al. (2021b). Based on these properties, two theorems were proposed which provide insight into the behaviour of the estimator as the subsample increases. Specifically, as \( N \to \infty \) and \( r \to \infty \), they showed that \( \tilde{\theta} \) is a consistent estimator of \( \hat{\theta}_{MLE} \), the maximum likelihood estimator of \( \theta \) based on \( F_N \). In addition, they showed that the approximation error, \( \tilde{\theta} - \hat{\theta}_{MLE} \), given \( F_N \), is approximately asymptotically Normally distributed with mean zero and variance \( V \).

When applying the general subsampling algorithm, it may not be clear how to appropriately choose \( \phi \) depending upon the goal of the analysis (e.g., parameter estimation, response prediction, etc.). To address this, Ai et al. (2021b) proposed determining \( \phi \) based on optimality criteria (e.g. \( A \)- and \( L \)-optimality), which leads to optimal subsampling probabilities that minimise the asymptotic mean squared error of \( \tilde{\theta} \) [or \( \text{tr}(V) \)] and \( J_X \tilde{\theta} \) [or \( \text{tr}(V_c) \)], respectively, where \( J_X \) is the observed information matrix and \( V_c \) is the variance of \( J_X \tilde{\theta} \). Based on each of these optimality criteria, subsampling probabilities can be determined for each data point, and these will be denoted as \( \phi^{mMSE} \) and \( \phi^{mV_c} \), respectively.

Unfortunately, the optimal subsampling probabilities \( \phi^{mMSE} \) and \( \phi^{mV_c} \) cannot be determined directly as they depend on \( \hat{\theta}_{MLE} \). To address this, Ai et al. (2021b) proposed a two-stage subsampling strategy where an initial random sample of the big data is used to estimate \( \hat{\theta}_{MLE} \). This estimate is then used to approximate the optimal subsampling probabilities. Such an approach is thus termed a two-stage subsampling algorithm. A full description of this algorithm is given in Ai et al. (2021b).

### 2.3 Limitations of the optimal subsampling algorithm

The above optimal subsampling approach has a number of limitations. One such limitation is the computational expense involved in obtaining the optimal subsampling probabilities, as these need to be determined for each data point in the big data set. To address this, Lee et al. (2021) introduced a faster two-stage subsampling procedure for GLMs using the Johnson–Lindenstrauss Transform (JLT) and subsampled Randomised Hadamard Transform (SRHT), which are techniques to downsize matrix volume. Another limitation is that the approximation for the optimal subsampling probabilities can lead to some data points having zero probability of being selected. Ai et al. (2021b) proposed to resolve this by setting these subsampling probabilities to a small value to ensure such data points have some (non-zero) probability of being selected. Lastly, one of the major limitations of the approach that has not been addressed previously is the inherent assumption that the big data can be appropriately described by a given model. That is, the subsampling probabilities are evaluated based
on an assumed model, and they are generally only optimal for this model. We suggest that this is a substantial limitation as specifying such a model in practice can be difficult. This motivates the development of methods that yield subsampling probabilities that are robust to the choice of model, and our proposed approach for this is outlined next.

3 Model robust subsampling method

In order to apply the two-stage subsampling approach from Ai et al. (2021b), optimal subsampling probabilities need to be evaluated, and these are based on a model that is assumed to appropriately describe the data. In practice, determining such a model may be difficult, and there could be a variety of models that appropriately describe the data. Hence, a subsampling approach that provides robustness to the choice of model is desirable. For this, we propose to consider a set of \( Q \) models which can be constructed to encapsulate a variety of scenarios that may be observed within the data. For each model in this set, define model probabilities \( \alpha_q \) for \( q = 1, \ldots, Q \) such that \( \sum_{q=1}^Q \alpha_q = 1 \), which represents our \textit{a priori} belief about the appropriateness of each model. Denote the model matrix for the \( q \)th model as \( X_q = h_q(X_0) \), i.e., some function of the data matrix \( X_0 \). To apply a subsampling approach for this model set, subsampling probabilities are needed, and they should be constructed such that the resulting data is expected to address the analysis aim, regardless of which model is actually preferred for the big data. For this purpose, we propose to form these subsampling probabilities via a weighted average (based on \( \alpha_q \)) of the subsampling probabilities that would be obtained for each model (singularly). This is the basic approach of our model robust subsampling algorithm, for which further details are provided below.

3.1 Properties of model robust subsampling algorithm

Updating the notation of the model matrix from \( X \) to \( X_q \) for the \( q \)th model subsequently leads to analogous definitions for \( x_{qi}, \theta_q, \hat{\theta}_q^{MLE}, J_X, V_q \) and \( V_{qc} \). In addition, let \( \dot{\psi}(u(\theta^T x_i)) \) and \( \dot{u}(\theta^T x_i) \) denote the first-order derivatives of \( \psi(u(\theta^T x_i)) \) and \( u(\theta^T x_i) \) (with respect to \( \theta \)), and denote the Euclidean norm of a vector \( a \) as \( \|a\| = (a^T a)^{1/2} \). Extending the ideas from Sect. 2.2, optimal subsampling probabilities can be selected based on certain optimality criteria to, for example, ensure efficient estimates of parameters across the \( Q \) models. To inform the choice of subsampling probabilities, consider the following theorems.

\textbf{Theorem 1} For a set of \( Q \) models with model probability \( \alpha_q \) for the \( q \)th model, \( q = 1, \ldots, Q \), if the subsampling probabilities are selected as follows:

\[ \phi_{q_i}^{mMSE} = \frac{|y_i - \hat{\psi}(u(\hat{\theta}_q^{MLE} x_{qi}))| \| J^{-1}_X \hat{u}(\hat{\theta}_q^{MLE} x_{qi}) x_{qi} \|}{\sum_{j=1}^N |y_j - \hat{\psi}(u(\hat{\theta}_q^{MLE} x_{qj}))| \| J^{-1}_X \hat{u}(\hat{\theta}_q^{MLE} x_{qj}) x_{qj} \|}, \]
i = 1, . . . , N, and \( \sum_{q=1}^{Q} \alpha_q = 1 \), then \( \sum_{q=1}^{Q} \alpha_q tr(V_q) \) attains its minimum.

The proof of Theorem 1 is available in “Appendix A”, which is an extension of the proof of Theorem 3 from Ai et al. (2021b).

**Theorem 2** For a set of Q models with model probability \( \alpha_q \) for the qth model, \( q = 1, . . . , Q \), if the subsampling probabilities are selected as follows:

\[
\phi_{qi} = \frac{|y_i - \hat{\psi}(u(\hat{\theta}_{q, MLE}^T x_{qi}))| \| \hat{u}(\hat{\theta}_{q, MLE}^T x_{qi}) x_{qi} \|}{\sum_{j=1}^{N} |y_j - \hat{\psi}(u(\hat{\theta}_{q, MLE}^T x_{qj}))| \| \hat{u}(\hat{\theta}_{q, MLE}^T x_{qj}) x_{qj} \|},
\]

\( i = 1, . . . , N \), and \( \sum_{q=1}^{Q} \alpha_q = 1 \), then \( \sum_{q=1}^{Q} \alpha_q tr(V_q c) \) attains its minimum.

The proof of Theorem 2 is similar to the proof of Theorem 1. Based on the above results, we propose that model robust subsampling probabilities can be obtained as follows: \( \phi_i = \sum_{q=1}^{Q} \alpha_q \phi_{qi} \), for \( i = 1, . . . , N \), and \( \sum_{q=1}^{Q} \alpha_q = 1 \), i.e. a model averaged approach based on the optimal subsampling probabilities from each model \( q \). Similar to the work of Ai et al. (2021b), these model robust subsampling probabilities depend on the maximum likelihood estimator found by considering the whole big data set. As this is not available in big data settings, a two-stage approach is proposed for model robust subsampling, and this is outlined next.

### 3.2 Model robust subsampling algorithm for GLMs

The two-stage model robust subsampling algorithm for GLMs is presented in Algorithm 2, where subsamples are drawn based on model averaged subsampling probabilities.

The first phase of Algorithm 2 entails randomly subsampling \( F_N \) (with replacement), and estimating model parameters for each of the \( Q \) models. Based on these estimated parameters, model specific subsampling probabilities are obtained, and these are combined based on \( \alpha_q \) to form model robust subsampling probabilities. Subsequently, \( r \geq r_0 \) data points are sampled from \( F_N \). The two subsamples are then combined, and each of the \( Q \) models are fitted (separately) based on the weighted log-likelihood function, which should yield, for example, efficient estimates of parameters across each of the models. In the following section, our proposed model robust subsampling approach is assessed via simulation and in two real-world scenarios.

### 4 Applications of optimal subsampling algorithms

In this section, a simulation study and two real-world applications are used to assess the performance of our proposed model robust subsampling algorithm (Algorithm 2) compared to: (1) the approach of Ai et al. (2021b) based on a single model, and to; (2) sampling completely at random. The main results are presented in this section with some results presented in the Supplementary Material. The simulation study and
Algorithm 2 Two-stage model robust subsampling algorithm for GLMs.

Stage 1
1: Random Sampling : Assign $\phi = (\phi_1, \ldots, \phi_N)$ for $F_N$. For example, in logistic regression $\phi_i = \phi_i^{prop} = c_0^{-1} y_i c_1 y_i$, where $c_0 = (2N - 2 \sum_{i=1}^{N} y_i)^{-1}$ and $c_1 = (2 \sum_{i=1}^{N} y_i)^{-1}$ are proportional subsampling probabilities that are based on the response composition, and $\phi_i = N^{-1}$ which is uniform subsampling probabilities for Poisson regression. According to $\phi$ draw random subsamples of size $r_0$, such that $S_{r_0} = \{y_i, x_{i0}^{(q)}, y_i^{(0)}, \phi_i^{(0)}\}_{i=1}^{N} = \{y_i(x_{i0}^{(q)}), y_i^{(0)}, \phi_i^{(0)}\}$ for $q = 1, \ldots, Q$.

2: Estimation: For $q = 1, \ldots, Q$ and $S_{r_0}$, find:

$$\hat{\theta}_{q}^{r_0} = \arg\max_{\theta_q} \log L(\theta_q | X_{q}^{r_0}, y^{(0)}, \phi^{(0)})$$

$$= \arg\max_{\theta_q} \frac{1}{r_0} \sum_{i=1}^{r_0} \left[ \frac{y_i^{(0)} u(\theta_q^{r_0} x_{q}^{r_0}) - \psi(u(\theta_q^{r_0} x_{q}^{r_0}))}{\phi_i^{r_0}} \right]$$

Stage 2
1: Optimal subsampling probability: Estimate optimal subsampling probabilities $\phi^{MSE} = \sum_{q=1}^{Q} \alpha_q \hat{\phi}_q^{MSE}$ or $\phi^{MSE} = \sum_{q=1}^{Q} \alpha_q \hat{\phi}_q^{MSE}$ with $\hat{\theta}_{q}^{MLE}$ replaced by $\hat{\theta}_{q}^{0}$.

2: Optimal subsampling and estimation: Based on $\phi^{MSE} = \phi^{MSE}$, draw subsample of size $r$ from $F_N$, such that $S_r = \{y_i(x_{i0}^{(q)}, y_i^{(q)}, \phi_i^{(q)})\}_{i=1}^{r} = \{y_i(x_{i0}^{(q)}, y_i^{(q)}, \phi_i^{(q)})\}$ for $q = 1, \ldots, Q$. Combine $S_{r_0}$ and $S_r$ to form $S_{(r_0+r)}$, and obtain :

$$\hat{\theta}_{q} = \arg\max_{\theta_q} \log L(\theta_q | X_{q}^{r_0+r}, y^{(r)}, \phi^{(r)}, X^{r} | y^{(r)}, \phi^{(r)})$$

$$= \arg\max_{\theta_q} \frac{1}{r_0 + r} \left[ \sum_{k \in \{r_0, r\}} \sum_{l=1}^{k} \frac{y_l^{(k)} u(\theta_q^{r_0} x_{q}^{r_0}) - \psi(u(\theta_q^{r_0} x_{q}^{r_0}))}{\phi_i^{r_0}} \right]$$

3: Output: $\hat{\theta}_{q}$ and $S_{(r_0+r)}$ for $q = 1, \ldots, Q$.

real-world applications were coded in the R statistical programming language (R Core Team 2021) with the help of RStudio IDE (2020), and code to reproduce our results is available through GitHub. Supplementary Material S2 provides specific GitHub hyperlinks to the code repositories.

4.1 Simulation study design

To explore the performance of our model robust subsampling approach, a simulation study was constructed based on the logistic and Poisson regression models. For each case, a set of $Q = 4$ models were assumed based on Shi and Tang (2021), and this set is summarised in Table 1. For each model, $F_N$ was constructed by assuming a distribution for the covariates and the corresponding response. The performance of the three sampling methods were then compared for each $F_N$ through evaluating six scenarios: (1) random sampling to estimate the parameters of the data generating model; (2) optimal subsampling under the data generating model—this simulates the case where an appropriate model was assumed for describing the big data; (3)–(5) optimal subsampling under alternative models (all models in Table 1 except the data generating model)—this simulates optimal subsampling under a ‘wrong’ model; (6) model robust subsampling to estimate the parameters of the data generating model...
under the assumption that each of the $q$ models are equally likely \textit{a priori} (i.e., assuming the model set as given in Table 1 with $\alpha_q = 1/Q$). For each simulation, $N = 10,000$, $r_0 = 100$, $r = 100, 200, \ldots, 1400$ and $M = 1000$.

To compare the results from each of the six scenarios, the simulated mean squared error (SMSE) of the estimated model parameters was evaluated as follows:

$$SMSE(\hat{\theta}, \iota) = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{p+t} (\hat{\theta}_{nm} - \iota_n)^2,$$

where $M$ is the number of simulations, $p + t$ is the number of parameters in the data generating model and $\hat{\theta}$ is an $M \times (p + t)$ matrix of the estimated model parameters from each simulation. Thus, to evaluate the SMSE to compare $\hat{\theta}$ with $\theta$ and $\hat{\theta}_{MLE}$, we set $\iota = \theta$ and $\hat{\theta}_{MLE}$, respectively. In addition, the average model information (i.e., the mean of $M$ determinants of the Fisher information matrix based on $\theta$) was also evaluated for comparison across the different subsampling scenarios.

Wang et al. (2018) and Ai et al. (2021b) showed that SMSE values based on comparisons with $\hat{\theta}_{MLE}$ were generally smaller when using $\phi^{mMSE}$ compared to $\phi^{mVc}$. Therefore, the results from the simulation study when using $\phi^{mVc}$ are included in Supplementary Material S3.1 and S3.2. However, results from both optimality criteria are discussed in the main text in the two real-world applications. In addition, in Supplementary Material S4, we have extended our simulation study to explore the effect of considering different values for $r_0$. Further, in Supplementary Material S5, we have extended our simulation study to explore the effect of considering non-uniform model probabilities and a model set that includes models with different covariates.

### 4.1.1 Logistic regression

Following Wang et al. (2018), covariate data ($x_1, x_2$) for the logistic regression model were simulated based on two distributions: Exponential ($\lambda$) and Multivariate Normal ($\mu, \Sigma$). The values of ($\lambda, \mu, \Sigma$) and $\theta$ are given in Table 2 for each data generating model. For all models, the first element of $\theta$ is the value of the parameter for the intercept with other values denoting values for slope parameters. While $\lambda, \mu$ and $\Sigma$ were selected arbitrarily, $\theta$ was determined so that the data generating model would be preferred (based on the Akaike Information Criterion 1974) over each rival model if the whole big data set was analysed. This was considered to avoid situations where an
alternative model might be preferred over the data generating model as $r$ approached $N$ and thus potentially hindering the interpretation of results.

For logistic regression, through applying Theorems 1 and 2, model robust subsampling probabilities are obtained as follows:

$$
\phi_i^{mMSE} = \sum_{q=1}^{Q} \frac{\alpha_q |y_i - \pi_{q,i}| \|J_{X_q}^{-1}x_{qi}\|}{\sum_{j=1}^{N} |y_j - \pi_{q,j}| \|J_{X_q}^{-1}x_{qj}\|}, \quad \phi_i^{mVc} = \sum_{q=1}^{Q} \frac{\alpha_q |y_i - \pi_{q,i}| \|x_{qi}\|}{\sum_{j=1}^{N} |y_j - \pi_{q,j}| \|x_{qj}\|}
$$

with $y_i \in \{0, 1\}$, $\pi_{q,i} = \exp(\hat{\theta}_{qMLE}^T x_{qi})/(1 + \exp(\hat{\theta}_{qMLE}^T x_{qi}))$, $J_{X_q} = N^{-1} \sum_{h=1}^{N} \pi_{q,h}(1 - \pi_{q,h})x_{qh}(x_{qh})^T$ and $i = 1, \ldots, N$.

The general model robust approach applied in this example is given in Algorithm 2, and this algorithm specifically based on logistic regression is given in Supplementary Material S1 as Algorithms S3.

Figure 1 provides summaries of the logarithm-scaled SMSE and average model information when the covariate data is generated from an Exponential distribution. The SMSE and average model information indicate that, under optimal subsampling for $\phi_i^{mMSE}$, the data generating model is typically preferred within the model set. This is expected as it is the case where the appropriate data generating model was correctly assumed to describe the big data. Of note, the proposed model robust approach performs similarly to the optimal subsampling approach. Notable increases in the SMSE and decreases in the average model information are observed when the incorrect model is assumed within optimal subsampling. For average model information, there are occasions where assuming a model that is more complex than the data generating model performs well when compared to alternative approaches (e.g. when both quadratic terms are assumed to be included in the model but only one is present). However, our model robust approach is still preferred when all scenarios are considered. In addition, random sampling tends to have the worst performance. Similar results were obtained when the covariate data were generated from a Multivariate Normal distribution with $L$-optimality, see Supplementary Material S3.1. Also, similar results were observed when SMSE was evaluated based on $\hat{\theta}_{MLE}$, see Supplementary Material S3.2.
Fig. 1 Logarithm of (a) average model information and (b) SMSE for the subsampling methods for the logistic regression model under $\phi^{mMSE}$. Covariate data were generated from the Exponential distribution.
Table 3  \( \mu, \Sigma \) and \( \theta \) values are given to generate \( x_1, x_2 \) through uniform and Multivariate Normal distributions and form \( F_N \) for each data generating Poisson regression model.

| Covariates | Distributions | Uniform | Normal (\( \mu = [0, 0], \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \)) |
|------------|---------------|---------|---------------------------------|
| \( x_1, x_2 \) | \( \theta = [1, 0.3, 0.3] \) | \( \theta = [1, 0.5, 0.1] \) |
| \( x_1, x_2, x_1^2 \) | \( \theta = [1, 0.7, -0.5, 0.4] \) | \( \theta = [1, -0.3, 0.3, 0.1] \) |
| \( x_1, x_2, x_2^2 \) | \( \theta = [1, -0.4, 0.5, 0.3] \) | \( \theta = [1, 0.3, -0.3, 0.1] \) |
| \( x_1, x_2, x_1^2, x_2^2 \) | \( \theta = [1, 0.5, 0.5, -0.3, 0.5] \) | \( \theta = [1, 0.3, 0.3, -0.3, 0.1] \) |

4.1.2 Poisson regression

A simulation study based on Poisson regression was constructed similar to the logistic regression case. In terms of generating covariate values, uniform and Multivariate Normal distributions were considered. Values for \( \theta \) were selected as described above, and are given in Table 3. Model robust subsampling probabilities for Poisson regression can be obtained by applying Theorems 1 and 2 which yields the following:

\[
\phi_m^{MSE} = \sum_{q=1}^{Q} \frac{\sum_{j=1}^{N} |y_j - \hat{\lambda}_q| ||J_q^{-1}x_q||}{\sum_{j=1}^{N} |y_j - \hat{\lambda}_q| ||J_q^{-1}x_q||}, \quad \phi_m^{Vc} = \sum_{q=1}^{Q} \frac{\sum_{j=1}^{N} |y_j - \hat{\lambda}_q||x_q||}{\sum_{j=1}^{N} |y_j - \hat{\lambda}_q||x_q||}
\]

with \( y_i \in N_0 \) or non-negative integers, \( \lambda_{q,i} = \exp(\hat{\theta}_{qMLE}^T x_{qi}) \), \( J_q = N^{-1} \sum_{h=1}^{N} \hat{\lambda}_q x_{qh}(x_{qh})^T \) and \( i = 1, \ldots, N \). The specific version of Algorithm 2 for Poisson regression is given in Supplementary Material S1 as Algorithm S4.

SMSE and average model information when the covariate data was generated from a uniform distribution and the response formed through a Poisson regression model are shown in Fig. 2. On average, random sampling performs worst, while the proposed model robust approach and the optimal subsampling method based on the data generating model perform the best, and have similar SMSE and average model information values. As for logistic regression, our model robust approach is preferable overall even though the complex model performs well in some instances. Again, the use of the optimal subsampling algorithm can lead to notable increases in SMSE when the assumed model is incorrect. Similar results were obtained when the covariate data was generated from a uniform distribution with \( L \)-optimality, see Supplementary Material S3.1. In addition, similar results were observed when SMSE was evaluated based on \( \hat{\theta}_{MLE} \), see Supplementary Material S3.2.

4.2 Real world applications

The three subsampling methods are applied to analyse the “Skin segmentation” and “New York City taxi fare” data under logistic and Poisson regression, respectively. In the simulation study, the parameters were specified for the data generating model.
Fig. 2 Logarithm of (a) average model information and (b) SMSE for the subsampling methods for the Poisson regression model under $\phi^{mMSE}$. Covariate data were generated from the uniform distribution.
However, in real-world applications these are unknown. In such cases, the subsampling methods cannot be compared as in Sect. 4.1. Instead, for the $Q$ model set, the Summed SMSE (SSMSE) under each subsampling methods can be evaluated as follows:

$$SSMSE(\hat{\theta}_{MLE}) = \sum_{q=1}^{Q} \alpha_q SMSE(\tilde{\theta}_q, \hat{\theta}_{q,MLE}),$$

where $\tilde{\theta}_q$ is a matrix with the estimated parameters for the $q$th model over $M$ simulations and $\hat{\theta}_{q,MLE}$ denotes $\hat{\theta}_{MLE}$ for the $q$th model.

In the following real-world examples, the set of $Q$ models includes the main effects model, with intercept, and all possible combinations of quadratic terms for continuous covariates with $\alpha_q$ set to $1/Q$. Again, these were constructed based on the work of Shi and Tang (2021).

### 4.2.1 Identifying skin from colours in images

Rajen and Abhinav (2012) considered the problem of identifying skin-like regions in images as part of the complex task of performing facial recognition. For this purpose, Rajen and Abhinav (2012) collated the “Skin segmentation” data set, which consists of RGB (R-red, G-green, B-blue) values of randomly sampled pixels from $N = 245,057$ face images (out of which 50,859 are skin samples and 194,198 are non-skin samples) from various age groups, race groups and genders. Bhatt et al. (2009) and Binias et al. (2018) applied multiple supervised machine learning algorithms to classify if images are skin or not based on the RGB colour data. In addition, Abbas and Farooq (2019) conducted the same classification task for two different colour spaces, HSV (H-hue, S-saturation, V-value) and YCbCr (Y-luma component, Cb-blue difference chroma component, Cr-red difference chroma component), by transforming the RGB colour space.

We consider the same classification problem but use a logistic regression model. Skin presence is denoted as one and skin absence is denoted as zero. Each colour vector is scaled to have a mean of zero and a variance of one (initial range was between 0 and 255). To compare subsampling methods, we set $r_0 = 200$, $r = 200, 300, \ldots, 1800$ for the subsamples, and construct a set of $Q$ models by considering an intercept with main effects model with all covariates (scaled colours red, green and blue) as the base model, and form all alternative models by including different combinations of quadratic terms of all covariates. This leads to $Q = 8$. Each of these models is considered equally likely a priori.

Figure 3 shows the SSMSE values over the $Q$ models for various subsample sizes obtained by applying logistic regression to the “Skin segmentation” data. The proposed model robust approach performs similarly to the optimal subsampling method for $r = 200$ and 300. However, when the sample size increases the optimal subsampling method performs poorly, and after $r = 1300$ random sampling actually has lower SSMSE values than optimal subsampling. The same is not true for our proposed model robust approach which has the lowest values of SSMSE throughout the selected
sample sizes under $\phi^m_{MSE}$ and $\phi^m_{Vc}$ optimal subsampling criteria. Of note, the optimal subsampling approach based on $\phi^m_{MSE}$ performed worse than random sampling in some cases. That is, minimal decrease in SSMSE was observed for optimal subsampling beyond $r=900$ compared to random sampling. Upon investigating this, it was found that this was likely due to one of the models being particularly poor (leading to higher SMSE with increasing $r$ values) for describing the data, and therefore led to inflated SSMSE values. Despite this, we note that the model robust approach appears to perform well in general.

4.2.2 New York City taxi cab usage

New York City (NYC) taxi trip and fare information from 2009 onward, consisting of over 170 million records each year, are publicly available courtesy of the New York City Taxi and Limousine Commission. Some analyses of interest of these NYC taxi data include: a taxi driver’s decision process to pick up a fair or cruise for customers which was modelled via a logistic regression model (Yazici et al. 2013); taxi demand and how it is impacted by location, time, demographic information, socioeconomic status and employment status which was modelled via a multiple linear regression model (Yang and Gonzales 2014); and the dependence of taxi supply and demand on location and time considered via the Poisson regression model (Yang and Gonzales 2017).

In our application, we are interested in how taxi usage varies with day of the week (weekday/weekend), season (winter/spring/summer/autumn), fare amount, and cash payment. The data used in our application is the “New York City taxi fare” data for
Fig. 4 Logarithm of SSMSE over the available models for Poisson regression applied on the 2013 NYC taxi usage data.

the year 2013, hosted by the University of Illinois Urbana Champaign (Donovan and Work 2016).

Each data point includes the number of rides recorded against the medallion number (a license number provided for a motor vehicle to operate as a taxi) \( y \), weekday or not \( x_1 \), winter or not \( x_2 \), spring or not \( x_3 \), summer or not \( x_4 \), summed fare amount in dollars \( x_5 \), and the ratio of cash payment trips in terms of all trips \( x_6 \). The continuous covariate \( x_5 \) was scaled to have a range of zero and one. Poisson regression was used to model the relationship between the number of rides per medallion and these covariates.

For our study, three subsampling methods are compared for the analysis of the taxi fare data, assigning \( r_0 = 100 \) and \( r = 100, 200, \ldots, 1900 \) for subsamples. The model set consists of the main effects model \( x_1, x_2, x_3, x_4, x_5, x_6 \) and all possible combinations of the quadratic terms of the continuous covariates \( x_5^2, x_6^2 \) which leads to \( Q = 4 \). Each of these models were considered equally likely \textit{a priori}.

The SSMSE over the four models is shown in Fig. 4. Our proposed model robust approach outperforms the optimal subsampling method for almost all sample sizes for both \( \phi^MSE \) and \( \phi^VC \) subsampling probabilities. Under \( \phi^VC \), the model robust approach actually initially performs worse than the optimal subsampling approach but this is quickly reserved as \( r \) is increased. Random sampling performs the worst, suggesting that there is benefit in using targeted sampling approaches (as proposed here) over random selection.
5 Discussion

In this article, we proposed a model robust subsampling approach for GLMs. This new approach extends the current optimal subsampling approach by considering a set of models (rather than a single model) when determining the subsampling probabilities. The exact formulation of these probabilities was via model averaging which was proposed based on results from Theorems 1 and 2. The robustness properties of this proposed approach were demonstrated in a simulation study, and in two real-world analysis problems, where it outperformed optimal subsampling and random sampling. Accordingly, we suggest that our model robust approach could be considered in future big data analysis problems.

The main limitation of our proposed approach is that the specified model set could become quite large as the number of covariates increases. We note, however, that this is also an issue for big data analysis (as each model would still need to be fitted and compared). Such an issue could potentially be addressed here in future work by extending the specification of the model set to a flexible class of models. This could be, for example, through a generalised additive model (Hastie and Tibshirani 1986; De Silva et al. 2022) or the inclusion of a discrepancy term in the linear predictor (Krishna et al. 2021). Such an approach would seem efficient as it would reduce the number of models that need to be considered in the model set. In a similar vein, one could consider determining subsampling probabilities based on the work of Adewale and Wiens (2009) and Adewale and Xu (2010) aimed at obtaining robust designs under a misspecified model. Another avenue of interest that could also be explored is determining $\alpha_q$ and/or reducing the model set based on the initial subsample of $F_N$ where, for example, models that clearly do not support the data could be dropped. Such extensions are planned as future research.

Supplementary information

The proof for Theorem 1 is included in the Appendix. Specific subsampling algorithms for logistic and Poisson regression models, code for the simulation study and real-world applications and finally some additional figures are available in the online Supplementary Material.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00362-023-01446-9.

Acknowledgements The authors would like to thank both reviewers and the Handling Editor for their comments which helped to improve our paper. Amalan Mahendran was supported by the Australian Government Research Training Program (RTP) Stipend (International) and RTP Fees Offset (International) Scholarship from the Queensland University of Technology. Computational resources and services used in this work were provided by the HPC and Research Support Group, Queensland University of Technology, Brisbane, Australia.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions
Proof

\[ \sum_{q=1}^{Q} \alpha_q \text{tr}(V_q) = \sum_{q=1}^{Q} \alpha_q \text{tr} \left( J_{X_q}^{-1} V_q c J_{X_q}^{-1} \right) \]

\[ = \frac{1}{N^2} \sum_{q=1}^{Q} \sum_{i=1}^{N} \left[ \frac{1}{\phi_{q_i}} \{ y_i - \dot{\psi}(u(\hat{\theta}_{q,MLE}^{T} x_{q_i})) \}^2 J_{X_q}^{-1} \right] \]

\[ = \frac{1}{N^2} \sum_{q=1}^{Q} \sum_{i=1}^{N} \left[ \frac{1}{\phi_{q_i}} \{ y_i - \ddot{\psi}(u(\hat{\theta}_{q,MLE}^{T} x_{q_i})) \}^2 \right] J_{X_q}^{-1} \]

\[ = \frac{1}{N^2} \sum_{q=1}^{Q} \sum_{i=1}^{N} \left[ \frac{1}{\phi_{q_i}} \{ y_i - \ddot{\psi}(u(\hat{\theta}_{q,MLE}^{T} x_{q_i})) \}^2 \right] J_{X_q}^{-1} \]

\[ \geq \frac{1}{N^2} \sum_{q=1}^{Q} \sum_{i=1}^{N} \left[ \frac{1}{\phi_{q_i}} \{ y_i - \ddot{\psi}(u(\hat{\theta}_{q,MLE}^{T} x_{q_i})) \} \right] J_{X_q}^{-1} \]

\[ \geq \frac{1}{N^2} \sum_{q=1}^{Q} \sum_{i=1}^{N} \left[ \frac{1}{\phi_{q_i}} \{ y_i - \ddot{\psi}(u(\hat{\theta}_{q,MLE}^{T} x_{q_i})) \} \right] J_{X_q}^{-1} \]

\[ = \frac{1}{N^2} \sum_{q=1}^{Q} \sum_{i=1}^{N} \left[ \frac{1}{\phi_{q_i}} \{ y_i - \ddot{\psi}(u(\hat{\theta}_{q,MLE}^{T} x_{q_i})) \} \right] J_{X_q}^{-1} \]

where the last inequality follows from the Cauchy–Schwarz inequality, and the equality in it holds if and only if

\[ \phi_{q_i} \propto |y_i - \ddot{\psi}(u(\hat{\theta}_{q,MLE}^{T} x_{q_i}))| \| J_{X_q}^{-1} \ddot{\psi}(u(\hat{\theta}_{q,MLE}^{T} x_{q_i})) x_{q_i} \| . \]
Then, for the $q$th model, the subsampling probabilities would be:

$$
\phi_{q_i}^{MSE} = \frac{|y_i - \hat{\psi}(u(\hat{\theta}_q^{MLE} x_{qi}))| \| J_{X_q}^{-1} \hat{u}(\hat{\theta}_q^{MLE} x_{qi}) x_{qi} \|}{\sum_{j=1}^{N} |y_j - \hat{\psi}(u(\hat{\theta}_q^{MLE} x_{qj}))| \| J_{X_q}^{-1} \hat{u}(\hat{\theta}_q^{MLE} x_{qj}) x_{qj} \|},
$$

(A2)

for $i = 1, \ldots, N$, and $\sum_{q=1}^{Q} \alpha_q \text{tr}(V_q)$ attains its minimum. □

References

Abbas AR, Farooq AO (2019) Skin detection using improved ID3 algorithm. Iraqi J Sci 60(2):402–410

Adewale AJ, Wiens DP (2009) Robust designs for misspecified logistic models. J Stat Plan Inference 139(1):3–15. https://doi.org/10.1016/j.jspi.2008.05.022

Adewale AJ, Xu X (2010) Robust designs for generalized linear models with possible overdispersion and misspecified link functions. Comput Stat Data Anal 54(4):875–890. https://doi.org/10.1016/j.csda.2010.09.032

Ai M, Wang F, Yu J et al (2021) Optimal subsampling for large-scale quantile regression. J Complex 62(101):512. https://doi.org/10.1016/j.jco.2020.101512

Ai M, Yu J, Zhang H et al (2021) Optimal subsampling algorithms for big data regressions. Stat Sin 31:749–772. https://doi.org/10.5705/ss.202018.0439

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723. https://doi.org/10.1109/TAC.1974.1100705

Atkinson A, Donev A, Tobias R (2007) Optimum experimental designs, with SAS. Oxford statistical science series. OUP, Oxford

Bhatt RB, Sharma G, Dhall A et al (2009) Efficient skin region segmentation using low complexity fuzzy decision tree model. In: 2009 Annual IEEE India conference. IEEE, pp 1–4. https://doi.org/10.1109/INDCON.2009.5409447

Binias B, Frąckiewicz M, Jaskot K et al (2018) Pixel classification for skin detection in color images. In: Advanced technologies in practical applications for national security. Springer, pp 87–99. https://doi.org/10.1007/978-3-319-64674-9_6

Chang X, Lin SB, Wang Y (2017) Divide and conquer local average regression. Electron J Stat 11(1):1326–1350. https://doi.org/10.1214/17-EJS1265

Cheng Q, Wang H, Yang M (2020) Information-based optimal subdata selection for big data logistic regression. J Stat Plan Inference 209:112–122. https://doi.org/10.1016/j.jspi.2020.03.004

Cleveland SW, Hafen R (2014) Divide and recombine (D&R): data science for large complex data. Stat Anal Data Min ASA Data Sci J 7(6):425–433. https://doi.org/10.1002/sam.11242

De Silva D, Fisher R, Radford B et al (2022) Model-robust Bayesian design through generalised additive models for monitoring submerged shoals. arXiv preprint. arXiv:2208.13980

Deldossi L, Tommassi C (2022) Optimal design subsampling from big datasets. J Qual Technol 54(1):93–101. https://doi.org/10.1080/00224065.2021.1889418

Donovan B, Work D (2016) New York City taxi trip data (2010–2013). https://doi.org/10.13012/J8PN93H8

Drovandi CC, Holmes C, McGree JM et al (2017) Principles of experimental design for big data analysis. Stat Sci 32(3):385–404. https://doi.org/10.1214/16-STS604

Guha S, Hafen R, Rounds J et al (2012) Large complex data: divide and recombine (D&R) with RHIPE. Stat Anal Data Min ASA Data Sci J 5(6):79–97. https://doi.org/10.1002/sam.10687

Hastie T, Tibshirani R (1986) Generalized additive models. Stat Sci 1(3):297–310

Karmakar B, Mukhopadhyay I (2020) Statistical validity and consistency of big data analytics: a general framework. Stat Appl 18(2):369–381

Kleiner A, Talwalkar A, Sarkar P et al (2014) A scalable bootstrap for massive data. J R Stat Soc B 76(4):795–816. https://doi.org/10.1111/rssb.12050

Krishna A, Joseph VR, Ba S et al (2021) Robust experimental designs for model calibration. J Qual Technol. https://doi.org/10.1080/00224065.2021.1930618
Lee J, Schifano ED, Wang H (2021) Fast optimal subsampling probability approximation for generalized linear models. Econom Stat. https://doi.org/10.1016/j.ecosta.2021.02.007

Li C, Hung Y, Xie M (2020) A sequential split-and-conquer approach for the analysis of big dependent data in computer experiments. Can J Stat 48(4):712–730. https://doi.org/10.1002/cjs.11559

Lin N, Xi R (2011) Aggregated estimating equation estimation. Stat Interface 4(1):73–83. https://doi.org/10.4310/SII.2011.v4.n1.a8

Ma P, Sun X (2015) Leveraging for big data regression. Wiley Interdiscip Rev Comput Stat 7(1):70–76. https://doi.org/10.1002/wics.1324

Ma P, Mahoney WM, Yu B (2015) A statistical perspective on algorithmic leveraging. J Mach Learn Res 16(27):861–911

Martinez-Mosquera D, Navarrete R, Lujan-Mora S (2020) Modeling and management big data in databases—a systematic literature review. Sustainability 12(2):634. https://doi.org/10.3390/su12020634

Meng C, Xie R, Mandal A et al (2021) LowCon: a design-based subsampling approach in a misspecified linear model. J Comput Graph Stat 30(3):694–708. https://doi.org/10.1080/10618600.2020.1844215

Nelder JA, Wedderburn RW (1972) Generalized linear models. J R Stat Soc A 135(3):370–384. https://doi.org/10.2307/2344614

Nguyen T, Gosine RG, Warrian P (2020) A systematic review of big data analytics for oil and gas industry 4.0. IEEE Access 8:61183–61201. https://doi.org/10.1109/ACCESS.2020.2979678

R Core Team (2021) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Rajen B, Abhinav D (2012) Skin segmentation. UCI Machine Learning Repository

Rehman A, Naz S, Razzaq I (2021) Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. Multimed Syst. https://doi.org/10.1007/s00530-020-00736-8

RStudio Team (2020) RStudio: integrated development environment for R. RStudio, PBC, Boston

Schifano DE, Wu J, Wang C et al (2016) Online updating of statistical inference in the big data setting. Technometrics 58(3):393–403. https://doi.org/10.1080/00401706.2016.1142900

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Shi C, Tang B (2021) Model-robust subdata selection for big data. J Stat Theory Pract 15(4):1–17. https://doi.org/10.1007/s42519-021-00217-9

Vaughan G (2020) Efficient big data model selection with applications to fraud detection. Int J Forecast 36(3):1116–1127. https://doi.org/10.1016/j.ijforecast.2018.03.002

Wang H (2019) More efficient estimation for logistic regression with optimal subsamples. J Mach Learn Res 20(132):1–59

Wang C, Chen MH, Schifano E et al (2016) Statistical methods and computing for big data. Stat Interface 9(4):399–414. https://doi.org/10.4310/SII.2016.v9.n4.a1

Wang H, Zhu R, Ma P (2018) Optimal subsampling for large sample logistic regression. J Am Stat Assoc 113(522):829–844. https://doi.org/10.1080/01621459.2017.1292914

Wang H, Yang M, Stufken J (2019) Information-based optimal subdata selection for big data linear regression. J Am Stat Assoc 114(525):393–405. https://doi.org/10.1080/01621459.2017.1408468

Xue Y, Wang H, Yan J et al (2020) An online updating approach for testing the proportional hazards assumption with streams of survival data. Biometrics 76(1):171–182. https://doi.org/10.1111/biom.13137

Yang C, Gonzales EJ (2014) Modeling taxi trip demand by time of day in New York City. Transp Res Rec 2429(1):110–120. https://doi.org/10.3141/2429-12

Yang C, Gonzales EJ (2017) Modeling taxi demand and supply in New York City using large-scale taxi GPS data. Springer, pp 405–425. https://doi.org/10.1007/978-3-319-40902-3_22

Yao Y, Wang H (2019) Optimal subsampling for softmax regression. Stat Pap 60(2):585–599. https://doi.org/10.1007/s00362-018-01068-6

Yao Y, Wang H (2021) A review on optimal subsampling methods for massive datasets. J Data Sci 19(1):151–172. https://doi.org/10.6339/21-JDS999

Yazici MA, Kamga C, Singhal A (2013) A big data driven model for taxi drivers’ airport pick-up decisions in New York City. In: 2013 IEEE international conference on big data. IEEE, pp 37–44. https://doi.org/10.1109/BigData.2013.6691775

Yi SY, Zhou YD (2023) Model-free global likelihood subsampling for massive data. Stat Comput 33(1):1–16. https://doi.org/10.1007/s11222-022-10185-0
Yu J, Wang H (2022) Subdata selection algorithm for linear model discrimination. Stat Pap. https://doi.org/10.1007/s00362-022-01299-8
Zhang Y, Zhao Y (2015) Astronomy in the big data era. Data Sci J. https://doi.org/10.5334/dsj-2015-011

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.