Statistica Sinica Preprint No: SS-2025-0057	
Title	Robust Jackknife Model Averaging
Manuscript ID	SS-2025-0057
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0057

Complete List of Authors	Kang You, Miaomiao Wang and Guohua Zou
Corresponding Authors	Guohua Zou
E-mails	ghzou@amss.ac.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

Statistica Sinica

Robust jackknife model averaging

Kang You[a,b], Miaomiao Wang[c] and Guohua Zou[a,¶]

[a] *Capital Normal University*

[b] *University of Kent*

[c] *Beijing University of Chinese Medicine*

Abstract: In the age of big data, model averaging has been proved to be a powerful tool for data analysis, which helps to mitigate bias and reduce overfitting that can arise from relying on a single model. However, outliers in large-scale datasets like image recognition and fraud detection can severely degrade traditional model averaging built on least squares or maximum likelihood. To address this challenge, we propose a robust jackknife model averaging (RJMA) approach, where the weights are selected by minimizing a leave-one-out cross-validation criterion. This framework is adaptable to situations where the dimensions of candidate models increase with the sample size. We establish the asymptotic optimality of the RJMA estimator, demonstrating its ability to minimize out-of-sample final prediction errors. We also present the consistency of the proposed weight estimator to the theoretically optimal weight vector. Furthermore, in the scenario where one or more correct models are present in the candidate model set, we show that RJMA assigns all weights to the correct models, leading to a consistent model averaging estimator. Additionally, we derive the influence function of the RJMA estimator and introduce the empirical prediction influence function to quantitatively evaluate its robustness. To illustrate the efficacy of the proposed methodology, we conduct numerical studies including Monte Carlo simulations

¶Corresponding author: Guohua Zou. Email: *ghzou@amss.ac.cn*.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

and a real data analysis, which confirm the practical applicability and robustness of the RJMA approach.

Key words and phrases: Cross-validation, generalised M estimators, influence function, model averaging, robustness.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

1. Introduction

With the rapid advancement of technologies such as the internet, biology, medicine, and social networks, vast amounts of data are continuously being generated, driving the evolution of machine learning. Model averaging, which can deal with the uncertainty from model selection process and combine the information from various candidate models by certain weights, has proven to be a powerful forecasting tool in machine learning (see, for example, Schomaker and Heumann (2020); Le and Clarke (2022); Hu and Zhang (2023)). An important development in this field is Bayesian model averaging that assigns posterior model probabilities to candidate models (Hoeting et al., 1999). In this paper, we focus on the frequentist model averaging (FMA) methods (Buckland et al., 1997; Yang, 2001; Hjort and Claeskens, 2003; Hansen, 2007). From the frequentist perspective, a major challenge in model averaging lies in selecting the optimal weights for each candidate model. Numerous criteria have been proposed to tackle this problem. Under the framework of parametric model, examples include Mallows model averaging (MMA) (Hansen, 2007; Wan et al., 2010), jackknife model averaging (JMA) (Hansen and Racine, 2012; Lu and Su, 2015), optimal mean squared error averaging (Liang et al., 2011; Wan et al., 2014), Kullback-Leibler model averaging (Zhang et al., 2016), and model averaging for high-dimensional data (Ando and Li, 2014; Wang et al., 2023). Many FMA approaches are also suggested

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

for semiparametric and nonparametric models. For instance, Li et al. (2015) proposed a semiparametric "model averaging marginal regression" (MAMAR) which can approximate a multivariate regression function by an affine combination of one-dimensional marginal regression functions. Further, the MAMAR method was extended to ultra-high dimensional time series by Chen et al. (2018). Gao et al. (2016) developed a leave-subject-out cross-validation model averaging procedure for longitudinal data and time series. Kitagawa and Chris (2016) suggested a data-driven model averaging method for semiparametric estimation of treatment effects. Zhu et al. (2019) developed a Mallows-type model averaging estimator for the varying-coefficient partially linear model. Li et al. (2022) proposed a semiparametric model averaging prediction approach for multi-category outcomes, coupled with the AdaBoost algorithm. Fang et al. (2022) suggested a semiparametric model averaging prediction method for a dichotomous response. However, most of the aforementioned model averaging methods are based on least squares or maximum likelihood approaches, both of which can be heavily influenced by even a small proportion of erroneous observations in the sample.

In the era of big data, the scale and diversity of data often make outliers inevitable across various fields, such as finance and image recognition. To reduce the influence of outliers on model selection, many valid robust statistical methods have been proposed including Ronchetti and Staudte (1994), Sommer and

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

Staudte (1995), Burman and Nolan (1995) and Ronchetti (1997), among others. These proposals are built on robust versions of classical selection criteria such as robust AIC and C_p. On the other hand, there are some robust model selection methods constructed by resampling technique like Ronchetti et al. (1997) and Wisnowski et al. (2003). Under the high-dimensional regression setting, several authors proposed penalized robust estimation methods which can perform parameter estimation and variable selection simultaneously (Li et al., 2011; Fan et al., 2014; Lozano et al., 2016). Furthermore, He et al. (2000) and others developed the generalized M (GM) estimators, which exhibit resistance to outliers both in the design and response variables. There have also been some efforts devoted to studying robust model averaging in the presence of outliers. For quantile regerssion, Lu and Su (2015) developed a jackknife model averaging criterion which selects the weights by minimizing a leave-one-out CV criterion, and Wang et al. (2023) extended this work to the case of high-dimension. For mean regression, Du et al. (2018) and Guo and Li (2021) proposed robust model averaging methods by adjusting the existing information criterion scores used in model selection. However, the weight selection criteria proposed in these two papers rely on intuitive considerations, and they do not demonstrate any optimality properties. Wang et al. (2024) suggested a robust model averaging approach by Mallows-type criterion which is based on the GM-type loss function. In their

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

paper, the authors demonstrated the asymptotic optimality and weight consistency of the proposed method.

In this context, we aim at developing a robust jackknife model averaging (RJMA) approach with optimality, which is robust to the outliers occurring in the response and/or covariates. To reduce the influence of outliers, we estimate each candidate model's parameters by minimizing the GM-type loss function. Furthermore, we opt for the weight vector by minimizing the GM-type leave-one-out CV criterion, which is a convex function of the weight vector. It is noteworthy that in Wang et al. (2024), the proposed robust weight selection criterion is non-convex. Additionally, unlike Wang et al. (2024), the proposed RJMA method is suitable for cases where the error term exhibits heteroscedasticity. Our method extends Hansen and Racine's (2012) JMA from the specific case of the quadratic loss function to a more general framework of convex loss function. This extension is nontrivial, as closed-form solutions are not available for the robust loss functions such as the absolute loss or Huber's function. As a result, we cannot apply the proof techniques used in Hansen and Racine (2012) to establish the asymptotic properties of the RJMA estimators. To address this challenging issue, we generalize Knight's (1998) identity, and demonstrate that the resultant RJMA estimator achieves asymptotic optimality in terms of minimizing the out-of-sample final prediction error (FPE). We also establish the consistency of the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

RJMA-based weight estimator to the theoretically optimal weight vector. It is noteworthy that our proof method diverges significantly from those in current literature on model averaging, as our approach relies on the robust loss function. Specifically, we utilize the Lagrange multiplier method to offer a more concise proof in establishing weight-consistency. On the basis of consistency of weight estimator, we extend our analysis to demonstrate that the RJMA estimator is asymptotically optimal in the sense of obtaining the lowest population risk function. Since the population risk function involves the expectation of the RJMA-based weight estimator $\widehat{\boldsymbol{w}}$, this type of asymptotic optimality differs from that based on the loss function or the out-of-sample FPE, and imparts a more meaningful theoretical property to RJMA estimator. The similar results are established only in a few papers like Liao et al. (2021). In this paper, we also consider an interesting situation where one or more candidate models are correct. In this case, we can prove that the RJMA-based weights will concentrate on all the correct models, and the resultant model averaging estimator is consistent.

Since the influence function of an estimator was defined by Hampel (1968), it has become an important measurement of robustness and been well studied. For a model averaging estimator, how to derive its influence function is a more complex problem, because model averaging involves both parameter estimation and weight choice. In this paper, we provide a solution for the RJMA estimator,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

and in order to characterize the quantitative robustness of the proposed model averaging estimator, we define the empirical prediction influence function (EPIF) which is totally dependent on the sample.

The rest of the paper is organized as follows. Section 2 introduces the model framework and develops a robust jackknife model averaging estimator. In Section 3, we present some theoretical properties on the RJMA estimator and the RJMA-based weight estimator under some regularity conditions. Sections 4 and 5 report the results from simulation studies and a real data example, respectively. Section 6 concludes. The robustness property of the RJMA estimator, the mathematical proofs of theorems and the additional simulation studies are given in the Supplementary Material.

2. Model framework and robust jackknife model averaging

Suppose that $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$ is an independent sample from the following data generating process

$$y_i = \mu_i + \varepsilon_i = \mu(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, ..., n,$$

where $\mu(\boldsymbol{x}_i) = \sum_{j=1}^{p} x_{ij}\theta_j$ with $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ being the covariates and θ_j $(j = 1, ..., p)$ being the corresponding coefficients, and $\varepsilon_1, ..., \varepsilon_n$ are the independent error terms. As in Hansen and Racine (2012), the error term ε_i is allowed

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

to be dependent on $\boldsymbol{x}_i$ for $i = 1, ..., n$. The purpose of this paper is to robustly estimate $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$ by model averaging method.

We consider M candidate models and the m^{th} one has the form of

$$y_i = \boldsymbol{x}_{i(m)}^T \Theta_{(m)} + \varepsilon_{i(m)} = \sum_{j=1}^{k_m} \theta_{j(m)} x_{ij(m)} + \varepsilon_{i(m)}, \quad i = 1, ..., n, \tag{2.1}$$

where $\Theta_{(m)} = (\theta_{1(m)}, \ldots, \theta_{k_m(m)})^T$ and $\boldsymbol{x}_{i(m)} = (x_{i1(m)}, \ldots, x_{ik_m(m)})^T$ with $x_{ij(m)}$ being a variable in $\boldsymbol{x}_i$ that appears as a regressor in the m^{th} candidate model, $\theta_{j(m)}$ being the corresponding coefficient, $j = 1, \ldots, k_m$, and k_m being the number of covariates. In this paper, we permit M and k_m ($m = 1, ..., M$) to diverge with the sample size n. To reduce the influence of the potential outliers, under the m^{th} candidate model, we suggest a GM-estimator $\widehat{\Theta}_{(m)}$ obtained by minimizing the objective function

$$Q_n(\Theta_{(m)}) = \sum_{i=1}^{n} h(\boldsymbol{x}_{i(m)}) \rho \left(y_i - \boldsymbol{x}_{i(m)}^T \Theta_{(m)} \right), \tag{2.2}$$

where ρ is a robust convex function which downweights the outliers in the response, and $h(\cdot)$ is a bounded function which protects against the leverage points by assigning them small weights (see Section 4 for a specific example of the weight function). Note that for the m^{th} model, the weight function should be defined as $h_m : \boldsymbol{x}_m \mapsto \mathbb{R}$. For notational simplicity, we use the same symbol h in this paper. Suppose that ρ has a derivative ψ, then $\widehat{\Theta}_{(m)}$ is a solution of the following equation: $\sum_{i=1}^{n} h(\boldsymbol{x}_{i(m)}) \psi \left(y_i - \boldsymbol{x}_{i(m)}^T \Theta_{(m)} \right) \boldsymbol{x}_{i(m)} = 0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

Let the weight vector $\boldsymbol{w} = (w_1, ..., w_M)^T$ belong to a unit simplex of $\mathbb{R}^M$: $\mathcal{W} = \{\boldsymbol{w} \in [0,1]^M : \sum_{m=1}^M w_m = 1\}$. Then the model averaging estimator of μ_i can be expressed as $\widehat{\mu}_i(\boldsymbol{w}) = \sum_{m=1}^M w_m \boldsymbol{x}_{i(m)}^T \widehat{\Theta}_{(m)}, \quad i = 1, ..., n.$

To choose the weights in the above estimator, we develop a jackknife criterion. For $m = 1, ..., M$ and $i = 1, ..., n$, denote $\widehat{\Theta}_{(m)}^{[-i]}$ as the robust jackknife estimator of $\Theta_{(m)}$ in model m with the i^{th} observation deleted. Thus, the robust jackknife weight selection criterion can be defined as

$$\mathrm{RCV}_n(\boldsymbol{w}) = \sum_{i=1}^n h(\breve{\boldsymbol{x}}_i)\rho\left(y_i - \sum_{m=1}^M w_m \boldsymbol{x}_{i(m)}^T \widehat{\Theta}_{(m)}^{[-i]}\right), \tag{2.3}$$

where the elements of $\breve{\boldsymbol{x}}_i$ consist of all the candidate models' covariates. The weight function $h(\cdot)$ in (2.3) is dependent on $\breve{\boldsymbol{x}}_i$ because the model averaging estimator is based on the covariates of all candidate models.

The optimal robust jackknife weight estimator $\widehat{\boldsymbol{w}} = (\widehat{w}_1, ..., \widehat{w}_M)^T$ is obtained by minimizing $\mathrm{RCV}_n(\boldsymbol{w})$ over the weight set $\mathcal{W}$, that is $\widehat{\boldsymbol{w}} = \arg\inf_{\boldsymbol{w}\in\mathcal{W}} \mathrm{RCV}_n(\boldsymbol{w})$. The resultant RJMA estimator of $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$ thus takes the form of $\widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{w}}) = (\widehat{\mu}_1(\widehat{\boldsymbol{w}}), ..., \widehat{\mu}_n(\widehat{\boldsymbol{w}}))^T$.

Remark 1. When the unit vector $\boldsymbol{\tau}_m$ is considered, where only the m^{th} element is one and all others are zero, $\mathrm{RCV}_n(\boldsymbol{\tau}_m)$ corresponds to the standard robust jackknife criterion used for selecting regression models. The model selected by minimizing this criterion, denoted as $\hat{\boldsymbol{\tau}}_m$, represents the standard robust jackknife choice (Ronchetti et al., 1997).

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

Remark 2. By setting $h(\cdot) = 1$ and $\rho(t) = t^2$, the proposed criterion $\mathrm{RCV}_n(\boldsymbol{w})$ is clearly equivalent to the jackknife model averaging criterion introduced by Hansen and Racine (2012).

Remark 3. In this paper, we relax the limitations imposed in Wang et al. (2024) that the dimensions and the number of candidate models are fixed.

3. Asymptotic theory

In this section, we focus on exploring the theoretical properties of the RJMA estimator and the RJMA-based weight estimator.

3.1 Asymptotic optimality of RJMA estimator

This subsection is devoted to demonstrating that RJMA estimator is asymptotically optimal in the sense that it minimizes the following out-of-sample FPE:

$$\mathrm{FPE}_n(\boldsymbol{w}) = \sum_{i=1}^{n} \mathrm{E}\left\{ h(\breve{\boldsymbol{x}}_i)\rho\left(\widetilde{y}_i - \sum_{m=1}^{M} w_m \boldsymbol{x}_{i(m)}^T \hat{\Theta}_{(m)} \right) \Big| \mathcal{D}_n \right\},$$

where $\widetilde{y}_i = \mu(\boldsymbol{x}_i) + \widetilde{\varepsilon}_i$ with $\widetilde{\varepsilon}_i$ being independent of and identically distributed with ε_i, and $\mathcal{D}_n = \{(y_i, \boldsymbol{x}_i) : i = 1, ..., n\}$.

To proceed, we first give some notations. Let $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^T$ be an $n \times p$ design matrix. Under the m^{th} candidate model, we define the pseudo-true

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

parameter $\Theta^*_{(m)}$ as a minimizer of the following objective function

$$Q(\Theta_{(m)}) = \sum_{i=1}^{n} \mathrm{E}\left\{h(\boldsymbol{x}_{i(m)})\rho\left(y_i - \boldsymbol{x}^T_{i(m)}\Theta_{(m)}\right)\right\}.$$

Assume that $\Theta^*_{(m)}$ is an interior point of $\mathcal{B}_m$ for $m = 1, ..., M$, where $\mathcal{B}_m \subseteq \mathbb{R}^{k_m}$ is the parameter space of $\Theta_{(m)}$. Let $\varepsilon^*_i(\boldsymbol{w}) = y_i - \sum_{m=1}^{M} w_m \boldsymbol{x}^T_{i(m)}\Theta^*_{(m)}$, $\xi^*_i(\boldsymbol{w}) = \mu_i - \sum_{m=1}^{M} w_m \boldsymbol{x}^T_{i(m)}\Theta^*_{(m)}$, $\varepsilon^*_{i(m)} = y_i - \boldsymbol{x}^T_{i(m)}\Theta^*_{(m)}$ and $S_{n(m)} = \sum_{i=1}^{n} h(\boldsymbol{x}_{i(m)})\boldsymbol{x}_{i(m)}\boldsymbol{x}^T_{i(m)}$. For a vector $\boldsymbol{\pi} = (\pi_1, ..., \pi_p)^T$, denote its L_2 norm by $\|\boldsymbol{\pi}\| = (\pi_1^2 + \cdots + \pi_p^2)^{1/2}$. Denote $\mathcal{H}$ as a neighborhood of zero and $\bar{k} = \max_{1\le m\le M} k_m$. Throughout the paper, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the smallest and largest eigenvalues of a real matrix A, respectively.

The establishment of the asymptotic optimality needs the following regularity assumptions, and we suppose that these assumptions hold almost surely. In what follows, we use c and C to denote two generic positive constants that can vary from line to line.

Assumption 1. The loss function ρ is convex on $\mathbb{R}$ with the right and left derivatives $\psi_+(\cdot)$ and $\psi_-(\cdot)$ respectively, where $\psi(\cdot)$ is a bounded function such that $\psi_-(t) \le \psi(t) \le \psi_+(t)$ for all $t \in \mathbb{R}$. Further, $\mathrm{E}\{\psi(\varepsilon_i)|\boldsymbol{x}_i\} = 0$ and $\mathrm{P}_i(\mathcal{D}) = 0$, where $\mathcal{D}$ is the set of points at which ρ is not differentiable and P_i denotes the conditional probability measure of ε_i given $\boldsymbol{x}_i$.

Assumption 2. For any $\boldsymbol{w} \in \mathcal{W}$, $\max_{1\le i\le n} \mathrm{E}\left[\left[\psi\{\varepsilon^*_i(\boldsymbol{w}) + t\} - \psi\{\varepsilon^*_i(\boldsymbol{w})\}\right]^2 |X\right] \to$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

0 as $t \to 0$.

Assumption 3. $R_i(t) := \mathrm{E}\{\psi(\varepsilon_i + t)|X\}$ has a derivative $R_{i1}(t)$. In addition, for any $\boldsymbol{w} \in \mathcal{W}$ and $u \in \mathcal{H}$, $\min_{1\le i\le n} R_{i1}\{\xi_i^*(\boldsymbol{w}) + u\} \ge c$.

Assumption 4. For a large enough n and $m = 1, ..., M$, $\lambda_{\min}(n^{-1}S_{n(m)}) \ge c$.

Assumption 5. $\max_{1\le i\le n} |\mu_i| \le C$; $\sup_{\boldsymbol{x}} \|h(\boldsymbol{x})\boldsymbol{x}\| \le C$; and $\|\Theta^*_{(m)}\| = O(\bar{k}^{1/2})$ uniformly in m.

Assumption 6. $\max_{1\le i\le n} \max_{1\le m\le M} \|\boldsymbol{x}_{i(m)}\| = O(\bar{k}^{1/2})$ and $\bar{k}Mn^{-1} = o(1)$.

In order to make sure that the robust estimator has a bounded influence function, the assumption on the boundedness of $\psi(t)$ is commonly used in the context of robust estimation (see, for example, Avella-Medina and Ronchetti (2018)). The remaining parts of Assumption 1 are standardly imposed in the M-estimation literature (see also Bai et al. (1992); Rao and Zhao (1992); Burman and Nolan (1995); Wu (2007)). It is noteworthy that in Assumption 1, the error term ε_i is allowed to come from the Huber's ϵ-contamination model or heavy-tailed distributions, which may be important in both practice and theory for robust forecasting. $\psi(t)$ is continuous almost surely because of the convexity of the function $\rho(t)$. Consequently, when t tends to zero, $\psi\{\varepsilon_i^*(\boldsymbol{w}) + t\} - \psi\{\varepsilon_i^*(\boldsymbol{w})\}$ is often close to zero, and Assumption 2 is thus easily satisfied. Since $\rho(t)$ is a convex function, $R_i(t)$ often has a continuous and nonnegative first-order derivative

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

$R_{i1}(t)$. For example, let $\rho(t)$ be the absolute loss function and ε_i have a continuous conditional probability density function $f_i(\cdot|\boldsymbol{x}_i)$, then by some elementary calculations, we obtain $R_{i1}(t) = 2f_i(-t|X)$. In the case where $\rho(t)$ is a Huber's function, i.e., $\rho(t) = (t^2\mathrm{I}_{\{|t|\le c\}})/2 + (c|t| - c^2/2)\mathrm{I}_{\{|t|>c\}}$ with I being an indicator function and c being a bending constant, some tedious calculations yield that $R_{i1}(t) = 2\mathrm{P}(|\varepsilon_i + t| \le c|X)$. On the other hand, $\max_{1\le i\le n}|\mu_i|$ is usually bounded, especially in nonparametric models (see, for example, Chen et al. (2018); Yang (2001)). By the definition of $\Theta^*_{(m)}$, it is seen that the value of $\boldsymbol{x}^T_{i(m)}\Theta^*_{(m)}$ is the one close to μ_i and thus $\boldsymbol{x}^T_{i(m)}\Theta^*_{(m)}$ is often bounded. Hence, $|\xi^*_i(\boldsymbol{w})|$ is often bounded and the second part of Assumption 3 is reasonable. Assumption 4 is a classical condition that has been made in the linear model literature (Fan and Peng, 2004; Li et al., 2011). The first part of Assumption 5 has been explained previously and holds in most real applications. The second part of Assumption 5 guarantees that the parameter estimator has a bounded influence function, which can be found in the GM-estimation literature (see, for example, Coakley and Hettamansperger (1993) and He et al. (2000)). Since $\Theta^*_{(m)}$ is a k_m-dimensional vector, the third part of Assumption 5 is mild. Assumption 6 imposes a mild restriction on $\boldsymbol{x}_{i(m)}$ and can be easily satisfied

Theorem 1. *Suppose that Assumptions 1-6 hold. If* $M^3\overline{k}\log^2(\overline{k}^{1/2}\log n)/n^{1-\delta} =$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

$o(1)$ *for some* $0 < \delta < 1$, *then*

$$\frac{\mathrm{FPE}_n(\widehat{\boldsymbol{w}})}{\inf_{\boldsymbol{w}\in\mathcal{W}} \mathrm{FPE}_n(\boldsymbol{w})} \to 1, \tag{3.1}$$

where the convergence is in probability.

Proof. See Section S3 of the Supplementary Material. □

Theorem 1 implies that RJMA estimator is asymptotically optimal in the sense that it is asymptotically identical to the infeasible best possible model averaging estimator, which minimizes the out-of-sample FPE $\mathrm{FPE}_n(\boldsymbol{w})$.

3.2 Consistency of the estimated weight vector

Denote $\overline{\mathrm{FPE}}_n(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[h(\breve{\boldsymbol{x}}_i)\rho\left\{\varepsilon_i^*(\boldsymbol{w})\right\}\right]$, which can be regarded as the population risk function. The theoretically optimal weight vector is defined as $\boldsymbol{w}^0 = \arg\inf_{\boldsymbol{w}\in\mathcal{W}} \overline{\mathrm{FPE}}_n(\boldsymbol{w})$. In this subsection, we devote to an investigation of the consistency of the proposed jackknife weight estimator $\widehat{\boldsymbol{w}}$ in two practical scenarios: (i) $\boldsymbol{w}^0$ is an interior point of the weight set $\mathcal{W}$ and (ii) one or more correct models are included in the set of the candidate models (i.e., $\boldsymbol{w}^0$ is not an interior point of $\mathcal{W}$). Here a correct model is defined as the one that contains all the truly relevant covariates.

We first consider the scenario (i). Denote $B_n = \sum_{i=1}^n h(\breve{\boldsymbol{x}}_i)\boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)\boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T$ and $D_n = \sum_{i=1}^n h^2(\breve{\boldsymbol{x}}_i)\boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)\boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T$, where $\boldsymbol{F}(\boldsymbol{x}_i,\Theta^*) = (\boldsymbol{x}_{i(1)}^T\Theta_{(1)}^*, ..., \boldsymbol{x}_{i(M)}^T\Theta_{(M)}^*)^T$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

with $\Theta^* = (\Theta_{(1)}^{*T}, ..., \Theta_{(M)}^{*T})^T$. We need the following additional regularity assumptions.

Assumption 7. $\max_{1\le i\le n}\max_{1\le m\le M}\|\boldsymbol{x}_{i(m)}\| = o(M^{-3/4}\overline{k}^{-1/2}n^{1/4})$ almost surely.

Assumption 8. There exists a sufficiently large integer N such that for $n > N$, $\lambda_{\min}(n^{-1}B_n) > c$ and $\lambda_{\max}(n^{-1}D_n) \le CM$ almost surely.

Assumption 7 is mild and can be easily satisfied. When n is large enough, $n^{-1}B_n$ is usually a positive definite matrix, and thus the first part of Assumption 8 is reasonable. It is seen that each entry b_{mk} of $n^{-1}D_n$ takes the form of $n^{-1}\sum_{i=1}^n h^2(\breve{\boldsymbol{x}}_i)\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^* \cdot \boldsymbol{x}_{i(k)}^T\Theta_{(k)}^*$ for $1 \le m, k \le M$. As we discussed in Section 3.1, $|\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*|$ is often bounded almost surely uniformly in i and m. This indicates that b_{mk} is often bounded for $1 \le m, k \le M$. So the second part of Assumption 8 is mild.

Theorem 2. *Under Assumptions 1-8, if $\boldsymbol{w}^0$ is an interior point of $\mathcal{W}$, then there is a global minimizer $\hat{\boldsymbol{w}}$ of $\mathrm{RCV}_n(\boldsymbol{w})$ such that*

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}^0\| = O_p(n^{-1/4}M^{1/4}). \tag{3.2}$$

Proof. See Section S4 of the Supplementary Material. □

Theorem 2 displays that the weight estimator $\hat{\boldsymbol{w}}$ tends to the theoretically optimal weight vector $\boldsymbol{w}^0$ at the rate of $n^{-1/4}M^{1/4}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

Remark 4. Unlike Wang et al. (2024), who derived a non-convex Mallows-type weight selection criterion, we demonstrate in the proof of Theorem 2 that if the loss function ρ is convex, then $\mathrm{RCV}_n(\boldsymbol{w})$ is also a convex function of $\boldsymbol{w}$. In the numerical studies, the "solnp" function in R software is utilized to find the minimizer of $\mathrm{RCV}_n(\boldsymbol{w})$.

According to (3.2), it is reasonable to require that

$$\mathrm{E}(\|\hat{\boldsymbol{w}} - \boldsymbol{w}^0\|) = O(n^{-1/4}M^{1/4}). \tag{3.3}$$

In this case, we have

Corollary 1. *Suppose that the assumptions of Theorem 2 hold. If (3.3) is satisfied, then*

$$|\overline{\mathrm{FPE}}_n(\widehat{\boldsymbol{w}}) - \overline{\mathrm{FPE}}_n(\boldsymbol{w}^0)| = O(\overline{k}^{1/2}M^{3/4}n^{-1/4}).$$

Proof. See Section S5 of the Supplementary Material. □

Corollary 1 shows that $\overline{\mathrm{FPE}}_n(\widehat{\boldsymbol{w}}) - \overline{\mathrm{FPE}}_n(\boldsymbol{w}^0)$ converges to zero as long as $\overline{k}^{1/2}M^{3/4}n^{-1/4} = o(1)$, which can be easily satisfied. Note that from the definition of $\boldsymbol{w}^0$, it is clear that the minimum of $\overline{\mathrm{FPE}}_n(\boldsymbol{w})$ over the weight set $\mathcal{W}$ is $\overline{\mathrm{FPE}}_n(\boldsymbol{w}^0)$. This indicates that the weight estimator $\widehat{\boldsymbol{w}}$ asymptotically minimizes $\overline{\mathrm{FPE}}_n(\boldsymbol{w})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

Theorem 3. *Suppose that the assumptions of Corollary 1 hold. If* $\overline{k}^{1/2}M^{3/4}n^{-1/4} = o(1)$*, then*

$$\frac{\overline{\mathrm{FPE}}_n(\widehat{\boldsymbol{w}})}{\inf_{\boldsymbol{w}\in\mathcal{W}}\overline{\mathrm{FPE}}_n(\boldsymbol{w})} = 1 + o(1).$$

Proof. See Section S6 of the Supplementary Material. □

Theorem 3 indicates that the weight vector which is selected by the RJMA criterion is asymptotically optimal in terms of minimizing $\overline{\mathrm{FPE}}_n(\boldsymbol{w})$.

In the following, we consider the scenario (ii). Assume that there are s_0 $(1 \leq s_0 \leq M)$ correct models in the candidate model set. Denote M_m $(m = 1, ..., M)$ as the m^{th} candidate model. To simplify the proof, we let $M_1, ..., M_{s_0}$ be the correct (or overfitted) models, and $M_{s_0+1}, ..., M_M$ be the misspecified (or underfitted) models. Let Ω^* be the subset of $\{1, ..., M\}$ which consists of all the correct models. For the candidate model set Ω^*, we define the weight set in $\mathbb{R}^M$ as $\widetilde{\mathcal{W}} \equiv \{\boldsymbol{w} \in \mathcal{W} : \sum_{m=1}^{s_o} w_m = 1\}$. Define the index set of active covariates as $\mathcal{K} = \{k = 1, ..., p : \theta_k \neq 0\}$. Let the cardinality of $\mathcal{K}$ be k_0. We assume $\breve{\boldsymbol{x}}_i$ to be a d-dimensional vector of covariates with $k_0 \leq d \leq p$, and the true parameter to be $\Theta_0 = (\theta_1, ..., \theta_{k_0}, 0, ..., 0)$ which is a d-dimensional vector. For the m^{th} $(m = 1, ..., M)$ candidate model, denote $\overline{\Theta}_{(m)}$ as a d-dimensional vector which consists of all the elements of the pseudo-true parameter $\Theta^*_{(m)}$ and $d - k_m$ zeros. For example, if we let $M_1 = \{x_1, x_3, x_4\}$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

the corresponding pseudo-true parameter be $\Theta^*_{(1)} = (\theta^*_1, \theta^*_3, \theta^*_4)^T$, then $\overline{\Theta}_{(1)}$ has the form of $\overline{\Theta}_{(1)} = (\theta^*_1, 0, \theta^*_3, \theta^*_4, 0, ..., 0)^T$. By the above notations, it is clear that $\boldsymbol{x}^T_{i(m)}\Theta^*_{(m)} = \breve{\boldsymbol{x}}^T_i\overline{\Theta}_{(m)}$ for $i = 1, ..., n$ and $m = 1, ..., M$.

To prove the weight selection consistency, we need some additional assumptions.

Assumption 9. $Q(\Theta_{(m)})$ has a unique minimizer at $\Theta^*_{(m)}$ for $m = 1, ..., M$.

Assumption 10. For any τ_i that lies between 0 and $\xi^*_i(\boldsymbol{w})$ with $\boldsymbol{w} \in \mathcal{W}\backslash\widetilde{\mathcal{W}}$, $\min_{1\le i\le n} R_{i1}(\tau_i) \ge c$ almost surely.

Assumption 11. When n is large enough, $\lambda_{\min}\left\{\frac{1}{n}\sum_{i=1}^n h(\breve{\boldsymbol{x}}_i)\breve{\boldsymbol{x}}_i\breve{\boldsymbol{x}}^T_i\right\} > c$ almost surely.

Assumption 12. $\inf_{\boldsymbol{w}\in\mathcal{W}\backslash\widetilde{\mathcal{W}}}\left\|\sum_{m=s_0+1}^M w_m(\overline{\Theta}_{(m)} - \Theta_0)\right\|^2 \ge c.$

Assumption 9 is a commonly used identification condition and the similar assumption can be found in White (1982). By Assumption 9, we conclude that when the m^{th} candidate model is correct, $\Theta^*_{(m)}$ takes the true parameter value and this yields that $x^T_{i(m)}\Theta^*_{(m)} = \mu_i$ for $i = 1, ..., n$. Assumptions 10 and 11 are similar to Assumptions 3 and 4, respectively, and the detailed explanations are provided in Section 3.1. Since $M_{s_0+1}, ..., M_M$ are the misspecified models, it is often true that for any $\boldsymbol{w} \in \mathcal{W}\backslash\widetilde{\mathcal{W}}$, $\sum_{m=s_0+1}^M w_m\overline{\Theta}_{(m)}$ cannot be equal to $\sum_{m=s_0+1}^M w_m\Theta_0$. Consequently, Assumption 12 is reasonable.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

Theorem 4. *Suppose that Assumptions 1-6 and 9-12 hold. Then,*

$$\mathrm{P}(\widehat{\boldsymbol{w}} \in \widetilde{\mathcal{W}}) \to 1. \tag{3.4}$$

Proof. See Section S7 of the Supplementary Material. □

Theorem 4 indicates that the proposed weight estimator $\widehat{\boldsymbol{w}}$ is located in $\widetilde{\mathcal{W}}$ with probability tending to one when there are correct models in the candidate model set, i.e., the weights will concentrate on all the correct models. The similar result was also obtained by Fang et al. (2022) and He et al. (2023).

Theorem 5. *Suppose that the assumptions of Theorem 4 hold. Then,*

$$n^{-1/2}\|\widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{w}}) - \boldsymbol{\mu}\| = o_p(1).$$

Proof. See Section S8 of the Supplementary Material. □

Theorem 5 indicates that the proposed RJMA estimator $\widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{w}})$ tends to the true value $\boldsymbol{\mu}$ in probability if there are correct models in the candidate model set.

In this paper, we also derive the influence function of the RJMA estimator, and establish its robustness property by demonstrating that its influence function is bounded. Furthermore, an empirical prediction influence function is proposed to quantitatively evaluate the robustness of the RJMA estimator. The details can be found in Section S1 of the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

4. Simulation studies

In this section, we evaluate the finite-sample performance of the proposed RJMA method via various simulations. Our two RJMA estimators are denoted as RJMA_A and RJMA_H when the loss functions are chosen to be the absolute deviation and Huber's functions, respectively. To make a comparison with existing competitors, we consider the following eleven model selection and averaging methods. Ronchetti and Staudte (1994) presented a robust version of Mallows' C_p for regression models, and the corresponding two model selection estimators are denoted as MC_p and HC_p when the weight functions of Mallows' and Huber's types are used, respectively. We denote the general Akaike-type model selection methods (Burman and Nolan, 1995) as MS_A and MS_H, the S_p-type robust model averaging methods (Guo and Li, 2021) as SMA_A and SMA_H, and the Mallows-type robust model averaging methods (Wang et al., 2024) as MA_A and MA_H, corresponding to the choices of absolute deviation and Huber's functions as the loss functions, respectively. Liu and Okui (2013) developed a heteroscedasticity-robust model averaging criterion which is denoted as HRC_p. Further, comparisons are drawn with the traditional model averaging methods MMA (Hansen, 2007) and JMA (Hansen and Racine, 2012). In total, we compare the performance of thirteen methods including RJMA_A, MS_A, MA_A, SMA_A, RJMA_H, MS_H, MA_H, SMA_H, MC_p, HC_p, MMA, JMA and HRC_p. The

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

evaluation of the performance for different estimators is based on the following out-of-sample mean absolute error (MAE) across $R = 300$ replications:

$$\mathrm{MAE} = \frac{1}{nR}\sum_{r=1}^{R}\sum_{i=1}^{n}\left|\mu_i^{(r)} - \widehat{\mu}_i^{(r)}\right|,$$

where $\mu_i^{(r)}$ is calculated using the clean testing dataset $\{\boldsymbol{x}_i^{(r)}\}_{i=1}^n$, $\widehat{\mu}_i^{(r)}$ is its estimation value, and r indexes the r^{th} simulation trial.

For the weight function involved in the GM-type loss function, in the simulation studies and real data example, we let it take the form of $h(\boldsymbol{x}_{i(m)}) = \varphi_b(q_{i(m)})/q_{i(m)}$, where $q_{i(m)}$ is the i^{th} diagonal element of the "hat matrix" $Q_{(m)} = \boldsymbol{x}_{(m)}(\boldsymbol{x}_{(m)}^T\boldsymbol{x}_{(m)})^{-1}\boldsymbol{x}_{(m)}^T$ with $\boldsymbol{x}_{(m)} = (\boldsymbol{x}_{1(m)}, ..., \boldsymbol{x}_{n(m)})^T$, and $\varphi_b(q_{i(m)}) = q_{i(m)}\mathrm{I}_{\{q_{i(m)}\le b\}} + b\mathrm{I}_{\{q_{i(m)}>b\}}$ with b being the bending constant. As in Sommer and Staudte (1995), we let $b = 1.5k_m/n$, then $\boldsymbol{x}_{i(m)}$ will be downweighted when $q_{i(m)}$ is larger than 1.5 times of the average leverage. Further, the tuning parameter corresponding to the Huber's function is set to be 1.345 (see also Ronchetti and Staudte (1994)).

As in Hansen and Racine (2012), we consider the following data generating process

$$y_i = \sum_{j=1}^{1000}\theta_j x_{ij} + \varepsilon_i, \qquad i = 1, 2, \cdots, n, \tag{4.1}$$

where $x_{i1} = 1$ is an intercept term and the remaining covariates x_{ij}, $j = 2, ..., 1000$, are independent and identically distributed; the parameters $\theta_j = c(2\nu)^{1/2}j^{-\nu-0.5}$ with $\nu = 0.5$, and c varied so that $R^2 = 2\nu Z(1+2\nu)c^2/\{1+$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

$Z(1+2\nu)c^2\} = 0.1, 0.3, \cdots, 0.9$ with $Z(k)$ being the Zeta function; and ε_i, $i = 1, ..., n$, are mutually independent. For the random covariates x_{ij}, $j = 2, ..., 1000$, and the error term ε_i, we consider the following six cases:

Case 1 The random covariates x_{ij}, $j = 2, ..., 1000$, and the error term ε_i follow the standard normal distribution $\mathcal{N}(0, 1)$. In this case, there are no outliers in the sample.

Case 2 The random covariates x_{ij}, $j = 2, ..., 1000$, follow the mixture distribution $(1-\varrho)\mathcal{N}(0, 1) + \varrho t(1)$ with ϱ being outlier proportion, and the error term ε_i follows the standard normal distribution $\mathcal{N}(0, 1)$. In this case, outliers occur only in the covariates.

Case 3 The random covariates x_{ij}, $j = 2, ..., 1000$, follow the standard normal distribution $\mathcal{N}(0, 1)$, and the error term ε_i follows the distribution $t(1)$. In this case, outliers occur only in the response.

Case 4 The random covariates x_{ij}, $j = 2, ..., 1000$, follow the mixture distribution $(1-\varrho)\mathcal{N}(0, 1) + \varrho t(1)$, and the error term ε_i follows the distribution $t(1)$. In this case, outliers occur in both the covariates and response.

Case 5 The random covariates x_{ij}, $j = 2, ..., 1000$, follow the standard normal distribution $\mathcal{N}(0, 1)$, and the error term ε_i follows the normal distribution with mean 0 and variance $\sum_{j=2}^{6} x_{ij}^4 + 0.01$. In this case, the observations are absent from outliers.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

Case 6 The random covariates x_{ij}, $j = 2, ..., 1000$, follow the mixture normal distribution $(1-\varrho)\mathcal{N}(0,1) + \varrho\mathcal{N}(5, 5^2)$, and the error term ε_i follows the distribution $(\sum_{j=2}^{6} x_{ij}^2 + 0.01)^{1/2} \times t(1)$. In this case, there are outliers in both the covariates and response.

Cases 1-4 and 5-6 correspond to the situations where the error terms are homoskedastic and heteroskedastic, respectively. In the simulation studies, the M candidate models are constructed as the nested regression models with variables $\{1, x_{i2}\}$, $\{1, x_{i2}, x_{i3}\}$, $\{1, x_{i2}, x_{i3}, x_{i4}\}$,..., respectively, where $M = [3n^{1/3}]$ with $n = 50$ and 100. In the implementation of simulation studies, for simplicity, the random covariates in Cases 2, 4 and 6 are generated in the following way: $100(1-\varrho)\%$ of the random covariates come from $\mathcal{N}(0,1)$ and the remaining $100\varrho\%$ come from $t(1)$ and $\mathcal{N}(5, 5^2)$, respectively. The simulation results are displayed in Tables 1-6. In the main text, we only report the simulation results for $\varrho = 0.2$. To facilitate comparisons, the best is shown in bold and the second-best is marked with a "†" symbol.

From Table 1, we observe that for Case 1, JMA usually has the best performance, and the performance difference among JMA, HRC$_p$ and MMA is not obvious. This indicates that when the sample is absent from outliers, the traditional model averaging methods can have satisfactory performance. For the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

comparison among the robust methods, we find that RJMA$_H$ usually performs better than MS$_A$, MA$_A$, SMA$_A$, MS$_H$, MA$_H$, SMA$_H$, MC$_p$ and HC$_p$, and the MAEs of RJMA$_A$ are often smaller than those of MS$_A$, SMA$_A$, MC$_p$ and HC$_p$.

It can be observed from Table 2 that when outliers occur only in the covariates, the proposed estimator RJMA$_H$ has the best performance and followed by RJMA$_A$. In contrast, traditional model averaging methods such as MMA, JMA, and HRCp perform poorly compared to other robust model selection and averaging techniques, highlighting the need for robust model averaging approaches. The performance of MA$_A$ is often better than that of SMA$_A$ and MS$_A$. As for MA$_H$, it usually outperforms SMA$_H$ and MS$_H$. Additionally, MC$_p$ tends to achieve better results than MS$_A$, SMA$_A$, MS$_H$ and SMA$_H$. From Table 3, we see that when the error term comes from a heavy-tailed distribution $t(1)$ and the covariates contain no outliers, the performance of RJMA$_H$ is often the best and followed by MA$_A$. RJMA$_A$ is only inferior to MS$_A$, MA$_A$, RJMA$_H$ and MA$_H$. It is found that the MAEs of MS$_A$ are often smaller than those of SMA$_A$ and SMA$_H$, but SMA$_A$ and SMA$_H$ usually dominate MC$_p$ and HC$_p$. It is also clear that in Case 3, the MAEs of MMA, JMA and HRC$_p$ are larger than those of other robust methods. When the sample is contaminated by outliers in both the response and covariates, it can be seen from Table 4 that the proposed estimators RJMA$_A$ and RJMA$_H$ usually occupy the top two, and followed by MA$_A$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

MA_H. We observe that MS_A and MS_H yield the better results than SMA_A and SMA_H, respectively. Moreover, the performance of MC_p is superior to that of SMA_H and SMA_A.

It can be observed from Table 5 that for the heteroskedastic data generating process, RJMA_H usually has the best performance, which implies that the proposed methodology is also heteroscedasticity-robust. As for RJMA_A, it often has better performance than MS_A, SMA_A, MS_H, SMA_H, MC_p, HC_p, MMA, JMA and HRC_p. It is also clear that MA_A and MA_H perform better than SMA_A and SMA_H, respectively. From Table 6, we find that when outliers occur in both the response and covariates and the error term is heteroskedastic, RJMA_H often has the best performance. In the case where the sample size $n = 50$, RJMA_A is only inferior to RJMA_H. However, when the sample size increases to $n = 100$ and $R^2 = 0.1, 0.3$ and 0.5, MA_A performs better than RJMA_A. For the case of $R^2 = 0.7$ and 0.9, RJMA_A outperforms MA_A. It is also seen that MA_H performs better than SMA_A and SMA_H. Furthermore, MC_p usually dominates MS_H, SMA_A and SMA_H.

In summary, for Case 1 where the dataset contains no outliers, the commonly used model averaging methods MMA, JMA and HRC_p usually perform better than other robust methods, but their performance becomes the worst in the presence of outliers. This result indicates that the model averaging method built

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

on the least squares is not stable when there are outliers in the sample. So it is necessary to develop a robust model averaging method which can be resistant to the leverage points and outliers in the response. It is observed from Tables 2-6 that in the presence of outliers and/or heteroscedasticity, the proposed methods RJMA$_A$ and RJMA$_H$ often rank in the top two. Further, in the absence of outliers in the dataset, from Table 1, we find that the performance difference between JMA and the proposed methods RJMA$_A$ and RJMA$_H$ is not obvious. Therefore, we conclude that RJMA method is a good choice for prediction.

We extend the simulation studies to the data generating process specified in (4.1), where the covariates x_{ij}, $j = 2, 3, \cdots, 1000$, are drawn from a mixed multidimensional distribution $0.8\mathcal{N}(0, \Sigma) + 0.2t_1(0, \Sigma)$ with the diagonal elements of Σ being 1, and its off-diagonal elements being $0.5^{|i-j|}$. The results for this scenario are detailed in Section S10.1 of the Supplementary Material. Additionally, we replicate the simulation setup from Burman and Nolan (1995) in Section S10.2 of Supplementary Material, which focuses on a non-linear model. In conclusion, RJMA$_A$ and RJMA$_H$ exhibit good performance across the examined scenarios, including dependent and complex structured data, when the outliers occur in the sample.

We also consider 10% and 30% outlier proportions for Cases 2, 4 and 6, and the covariate-dependent setting in Section S10.3. Overall, the proposed methods

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

Table 1: MAEs of various estimators for Case 1

n	R^2	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
50	0.1	0.347	0.350	0.298	0.417	0.250	0.315	0.272	0.278	0.451	0.450	0.245†	**0.243**	0.248
	0.3	0.375	0.407	0.348	0.446	0.307	0.381	0.323	0.328	0.472	0.472	0.304†	**0.302**	0.306
	0.5	0.441	0.500	0.432	0.496	0.388	0.459	0.398	0.400	0.501	0.501	0.387†	**0.385**	0.388
	0.7	0.521	0.602	0.523	0.569	0.478	0.540	0.482	0.483	0.554	0.555	0.474†	**0.473**	0.475
	0.9	0.794	0.855	0.798	0.804	0.745	0.767	0.737	**0.734**	0.758	0.758	0.737	0.737	0.735†
100	0.1	0.259	0.231	0.214	0.289	0.193	0.230	0.202	0.206	0.344	0.344	**0.188**	**0.188**	0.190†
	0.3	0.299	0.318	0.275	0.321	0.252	0.305	0.258	0.260	0.356	0.356	0.251†	**0.250**	0.252
	0.5	0.359	0.400	0.345	0.372	0.313	0.363	0.316	0.318	0.386	0.386	0.310†	**0.309**	0.310†
	0.7	0.436	0.486	0.432	0.451	0.398†	0.433	0.398†	0.398†	0.432	0.431	**0.395**	**0.395**	**0.395**
	0.9	0.661	0.689	0.652	0.653	0.616	0.628	0.613	0.612†	0.621	0.621	**0.611**	0.612†	**0.611**

Table 2: MAEs of various estimators for Case 2: $\varrho = 0.2$

n	R^2	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
50	0.1	0.303†	0.514	0.322	0.572	**0.262**	0.527	0.346	0.571	0.480	0.619	2.062	1.554	1.917
	0.3	0.381†	0.883	0.460	0.986	**0.352**	0.914	0.487	1.127	0.611	1.154	2.656	2.025	2.329
	0.5	0.461†	0.850	0.514	0.894	**0.434**	0.900	0.541	1.305	0.676	1.155	4.313	2.987	3.891
	0.7	0.580†	1.205	0.672	1.314	**0.564**	1.229	0.695	1.810	0.867	1.592	6.175	4.887	5.612
	0.9	0.936†	1.941	1.060	2.184	**0.935**	2.206	1.106	3.005	1.394	2.753	10.240	13.040	9.264
100	0.1	0.209	0.362	0.208†	0.395	**0.190**	0.381	0.221	0.406	0.304	0.432	1.851	1.088	1.603
	0.3	0.286†	0.547	0.295	0.559	**0.274**	0.585	0.320	0.636	0.397	0.642	3.587	2.105	3.114
	0.5	0.351†	0.717	0.394	0.731	**0.348**	0.775	0.416	0.997	0.522	0.979	3.940	2.509	3.370
	0.7	0.458†	0.819	0.481	0.838	**0.457**	0.966	0.512	1.092	0.600	1.021	5.489	7.780	4.614
	0.9	0.745†	1.361	0.771	1.407	**0.744**	1.604	0.804	2.209	0.902	2.067	8.741	6.126	7.567

RJMA$_A$ and RJMA$_H$ perform well and often rank in the top two.

5. Real data example

To demonstrate the proposed method, we apply it to analyze human immunodeficiency virus (HIV) data from the AIDS Clinical Trials Group (ACTG) protocol 175 study (Hammer et al., 1996), which is available in the `R` package `speff2trial`. The ACTG 175 trial evaluated the efficacy of treatments in-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

Table 3: MAEs of various estimators for Case 3

n	R^2	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
50	0.1	0.527	0.450	**0.410**	0.777	0.437†	0.587	0.550	0.876	1.111	1.118	4.801	4.095	4.254
	0.3	0.527	0.491†	**0.442**	0.717	**0.442**	0.564	0.525	0.807	1.030	1.035	5.231	4.300	4.524
	0.5	0.594	0.632	0.539†	0.782	**0.538**	0.732	0.621	0.886	1.054	1.063	4.476	3.590	3.786
	0.7	0.681	0.783	0.669†	0.872	**0.645**	0.841	0.718	0.952	1.108	1.111	4.234	3.541	3.752
	0.9	1.014	1.135	1.010	1.080	**0.964**	1.153	0.999†	1.179	1.244	1.248	3.951	3.405	3.553
100	0.1	0.341	0.260†	**0.245**	0.395	0.289	0.360	0.325	0.508	0.705	0.705	3.884	3.285	3.368
	0.3	0.397	0.355	**0.323**	0.465	0.351†	0.451	0.386	0.549	0.729	0.730	5.377	4.403	4.479
	0.5	0.457	0.455	**0.405**	0.500	0.412†	0.516	0.437	0.581	0.737	0.736	4.066	3.478	3.549
	0.7	0.535	0.589	0.517†	0.575	**0.512**	0.632	0.531	0.640	0.758	0.758	3.556	3.039	3.125
	0.9	0.797	0.885	0.810	0.811	**0.777**	0.913	0.781†	0.865	0.904	0.903	4.791	4.189	4.250

Table 4: MAEs of various estimators for Case 4: $\varrho = 0.2$

n	R^2	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
50	0.1	0.417†	0.752	0.426	0.974	**0.404**	0.885	0.561	1.192	0.969	1.199	3.518	2.853	3.456
	0.3	0.495†	0.997	0.538	1.228	**0.464**	1.074	0.680	1.624	1.152	1.611	4.301	4.458	4.086
	0.5	0.584†	1.155	0.685	1.498	**0.579**	1.285	0.816	1.939	1.273	1.788	7.703	6.556	7.294
	0.7	0.731†	1.582	0.836	1.818	**0.724**	1.719	1.011	2.988	1.509	2.603	8.305	6.584	8.018
	0.9	1.188†	2.448	1.451	2.942	**1.145**	2.814	1.634	4.638	2.234	3.869	13.620	11.113	12.980
100	0.1	0.281	0.521	**0.258**	0.620	0.280†	0.673	0.346	0.792	0.610	0.809	4.091	3.443	3.894
	0.3	**0.355**	0.721	0.361	0.805	0.360†	0.862	0.475	1.071	0.746	1.029	6.174	4.306	5.544
	0.5	**0.434**	0.911	0.448	1.004	0.440†	1.083	0.551	1.242	0.821	1.204	5.746	3.945	5.190
	0.7	**0.569**	1.181	0.583	1.245	0.577†	1.419	0.673	1.676	0.959	1.539	7.799	6.194	7.054
	0.9	**0.868**	1.713	0.947	1.819	0.882†	2.083	1.023	2.509	1.347	2.297	11.096	7.706	10.266

Table 5: MAEs of various estimators for Case 5

n	R^2	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
50	0.1	0.911	0.981	0.807†	1.336	**0.739**	1.037	0.980	1.133	1.560	1.582	1.078	0.944	0.998
	0.3	0.946	1.015	0.831†	1.367	**0.763**	1.007	0.991	1.134	1.583	1.603	1.041	0.925	0.974
	0.5	0.957	1.031	0.848†	1.312	**0.789**	1.035	0.996	1.137	1.567	1.588	1.081	0.972	1.017
	0.7	1.003	1.153	0.977†	1.382	**0.887**	1.127	1.077	1.194	1.582	1.601	1.129	1.037	1.077
	0.9	1.332	1.544	1.298†	1.550	**1.221**	1.490	1.330	1.419	1.687	1.711	1.445	1.379	1.406
100	0.1	0.650	0.553	**0.481**	0.808	0.491†	0.649	0.607	0.756	1.117	1.131	0.742	0.644	0.665
	0.3	0.685	0.607	0.550†	0.836	**0.544**	0.702	0.652	0.796	1.139	1.150	0.795	0.704	0.724
	0.5	0.718	0.692	0.605†	0.874	**0.597**	0.765	0.686	0.812	1.112	1.124	0.839	0.765	0.780
	0.7	0.805	0.831	0.715†	0.918	**0.691**	0.881	0.761	0.870	1.134	1.145	0.869	0.821	0.831
	0.9	1.051	1.197	1.030†	1.122	**0.988**	1.181	1.011	1.098	1.252	1.263	1.171	1.146	1.153

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

Table 6: MAEs of various estimators for Case 6: $\varrho = 0.2$

n	R^2	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
50	0.1	1.169†	3.327	1.220	4.457	**1.110**	3.311	1.877	7.384	3.321	6.178	22.249	12.465	20.943
	0.3	1.368†	3.310	1.432	4.385	**1.354**	3.506	2.038	6.560	3.341	5.928	22.259	12.819	21.365
	0.5	1.693†	3.684	1.805	4.646	**1.648**	3.938	2.362	6.518	3.557	5.872	21.244	11.540	20.165
	0.7	2.199†	4.129	2.357	4.758	**2.119**	4.523	2.717	6.639	3.723	5.962	23.096	13.546	21.919
	0.9	3.450†	5.455	3.669	5.972	**3.404**	6.002	3.884	9.379	4.556	8.040	22.818	15.083	21.278
100	0.1	0.785	1.913	**0.713**	2.520	0.751†	2.164	1.070	3.467	2.048	3.339	18.150	8.808	15.731
	0.3	1.080	2.136	**1.028**	2.582	1.056†	2.450	1.314	3.493	2.107	3.316	20.831	10.808	17.914
	0.5	1.300	2.445	1.288†	2.706	**1.275**	2.795	1.523	3.649	2.234	3.466	17.908	9.230	14.672
	0.7	1.689†	2.733	1.721	2.892	**1.672**	3.128	1.864	3.669	2.422	3.507	15.627	9.241	12.693
	0.9	2.718†	3.772	2.747	3.775	**2.611**	4.347	2.795	4.519	3.189	4.309	19.504	11.624	17.228

volving either a single nucleoside or a combination of two nucleosides in adults infected with HIV type 1 (HIV-1), whose baseline CD4 cell counts ranged from 200 to 500 per cubic millimeter. Participants were divided into two groups based on the treatment regimen: one received zidovudine, ZDV, monotherapy (ZDV only), while the other received one of three alternative therapies (ZDV + didanosine, ddI; ZDV + zalcitabine, ddC; and ddI only). In total, 2139 subjects were enrolled in the study across both groups, in which there are 368 female patients.

In this example, we only analyze the 368 female patients. We select the CD4 cell count at 96 ± 5 weeks post baseline (CD4$_{96}$) as the response variable (see also Han et al. (2019)). The chosen covariates include nine factors: treatment indictor (trt; 0=ZDV only), CD4 cell count at baseline (CD4$_0$), age in years at baseline (age), weight in kg at baseline (weight), race (race; 0=white), history of intravenous drug use (drug; 0=no), indicator of off-treatment before 96 ± 5

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

weeks (offtrt; 0=no), CD4 cell count at 20 ± 5 weeks (CD4_{20}) and CD8 cell count at 20 ± 5 weeks (CD8_{20}). After excluding subjects with missing CD4_{96} values, the final sample consists of $n = 218$ observations. The covariates are ranked according to their absolute correlation with the response variable in the following order: x_1 (CD4_{20}), x_2 (CD4_0), x_3 (offtrt), x_4 (trt), x_5 (weight), x_6 (race), x_7 (CD8_{20}), x_8 (age) and x_9 (drug). We construct a sequence of nine nested candidate models, beginning with $\{1, x_1\}$, and incrementally adding covariates up to $\{1, x_1, x_2, ..., x_9\}$. All the non-dummy variables (CD4_{96}, CD4_{20}, CD4_0, weight, CD8_{20}, and age) are standardized to have zero mean and unit variance for the sake of avoiding the influence of different scales.

For each repetition, we randomly split the $n = 218$ observations into training and testing sets, denoted as $\{\boldsymbol{x}_s, y_s\}_{s=1}^{n_1}$ and $\{\boldsymbol{x}_t, y_t\}_{t=1}^{n_2}$, respectively. Denote $\widehat{\mu}_t$ as the predictive value of the response variable calculated by a given averaging/selection method. To evaluate the performance of different methods, we calculate the absolute prediction error (APE) and mean absolute prediction error (MAPE), defined as:

$$\text{APE}^{(r)} = \frac{1}{n_2}\sum_{t=1}^{n_2}\left|y_t^{(r)} - \widehat{\mu}_t^{(r)}\right| \text{ and } \text{MAPE} = \frac{1}{R}\sum_{r=1}^{R}\text{APE}^{(r)}$$

respectively, where r denotes the index for the r^{th} repetition.

We consider the training sample sizes of $n_1 = 50$, 100 and 150, respectively. To seed outliers in the sample, we randomly choose 20% of the values for

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

$CD4_{20}$, $CD4_0$, weight and $CD8_{20}$ in the training sample and replace them with samples drawn from a heavy-tailed distribution $t(1)$. All the simulation results are reported in Figure 1 and Table 7.

It can be seen from Figure 1 and Table 7 that when the training sample sizes $n_1 = 50$ and 100, the proposed methods RJMA_A and RJMA_H exhibit the best performance, and followed by MA_A and MA_H. As the training sample size increases to $n_1 = 150$, the most favored is RJMA_A and followed by MA_A. RJMA_H performs slightly worse than MA_A, with performance comparable to that of MA_H. Due to the presence of outliers, the performance of MMA, J-MA and HRC_p is not good. This example suggests that the proposed RJMA_A and RJMA_H are good candidates for prediction when outliers are present in the sample.

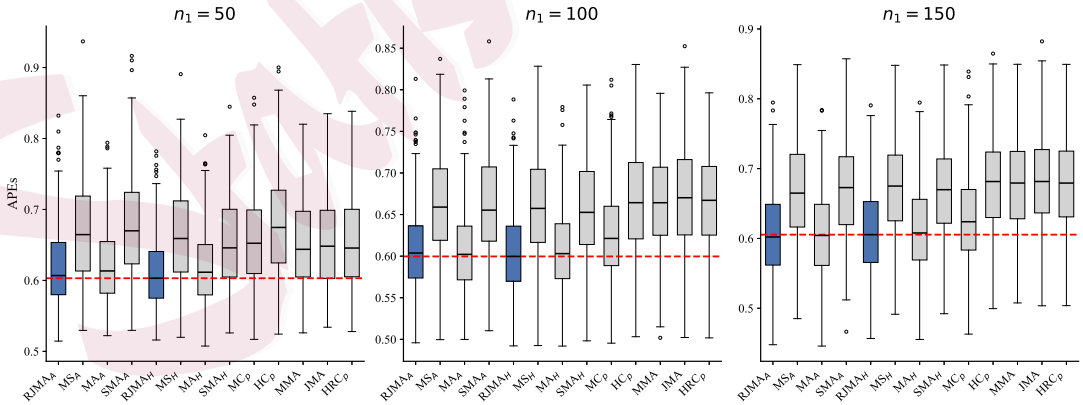

Figure 1: Boxplots for the APEs of various forecasts based on $R = 300$ random divisions: HIV data

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

Table 7: MAPEs of various forecasts for HIV data based on $R = 300$ random divisions

n	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
50	0.619†	0.670	0.622	0.676	**0.612**	0.667	0.620	0.654	0.659	0.682	0.653	0.656	0.655
100	0.607†	0.663	0.607†	0.661	**0.605**	0.660	0.607†	0.656	0.625	0.668	0.663	0.670	0.665
150	**0.606**	0.668	0.607†	0.669	0.610	0.673	0.611	0.669	0.627	0.680	0.676	0.680	0.677

6. Concluding remarks

This paper developed a robust jackknife model averaging method which can be regarded as a robust version of the jackknife model averaging suggested by Hansen and Racine (2012). The RJMA method aims to provide robust parameter estimators by utilizing specific GM-type loss functions and selecting weights through minimizing the GM-type leave-one-out CV criterion. This approach ensures that the RJMA estimator remains reliable even when the dataset is contaminated by outliers in response and/or covariates. We demonstrated that the RJMA estimator is asymptotically optimal in terms of minimizing the out-of-sample FPE. In the case where the theoretically optimal weight $\boldsymbol{w}^0$ is an interior point of the weight set, we derived the rate of RJMA-based weight estimator converging to $\boldsymbol{w}^0$. If there is one or more correct models in the candidate model set, we showed that all the RJMA-based weights are assigned to these correct models asymptotically. Furthermore, we explored the influence function under the model averaging framework, and defined EPIF to evaluate the quantitative

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

robustness of RJMA estimator. Both simulation study and real data analysis display the satisfactory and robust performance of the proposed RJMA method in the presence of outliers.

It is worth noting that many other traditional model averaging methods like optimal model averaging (Liang et al., 2011) and Kullback-Leibler model averaging (Zhang et al., 2015) are also sensitive to outliers. So, it is meaningful to develop the robust versions of these weight selection criteria. More recently, extensive research has focused on developing model averaging methods for nonparametric and semiparametric mean regression models because these models are more flexible than the tightly specified parametric models. Thus, it is also necessary to investigate the robust model averaging method in the nonparametric or semi-parametric situation. In this paper, we consider the case where the observations are mutually independent. How to extend the proposed method to the autoregressive model is an interesting topic. The study of robust model averaging inference is an important problem. However, since the closed-form expressions of the robust estimators are unknown, it would be hard to derive the asymptotic distribution of robust model averaging estimator, which warrants our future work.

Supplementary Material

The Supplementary Material contains the robustness property of the RJMA

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

estimator, proofs of theorems and additional simulation studies.

Acknowledgments

The authors thank the editor, the associate editor, and two referees for their careful reviews and helpful suggestions. Zou and Wang's work was supported by the National Natural Science Foundation of China (Grant Nos. 12531012, 12031016, 12426308 and 12401335). Zou's work was also supported by the Beijing Outstanding Young Scientist Program (Grant No. JWZQ20240101027). You's work was partially supported by the Engineering and Physical Sciences Research Council of United Kingdom (Grant No. EP/X038297/1).

References

Ando, T. and K.-C. Li (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association* **109**, 254–265.

Avella-Medina, M. and E. Ronchetti (2018). Robust and consistent variable selection in high-dimensional generalized linear models. *Biometrika* **105**, 31–44.

Bai, Z. D., C. R. Rao, and Y. Wu (1992). M-estimation of multivariate linear regression parameters under a convex discrepancy function. *Statistica Sinica* **2**, 237–254.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

Burman, P. and D. Nolan (1995). A general Akaike-type criterion for model selection in robust regression. *Biometrika* **82**, 877–886.

Chen, J., D. G. Li, O. Linton, and Z. D. Lu (2018). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of the American Statistical Association* **113**, 919–932.

Coakley, C. W. and T. P. Hettamansperger (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association* **88**, 872–880.

Du, J., Z. Z. Zhang, and T. F. Xie (2018). Model averaging for M-estimation. *Statistics* **52**, 1417–1432.

Fan, J. Q., Y. Y. Fan, and E. Barut (2014). Adaptive robust variable selection. *Annals of Statistics* **42**, 324–351.

Fan, J. Q. and H. Peng (2004). On nonconcave penalized likelihood with diverging number of parameters. *Annals of Statistics* **32**, 928–961.

Fang, F., J. L. Li, and X. C. Xia (2022). Semiparametric model averaging prediction for dichotomous response. *Journal of Econometrics* **229**, 219–245.

Gao, Y., X. Y. Zhang, S. Y. Wang, and G. H. Zou (2016). Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics* **192**, 139–151.

Guo, Y. F. and Z. H. Li (2021). Outlier robust model averaging based on S_p criterion. *Stat* **10**, 1–10.

Hammer, S. M., D. A. Katzenstein, M. D. Hughes, H. Gundaker, R. T. Schooley, R. H. Haubrich, et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine* **335**,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

1081–1090.

Hampel, F. R. (1968). "Contributions to the theory of robust estimation", Ph. D. thesis. *University of California Berkeley*. URL: https://books.google.co.uk/books?id=nh_MvwEACAAJ.

Han, P., L. Kong, J. Zhao, and X. Zhou (2019). A general framework for quantile estimation with incomplete data. *Journal of the Royal Statistical Society B* **81**, 305–333.

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

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

He, B. H., S. G. Ma, X. Y. Zhang, and L. X. Zhu (2023). Rank-based greedy model averaging for high-dimensional survival data. *Journal of the American Statistical Association* **118**, 2658–2670.

He, X. M., D. G. Simpson, and G. Y. Wang (2000). Breakdown points of t-type regression estimators. *Biometrika* **87**, 675–687.

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

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science* **14**, 382–401.

Hu, X. N. and X. Y. Zhang (2023). Optimal parameter-transfer learning by semiparametric model averaging. *Journal of Machine Learning Research* **24**, 1–53.

Kitagawa, T. and M. Chris (2016). Model averaging in semiparametric estimation of treatment effects. *Journal of Econometrics* **193**, 271–289.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *Annals of Statistics* **26**, 755–770.

Le, T. M. and B. S. Clarke (2022). Model averaging is asymptotically better than model selection for prediction. *Journal of Machine Learning Research* **23**, 1–53.

Li, D. G., O. Linton, and Z. D. Lu (2015). A flexible semiparametric forecasting model for time series. *Journal of Econometrics* **187**, 345–357.

Li, G. R., H. Peng, and L. Zhu (2011). Nonconcave penalized M-estimation with a diverging number of parameters. *Statistica Sinica* **21**, 391–419.

Li, J. L., J. Lv, A. T. K. Wan, and J. Liao (2022). Adaboost semiparametric model averaging prediction for multiple categories. *Journal of the American Statistical Association* **117**, 495–509.

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

Liao, J., G. H. Zou, Y. Gao, and X. Y. Zhang (2021). Model averaging prediction for time series models with a diverging number of parameters. *Journal of Econometrics* **223**, 190–221.

Liu, Q. F. and R. Okui (2013). Heteroscedasticity-robust C_p model averaging. *Econometrics Journal* **16**, 463–472.

Lozano, A., N. Meinshausen, and E. Yang (2016). Minimum distance lasso for robust high-dimensional regression. *Electronic Journal of Statistics* **10**, 1296–1340.

Lu, X. and L. J. Su (2015). Jackknife model averaging for quantile regressions. *Journal of Econometric-*

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

s **188**, 40–58.

Rao, C. R. and L. C. Zhao (1992). Approximation to the distribution of M-estimates in linear models by randomly weighted bootstrap. *Sankhyā A* **54**, 323–331.

Ronchetti, E. (1997). Robustness aspects of model choice. *Statistica Sinica* **7**, 327–338.

Ronchetti, E., C. Field, and W. Blanchard (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association* **92**, 1017–1023.

Ronchetti, E. and R. G. Staudte (1994). A robust version of Mallows' C_p. *Journal of the American Statistical Association* **89**, 550–559.

Schomaker, M. and C. Heumann (2020). When and when not to use optimal model averaging. *Statistical Papers* **61**, 2221–2240.

Sommer, S. and R. G. Staudte (1995). Robust variable selection in regression in the presence of outliers and leverage points. *Australian Journal of Statistics* **37**, 323–336.

Wan, A. T. K., X. Y. Zhang, and S. Wang (2014). Frequentist model averaging for multinomial and ordered logit models. *International Journal of Forecasting* **30**, 118–128.

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

Wang, M. M., K. You, L. X. Zhu, and G. H. Zou (2024). Robust model averaging approach by Mallows-type criterion. *Biometrics* **80**, ujae128.

Wang, M. M., X. Y. Zhang, A. T. K. Wan, K. You, and G. H. Zou (2023). Jackknife model averaging for

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0057

high-dimensional quantile regression. *Biometrics* **79**, 178–189.

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

Wisnowski, J. W., J. R. Simpson, D. C. Montgomery, and G. C. Runger (2003). Resampling methods for variable selection in robust regression. *Computational Statistics and Data Analysis* **43**, 341–355.

Wu, W. B. (2007). M-estimation of linear models with dependent errors. *Annals of Statistics* **35**, 495–521.

Yang, Y. H. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association* **96**, 574–588.

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

Zhang, X. Y., G. H. Zou, and R. Carroll (2015). Model averaging based on Kullback-Leibler distance. *Statistica Sinica* **25**, 1583–1598.

Zhu, R., A. T. K. Wan, X. Y. Zhang, and G. H. Zou (2019). A Mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association* **114**, 882–892.