A Subsampling Strategy for AIC-based Model Averaging with Generalized Linear Models

Jun Yu
School of Mathematics and Statistics, Beijing Institute of Technology
HaiYing Wang
Department of Statistics, University of Connecticut
Mingyao Ai
LMAM, School of Mathematical Sciences and Center for Statistical Science, Peking University

August 7, 2024

Abstract

Subsampling is an effective approach to address computational challenges associated with massive datasets. However, existing subsampling methods do not consider model uncertainty. In this paper, we investigate the subsampling technique for the Akaike information criterion (AIC) and extend the subsampling method to the smoothed AIC model-averaging framework in the context of generalized linear models. By correcting the asymptotic bias of the maximized subsample objective function used to approximate the Kullback–Leibler divergence, we derive the form of the AIC based on the subsample. We then provide a subsampling strategy for the smoothed AIC model-averaging estimator and study the corresponding asymptotic properties of the loss and the resulting estimator. A practically implementable algorithm is developed, and its performance is evaluated through numerical experiments on both real and simulated datasets.

Keywords: Big Data, Information Criterion, Nonuniform Subsampling, Smoothed AIC

1 Introduction

Subsampling is a popular method to address big data challenges imposed by exponentially growing data volumes. In many areas of analysis, it successfully alleviates the computational burden brought by large-scale datasets. There are two basic approaches in current research investigations. One approach is to find the most representative data points for the entire dataset, which is model-free. Typical examples include Latin-hypercube-design-based subsampling (Zhao et al., 2018; He et al., 2024), uniform-design-based subsampling (Shi and Tang, 2021; Zhang et al., 2023; Zhou et al., 2023), and support-points-based subsampling (Mak and Joseph, 2018; Joseph and Mak, 2021; Joseph and Vakayil, 2022). Another approach is model-assisted subsampling, which aims to find the most informative data points to improve estimation efficiency for specific models. Important works include, but are not limited to, leverage score subsampling (Ma et al., 2015, 2022), Lowcon (Meng et al., 2021), and information-based optimal subsampling (Wang et al., 2019; He et al., 2024) for linear models; local case-control subsampling (Fithian and Hastie, 2014; Han et al., 2020) and optimal subsampling motivated by the A-optimality criterion (OSMAC, Wang et al., 2018) for logistic regression; and optimal subsampling methods for other more complicated models (Wang and Ma, 2021; Ai et al., 2021; Yu et al., 2022, 2024; Ye et al., 2024).

The aforementioned investigations focus on estimating the unknown parameters with a given model. In practice, the true data-generating model is always unknown, and multiple candidate models are often plausible. For example, in high-energy physics, scientists are interested in determining if a process produces supersymmetric particles or not (Baldi et al., 2014). The supersymmetric benchmark dataset[1] in the UCI machine learning repository was created to study the two classes of processes. Each record in the dataset represents a hypothetical collision between particles with eight kinematic properties features such as energy levels and momenta, along with some high-level features derived by physicists to help distinguish the two classes. Researchers may build multiple candidate models with the eight kinematic features, together with higher-order features, and possibly additional features such as interactions among the eight kinematic features. Model averaging is usually regarded as a powerful tool to achieve the smallest risk in estimation among the candidate models. See

[1] `https://doi.org/10.24432/C54606`

Buckland et al. (1997); Hjort and Claeskens (2003); Hansen (2014); Yuan and Yang (2005); Claeskens et al. (2008); Liang et al. (2011); Zhang (2015); Peng and Yang (2022), among others, for the advantages of model averaging. Finding model averaging estimators with massive data can be daunting due to the computing costs in both parameter estimation and weight determination for all candidate models. To alleviate the computation burden, we investigate the subsampling strategy for model averaging.

Compared to existing approaches, designing an efficient subsampling strategy for model averaging estimators meets the following three challenges. Firstly, as shown in Wang (2019), if the model is misspecified, then "optimal" subsampling probabilities are no longer optimal and may even reduce the estimation efficiency. Thus, the basic question becomes how to design subsampling probabilities that benefit the estimation of the candidate models with larger model weights. This is unknown in the literature of subsampling. Secondly, due to the non-uniform and data-dependent sampling approach, the selected subdata and the entire data often have different distributions. Consequently, a model that is good for describing the selected subdata may fail to summarize the entire data well. Subsample-based model weights should reflect the model information distilled for the entire data. Thirdly, one may want to explore a larger number of candidate models with a larger sample size, so it is necessary to let the number of predictors and the number of candidate models grow with the subsample size. In the language of asymptotic analysis, they are allowed to diverge as the subsample size increases. Although some investigations, such as Wang et al. (2019); Ma et al. (2022), have tried to address the challenges caused by a diverging number of predictors, their studies are on linear models using least squares estimators with explicit expressions. Their results cannot be easily extended to generalized linear models due to multiple technical difficulties, e.g., no explicit forms of the estimators and multiple candidate models to consider simultaneously.

We address the aforementioned issues and study the subsampling strategy of the AIC-based model averaging approach for generalized linear models. We opt to use smoothed AIC (S-AIC) weights (Buckland et al., 1997) because they are computationally more efficient than other weighted averaging methods, such as Mallows model averaging (Hansen, 2007; Wan et al., 2010), optimal mean squared error averaging (Liang et al., 2011; Zhang et al.,

2016), and the jackknife model averaging (Hansen and Racine, 2012). In addition, the AIC and S-AIC enjoy the asymptotic efficiency property that achieves the smallest estimation loss/risk among all the candidate models (Claeskens et al., 2008, Chapter 4). To improve the performance of the model averaging estimator, we propose a **m**ini-max **a**symptotic uncertainty **s**ubsampling **s**trategy (MASS). We derive the form of the subsampled AIC by correcting the asymptotic bias in approximating a Kullback-Leibler type divergence caused by non-uniform subsampling (9), and use it to define the subsample smoothed AIC model averaging estimator. We also establish the uniform consistency of the subsample-based estimators to the full-data-based estimator across candidate models with diverging dimensions for generalized linear models (Proposition 1 and Theorem 4). The relative information loss of the subsample-based estimator to the full-data estimator is studied (Theorem 3). To the best of our knowledge, this has not been studied in the literature.

The rest of the paper is organized as follows. Section 2 describes the model setup of our investigation. Section 3 derives the expression of the subsample-based AIC and shows its asymptotic property in model selection. We introduce the subsample model averaging estimator together with a subsampling strategy in Section 4, and derive its theoretical properties. In Section 5, we present numerical studies on both simulated and real datasets. Technical proofs are relegated to the Supplementary Material.

2 Preliminaries

2.1 Model Setup and Notations

Consider response distributions from the one-parameter natural exponential family with the following density:

$$f(y|\theta) = h(y)\exp(y\theta - \psi(\theta))d\mu(y), \tag{1}$$

where θ satisfies that $\int h(y)\exp(y\theta - \psi(\theta))d\mu(y) < \infty$ under the dominating measure μ. Suppose we have n independent observations $\{(y_i, \boldsymbol{x}_i^{\mathrm{T}})^{\mathrm{T}}, i = 1, \ldots, n\}$, where y_i's $\in \mathbb{R}$ are the responses and $\boldsymbol{x}_i$'s $\in \mathbb{R}^q$ are the covariates. The conditional distribution of y_i given $\boldsymbol{x}_i$ is linked in the working model through the natural parameter θ in (1) by

$$\theta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \quad \text{for} \quad i = 1, \ldots, n. \tag{2}$$

Consider a set of m candidate models $\mathcal{M}_1, \ldots, \mathcal{M}_m$ which are used to capture the relationship between $\boldsymbol{x}$ and y through (2). Here, the kth candidate model $\mathcal{M}_k$ includes some (or all) of variables in $\boldsymbol{x}$.

To facilitate the presentation, let $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}$, $\boldsymbol{Y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathcal{F}_n = (X, \boldsymbol{Y})$, and q_k be the number of parameters in model $\mathcal{M}_k$. Let $P_k \in \mathbb{R}^{q_k \times q}$ be a selection (projection) matrix associated with $\mathcal{M}_k$ such that $P_k = (\boldsymbol{e}_{j_1}, \ldots, \boldsymbol{e}_{j_{q_k}})^{\mathrm{T}}$, where $1 \leq j_1 < \cdots < j_{q_k} \leq q$ are a subset of the column indices of the model matrix X and $\boldsymbol{e}_j \in \mathbb{R}^q$ is a unit vector with the jth element being one. With this notation, we can write $\boldsymbol{\beta}_k = P_k \boldsymbol{\beta}$. Motivated by the "bet on sparsity" principle (Hastie et al., 2009), the largest number of features to consider in a candidate model is not necessarily q. To distinguish the largest number of parameters for the models in the candidate pool and the number of the features in X, we use $q_{(L)}$ to denote the largest dimension of the candidate models among $\mathcal{M}_1, \ldots, \mathcal{M}_m$.

Using the above notations, the kth candidate model $\mathcal{M}_k$ can be written as

$$f_k(y|\boldsymbol{\beta}_k, \boldsymbol{x}) = h(y) \exp\left(y \boldsymbol{\beta}_k^{\mathrm{T}} P_k \boldsymbol{x} - \psi(\boldsymbol{\beta}_k^{\mathrm{T}} P_k \boldsymbol{x})\right), \tag{3}$$

and the full-data-based maximum likelihood estimator $\hat{\boldsymbol{\beta}}_k$ with $\mathcal{F}_n$ under model $\mathcal{M}_k$ is the maximizer of the log-likelihood function

$$\ell_k(\boldsymbol{\beta}_k) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i \boldsymbol{\beta}_k^{\mathrm{T}} P_k \boldsymbol{x}_i - \psi(\boldsymbol{\beta}_k^{\mathrm{T}} P_k \boldsymbol{x}_i)\right). \tag{4}$$

2.2 General Subsampling Framework

Let π_i be the sampling probability for the ith data point in one sampling draw and denote $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)$. Here the $\boldsymbol{\pi}$ may depend on the observed data. The subsample $\{(y_i^*, \boldsymbol{x}_i^{*\mathrm{T}}, \pi_i^*)^{\mathrm{T}}, i = 1, \ldots, r\}$ is constructed by sampling with replacement for r times according to the sampling distribution $\boldsymbol{\pi}$. Here y_i^*, $\boldsymbol{x}_i^*$, and π_i^* denote the response, predictor, and sampling probability of the ith data point in the subsample, respectively. Based on the subsample, the quasi-likelihood estimator $\tilde{\boldsymbol{\beta}}_k$ under model $\mathcal{M}_k$ is the maximizer of the following objective function:

$$\ell_k^*(\boldsymbol{\beta}_k) = \frac{1}{nr} \sum_{i=1}^{r} \frac{1}{\pi_i^*} \left(y_i^* \boldsymbol{\beta}_k^{\mathrm{T}} P_k \boldsymbol{x}_i^* - \psi(\boldsymbol{\beta}_k^{\mathrm{T}} P_k \boldsymbol{x}_i^*)\right). \tag{5}$$

For ease of presentation, we call (5) a subsample-based log-likelihood function throughout this paper, since it is an unbiased estimator of the full-data-based log-likelihood function under model $\mathcal{M}_k$.

To ensure the consistency and asymptotic normality of the resultant estimator $\tilde{\boldsymbol{\beta}}_k$ with respect to the full-data-based estimator under each candidate model, we assume the following regularity conditions.

Assumption 1. *For each candidate model $\mathcal{M}_k$, the parameter $\boldsymbol{\beta}_k$ lies in $\Lambda_k = \{\boldsymbol{\beta}_k : \|\boldsymbol{\beta}_k\| \leq C\}$, and the full-data-based estimator $\hat{\boldsymbol{\beta}}_k$ is an inner point of Λ_k with probability one. Here C is a constant and $\|\cdot\|$ denotes the l_2 norm for a vector.*

Assumption 2. *Let $\dot{\psi}$, $\ddot{\psi}$, and $\dddot{\psi}$ be the first, second, and third derivatives of ψ, respectively. There exist integrable functions $g_l(\boldsymbol{x})$ for $l = 0, \ldots, 3$, such that $\psi^2(\sum_{k=1}^m \omega_k \boldsymbol{\beta}_k^{\mathrm{T}} P_k \boldsymbol{x}) < g_0(\boldsymbol{x})$, $\dot{\psi}^6(\sum_{k=1}^m \omega_k \boldsymbol{\beta}_k^{\mathrm{T}} P_k \boldsymbol{x}) < g_1(\boldsymbol{x})$, $\ddot{\psi}^6(\sum_{k=1}^m \omega_k \boldsymbol{\beta}_k^{\mathrm{T}} P_k \boldsymbol{x}) < g_2(\boldsymbol{x})$, and $\dddot{\psi}^2(\sum_{k=1}^m \omega_k \boldsymbol{\beta}_k^{\mathrm{T}} P_k \boldsymbol{x}) < g_3(\boldsymbol{x})$. Further assume that $\sup_{\|\boldsymbol{u}\|=1} E(\|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}\|^9) < \infty$ and $E(y^6) < \infty$. Here $\omega_k \in [0, 1]$ denotes the weight of the kth model, and $\sum_{k=1}^m \omega_k = 1$.*

Assumption 3. *Denote $\lambda_{min}(\cdot)$ as the smallest eigenvalue and $\|A\|_s$ as the spectral norm of a matrix A (the largest eigenvalue for a non-negative definite matrix). Let $A(\boldsymbol{\beta}_k) = n^{-1}\sum_{i=1}^n \ddot{\psi}(\boldsymbol{\beta}_k^{\mathrm{T}} P_k \boldsymbol{x}_i) P_k \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} P_k^{\mathrm{T}}$, and $B(\boldsymbol{\beta}_k) = n^{-1}\sum_{i=1}^n (y_i - \dot{\psi}(\boldsymbol{\beta}_k^{\mathrm{T}} P_k \boldsymbol{x}_i))^2 P_k \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} P_k^{\mathrm{T}}$. With probability one, it holds that $0 < \lim_{n\to\infty} \inf_{k,\boldsymbol{\beta}_k} \lambda_{min}(A(\boldsymbol{\beta}_k)) \leq \lim_{n\to\infty} \sup_{k,\boldsymbol{\beta}_k} \|A(\boldsymbol{\beta}_k)\|_s < \infty, 0 < \lim_{n\to\infty} \inf_{k,\boldsymbol{\beta}_k} \lambda_{min}(B(\boldsymbol{\beta}_k)) \leq \lim_{n\to\infty} \sup_{k,\boldsymbol{\beta}_k} \|B(\boldsymbol{\beta}_k)\|_s < \infty$.*

Assumption 4. *For $\delta \in (0, 1/2)$, the subsampling probabilities satisfy $\sum_{i=1}^n (n^{2+\delta}\pi_i^{1+\delta})^{-1} y_i^6 = O_P(1)$, $\sup_{\|\boldsymbol{u}\|=1} \sum_{i=1}^n (n^{2+\delta}\pi_i^{1+\delta})^{-1} \|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}_i\|^9 = O_P(1)$, and $\sum_{i=1}^n (n^{2+\delta}\pi_i^{1+\delta})^{-1} g_l(\boldsymbol{x}_i) = O_P(1)$, for $l = 0, \ldots, 3$, where $g_l(\boldsymbol{x}_i)$'s are defined in Assumption 2 and $O_P(1)$ means bounded in probability.*

Assumption 5. *For some $\kappa \in (0, \infty)$,*

$$\sup_{\|\boldsymbol{u}\|=1} \max_{1\leq i\leq n} \frac{|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}_i|^6 \vee 1}{n\log^{\kappa}(n)\pi_i} = O_P(1), \qquad \sup_{k} \max_{1\leq i\leq n} \frac{\psi(\hat{\boldsymbol{\beta}}_k^{\mathrm{T}} P_k \boldsymbol{x}_i)}{n\log^{\kappa}(n)\pi_i} = O_P(1),$$

where $a \vee b = \max(a, b)$.

Assumption 1 is often assumed for the maximum likelihood estimator such as in White (1982). Assumption 2 imposes some moment conditions. Similar conditions are also assumed in Ando et al. (2017). Assumption 3 indicates that the log-likelihood function is convex and ensures that the maximum likelihood estimator is unique (Lv and Liu, 2014). Some tail behaviors of the data are required in Assumptions 4 and 5 which mitigate the inflation of the sampling variance. More precisely, it is used to ensure that the Hessian matrix of (5) concentrates around $-A(\boldsymbol{\beta}_k)$ (Chen et al., 2012), which implies that the $\ell_k^*(\boldsymbol{\beta}_k)$ is concave and the resultant estimator $\tilde{\boldsymbol{\beta}}_k$ is unique for $\mathcal{M}_1, \ldots, \mathcal{M}_m$. These assumptions are not restrictive. Taking the logistic regression as an example, Assumptions 2, 4 and 5 are naturally satisfied when the covariate distribution is sub-Gaussian for the proposed subsampling method and the uniform subsampling method.

To capture the uniform convergence rate of the subsample-based estimator, we derive the following proposition.

Proposition 1. *If Assumptions 1–5 hold and* $(\log(m) + q_{(L)} \log(q)) \log^{2\kappa}(n)/r \to 0$ *as* $n, r \to \infty$*, then for any* $\epsilon > 0$*, there exists a finite* Δ_ϵ *and* r_ϵ*, such that for all* $r > r_\epsilon$*,*

$$\mathrm{pr}\left(\sup_k \|\tilde{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_k\| \geq \sqrt{q_{(L)} \log^{\kappa}(n) \log(q)/r}\, \Delta_\epsilon \Big| \mathcal{F}_n \right) < \epsilon, \tag{6}$$

with probability approaching one.

3 Subsample-based Information Criteria

In this section, we propose an appropriate definition of the AIC in the subsampling framework. Let $f_{\mathrm{true}}(y|\boldsymbol{x})$ be the true data generating conditional density of y given $\boldsymbol{x}$ and $f_k(y|\boldsymbol{\beta}_k, \boldsymbol{x})$ be a parametric approximation under model $\mathcal{M}_k$. We assume that the distribution of $\boldsymbol{x}$ is ancillary to the regression parameter. The Kullback–Leibler (KL) divergence between the true model $f_{\mathrm{true}}(y|\boldsymbol{x})$ and candidate model $\mathcal{M}_k$ with $f_k(y|\boldsymbol{\beta}_k, \boldsymbol{x})$ is

$$\begin{aligned} &\mathrm{KL}\left(f_{\mathrm{true}}(y|\boldsymbol{x}), f_k(y|\boldsymbol{\beta}_k, \boldsymbol{x})\right) \\ &= \iint \log\left(f_{\mathrm{true}}(y|\boldsymbol{x})\right) f_{\mathrm{true}}(y|\boldsymbol{x}) dy dF_{\boldsymbol{x}} - \iint \log\left(f_k(y|\boldsymbol{\beta}_k, \boldsymbol{x})\right) f_{\mathrm{true}}(y|\boldsymbol{x}) dy dF_{\boldsymbol{x}}, \end{aligned} \tag{7}$$

where $dF_{\boldsymbol{x}}$ means the integration with respect to the marginal distribution of $\boldsymbol{x}$. Let $f_k(y|\boldsymbol{\beta}_{k,\mathrm{pop}}, \boldsymbol{x})$ with $\boldsymbol{\beta}_{k,\mathrm{pop}} = \arg\min_{\boldsymbol{\beta}_k} \mathrm{KL}(f_{\mathrm{true}}(y|\boldsymbol{x}), f_k(y|\boldsymbol{\beta}_k, \boldsymbol{x}))$ be the least false approximating model,

which achieves the smallest KL divergence under $\mathcal{M}_k$. As mentioned in Sin and White (1996), one primary purpose of information criteria is to select the model $\mathcal{M}_k$ with the smallest $\mathrm{KL}(f_{\text{true}}(y|\boldsymbol{x}), f_k(y|\boldsymbol{\beta}_{k,\text{pop}}, \boldsymbol{x}))$. We call this model the best model and denote it as $\mathcal{M}_B$. If there are multiple models that achieve the minimum KL divergence, we define $\mathcal{M}_B$ to be the model with the fewest parameters, and we assume that $\mathcal{M}_B$ is unique throughout this paper. When the true data-generating model is included in the candidate pool, $\mathcal{M}_B$ is the true model. We call a model an underfitted model if it does not include all the predictors of $\mathcal{M}_B$, and use $\mathcal{U}$ to denote the set of underfitted models. If the smallest model is the best model, then $\mathcal{U}$ is empty; if the largest model is the best model, then $\mathcal{U}$ contains $m-1$ models.

Since $\boldsymbol{\beta}_{k,\text{pop}}$ is unknown, it is estimated via the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_k$. The AIC aims to select the model $\mathcal{M}_k$ that minimizes $\mathrm{KL}(f_{\text{true}}(y|\boldsymbol{x}), f_k(y|\hat{\boldsymbol{\beta}}_k, \boldsymbol{x}))$, i.e., the KL divergence between the true model and the model estimated with the maximum likelihood Akaike (1998). In the definition of (7), the first term is a constant across all candidate models. The key to the success of model selection is to approximate the second term accurately. The law of large numbers tells us that for each fixed value of $\boldsymbol{\beta}_k$,

$$\ell_k(\boldsymbol{\beta}_k) \to E\ell_k(\boldsymbol{\beta}_k) = E_{(\boldsymbol{x},y)} \log f_k(y|\boldsymbol{\beta}_k, \boldsymbol{x}) = \iint \log(f_k(y|\boldsymbol{\beta}_k, \boldsymbol{x})) f_{\text{true}}(y|\boldsymbol{x}) dy dF_{\boldsymbol{x}}, \tag{8}$$

almost surely under appropriate integrability. However, since $\hat{\boldsymbol{\beta}}_k$ is the maximizer of $\ell_k(\boldsymbol{\beta}_k)$, $\ell_k(\hat{\boldsymbol{\beta}}_k)$ is not unbiased towards $E_{(\boldsymbol{x},y)} \log f_k(y|\hat{\boldsymbol{\beta}}_k, \boldsymbol{x})$. Akaike (1998) showed that $\ell_k(\hat{\boldsymbol{\beta}}_k)$ tends to overestimate $E_{(\boldsymbol{x},y)} \log f_k(y|\hat{\boldsymbol{\beta}}_k, \boldsymbol{x})$ and the asymptotic bias is q_k/n where q_k is the dimension of $\boldsymbol{\beta}_k$. The AIC uses q_k/n to correct the bias in $\ell_k(\hat{\boldsymbol{\beta}}_k)$ with the goal to select the estimated model that has the smallest KL divergence to the data-generating model.

In the subsampling framework with massive data, $\hat{\boldsymbol{\beta}}_k$ is hard to obtain due to the huge computational cost and hence $\boldsymbol{\beta}_{k,\text{pop}}$ is estimated by $\tilde{\boldsymbol{\beta}}_k$. To select a better working model, we need to accurately approximate the KL divergence, $\mathrm{KL}(f_{\text{true}}(y|\boldsymbol{x}), f_k(y|\tilde{\boldsymbol{\beta}}_k, \boldsymbol{x}))$. The key is to accurately approximate $E_{(\boldsymbol{x},y)} \log f_k(y|\tilde{\boldsymbol{\beta}}_k, \boldsymbol{x}) = E_{(\boldsymbol{x}_{\text{new}}, y_{\text{new}})} \log f_k(y_{\text{new}}|\tilde{\boldsymbol{\beta}}_k, \boldsymbol{x}_{\text{new}})$, where $(y_{\text{new}}, \boldsymbol{x}_{\text{new}})$ means a new observation generated from the unknown true distribution. The quantity $E_{(\boldsymbol{x},y)} \log f_k(y|\tilde{\boldsymbol{\beta}}_k, \boldsymbol{x})$ describes the goodness of the estimated model under $\mathcal{M}_k$ for predicting a future response (Konishi and Kitagawa, 2007).

Again, $\ell_k^*(\tilde{\boldsymbol{\beta}}_k)$ is biased towards $E_{(\boldsymbol{x},y)} \log f_k(y|\tilde{\boldsymbol{\beta}}_k, \boldsymbol{x})$ because the same subsample is used

to estimate both the parameter and the KL divergence. Since $\tilde{\boldsymbol{\beta}}_k$ is the maximizer of $\ell^*(\boldsymbol{\beta}_k)$, using $\ell^*(\tilde{\boldsymbol{\beta}}_k)$ directly tends to overestimate $E_{(\boldsymbol{x},y)} \log f_k(y|\tilde{\boldsymbol{\beta}}_k, \boldsymbol{x})$, which implies that $\ell_k^*(\tilde{\boldsymbol{\beta}}_k)$ overestimates the model's ability in prediction. If $\ell_k^*(\tilde{\boldsymbol{\beta}}_k)$ is naively used for model selection, it often ends up with a model that does not have the best prediction performance. The selected model tends to overfit the subsample but does not have the best representation for the full dataset.

To remove the influence of using the same subsample twice for estimating both the parameter and the KL divergence, we derive the asymptotic mean of $D_k := \ell_k^*(\tilde{\boldsymbol{\beta}}_k) - E_{(\boldsymbol{x},y)} \log f_k(y|\tilde{\boldsymbol{\beta}}_k, \boldsymbol{x})$, which provides a bias correction for estimating the KL divergence. Under Assumptions 1–5, as $r, n \to \infty$, if $q_k \log^{\kappa}(n)/r \to 0$, then

$$\begin{aligned} D_k =& \ell_k^*(\hat{\boldsymbol{\beta}}_k) - \ell_k(\hat{\boldsymbol{\beta}}_k) - (\tilde{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_k)^{\mathrm{T}} E_{(\boldsymbol{x},y)} \left(\partial \log f_k(y|\hat{\boldsymbol{\beta}}_k, \boldsymbol{x}) / \partial \boldsymbol{\beta}_k \right) \\ &+ \ell_k(\hat{\boldsymbol{\beta}}_k) - E_{(\boldsymbol{x},y)} \left(\log f_k(y|\hat{\boldsymbol{\beta}}_k, \boldsymbol{x}) \right) + (\tilde{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_k)^{\mathrm{T}} A_k (\tilde{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_k) + o_{P|\mathcal{F}_n}(q_k/r), \end{aligned} \tag{9}$$

where $o_{P|\mathcal{F}_n}$ means convergence in conditional probability given the full data.

In D_k, the term $\ell_k^*(\hat{\boldsymbol{\beta}}_k) - \ell_k(\hat{\boldsymbol{\beta}}_k)$ has a mean zero and $(\tilde{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_k)^{\mathrm{T}} E_{(\boldsymbol{x},y)} (\partial \log f_k(y|\boldsymbol{\beta}_k, \boldsymbol{x}) / \partial \boldsymbol{\beta}_k)$ has an asymptotic mean zero conditional on $\mathcal{F}_n$, so they do not contribute to the asymptotic bias. The rest terms can be decomposed into two parts. The first part $\ell_k(\hat{\boldsymbol{\beta}}_k) - E_{(\boldsymbol{x},y)} \log f_k(y|\hat{\boldsymbol{\beta}}_k, \boldsymbol{x})$ is the generalization bias from the full data to the population, which has an unconditional asymptotic mean of q_k/n according to the classical AIC theory. The second part $(\tilde{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_k)^{\mathrm{T}} A_k (\tilde{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_k)$ describes the bias from the subsample-based estimator to the full-data-based estimator which has a conditional asymptotic mean of $\mathrm{tr}(V_{k,c} A_k^{-1})/r$ according to Proposition S.2. Therefore, conditionally on $\mathcal{F}_n$, the asymptotic bias of $\ell_k^*(\tilde{\boldsymbol{\beta}}_k)$ in approximating $E_{(\boldsymbol{x},y)} \log f_k(y|\tilde{\boldsymbol{\beta}}_k, \boldsymbol{x})$ is $\mathrm{tr}(V_{k,c} A_k^{-1})/r + q_k/n$. This becomes $\mathrm{tr}(V_{k,c} A_k^{-1})/r$ if $r = o(n)$.

Based on Proposition S.2 and (9), we define the subsample-based AIC value for model $\mathcal{M}_k$ as

$$\mathrm{AIC}_{\mathrm{sub}}(\mathcal{M}_k) = -2r\ell_k^*(\tilde{\boldsymbol{\beta}}_k) + 2\mathrm{tr}\left(V_{k,c} A_k^{-1}\right) + 2rq_k/n. \tag{10}$$

Remark 1. In the subsample-based AIC in (10), the first term describes the goodness of fit for model $\mathcal{M}_k$ on the subsample and the bias correction terms (the second and third terms)

penalize the model complexity. Here $2\text{tr}(V_{k,c}A_k^{-1})$ is the bias correction term for using $2r\ell_k^*(\tilde{\boldsymbol{\beta}}_k)$ to replace $2n\ell_k(\hat{\boldsymbol{\beta}}_k)/r$, and $2rq_k/n$ is the bias correction term for $2n\ell_k(\hat{\boldsymbol{\beta}}_k)/r$. For oversampling with $r \gg n$, the term $2rq_k/n$ dominates $2\text{tr}(V_{k,c}A_k^{-1})$. In this scenario, AIC_{sub} is just r/n times the classical AIC, implying that oversampling does not give additional benefits in terms of model selection. If r is of the same order as n, there is a clear trade-off between the epistemic bias, $2n\ell_k(\hat{\boldsymbol{\beta}}_k)/r - 2n\ell_k(\boldsymbol{\beta}_{k,\text{pop}})/r = O(rq_k/n)$, and the sampling variance, $2\text{tr}(V_{k,c}A_k^{-1}) = O_P(r^{-1})$. For the more practical scenario that the subsample size is much smaller than the full sample size, $\text{tr}(V_{k,c}A_k^{-1}) \gg rq_k/n$, so the bias term in subsample-based AIC mainly comes from sampling volatility. Consequently, improving the quality of the subsample-based estimator will also help identify the best model among the candidates. Although the relation between informative subsampling and model selection is not surprising, it has not been well studied in the literature.

Theorem 1. *Under Assumptions 1–5, if* $(\log(m) + q_{(L)}\log(q))\log^{2\kappa}(n)/r \to 0$ *and* $\lim r/n < \infty$*, then as* $r \to \infty$ *and* $n \to \infty$*, the* AIC_{sub} *defined in (10) selects an underfitted model* $\mathcal{M}_k \in \mathcal{U}$ *with probability going to zero, namely,*

$$\text{pr}\Big(\arg\min_{\mathcal{M}_k} \text{AIC}_{\text{sub}}(\mathcal{M}_k) \in \mathcal{U} \Big| \mathcal{F}_n\Big) \to 0, \tag{11}$$

in probability.

Although Theorem 1 is valid for the case that $0 < \lim r/n < \infty$, there is no essential computational benefits to consider a subsample size of the same order of the full data. Despite some insights on the variability of the AIC, this setting provides no significant improvement in computation or statistical inference compared with the vanilla AIC Shibata (1997). Therefore, we focus on the case $r/n \to 0$ in the rest of the paper.

4 Subsample Smoothed AIC Model Averaging

Besides using the information criteria to filter underfitted models, model averaging is usually adopted as an alternative and the corresponding estimator can often improve the estimation efficiency (Claeskens et al., 2006, 2008). The S-AIC is a popular weighting technique due to its simplicity of implementation. When subsampling for computational efficiency, the subsample

size is typically much smaller than the full data size, so we focus on this scenario and assume $r = o(n)$ in the following of the paper. In S-AIC, we construct a weighted average of the estimators in the candidate pool. For each candidate model, we compute the weight as

$$\tilde{\omega}_k = \frac{\exp(-\mathrm{AIC}_{\mathrm{sub}}(\mathcal{M}_k)/2)}{\sum_{l=1}^{m} \exp(-\mathrm{AIC}_{\mathrm{sub}}(\mathcal{M}_l)/2)}, \tag{12}$$

for $k = 1, \ldots, m$. The subsample-based S-AIC estimator is defined as $\tilde{\boldsymbol{\beta}} = \sum_{k=1}^{m} \tilde{\omega}_k P_k^{\mathrm{T}} \tilde{\boldsymbol{\beta}}_k$, where $\tilde{\boldsymbol{\beta}}_k$ is the subsample-based estimator under $\mathcal{M}_k$.

4.1 Model Averaging Subsampling Strategy

The key idea of the S-AIC estimator is to put more weight on candidate models that are estimated to have better performance in predicting future responses. Thus, it is ideal to find a subsample that can help better approximate the $E_{(\boldsymbol{x},y)} \log f_k(y|\tilde{\boldsymbol{\beta}}_k, \boldsymbol{x})$ for all candidate models. From (9) and the discussion below it, we see that the terms $\ell_k^*(\hat{\boldsymbol{\beta}}_k) - \ell_k(\hat{\boldsymbol{\beta}}_k)$ and $(\tilde{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_k)^{\mathrm{T}} E_{(\boldsymbol{x},y)}(\partial \log f_k(y|\hat{\boldsymbol{\beta}}_k, \boldsymbol{x})/\partial \boldsymbol{\beta}_k)$ are not used to define the subsample-based AIC in (10) because their asymptotic means that given the full data are zero so they do not contribute to the asymptotic bias. However, both terms are subject to the randomness of subsampling so they do contribute to the variation of using $\ell_k^*(\tilde{\boldsymbol{\beta}}_k)$ to define the AIC. An ideal subsampling strategy should try to reduce this variation. The term $\ell_k^*(\hat{\boldsymbol{\beta}}_k) - \ell_k(\hat{\boldsymbol{\beta}}_k)$ is of order $O_{P|\mathcal{F}_n}(r^{-1/2})$. Note that $E_{(\boldsymbol{x},y)}(\partial \log f_k(y|\hat{\boldsymbol{\beta}}_k, \boldsymbol{x})/\partial \boldsymbol{\beta}_k)$ is the population score function evaluated at the full-data-based estimator under $\mathcal{M}_k$, so its elements are of order $O_P(n^{-1/2})$. Thus Proposition S.1 indicates that this term is of order $O_{P|\mathcal{F}_n}(q_k^{1/2}/(nr)^{1/2})$ and it is a small term since q_k is much smaller than n. Recall that the asymptotic bias of $\ell_k^*(\tilde{\boldsymbol{\beta}}_k)$ is of order $O_{P|\mathcal{F}_n}(q_k/r)$. Combining the variance and bias, the overall uncertainty by the subsampling randomness is of order $O_{P|\mathcal{F}_n}(1/r + q_k^2/r^2)$. When $q_k = o(r^{3/4})$, the dominating term is $\ell_k^*(\hat{\boldsymbol{\beta}}_k) - \ell_k(\hat{\boldsymbol{\beta}}_k)$ and other terms are negligible regarding the randomness caused by subsampling. Therefore, we can focus on selecting an informative subsample that minimizes the conditional variance of $\ell_k^*(\hat{\boldsymbol{\beta}}_k) - \ell_k(\hat{\boldsymbol{\beta}}_k)$ given $\mathcal{F}_n$.

Thanks to Theorem 1, we know the weight assigned by the S-AIC weighting scheme in (12) to an underfitted model in $\mathcal{U}$ is asymptotically zero. Thus we can focus on minimizing the

asymptotic variance of $\ell_k^*(\hat{\boldsymbol{\beta}}_k)$ for $\mathcal{M}_k \in \mathcal{U}^c$ only, where $\mathcal{U}^c$ is the complement set of $\mathcal{U}$, i.e., the set of candidate models that includes all the predictors of the best model $\mathcal{M}_B$. Although the set $\mathcal{U}^c$ is unknown, the models in it can be embedded within the model that contains all the predictors of $\boldsymbol{x}$. We call this model the full model and denote it as $\mathcal{M}_{\text{full}}$. We recommend finding the subsampling strategy that minimizes the asymptotic variance of $\ell_{\text{full}}^*(\hat{\boldsymbol{\beta}}_{\text{full}})$ instead. When there are no redundant variables and the full model $\mathcal{M}_{\text{full}}$ is in the candidate pool, this is a natural choice according to Theorem 1. If $\mathcal{M}_{\text{full}}$ is not the best model, this is still a reasonable choice because the asymptotic variance of $\ell_{\text{full}}^*(\hat{\boldsymbol{\beta}}_{\text{full}})$ is an upper bound of the asymptotic variances of $\ell_k^*(\hat{\boldsymbol{\beta}}_k)$ for any $\mathcal{M}_k \in \mathcal{U}^c$. This is a type of **m**ini-max **a**symptotic uncertainty **s**ubsampling **s**trategy, and we call it MASS.

Theorem 2. *Assume that the maximum likelihood estimator under $\mathcal{M}_{\text{full}}$, say $\hat{\boldsymbol{\beta}}_{\text{full}}$, exists and Assumptions 1–2 also hold for the full model $\mathcal{M}_{\text{full}}$. The subsampling probabilities that achieve the minimum asymptotic variance of $\ell_{\text{full}}^*(\hat{\boldsymbol{\beta}}_{\text{full}})$ are*

$$\pi_i^{\text{MASS}} = \frac{|y_i \hat{\boldsymbol{\beta}}_{\text{full}}^{\text{T}} \boldsymbol{x}_i - \psi(\hat{\boldsymbol{\beta}}_{\text{full}}^{\text{T}} \boldsymbol{x}_i)|}{\sum_{l=1}^{n} |y_i \hat{\boldsymbol{\beta}}_{\text{full}}^{\text{T}} \boldsymbol{x}_i - \psi(\hat{\boldsymbol{\beta}}_{\text{full}}^{\text{T}} \boldsymbol{x}_i)|}, \tag{13}$$

for $i = 1, \ldots, n$.

Theorem 2 encourages us to select the data points with larger absolute values of the corresponding log-likelihood, i.e., $|y_i \hat{\boldsymbol{\beta}}_{\text{full}}^{\text{T}} \boldsymbol{x}_i - \psi(\hat{\boldsymbol{\beta}}_{\text{full}}^{\text{T}} \boldsymbol{x}_i)|$. Intuitively, data points with $|y_i \hat{\boldsymbol{\beta}}_{\text{full}}^{\text{T}} \boldsymbol{x}_i - \psi(\hat{\boldsymbol{\beta}}_{\text{full}}^{\text{T}} \boldsymbol{x}_i)|$ close to zero contribute less to the log-likelihood function, so it is reasonable to assign smaller sampling probabilities on them. There are some potential risks of sampling according to π_i^{MASS} directly. For example, relying on the large absolute values of the log-likelihood data points, the resultant estimator may be sensitive to outliers. In addition, if the data points with extremely small π_i^{MASS} are sampled, the subsample-based estimator will become unstable. To make the estimator more stable and robust, we adopt the technique of defensive importance sampling (Hesterberg, 1995; Owen and Associate, 2000). This approach is also known as shrinkage subsampling (Ma et al., 2015). To be specific, we recommend using the following subsampling probabilities

$$\pi_i^{\text{SMASS}} = (1 - \rho)\pi_i^{\text{MASS}} + \rho n^{-1}, \quad i = 1, \ldots, n, \tag{14}$$

where $\rho \in (0,1)$. Mixing the MASS probabilities with the uniform probability improves the stability of the subsample-based estimator. The empirical results suggest that the shrinkage subsampling method is not sensitive to the selection of ρ and works well when ρ is not very close to zero or one. In practice, it may not be feasible to obtain $\hat{\boldsymbol{\beta}}_{\text{full}}$ using the full data. We take a pilot subsample of size r_0 to explore the data and obtain a pilot estimator, say $\tilde{\boldsymbol{\beta}}_{\text{full},0}$, to be used for calculating the proposed sampling probabilities. We denote the resulting sampling probabilities by $\tilde{\boldsymbol{\pi}}^{\text{SMASS}}$. We then use $\tilde{\boldsymbol{\pi}}^{\text{SMASS}}$ to take a second subsample of size r according to the computational capacity.

With the specific $\tilde{\pi}_i^{\text{SMASS}}$, Assumption 4 is automatically satisfied under Assumptions 1–3, and Assumption 5 can be refined by a sufficient tail condition presented in Assumption 6.

Assumption 6. *For some $\kappa \in (0,\infty)$,*

$$\sup_{\|\boldsymbol{u}\|=1} \max_{1\le i\le n} \frac{|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}_i|^6}{\log^{\kappa}(n)} = O_P(1), \quad \sup_{\mathcal{M}_k} \max_{1\le i\le n} \frac{\psi(\hat{\boldsymbol{\beta}}_k^{\mathrm{T}} P_k \boldsymbol{x}_i)}{\log^{\kappa}(n)} = O_P(1).$$

4.2 Theoretical Properties

To measure the performance of the subsample-based S-AIC estimator $\tilde{\boldsymbol{\beta}}$ under the proposed subsampling procedure, we adopt the idea of Ando et al. (2017) and define the KL-divergence based loss (normalized by the sample size) as

$$\tilde{\mathcal{L}}(\boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^{n}\left\{ y_i\left(\theta_i - \sum_{k=1}^{m}\omega_k \tilde{\boldsymbol{\beta}}_k^{\mathrm{T}} P_k \boldsymbol{x}_i\right) - \left(\psi(\theta_i) - \psi\left(\sum_{k=1}^{m}\omega_k \tilde{\boldsymbol{\beta}}_k^{\mathrm{T}} P_k \boldsymbol{x}_i\right)\right)\right\}, \quad (15)$$

where θ_i is the true parameter that generate y_i through (1) and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)$ is a general weight. It is worth mentioning that $\tilde{\mathcal{L}}(\tilde{\boldsymbol{\omega}})$ with $\tilde{\boldsymbol{\omega}}$ calculated via (12) measures the generalization error of $\tilde{\boldsymbol{\beta}}$ from the subsample to the full data. This reflects how well $\tilde{\boldsymbol{\beta}}$ can be used to describe the full data set. The following theorem shows that the subsample S-AIC weight performs similarly to the full-data-based S-AIC weight in terms of the Kullback–Leibler loss.

Theorem 3. *Let $\zeta = \inf_{\boldsymbol{\omega}\in\mathcal{C}_m} \hat{\mathcal{L}}(\boldsymbol{\omega})$, where $\mathcal{C}_m = \{\boldsymbol{\omega} \in [0,1]^m : \sum_{k=1}^m \omega_k = 1\}$ and $\hat{\mathcal{L}}(\boldsymbol{\omega})$ has the same expression of (15) except that $\tilde{\boldsymbol{\beta}}_k$ is replaced by the full-data-based estimator $\hat{\boldsymbol{\beta}}_k$. Under Assumptions 1–3 and 6, if as $r \to \infty$, $n \to \infty$, $(\log(m) + \zeta^{-2} q_{(L)} \log(q)) \log^{2\kappa}(n)/r \to 0$ and $r/n \to 0$, then*

$$\frac{\tilde{\mathcal{L}}(\tilde{\boldsymbol{\omega}})}{\tilde{\mathcal{L}}(\hat{\boldsymbol{\omega}})} \to 1, \quad \textit{and} \quad \frac{\tilde{\mathcal{L}}(\tilde{\boldsymbol{\omega}})}{\hat{\mathcal{L}}(\hat{\boldsymbol{\omega}})} \to 1, \quad (16)$$

in probability, where $\tilde{\boldsymbol{\omega}} = (\tilde{\omega}_1, \ldots, \tilde{\omega}_m)$ *and* $\hat{\boldsymbol{\omega}} = (\hat{\omega}_1, \ldots, \hat{\omega}_m)$ *are the subsample and full sample S-AIC weights, respectively.*

Theorem 3 indicates that the subsample S-AIC weight is asymptotically as good as the full-data-based S-AIC weight in terms of the KL divergence loss. In the following, we show the consistency of $\tilde{\boldsymbol{\beta}}$ to the full-data-based S-AIC estimator $\hat{\boldsymbol{\beta}} = \sum_{k=1}^{m} \hat{\omega}_k P_k^{\mathrm{T}} \hat{\boldsymbol{\beta}}_k$.

Theorem 4. *Let* m_c *be the number of models in* $\mathcal{U}^c$*. Under Assumptions 1–3, and 6, if conditions* $m_c r/(n \log(q) \log^{\kappa}(n)) \to 0$ *and* $(\log(m) + q_{(L)} \log(q)) \log^{2\kappa}(n)/r \to 0$ *holds as* $n, r \to \infty$*, then the S-AIC estimator* $\tilde{\boldsymbol{\beta}}$ *is consistent to full-data-based S-AIC estimator* $\hat{\boldsymbol{\beta}}$ *in conditional probability given* $\mathcal{F}_n$*. More precisely, (i) when* $m_c = O(\log(q)\log^{\kappa}(n))$*, with probability approaching one, for any* $\epsilon > 0$*, there exists a finite* δ_ϵ *and* r_ϵ *such that for all* $r > r_\epsilon$*,*

$$\mathrm{pr}\left(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\| \geq \sqrt{m_c q_{(L)}/r}\, \delta_\epsilon \middle| \mathcal{F}_n \right) < \epsilon; \tag{17}$$

or (ii) when $m_c/(\log(q)\log^{\kappa}(n)) \to \infty$ *and* $m_c r/(n \log(q)\log^{\kappa}(n)) \to 0$*, with probability approaching one, for any* $\epsilon > 0$*, there exists a finite* δ_ϵ *and* r_ϵ *such that for all* $r > r_\epsilon$*,*

$$\mathrm{pr}\left(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\| \geq \sqrt{q_{(L)} \log(q)\log^{\kappa}(n)/r}\, \delta_\epsilon \middle| \mathcal{F}_n \right) < \epsilon. \tag{18}$$

Remark 2. In practice, prior information and subject knowledge are often helpful to identify plausible candidate models so that the size of the candidate model set is much smaller than 2^q. An exhaustive search may be directly implemented in this case. When such information is not available, an exhaustive search across $m = 2^q$ models is often computationally infeasible. To reduce the computational burden, forward selection usually serves as an alternative approach to an all-subset search. The forward selection procedure starts from the null model that includes the intercept term only, and then it sequentially adds one variable at a time to the model that yields the lowest value of the AIC. More precisely, in the first step, it adds the variable that yields the lowest value of AIC among models with only one variable. In the second step, it adds the variable that yields the lowest value of AIC when added to the previously selected model with one variable. This process stops when $q_{(L)}+1$ nested models are obtained. Here, the maximum model size $q_{(L)}$ may be determined by some prior knowledge or can be taken as $q_{(L)} = q$ when such knowledge is absent. After obtaining the $q_{(L)}+1$ candidate models, we calculate the corresponding S-AIC weights.

5 Numerical Studies

We conduct numerical experiments to evaluate the finite sample performance of the proposed method on two real datasets and two synthetic datasets. Further numerical results with more synthetic datasets are relegated to the Supplementary Material. Computations are performed in `R`.

5.1 Beijing Multi-site Air-quality Dataset

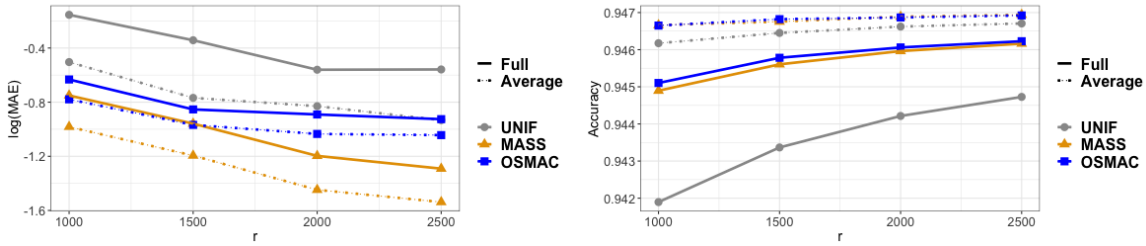
In the following, we experiment on a real dataset about Beijing's air quality. This dataset consists of hourly air pollutants records from twelve air-quality monitoring sites in Beijing from March 1st, 2013 to February 28th, 2017. There are 420,768 records in the data. The dataset is available in the UCI database at `https://archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data`, and more information about it can be found in Zhang et al. (2017). One research interest is predicting whether the air is currently polluted using the PM2.5 data from the past 23 hours. According to the ambient air quality standard in China, we call the air is polluted if the PM2.5 is greater than $75\mu g/m^3$. A logistic regression model with the PM2.5 values from the past 23 hours is used to predict the air quality. After removing the incomplete cases, a logistic regression is fitted.

Since the predictors are the PM2.5 values from the past 23 hours, we consider the candidate model set that consists of the 23 nested models, each with the PM2.5 values in the past j ($j = 1, ..., 23$) hours as predictors. More precisely, $\mathcal{M}_j$ is the model with the j predictors being the PM2.5 values in the past j hours.

We evaluate the performance of the AIC_{sub} in (10) for model averaging with the proposed MASS subsampling strategy. For comparison, we also implement the OSMAC subsampling for which $\pi_i \propto |y_i - \dot{\psi}(\tilde{\boldsymbol{\beta}}_{\text{full},0})|\|\boldsymbol{x}_i\|$ under the L-optimality, and the uniform subsampling (UNIF) for which $\pi_i = n^{-1}$. Here $\tilde{\boldsymbol{\beta}}_{\text{full},0}$ denotes the pilot-sample-based estimator for the full model. We use the L-optimality for OSMAC for the following two reasons. Firstly, the number of predictors is usually large in a model averaging problem. Thus we need to control the computational cost in calculating sampling probabilities within $O(nq)$ instead of $O(nq^2)$. Secondly, in order to achieve a consistent estimator of the full model's information matrix, we

need a much larger r_0, which implies a large $r_0/(r+r_0)$ when the sampling budgets is limited. As illustrated in Figure 3(b), a large $r_0/(r+r_0)$ may lead to an inefficient subsample-based estimator.

We measure the performance of a sampling strategy $\boldsymbol{\pi}$ via the empirical mean absolute error (MAE) which is the average l_1 distance between a subsample-based estimator $\tilde{\boldsymbol{\beta}}$ and a full-data-based estimator $\hat{\boldsymbol{\beta}}$. We repeated the simulation procedure for 500 times to calculate the empirical MAE. To further demonstrate the advantage of the model averaging approach over the full-model approach, the results of the full-model approach with MASS, OSMAC, and UNIF subsampling probabilities are also presented as benchmarks. We fix r_0 and ρ at 500 and 0.2, respectively. The empirical MAE, together with the accuracy on classifying the full data are presented in Figure 1.

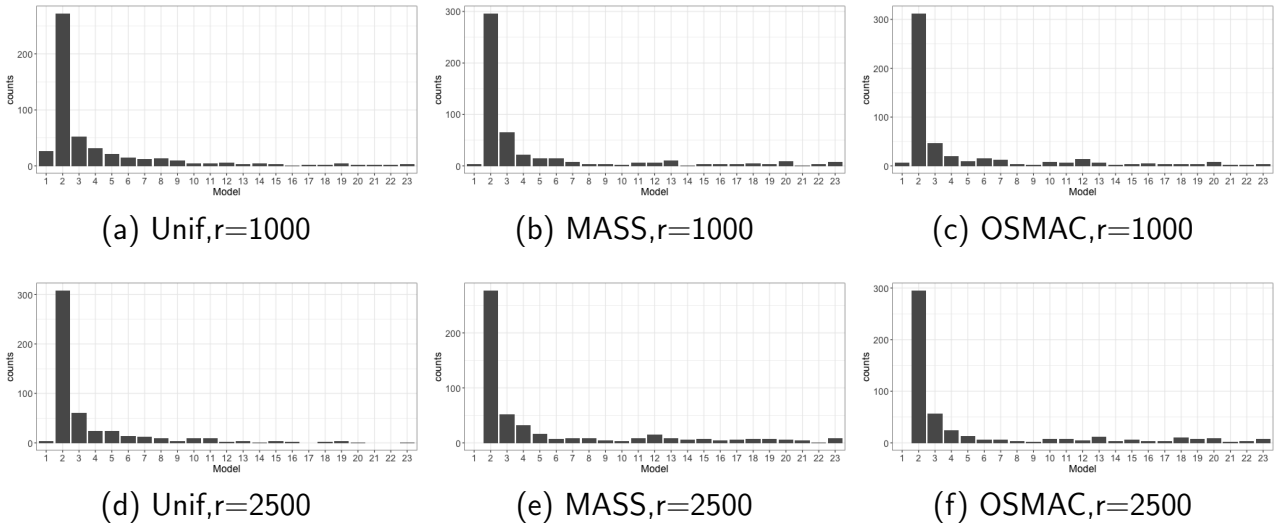


(a) log(MAE) (b) Classification accuracy

Figure 1: A graph showing the median of log MAE and prediction accuracy with different subsample size r for the Beijing multi-site air-quality dataset based on the UNIF (grey lines with circle), the MASS (yellow lines with triangle), and the OSMAC (blue lines with square) subsampling methods. Here the solid lines stand for the full-model approach, and the dotted lines stand for the averaging approach. The r_0 and ρ are fixed at 500 and 0.2, respectively.

From Figure 1, one can see that the model averaging method always results in a smaller MAE and a higher prediction accuracy compared with the full-model approach when the same sampling probabilities are adopted. Judging from the selection results reported in Figure 2, we believe this phenomenon comes from the fact that there are redundant variables in $\mathcal{M}_{\text{full}}$. The MAE for all subsampling methods increases as r increases, which confirms the theoretical result on the consistency of the subsampling methods.

Figure 2 reports the frequency that model $\mathcal{M}_j$ receives the highest weight. All methods tend to select $\mathcal{M}_2$ as the best model, which implies that the air quality can be well predicted



(a) Unif,r=1000 (b) MASS,r=1000 (c) OSMAC,r=1000

(d) Unif,r=2500 (e) MASS,r=2500 (f) OSMAC,r=2500

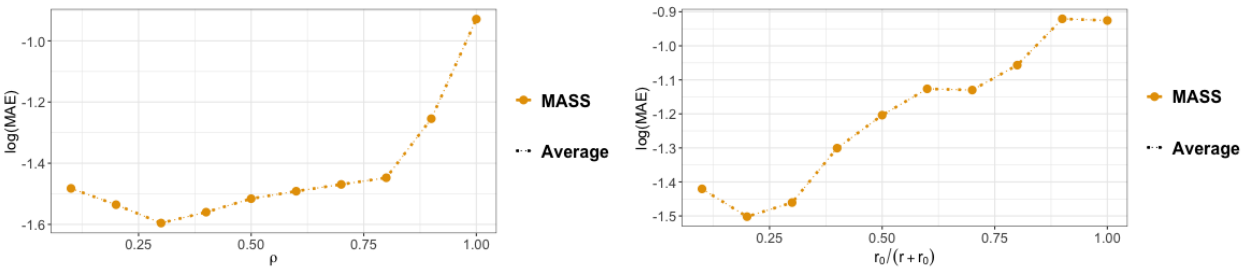
Figure 2: The times that model $\mathcal{M}_j$ enjoys the highest weight with $r = 1000$ (upper panel) and $r = 2500$ (lower panel). Here we fixed $r_0 = 500, \rho = 0.2$.

by the PM2.5 values in the last two hours. Compared with the OSMAC and the MASS, the UNIF has a higher chance to select $\mathcal{M}_1$ as the best model when $r = 1,000$. Comparing the results in (a)-(c) with those in (d)-(f), we see that $\mathcal{M}_1$ is an underfitted model as discussed in Theorem 1. This can be understood as using the PM2.5 value in the past one hour only is not sufficient enough to explain the current air quality. OSMAC and MASS are more likely to rule out the underfitted model compared with the uniform subsampling. This is a reason why the two methods outperform the uniform subsampling.

In the following, we evaluate the impact of the tuning parameter ρ in (14) and the pilot sample size r_0 on the performance of the MASS. We present the results with $r_0 = 500$ and $r = 2500$ for the sensitivity analysis on ρ and fix $r_0 + r = 3000$ for the sensitivity analysis on r_0. The log(MAE) against different ρ and $r_0/(r_0 + r)$ are reported in Figure 3 (a) and (b), respectively. It is seen that the proposed method performs well and are not very sensitive to ρ when it is between 0.2 and 0.5; the relative variation is less than 10%. With a fixed $\rho = 0.2$, one can see that MASS performs well when $r_0/(r_0 + r)$ is between 0.15 and 0.3.

5.2 The SUSY dataset

We experiment on a real dataset about supersymmetric particles available on `https://archive.ics.uci.edu/dataset/279/susy`. The task is to distinguish between a signal pro-



(a) ρ (b) $r_0/(r + r_0)$

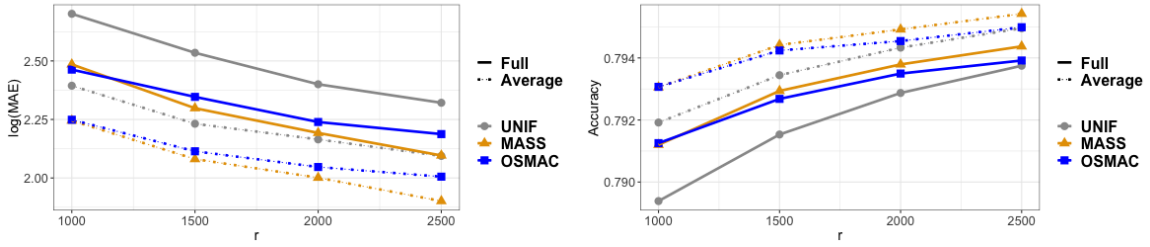
Figure 3: Median log MAE against different ρ values with $r_0 = 500, r = 2000$ (left panel) and median log MAE against different r_0 values with $r_0 + r = 3000, \rho = 0.2$ (right panel).

cess which produces supersymmetric particles and a background process which does not. There are eight features that are kinematic properties measured by the particle detectors in the accelerator, which are known as the low-level features. There are another ten features that are derived by physicists based on the low-level features to help discriminate between the two classes. More information about the data is available in Whiteson (2014). Here we consider a class of logistic regressions with 46 possible covariates (features), consisting of the original 18 features and 28 interactions of the eight low-level features.

Due to limited computational resources, it is infeasible for us to consider all the 2^{46} possible models. Thus, the forward selection method as discussed in Remark 2 is adopted. Again, we report the results for model averaging with the proposed MASS subsampling strategy together with OSMAC and uniform subsampling strategies. The r_0 and ρ are fixed at 500 and 0.2, respectively. Results for the full-model approaches are also reported for comparison.

Figure 4 shows that the model averaging method always leads to a smaller MAE compared with the full-model approach when the same sampling probabilities are adopted. As expected the MASS and OSMAC have better performances compared with uniform subsampling.

The S-AIC weights for models with less than 15 predictors are less than 10^{-38} when the forward regression is implemented on the full data. The extremely small weights imply that models with less than 15 predictors are likely to be underfitted models. We record the number of predictors in the best model selected by the smallest AIC_{sub}, say d_B, to reflect the model selection performance. The number of times that $d_B < 15$ for the UNIF, the OSMAC, and



(a) log(MAE) (b) Prediction accuracy

Figure 4: A graph showing the median of log MAE and prediction accuracy with different subsample size r for the SUSY dataset based on UNIF (grey lines with circle), MASS (yellow lines with triangle) and OSMAC (blue lines with square) subsampling methods. Here the solid lines stand for the full-model approach, and the dotted lines stand for the averaging approach.

the MASS, are 88, 73, and 68, respectively, out of the 500 replications when $r = 1000$. This implies that the MASS is more effective than the OSMAC in excluding underfitted models, and they are both better than the UNIF.

We close this section by evaluating the computational efficiency. We implemented all methods using the `R` programming language and recorded the computing times of the three subsampling strategies using the `Sys.time()` function. Computations were carried out on an iMac (Retina 5K, 2020) with a 10-Core Intel Core i9 processor. We also record the computing time on the full dataset as a benchmark. Results are presented in Table 1.

Table 1: Computational time (in seconds) of the S-AIC estimator on the Beijing multi-site air-quality and SUSY datasets.

	r	1000	1500	2000	2500	Full data
Air-quality dataset	UNIF	0.0817	0.1051	0.1277	0.1504	18.5777
	MASS	0.1037	0.1224	0.1432	0.2081	
	OSMAC	0.1139	0.1361	0.1609	0.1765	
SUSY dataset	UNIF	6.5255	8.3666	10.6676	12.2237	24469.62
	MASS	6.9350	9.2282	10.8142	12.5163	
	OSMAC	7.3816	8.9530	10.6676	12.2237	

It is seen that all subsampling methods are significantly faster than the full-data calculation for the S-AIC estimator. The UNIF is faster than the MASS and the OSMAC, but the difference is not significant. The main reason is that the computational time is mainly spent on calculating the AIC values of the candidate models. The time complexity for calculating $\tilde{\boldsymbol{\beta}}_k$

under $\mathcal{M}_k$ is $O(rq_k^2)$. For nested models as in the air-quality dataset, the time complexity of calculating the model averaging estimator based on a subsample is $O(r\sum_{j=1}^{q-1}(j+1)^2) = O(rq^3)$. When forward selection is adopted, $q+1-j$ models with $j+1$ covariates are calculated in the jth iteration, leading to a time complexity of $O(r\sum_{j=1}^{m}(q+1-j)(j+1)^2) = O(rq^4)$. The MASS and OSMAC only take $O(nq)$ time to calculate the sampling probabilities. Therefore, the additional time in calculating the subsampling probabilities may not be a leading order term in the computational complexity. Consequently, our method has comparable computational performance with the uniform subsampling method.

5.3 Simulation Results

It is known that model averaging estimators are impacted by candidate model specification. In the following, we further validate the proposed method on the synthetic dataset with different candidate models. The response is generated by a logistic regression with $q = 30$ potential covariates. The full data size is set to be $n = 500,000$. The nonzero components of $\boldsymbol{\beta}$ have decreasing sizes as suggested in Zheng et al. (2019). Specifically, $\beta_j = 2/j$ for $j = 1, \ldots, 6$, and $\beta_j = 0$ for the rest.

The following two distributions are used to generate covariates $\boldsymbol{x}_i$'s.

Case 1 Multivariate normal distribution $N(\mathbf{0}, \Sigma_1)$ with the (i, j)th entry of Σ_1 being $0.5^{|i-j|}$.

Case 2 The first 10 dimensions of the covariate come from $N(\mathbf{0}, \Sigma_1)$, and the rest dimensions consist of quadratic and cubic transformation of the first 10 dimensions.

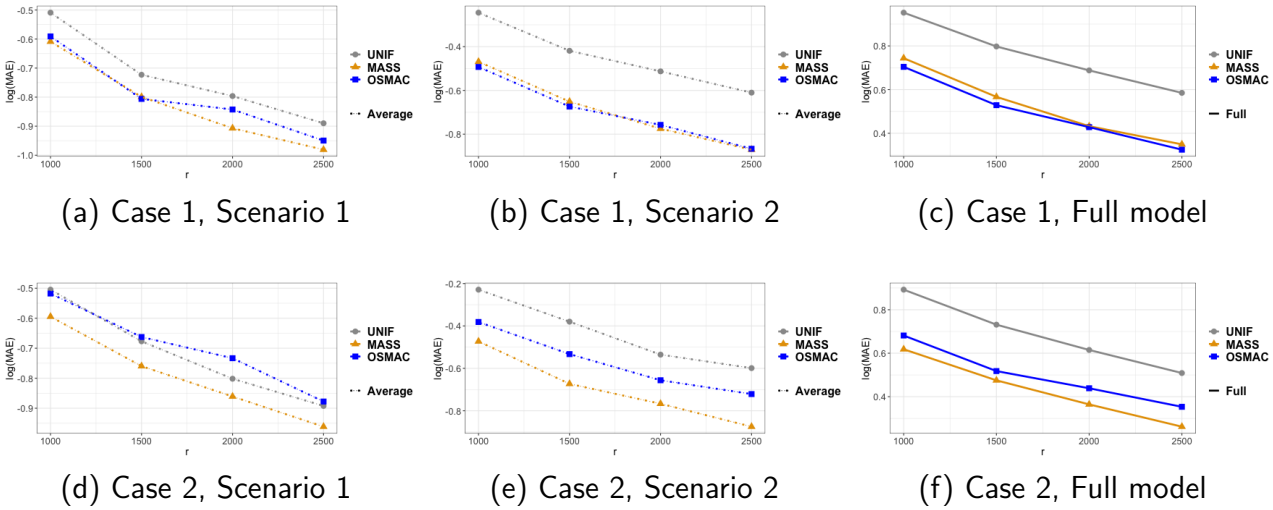
We consider the following two scenarios for the candidate model specification.

Scenario 1 The $\mathcal{M}_j$ contains the first j predictors. In this case, there are 29 models in the candidate set.

Scenario 2 The forward selection procedure is used to explore the candidate models with prior knowledge on the largest number of predictors. Here we assume the number to be eight where the largest model contains 30% more predictors than the best true model.

We fix $r_0 = 500$ and $\rho = 0.2$ and set r to 1000, 1500, 2000, and 2500. The uniform subsampling is implemented with a subsample size $r + r_0$ for fair comparisons. The simulation results are given in Figure 5. We opt to show the full-model approach and model averaging approach in different panels since the scaling of log MAE in the two methods is different.

We see that the MAE for all subsampling methods decreases as r increases, which confirms the theoretical consistency of the subsampling methods. As expected, the MASS always leads



(a) Case 1, Scenario 1 (b) Case 1, Scenario 2 (c) Case 1, Full model

(d) Case 2, Scenario 1 (e) Case 2, Scenario 2 (f) Case 2, Full model

Figure 5: A graph showing the median of the log MAE with different subsample size r for different distributions of covariates and different candidate models. Here we opt to show the full-model approach and model averaging approach in different panels since the scaling of log MAE in the two methods is different. The full-model approach is the same under Scenarios 1 and 2.

to a smaller MAE compared with the UNIF. Although the OSMAC outperforms the UNIF with the full model, Figure 5(d) shows that it does not necessarily outperform the UNIF in the model averaging framework due to model uncertainty. Similar phenomenon is observed in Figures 5(a) and (c) that OSAMC outperforms MASS with the full-model approach while MASS has a better performance under the model averaging framework.

6 Conclusion

In this paper, we have investigated the subsample-based S-AIC estimator and developed a MASS subsampling strategy to improve the subsample-based model averaging method. We have derived the asymptotic properties of the estimators under candidate models with diverging dimensions and derived the appropriate expression of the subsample AIC. We have also carried out numerical experiments on both simulated and real datasets to evaluate its practical performance. Both theoretical results and numerical results demonstrate the great potential of the proposed method in extracting useful information from massive datasets. Our investigations have focused on the subsample-based AIC model averaging, and the technical proofs are already complicated. We only considered averaging candidate models with different

covariates in the linear predictor as studied in Ando et al. (2017). More complicated scenarios, such as that when candidate models have different link functions and/or different distribution assumptions are also important and need to be investigated in future research. We hope this work will attract more attention to the promising technique of model averaging in subsampling big data.

Supplementary Material

Narrative Supplement The pdf file contains an algorithm, distributional results on the subsample-based S-AIC estimator, all the technical proofs, and additional simulation results.

Code Supplement The zip file contains the R codes that were used for the numerical results of the paper.

Acknowledgments

The authors would like to thank the Editor, an Associate Editor, and two reviewers for their insightful comments which helped substantially improve the manuscript. Correspondence should be addressed to HaiYing Wang or Mingyao Ai.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Ai's work was supported by NSFC grants 12071014 and 12131001, National Key R&D Program of China 2020YFE0204200, and LMEQF. Wang's work was supported by NSF grant 2105571 and UConn CLAS Research Funding in Academic Themes. Yu's work was supported by the Beijing Municipal Natural Science Foundation No.1232019, and the Beijing Institute of Technology research fund program for young scholars.

References

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

Akaike, H. (1998), "Information theory and an extension of the maximum likelihood principle," in *Selected papers of hirotugu akaike*, Springer, pp. 199–213.

Ando, T., Li, K.-C., et al. (2017), "A weight-relaxed model averaging approach for high-dimensional generalized linear models," *The Annals of Statistics*, 45, 2654–2679.

Baldi, P., Sadowski, P., and Whiteson, D. (2014), "Searching for exotic particles in high-energy physics with deep learning," *Nature communications*, 5, 1–9.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997), "Model selection: an integral part of inference," *Biometrics*, 53, 603–618.

Chen, R. Y., Gittens, A., and Tropp, J. A. (2012), "The masked sample covariance estimator: an analysis using matrix concentration inequalities," *Information and Inference: A Journal of the IMA*, 1, 2–20.

Claeskens, G., Croux, C., and Kerckhoven, J. V. (2006), "Variable selection for logistic regression using a prediction-focused information criterion," *Biometrics*, 62, 972–979.

Claeskens, G., Hjort, N. L., et al. (2008), *Model selection and model averaging*, Cambridge University Press.

Fithian, W. and Hastie, T. (2014), "Local case-control sampling: Efficient subsampling in imbalanced data sets," *The Annals of statistics*, 42, 1693–1724.

Han, L., Tan, K. M., Yang, T., Zhang, T., et al. (2020), "Local uncertainty sampling for large-scale multiclass logistic regression," *The Annals of Statistics*, 48, 1770–1788.

Hansen, B. E. (2007), "Least squares model averaging," *Econometrica*, 75, 1175–1189.

— (2014), "Model averaging, asymptotic risk, and regressor groups," *Quantitative Economics*, 5, 495–530.

Hansen, B. E. and Racine, J. S. (2012), "Jackknife model averaging," *Journal of Econometrics*, 167, 38–46.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning*, Springer.

He, L., Li, W., Song, D., and Yang, M.-S. (2024), "A systematic view of information-based optimal subdata selection: Algorithm development, performance evaluation, and application in financial data," *Statistica Sinica*, doi:10.5705/ss.202022.0019.

Hesterberg, T. (1995), "Weighted average importance sampling and defensive mixture distributions," *Technometrics*, 37, 185–194.

Hjort, N. L. and Claeskens, G. (2003), "Frequentist model average estimators," *Journal of the American Statistical Association*, 98, 879–899.

Joseph, V. R. and Mak, S. (2021), "Supervised compression of big data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14, 217–229.

Joseph, V. R. and Vakayil, A. (2022), "Split: An optimal method for data splitting," *Technometrics*, 64, 166–176.

Konishi, S. and Kitagawa, G. (2007), *Information Criteria and Statistical Modeling*, Springer.

Liang, H., Zou, G., Wan, A. T., and Zhang, X. (2011), "Optimal weight choice for frequentist model average estimators," *Journal of the American Statistical Association*, 106, 1053–1066.

Lv, J. and Liu, J. S. (2014), "Model selection principles in misspecified models," *Journal of the Royal Statistical Society: Series B*, 76, 141–167.

Ma, P., Chen, Y., , Zhang, X., Xing, X., Ma, J., and W.Mahoney, M. (2022), "Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms," *Journal of Machine Learning Research*, 23, 1–45.

Ma, P., Mahoney, M. W., and Yu, B. (2015), "A statistical perspective on algorithmic leveraging," *Journal of Machine Learning Research*, 16, 861–919.

Mak, S. and Joseph, V. R. (2018), "Support points," *The Annals of Statistics*, 46, 2562–2592.

Meng, C., Xie, R., Mandal, A., Zhang, X., Zhong, W., and Ma, P. (2021), "Lowcon: A design-based subsampling approach in a misspecified linear model," *Journal of Computational and Graphical Statistics*, 30, 694–708.

Owen, A. and Associate, Y. Z. (2000), "Safe and effective importance sampling," *Journal of the American Statistical Association*, 95, 135–143.

Peng, J. and Yang, Y. (2022), "On improvability of model selection by model averaging," *Journal of Econometrics*, 229, 246–262.

Shi, C. and Tang, B. (2021), "Model-robust subdata selection for big data," *Journal of Statistical Theory and Practice*, 15, 1–17.

Shibata, R. (1997), "Bootstrap estimate of kullback-leibler information for model selection," *Statistica Sinica*, 7, 375 – 394.

Sin, C.-Y. and White, H. (1996), "Information criteria for selecting possibly misspecified parametric models," *Journal of Econometrics*, 71, 207–225.

Wan, A. T., Zhang, X., and Zou, G. (2010), "Least squares model averaging by mallows criterion," *Journal of Econometrics*, 156, 277–283.

Wang, H. (2019), "More efficient estimation for logistic regression with optimal subsamples." *Journal of Machine Learning Research*, 20, 1–59.

Wang, H. and Ma, Y. (2021), "Optimal subsampling for quantile regression in big data," *Biometrika*, 108, 99–112.

Wang, H., Zhu, R., and Ma, P. (2018), "Optimal subsampling for large sample logistic regression," *Journal of the American Statistical Association*, 113, 829–844.

Wang, H. Y., Yang, M., and Stufken, J. (2019), "Information-based optimal subdata selection for big data linear regression," *Journal of the American Statistical Association*, 114, 393–405.

White, H. (1982), "Maximum likelihood estimation of misspecified models," *Econometrica*, 50, 1–25.

Whiteson, D. (2014), "SUSY," UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C54606.

Ye, Z., Yu, J., and Ai, M. (2024), "Optimal subsampling for multinomial logistic models with big data," *Statistica Sinica*, doi:10.5705/ss.202022.0277.

Yu, J., Ai, M., and Ye, Z. (2024), "A review on design inspired subsampling for big data," *Statistical Papers*, 65, 467–510.

Yu, J., Wang, H., Ai, M., and Zhang, H. (2022), "Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data," *Journal of the American Statistical Association*, 117, 265–276.

Yuan, Z. and Yang, Y. (2005), "Combining linear regression models: When and how?" *Journal of the American Statistical Association*, 100, 1202–1214.

Zhang, M., Zhou, Y., Zhou, Z., and Zhang, A. (2023), "Model-free subsampling method based on uniform designs," *IEEE Transactions on Knowledge and Data Engineering*, 1–13.

Zhang, S., Guo, B., Dong, A., He, J., Xu, Z., and Chen, S. X. (2017), "Cautionary tales on air-quality improvement in Beijing," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473, 20170457.

Zhang, X. (2015), "Consistency of model averaging estimators," *Economics Letters*, 130, 120–123.

Zhang, X., Yu, D., Zou, G., and Liang, H. (2016), "Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models," *Journal of the American Statistical Association*, 111, 1775–1790.

Zhao, Y., Amemiya, Y., and Hung, Y. (2018), "Efficient Gaussian process modeling using experimental design-based subagging," *Statistica Sinica*, 28, 1459–1479.

Zheng, C., Ferrari, D., and Yang, Y. (2019), "Model selection confidence sets by likelihood ratio testing," *Statistica Sinica*, 29, 827–851.

Zhou, Z., Yang, Z., Zhang, A., and Zhou, Y. (2023), "Efficient model-free subsampling method for massive data," *Technometrics*, doi:10.1080/00401706.2023.2271091.