A model robust design approach for optimally sub-sampling big data

Amalan Mahendran[1,2], Helen Thompson[1,2], James M. McGree[1,2]
[1] *School of Mathematical Sciences, Queensland University of Technology*
[2] *Centre for Data Science, Queensland University of Technology*

1. Introduction

Subsampling methods have been proposed as computationally efficient approaches to analyse big data. Here, we consider sampling big data $F_N = (\boldsymbol{X}_0, \boldsymbol{y})$, with data matrix $\boldsymbol{X}_0 = (\boldsymbol{x}_{01}, \ldots, \boldsymbol{x}_{0N})^T \in R^{N\times p}$ and response vector $\boldsymbol{y} = (y_1, \ldots, y_N)^T$, based on subsampling probabilities ϕ, where ϕ is based on an optimality criterion e.g. A-optimality (Ai et al. 2021). However, a key limitation of this approach is that ϕ depends on an assumed model through the optimality criterion, and such a model may be difficult to specify in practice. To overcome this limitation, we propose to consider a set of models and a model averaging approach for determining ϕ.

2. Generalised Linear Models (GLMs)

A GLM is defined via three components:

- distribution of $\boldsymbol{y}$, such that $f(y;\omega,\gamma) = \exp\left(\frac{y\omega-\psi(\omega)}{a(\gamma)} + b(y,\gamma)\right)$, for some functions $\psi(.), a(.)$ and $b(.)$, natural parameter ω and dispersion γ;
- linear predictor $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{p+t})^T$ is the parameter vector and $\boldsymbol{X} = h(\boldsymbol{X}_0) \in R^{N\times(p+t)}$ for some function $h(.)$, i.e., columns may be appended to $\boldsymbol{X}_0$ in big data set F_N corresponding to an intercept and/or higher-order terms to create $\boldsymbol{X}$; and
- link function $g(.)$, such that $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$ for $\boldsymbol{\mu} = E(\boldsymbol{y})$.

3. General subsampling algorithm

Algorithm 1 is a subsampling algorithm for GLMs, where $\boldsymbol{\theta}$ is estimated using a weighted log-likelihood with weights $\frac{1}{\phi}$.

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

1 **Sampling:** Assign ϕ_i to $\{\boldsymbol{x}_i, y_i\}$, for $i = 1, ..., N$, $\phi_i \in (0,1)$, $\sum_{i=1}^N \phi_i = 1$, and draw a subsample of size r from $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\phi})$ with replacement to yield $S = \{\boldsymbol{x}_l^*, y_l^*, \phi_l^*\}_{l=1}^r = (\boldsymbol{X}^*, \boldsymbol{y}^*, \boldsymbol{\phi}^*)$.

2 **Estimation:** Based on S, find $\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}|\boldsymbol{X}^*, \boldsymbol{y}^*, \boldsymbol{\phi}^*) \equiv \arg\max_{\boldsymbol{\theta}} \frac{1}{r}\sum_{l=1}^r \frac{y_l^* u(\boldsymbol{\theta}^T\boldsymbol{x}_l^*) - \psi(u(\boldsymbol{\theta}^T\boldsymbol{x}_l^*))}{\phi_l^*}$. **Output:** $\tilde{\boldsymbol{\theta}}$ and S.

To determine the values of ϕ, Ai et al. (2021) considered the A-optimality criterion, corresponding to minimisation of the asymptotic mean squared error of $\tilde{\boldsymbol{\theta}}$ (tr($\boldsymbol{V}$)), to obtain the subsampling probabilities:

$$\phi_i = \frac{|y_i - \dot{\psi}(u(\hat{\boldsymbol{\theta}}_{MLE}^T\boldsymbol{x}_i))|\,||\boldsymbol{J}_{\boldsymbol{X}}^{-1}\dot{u}(\hat{\boldsymbol{\theta}}_{MLE}^T\boldsymbol{x}_i)\boldsymbol{x}_i||}{\sum_{j=1}^N |y_j - \dot{\psi}(u(\hat{\boldsymbol{\theta}}_{MLE}^T\boldsymbol{x}_j))|\,||\boldsymbol{J}_{\boldsymbol{X}}^{-1}\dot{u}(\hat{\boldsymbol{\theta}}_{MLE}^T\boldsymbol{x}_j)\boldsymbol{x}_j||}, \quad (1)$$

where $\hat{\boldsymbol{\theta}}_{MLE}$ is an estimate of $\boldsymbol{\theta}$ based on the whole big data set and $\boldsymbol{J}_{\boldsymbol{X}}$ is the observed information matrix. As $\boldsymbol{\phi}$ depends on $\hat{\boldsymbol{\theta}}_{MLE}$, a two stage subsampling strategy was proposed where an initial subsample was used to provide an estimate of $\hat{\boldsymbol{\theta}}_{MLE}$.

4. Model robust optimal subsampling method

Following Ai et al. (2021), ϕ depends on a model that is assumed to appropriately describe the big data. Unfortunately, proposing such a model can be difficult in practice, and there may be a variety of models that appropriately describe the data. To overcome this, we consider a set of Q models that encapsulate a variety of scenarios that may be observed within the big data. For each model, we define model probabilities α_q for $q = 1, \ldots, Q$ such that $\sum_{q=1}^Q \alpha_q = 1$, which represents our *a priori* belief about the appropriateness of each model. The subsampling probabilities are then given by **Theorem 1**.

Theorem 1. *For a set of Q models with model probability α_q for the q-th model, $q = 1, \ldots Q$, if the subsampling probabilities are selected as follows:*

$$\phi_i = \textstyle\sum_{q=1}^Q \alpha_q \phi_{qi},$$

where $\sum_{q=1}^Q \alpha_q = 1$ and ϕ_{qi} is given by Eq. (1) for the q-th model, then $\sum_{q=1}^Q \alpha_q tr(\boldsymbol{V}_q)$ attains its minimum – $tr(\boldsymbol{V}_q)$ is the asymptotic mean squared error of $\tilde{\boldsymbol{\theta}}$ corresponding to model q.

5. I) Setup of simulation study - Logistic regression

Compare random sampling (RS), optimal subsampling (OS) and model robust optimal subsampling (MROS) for models in Table 1. Use $\alpha_q = 1/4$, $N = 10000$, $M = 1000$ simulations and subsample sizes $r = 100, 200, \ldots, 1400$.

Table 1: Model set for F_N

$-2 + 1.5\boldsymbol{x}_1 + 0.3\boldsymbol{x}_2$
$-2 + 1.7\boldsymbol{x}_1 - 1.2\boldsymbol{x}_2 + 0.2\boldsymbol{x}_1^2$
$-2 - 1.3\boldsymbol{x}_1 + 1.9\boldsymbol{x}_2 + 0.9\boldsymbol{x}_2^2$
$-2 + 1.9\boldsymbol{x}_1 + 1.9\boldsymbol{x}_2 + 0.9\boldsymbol{x}_1^2 + 0.7\boldsymbol{x}_2^2$

5. II) Simulation study - methodology and results

- For each model, F_N was constructed by using an exponential distribution to generate covariates (x_1, x_2) and a logistic regression model for the corresponding response (y). The performance of the sampling methods was then compared for each F_N through evaluating six scenarios:

  1. RS to estimate parameters of data generating model;
  2. OS under data generating model;
  3 – 5. OS under alternative models i.e., to estimate parameters of data generating model using samples obtained from alternative model;
  6. MROS to estimate parameters of data generating model.

- Simulated Mean Squared Error (SMSE) was used to compare each approach as follows: $SMSE(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^M \sum_{n=1}^{p+t}(\tilde{\theta}_{nm} - \theta_n)^2$, where θ_n is the n-th underlying parameter of the data generating model and $\tilde{\theta}_{nm}$ is the estimate of this parameter from the m-th simulation.
- Under OS, the data generating model (red) is typically preferred within the model set (smallest SMSE), as it is the case where the appropriate data generating model was correctly assumed to describe the big data (Fig. 1).
- Increases in the SMSE are observed when the incorrect model (pink) is considered for OS compared to the data generating model (red). The proposed model robust approach (green) performs similarly to the OS approach (red).

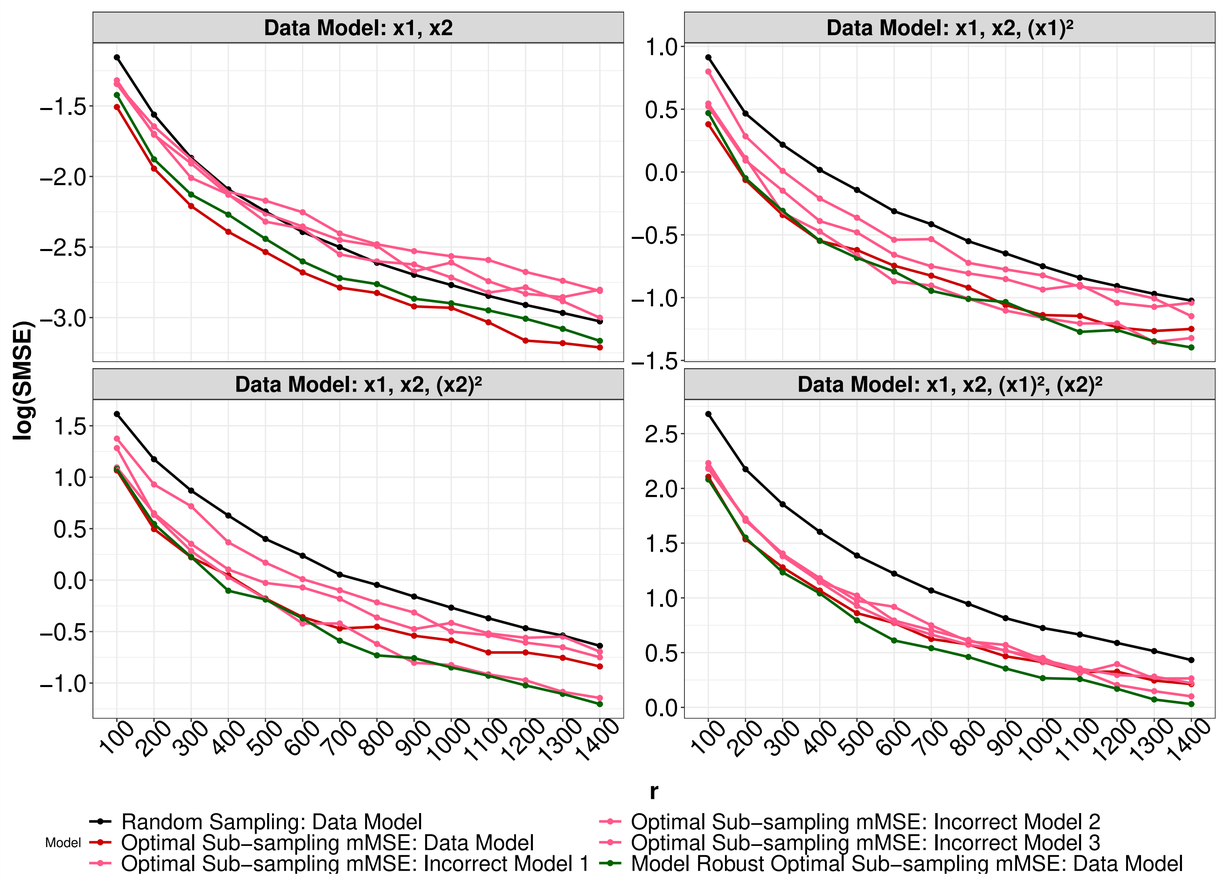

Figure 1: Logarithm of SMSE (smaller is better) for the subsampling methods for the logistic regression model.

6. Real world application

- "Skin segmentation" data, classify $N = 245,057$ face images (out of which $50,859$ are skin samples and $194,198$ are non-skin samples) based on RGB (R-red, G-green, B-blue) values of randomly sampled pixels.
- $Q = 8$ models, includes the main effects model, with intercept, and all possible combinations of quadratic terms for continuous covariates. SMSE under each subsampling method was evaluated for various r and $M = 1000$ simulations as $SSMSE(\hat{\boldsymbol{\theta}}) = \sum_{q=1}^Q SMSE_q(\hat{\boldsymbol{\theta}})$.
- MROS (green) consistently outperforms OS (red) and RS (black) (Fig. 2).

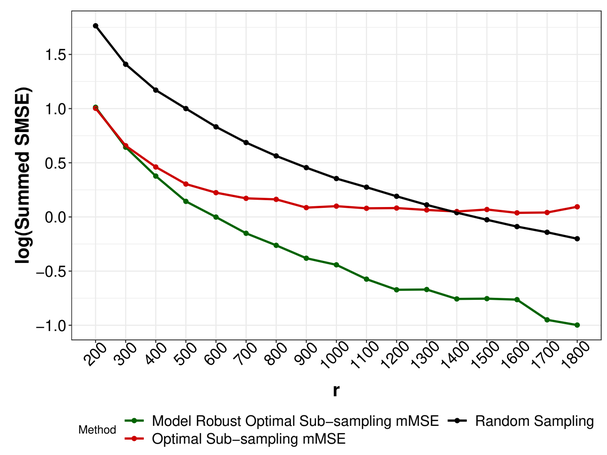

Figure 2: Logarithm of summed SMSE over the available models for logistic regression applied on the "Skin segmentation" data.

7. Conclusion

Our MROS approach, with supporting theoretical results, performs better than OS and RS for various subsample sizes. Future work is to extend our approach for a flexible set of models for example Generalised Additive Models or the inclusion of a discrepancy term in the linear predictor.

References

Ai, M., J. Yu, H. Zhang, and H. Wang (2021). "Optimal subsampling algorithms for Big data regressions". In: *Statistica Sinica* 31, pp. 749–772.