Computational Statistics and Data Analysis 178 (2023) 107598

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Visualization and assessment of model selection uncertainty

Yichen Qin [a], Linna Wang [b], Yang Li [c], Rong Li [c,*]

[a] Department of Operations, Business Analytics, and Information Systems, University of Cincinnati, Cincinnati, OH, USA
[b] Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH, USA
[c] School of Statistics and Center for Applied Statistics, Renmin University of China, Beijing, China

ARTICLE INFO

Article history:
Received 13 March 2021
Received in revised form 2 October 2021
Accepted 6 October 2021
Available online 27 August 2022

Keywords:
Bootstrap
Model selection deviation
Distribution of the selected model

ABSTRACT

Although model selection is ubiquitous in scientific discovery, the stability and uncertainty of the selected model is often hard to evaluate. How to characterize the random behavior of the model selection procedure is the key to understand and quantify the model selection uncertainty. To this goal, initially several graphical tools are proposed. These include the G-plots and H-plots, to visualize the distribution of the selected model. Then the concept of model selection deviation to quantify the model selection uncertainty is introduced. Similar to the standard error of an estimator, model selection deviation measures the stability of the selected model given by a model selection procedure. For such a measure, a bootstrap estimation procedure is discussed and its desirable performance is demonstrated through simulation studies and real data analysis.

1. Introduction

Model selection plays an important role in modern scientific discoveries. There has been exciting work on developing penalized model selection methods for linear regressions, such as Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), MCP (Zhang et al., 2010), and others. With many methods at our disposal, it is crucial to understand their stability and uncertainty before committing to one or a few methods. No matter which selection method is applied, selection uncertainty is a ubiquitous issue and has complex impact on the subsequent inference. Therefore, in this article, we propose a few graphical and numerical tools to understand and quantify the model selection uncertainty.

Similar to a point estimate of the population parameter in the classical statistics framework, the selected model given by a selection method can be viewed as a random "point estimate" of the true model or optimal model. To study its random behavior, it is natural to investigate its distribution, i.e., the distribution of the selected model or model selection distribution. Unfortunately, such a distribution is complex and sometimes mathematically intractable. Many classical tools for parameter estimation are no longer applicable in the model selection setting. This is partially because the support of such a distribution, i.e., all possible models, is discrete. Also, these models cannot be ordered or compared easily due to their complex relationship. In addition, the size of the support grows exponentially as the data dimension increases. For example, with p covariates, there can be 2^p possible regression models.

Such a distribution, if available, would give analysts a comprehensive understanding of the random behavior of the selected model. Among all the aspects of the random behavior, the stability is one of the most important ones because it measures the selection uncertainty and trustworthiness of the selected model. The issue of model selection uncertainty

[*] Corresponding author.
E-mail address: lr130911@163.com (R. Li).

https://doi.org/10.1016/j.csda.2022.107598
0167-9473/© 2022 Published by Elsevier B.V.

is two folds. First, given different samples drawn from a common population, the same selection method may identify different models. Second, different selection methods, when applying to the same data set, will result in different models. Although many methods claim to have achieved the optimal performance under specific settings, their selection results are often quite different.

Model selection uncertainty has always been an active area of research (Chatfield, 1995). Hansen et al. (2011) propose model confidence set (MCS) to yield information about the uncertainty surrounding the model selection, which has been frequently used to measure the estimation uncertainty (Bayer, 2018; Seri et al., 2020). Nan and Yang (2014) propose the variable selection deviation measure to evaluate the reliability and trustworthiness of the selected model based on a model averaging approach (Yang and Yang, 2017; Ye et al., 2018). To derive a suitable model in multivariate regression models, Sauerbrei et al. (2015) adopt bootstrap resampling to assess variable selection stability. Hennig and Sauerbrei (2019) propose a measure to explore variable selection instability by analyzing dissimilarities among the results from different bootstrap samples. Yu et al. (2022) propose to estimate F- and G-measures to compare different variable selection methods. Ferrari and Yang (2015) propose to construct the variable selection confidence set by a sequence of F-tests for linear regression model and likelihood ratio testings for generalized linear model (Zheng et al., 2019a,b). To increase the interpretability of the confidence set, Li et al. (2019b) propose model confidence bounds, which contain a pair of nested models that trap the true model with a certain probability. Wang et al. (2021) further extend the confidence bounds to graphical models and introduce some visualization tools. Alternatively, Liu et al. (2020) propose two simple measures of uncertainty for a model selection procedure. One is similar in spirit to confidence set and the other is focusing on the error in model selection. The aforementioned methods all focus on measuring the model selection uncertainty. Meanwhile, to reduce selection uncertainty and provide a better selection result, Meinshausen and Bühlmann (2010) propose stability selection to improve upon existing selection methods using a subsampling approach. Lim and Yu (2016) propose a model-free criterion for selecting the tuning parameter based on a new estimation stability metric. For a comprehensive review on model selection, please see Ding et al. (2018).

On the other hand, the concept of the distribution of the selected model, which is a broader topic containing selection uncertainty, is relatively less studied and has only started to gain more attention. Knight and Fu (2000) first provide the asymptotic distribution of the Lasso estimator in the low-dimensional setting. The distribution of parameter estimates from Lasso, SCAD, and thresholding are further investigated in Pötscher and Leeb (2009) for finite sample and large sample limit. Zhou (2014) propose Monte Carlo simulation based approach to estimate such distributions. Finally, Ewald et al. (2020) completely characterize the distribution of the Lasso estimator in finite samples for linear regressions and study the model selection properties of the Lasso. These existing works mostly focus on the distribution of the parameter estimate as opposed to the selected model. Although some theoretical results have been established for the distribution of selected model by Lasso, much less work is done for visualizing such a distribution, which is difficult but also useful in practice. To fill this gap in the literature, we propose to visualize the distribution of the selected model in this article and use the visualization to measure selection uncertainty.

In this article, we introduce new visualization tools for the distribution of the selected model. By grouping models of a similar structure together, we are able to visualize the distribution more efficiently and clearly, and reveal important patterns in the distribution that are not available through other types of analysis. The proposed visualization is useful in graphical comparison of different selection methods, giving analysts a good sense of level of randomness each method comes with. With the help of the proposed visualization, we further introduce the concept of model selection deviation (MSD) which can be considered as the standard deviation of the distribution of the selected model. Such a measure allows numerical comparison of various model selection methods in terms of their stability. Under appropriate transformation, we further develop a fast bootstrap estimation procedure for model selection deviation and demonstrate its desirable performance in simulation and real data analysis. Throughout the article, we have focused on linear regressions. However, we would like to point out that the methodology developed in this article can be extended to more complex settings, such as generalized linear models and graphical models, with minimum modifications.

Note that, under a consistent model selection procedure, the probability that the selected model equals to the true model converges to one. Therefore, the model selection uncertainty vanishes asymptotically given a fixed number of covariates. However, under the finite sample size, the model selection uncertainty is nonnegligible. Most of the analysis presented in this article focus on the moderate sample size.

This article contributes to the literature in the following aspects. To the best of our knowledge, this article presents the first attempt in the literature to visualize the distribution of selected model. Using the proposed visualization techniques, the random behavior of the different model selection procedures can be compared and studied. Such visualizations allow us to define various attributes of the distribution, such as the mode and the skewness, characterizing various aspects of the distribution. One of the most important numeric attributes is our model selection deviation, which is an extension of the traditional standard deviation of a univariate distribution to the distribution of the selected model. Therefore, the tools provided in this article allow analysts to both quantitatively and graphically compare various selection procedures in terms of their stability and other aspects.

The rest of the paper is organized as follows. In Section 2, we introduce the framework and notations. In Section 3, we propose several new graphical tools to visualize the distribution of the selected model and discuss their connections and distinctions. Based on these visualizations, in Section 4, we introduce a new numeric measure, model selection deviation, to quantify the model selection uncertainty, and further discuss a bootstrap estimation procedure. Lastly, we demonstrate the desirable performance of the proposed visualization and uncertainty measure in simulation in Section 5 and in real data analysis in Section 6. We conclude in Section 7 and provide additional results in the supplementary materials.

2. Preliminaries

In this article, we focus on linear regression models. Let $\mathbf{Y} = (y_1, ..., y_n)^T$ be an $n \times 1$ response vector. Suppose $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$ is an $n \times p$ design matrix of p predictors and $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T \in \mathbb{R}^p$. Without loss of generality, we assume $\mathbf{X}$ are column-wise standardized with zero means and unit variances. Let $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$ be the parameter vector and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. We further assume some elements in $\boldsymbol{\beta}$ are zeros, but we do not know which ones. Let $\boldsymbol{\beta}^0$ be the true coefficients and $m_0 = \{j : \beta_j^0 \neq 0, j = 1, ..., p\}$ represent the true model. For any model m, we denote its model complexity as $|m|$, i.e., the number of variables. Let $m_{\text{full}} = \{1, ..., p\}$ represent the full model and let $\mathcal{M} = \left\{m : \emptyset \subseteq m \subseteq m_{\text{full}}\right\}$ denote the set of all of the possible models. Let $\widehat{m}$ represent a model selected by a model selection procedure, e.g., Lasso, and $\emptyset \subseteq \widehat{m} \subseteq \{1, ..., p\}$. Under the data generating process specified above, we further define the most frequently selected model as the mode model, denoted as m^*,

$$m^* = \arg\max_{m \in \mathcal{M}} \mathbb{P}(\widehat{m} = m).$$

Asymptotically, with probability tending to one, the mode model from a consistent model selection becomes the true model. Under the finite sample size, the mode model may or may not be the true model. However, the mode model still gives us insightful information about the behavior of the selected model because it is the mostly likely outcome from the model selection procedure. In this article, the mode model plays an important role in the visualization and uncertainty assessment, because it is a measure of central tendency of the selected model.

In this article, we mostly focus on penalized regression methods for obtaining $\widehat{m}$. Specifically, $\widehat{m} = \{j : \widehat{\beta}_j \neq 0, 1 \leq j \leq p\}$ where the estimator $\widehat{\boldsymbol{\beta}}$ minimizes a penalized likelihood criterion with the form $C(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/(2n) + \sum_{j=1}^p J_\lambda(\beta_j)$. Here $\lambda \geq 0$ is a user-defined regularization parameter and $J_\lambda(\cdot)$ is a penalty function $J_\lambda : \mathbb{R}^p \mapsto \mathbb{R}^+$. For example, $J_\lambda(\beta_j) = \lambda|\beta_j|$ gives the L_1-norm, which corresponds to Lasso. When $J_\lambda(\beta_j) = \lambda\omega_j|\beta_j|$ where ω_j represents the weight, it corresponds to adaptive Lasso (Zou, 2006). Other choices of penalty functions include SCAD, MCP (Zhang et al., 2010), and elastice net (Zou and Hastie, 2005). Note that $|\cdot|$ represents the absolute value when applied to a scalar and the cardinality when applied to a set.

3. Model selection visualization

The selected model from a model selection procedure can be considered as a random "point estimate" for the true model. Therefore, it is important to understand its random behavior through its distribution, i.e., the distribution of the selected model (or model selection distribution). However, such a distribution is often complex and mathematically intractable in many settings. As the first step, we introduce several graphical tools to visualize such a distribution. Note that these visualizations are applicable to not only the penalized regression methods such as Lasso-type estimator, but also stepwise regressions, best subset regressions, and etc. For simplicity, we use Lasso as a running example.

3.1. Visualization of distribution of selected model by model groups

The distribution of the selected model is generally hard to visualize, because the support of the distribution is on all possible models, and these models have complex relationships among themselves. We first present a naive visualization of such a distribution to show its difficulty. We simulate $M = 100000$ data sets with $n = 300$ and $p = 10$ according to $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{I}_{p \times p})$, $\epsilon_i \sim N(0, 5^2)$, and $\boldsymbol{\beta}^0 = (3, 2.5, 2, 1.5, 1, 0, 0, 0, 0, 0)$. We obtain the probability of each possible model being selected by Lasso with $\lambda = 0.7$, i.e., model frequency, and visualize the distribution of the selected model in Fig. 1.

Each circle represents a unique model selected by Lasso with a positive probability. The color in the circle represents the model frequency. Note that we have $2^p = 1024$ possible models but only plot the ones with positive model frequencies. The vertical axis shows the model complexity. The models in the same row are arranged according to their model frequencies descendingly. There exists a line connecting two models if the large model m_2 includes the small model m_1 with one extra variable, i.e., $m_2 \supset m_1$ and $|m_2 \setminus m_1| = 1$.

Obviously, the figure is hard to read, indicating the complex nature of the distribution of the selected model. Many models in the figure are connected, meaning their structures are similar, but we cannot easily understand their relationships. Hence, such a figure often overwhelms analysts.

In order to focus on the important patterns in the distribution, we propose to visualize the distribution of the selected model by groups and call it G-plot. Under the same setting as before, we present its G-plot in Fig. 2. The motivation is that, since the model space is too large, we put models with similar structures into groups and visualize the group frequency, i.e., sum of model frequencies in the group. In the figure, each model group is represented by a circle while the group frequency is represented by the color intensity. The groups are placed in the figure according to their model complexities and their structural relationships. The vertical axis represents the model complexity.

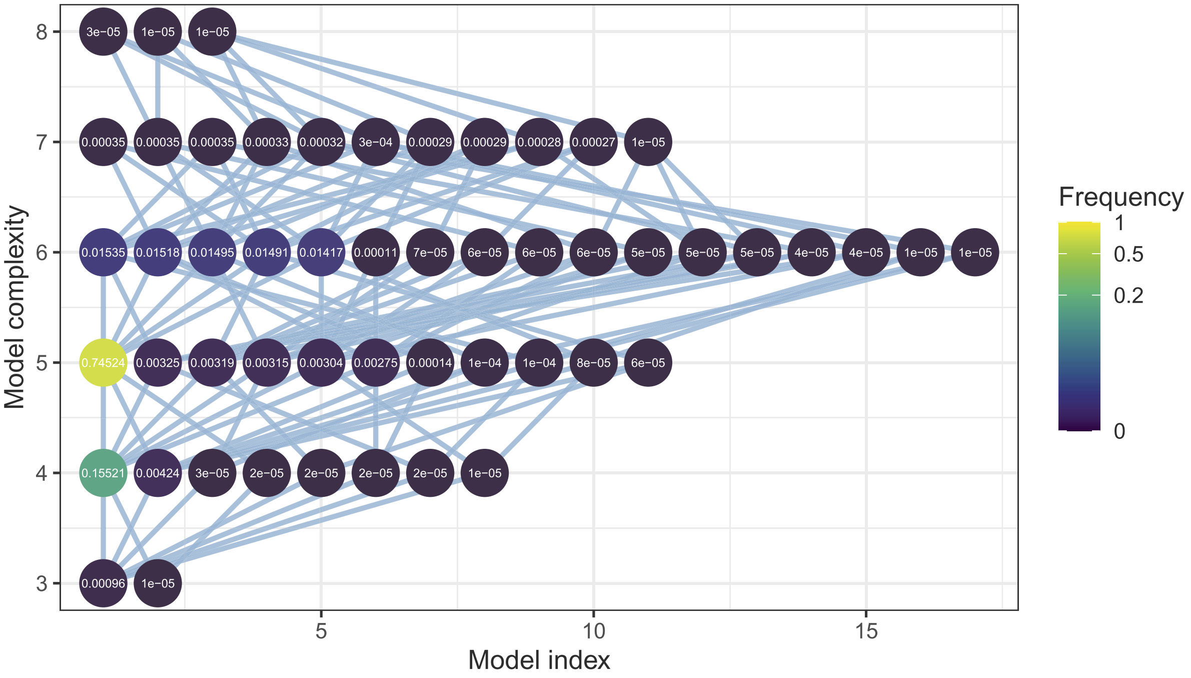


Fig. 1. This is a naive visualization of the distribution of selected model from Lasso with $\lambda = 0.7$. The data is simulated from a linear model with $p = 10, n = 300, \boldsymbol{\beta} = (3, 2.5, 2, 1.5, 1, 0, 0, 0, 0, 0)$. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

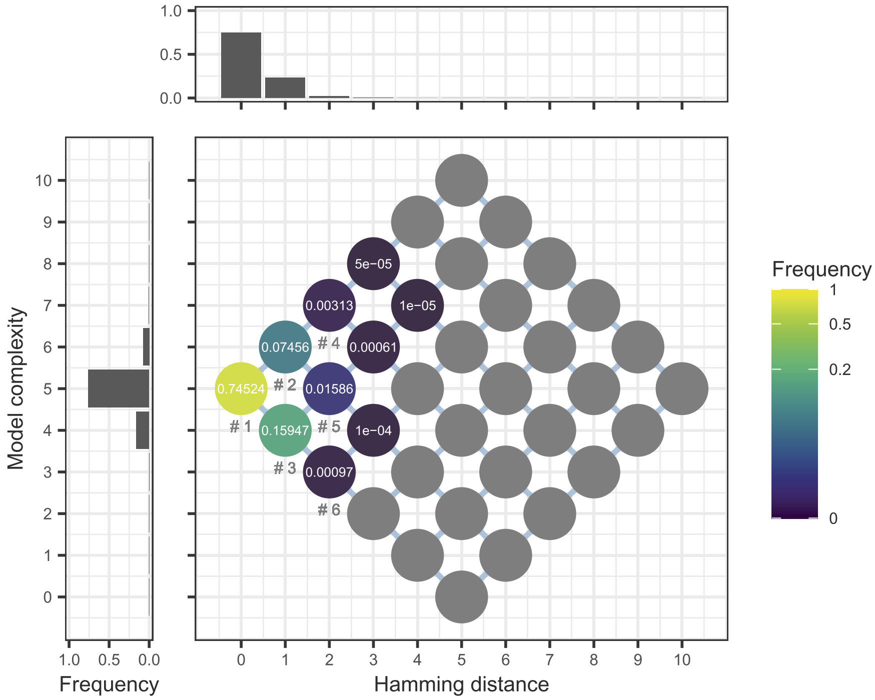


Fig. 2. This is an example of the G-plot. We use Lasso method to select the model with tuning parameter $\lambda = 0.7$ under the setting $p = 10, n = 300, \boldsymbol{\beta} = (3, 2.5, 2, 1.5, 1, 0, 0, 0, 0, 0)$, the mode model complexity is 5.

More specifically, Table 1 presents the composition of the groups. Group #1 contains only the mode model m^* and is placed on the xy-coordinates of $(0, |m^*|)$ or $(0, 5)$ as an “anchor”. Other models are grouped according to their model complexities and their Hamming distances to the mode model. The Hamming distance between two models m_1 and m_2 is defined as

Table 1
The composition of model groups of G-plot shown in Fig. 2. The numbers in variable list: 1 - variable included; 0 - variable excluded.

Group #	Group Cardinality	Variable list of models in the group										Model Complexity	Hamming Distance	Model Frequency	Group Frequency
		1	2	3	4	5	6	7	8	9	10				
1	1	1	1	1	1	1	0	0	0	0	0	5	0	0.746	0.746
2	5	1	1	1	1	1	1	0	0	0	0	6	1	0.015	0.075
		1	1	1	1	1	0	1	0	0	0			0.015	
		1	1	1	1	1	0	0	1	0	0			0.015	
		1	1	1	1	1	0	0	0	1	0			0.014	
		1	1	1	1	1	0	0	0	0	1			0.015	
3	5	1	1	1	1	0	0	0	0	0	0	4	1	0.155	0.159
		1	1	1	0	1	0	0	0	0	0			0.004	
		1	1	0	1	1	0	0	0	0	0			0.000	
		1	0	1	1	1	0	0	0	0	0			0.000	
		0	1	1	1	1	0	0	0	0	0			0.000	
4	10	1	1	1	1	1	1	1	0	0	0	7	2	0.000	0.003
		1	1	1	1	1	1	0	1	0	0			0.000	
							...							...	
		1	1	1	1	1	0	0	1	0	1			0.000	
		1	1	1	1	1	0	0	0	1	1			0.000	
5	25	1	1	1	1	0	1	0	0	0	0	5	2	0.003	0.019
		1	1	1	1	0	0	1	0	0	0			0.003	
							...							...	
		0	1	1	1	1	0	0	0	1	0			0.000	
		0	1	1	1	1	0	0	0	0	1			0.000	
6	10	1	1	1	0	0	0	0	0	0	0	3	2	0.001	0.001
		1	1	0	1	0	0	0	0	0	0			0.000	
							...							...	
		0	1	0	1	1	0	0	0	0	0			0.000	
		0	0	1	1	1	0	0	0	0	0			0.000	
...	...						...					...	...	...	...

$$H(m_1\|m_2) = |(m_1 \setminus m_2) \cup (m_2 \setminus m_1)|,$$

which also represents the number of different variables or the cardinality of the symmetric difference. So the models in the same group have the same complexity and similar structure. Conceptually, each group of models can be expressed as $\{m : H(m^*\|m) = i, |m| = j\}$ for different i and j, and the group is placed at xy-coordinates of (i, j). The left histogram displays the frequency of model complexity, while the top histogram displays the frequency of Hamming distance.

Based on this criterion, group #2 consists of all the models that contain the mode model and also have one extra variable. Group #2 is placed at xy-coordinates of $(1, 6)$ because its complexity is 6 and its members all have Hamming distances of 1 to the mode model. Similarly, group #3 consists of all sub-models of mode model with one less variable. Group #4 and group #6 are defined in a similar way. Group #5 is defined as all the models that miss one variable from the mode model but have one extra variable, so they have the same complexity as the mode model. The rest of groups are defined in the same fashion and numbered sequentially. Alternatively, every model in group #5 can be considered as a sub-model of at lease one model in group #2 or a super-model of at least one model in group #3.

Table 1 also reports model frequency, group frequency, model complexity, the Hamming distance to mode model, model complexity, and group cardinality (i.e., the number of all possible models in the group). In Fig. 2, within each circle, the white number is the group frequency. When two circles are connected by a line, it means every model in one group has at least one supermodel (or submodel) in the other group with exactly one more (or less) variable. The line indicates that the two groups are, roughly speaking, one Hamming distance away. Therefore, for each model in each group, its Hamming distance to the mode model is the number of edges on the shortest path between the model at hand and the mode model.

In the figure, we can see that the mode model visualizes the most likely model. If Lasso misses the mode model, it is highly likely that it will select a model with one less variable than the mode model (i.e., group #3), and relatively less likely to add a variable (i.e., group #2). However, it is rare that Lasso will miss two variables (i.e., group #4, #5, #6), and even if it does, Lasso favors smaller models (i.e., group #6) than larger models (i.e., group #4). Therefore, the proposed visualization allows us to see patterns which are hard to find in the naive visualization.

Obviously, the mode model plays an important role in the visualization because it indicates the “center” of the distribution. More importantly, the mode model functions as an anchor because everything else in our visualization, such as the groups and Hamming distance, is relative to this anchor. Among the popular choices of central tendency measure, such as mean, median, and mode, it is natural to use the mode in our setting because neither mean or median is clearly defined for models. In practice, the mode model is never known, so we could replace it with its estimate, e.g., bootstrap mode model. Estimating such a distribution is also an important task and we leave this topic as a future research direction.

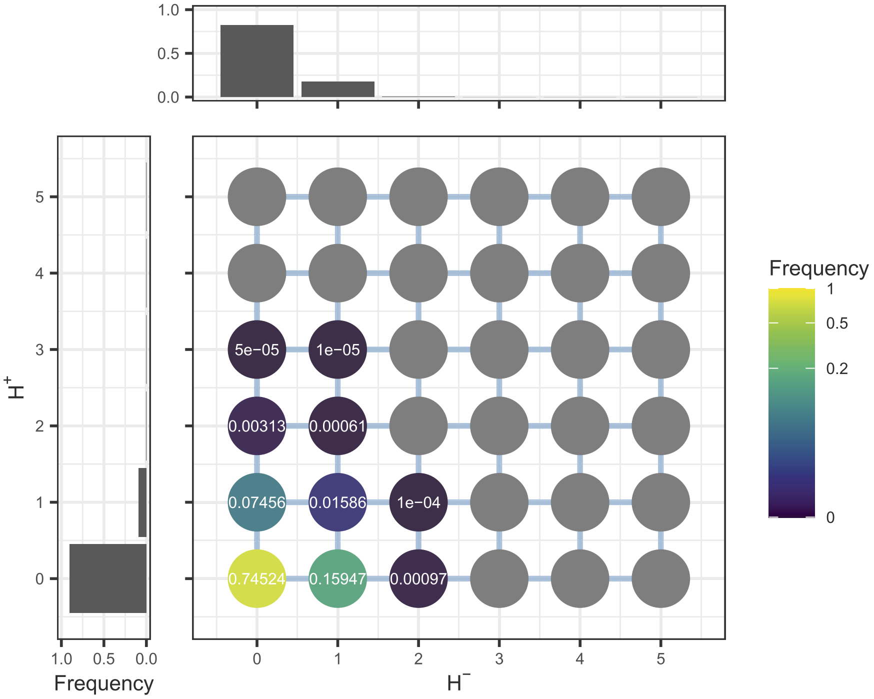


Fig. 3. This is an example of the H-plot. We use Lasso method to select the model with tuning parameter $\lambda = 0.7$ under the setting $p = 10, n = 300, \boldsymbol{\beta} = (3, 2.5, 2, 1.5, 1, 0, 0, 0, 0, 0)$, the complexity of mode model is 5.

3.2. Visualization of distribution of selected model by decomposed Hamming distance

Essentially, the G-plot divides the models into groups according to the model complexity and the Hamming distance. In this section, we further focus on the Hamming Distance. Note that the Hamming distance between the mode model m^* and any arbitrary model m can be decomposed into two parts, H^- and H^+, that is,

$$H^-(m^*\|m) = |m^*\backslash m|, \quad H^+(m^*\|m) = |m\backslash m^*|, \quad \text{and} \quad H(m^*\|m) = H^-(m^*\|m) + H^+(m^*\|m).$$

Here H^- represents the number of missing variables by m compared to the mode model. Meanwhile, H^+ represents the number of redundant variables in m compared to the mode model. In total, $H = H^- + H^+$ represents the number of different variables between m and mode model.

Therefore, we can form the model groups according to H^- and H^+. Specifically, each model group can be expressed as $\{m : H^-(m^*\|m) = i, H^+(m^*\|m) = j\}$ for different i and j. As an alternative to the G-plot, we can plot each group at xy-coordinates of (i, j), and we call this new visualization H-plot. Under the same setting as before, we present an example of the H-plot in Fig. 3. Each circle and its color represent the model group and its group frequency. The groups are placed in the figure according to H^- (horizontal axis) and H^+ (vertical axis). The mode model is on the xy-coordinates of $(0, 0)$ since $H^-(m^*\|m^*) = H^+(m^*\|m^*) = 0$. The left histogram displays the distribution of H^+, while the top histogram displays the distribution of H^-.

For the groups defined in the H-plot, we have the following result.

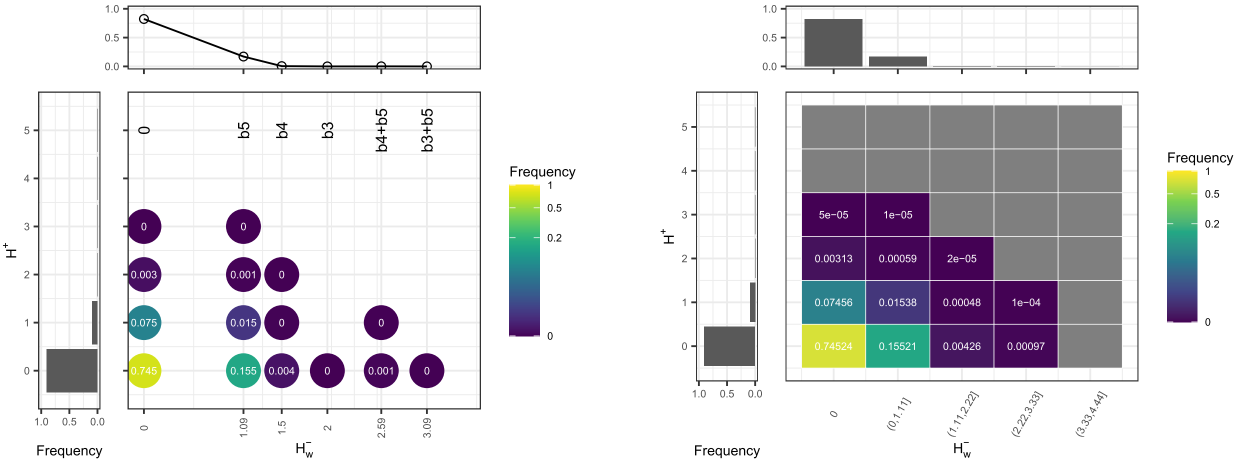
Proposition 1. *Each model group in the G-plot at the xy-coordinates of (i, j) is equal to one model group in the H-plot at the xy-coordinates of $((i - j + |m^*|)/2, (i + j - |m^*|)/2)$, that is,*

$$\left\{m : H(m^*\|m) = i, |m| = j\right\} = \left\{m : H^-(m^*\|m) = (i - j + |m^*|)/2, H^+(m^*\|m) = \max\{0, (i + j - |m^*|)/2\}\right\},$$

where the left hand side represents the group in the G-plot and the right hand side represents the group in the H-plot.

Therefore, interestingly, the model groups defined by H^- and H^+ in the H-plot are the same as the groups defined in the G-plot. In fact, the H-plot can be obtained by simply rotating the G-plot counterclockwise by 45 degrees (and recentering). Therefore, the group frequencies in both the H- ang G-plots are the same. The axes in the G-plot are Hamming distance and model complexity, while the axes in the H-plot are H^- and H^+.

Although the G- and H-plots are similar, their emphases are different. The G-plot shows the pattern of the entire distribution in the scale of model complexity. The H-plot highlights the difference between the selected model and the mode



(a) Scatter plot

(b) Heatmap

Fig. 4. These are the examples of weighted Hamming distance based scatter plot and heatmap. We use Lasso method to select the model with $\lambda = 0.7$ under the setting $p = 10, n = 300, \boldsymbol{\beta} = (3, 2.5, 2, 1.5, 1, 0, 0, 0, 0, 0)$, the mode model complexity is 5.

model such as missed variables or redundant variables. The histograms in both plots can intuitively display their different focuses. The two visualizations complement each other in describing the distribution of the selected model. So one should look at both to gain a comprehensive picture of the model selection randomness.

Using the proposed H-plot, the model selection uncertainty can be easily seen and assessed. When all the selected models are closer to the mode model, the uncertainty is low, and vice versa. Not only can we visualize the selection uncertainty, we can also visualize in what directions the uncertainty is most significant. For example, does a selection method tend to over-select variables or to omit variables? These can all be answered by looking at the x- and y-axes of the H-plot, i.e., the two histograms. Meanwhile, how do these two sources of selection errors interact with each other? This can be answered by looking at the circles and their colors in the H-plot, which is essentially the joint distribution of these two errors. In Section 4, we will come back to the G-plot and the H-plot and discuss some numeric measures of the model selection uncertainty.

3.3. Visualization of distribution of selected model by scatterplot and heatmap

The H-plot focuses on the difference between the selected and the mode model, in terms of missing variables and redundant variables. However, in the case of missing variables, the magnitude of the missing variable's coefficient is also important because it indicates the negative impact of missing such a variable. To take this information into consideration, we define the weighted Hamming distance as $H_w^-(m^*\|m) = \sum_{j\in\{m^*\setminus m\}} |\beta_j^0|$ where β_j^0 is the j-th element in $\boldsymbol{\beta}^0$. Similarly, we define $H_w^+(m^*\|m) = \sum_{j\in\{m\setminus m^*\}} |\widehat{\beta}_j^m|$, where $\widehat{\beta}_j^m$ is the j-th coefficient estimate from the model m in which the j-th variable is selected. Note that the distinction between H_w^- and H_w^+ is mainly in two aspects. First, H_w^- measures the cost of missing variables, and the loss of missing a variable with a large coefficient is more serious than missing a variable with a small coefficient. Whereas the size of H_w^+ has less explanatory because there is no difference between the importance of redundant variables, and the estimated coefficients of the redundant variables are usually small and indistinguishable compared with the number of redundant variables H^+. Second, H_w^- depends on three quantities: the true mode model, true coefficient, and the selected model, among which the selected model is the only random quantity. In contrast, H_w^+ depends on another three quantities: the true mode model, the selected model, and the estimated coefficients under the selected model, among which both the selected model and its estimated coefficients are random. Thus, H_w^- reflects the model selection uncertainty, while H_w^+ reflects the combination of the model selection uncertainty and the parameter estimation uncertainty, which is beyond the scope of this paper. This is why we do not show the H_w^+ in the following figure, but only define it in the sense of completeness. Based on weighted Hamming distance, we further propose another two types of visualizations, namely weighted Hamming distance based scatter plot and weighted Hamming distance based heatmap. We show examples of these two visualizations in Fig. 4.

We propose to generate the scatter plot as follows. We group all the selected models according to their $H_w^-(m^*\|m)$ and $H^+(m^*\|m)$, and plot the model group on the xy-coordinates at $(H_w^-(m^*\|m), H^+(m^*\|m))$. The color indicates the group frequency. Therefore, the x-axis of the scatter plot is the weighted Hamming distance H_w^- to the mode model, while the y-axis remains as H^+. Obviously, the mode model is at the xy-coordinates of $(0, 0)$. Such a scatter plot allows us to further understand the details of the distribution. For example, when a selected model misses a variable from the mode model, does the variable tend to have a large or small coefficient?

Lastly, we propose the heatmap as an alternative to the scatter plot. As the data dimension increases, the number of unique models increases exponentially. As a result, the scatter plot would become more difficult to read as many model groups overlap with each other. Therefore, we propose to divide the x- and y-axes of the scatter plot into equal spaced intervals and convert the scatter plot into the heatmap (or 2d histogram). The color of each rectangle in the heatmap represents the sum of the frequencies of models whose H_w^- and H^+ fall into the corresponding intervals.

In this section, we have proposed various visualizations, such as the G-plot, H-plot, scatterplot, and heatmap, for the distribution of the selected model. We will propose numeric measures/attributes using these visualizations to further quantify model selection uncertainty in the next section.

4. Model selection uncertainty assessment

Based on the visualization proposed above, we now introduce a new measure of model selection uncertainty. This section focuses on penalized regression methods such as Lasso-type estimator. We use Lasso as a running example.

4.1. Model selection deviation

Similar to the standard error of an estimator, we propose a new uncertainty measure for model selection procedures, namely, model selection deviation (MSD). Such a measure can be considered as the "standard deviation" of the distribution of the selected model. The smaller the measure, the lower the selection uncertainty. The motivation of this measure is to use the expected Hamming distance to the mode model to evaluate the selection stability. Note that the mode model represents the "center" of the distribution. Formally, we have

$$\mathrm{MSD}^- = \mathbb{E}|m^* \setminus \widehat{m}| = \mathbb{E}H^-(m^*\|\widehat{m}), \quad \mathrm{MSD}^+ = \mathbb{E}|\widehat{m} \setminus m^*| = \mathbb{E}H^+(m^*\|\widehat{m}), \quad \text{and} \quad \mathrm{MSD} = \mathrm{MSD}^- + \mathrm{MSD}^+.$$

Conceptually, MSD^- is the expected number of variables in the mode model that are missed by the selected model, while MSD^+ is the expected number of extra variables in the selected model compared to the mode model. Therefore, MSD is the expected number of variables by which the selected model differs from the mode model. These definitions are consistent with the H-plot and G-plot. The MSD^- is essentially the expectation from the top histogram in the H-plot, MSD^+ is the expectation from the left histogram in the H-plot, and MSD is the expectation from the top histogram in the G-plot.

Lastly, we can define the skewness of the distribution as $\mathrm{Skewness} = (P^+ - P^-)/(P^+ + P^-)$ where $P^+ = \sum_{m:|m|>|m^*|} \mathbb{P}(\widehat{m} = m)$ and $P^- = \sum_{m:|m|<|m^*|} \mathbb{P}(\widehat{m} = m)$. Here P^- (and P^+) represents the probability that a selected model has a lower (higher, respectively) complexity than the mode model.

The mode model plays an important role since both the model selection deviations and skewness depend on it. As an example, we plot the mode model as a function of the tuning parameter λ of Lasso in Fig. 5a under the setting in Section 3.1. Each bar represents whether the variable is selected or not in the mode model. Since β_1 has the largest value, the bar of X_1 is the longest, meaning the variable appears in the mode model as long as $\lambda < 2.9$. Generally, the mode model changes from the full model to the empty model as λ increases.

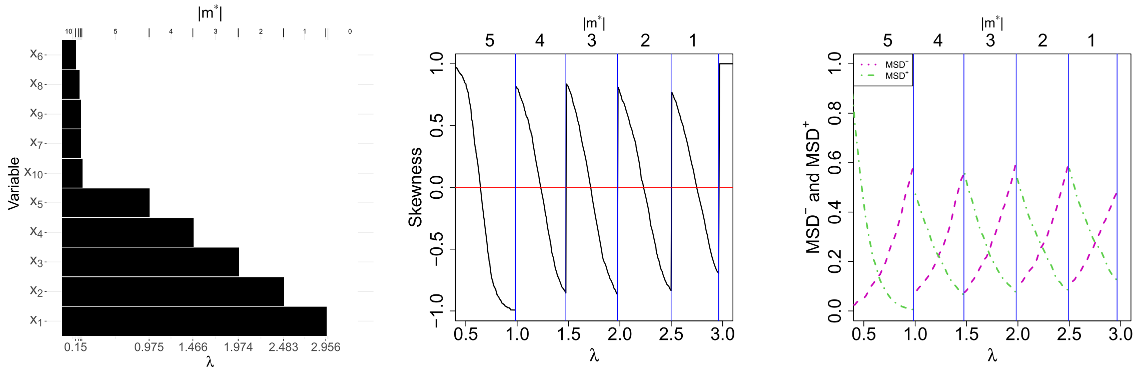
We further plot the skewness and the model selection deviations as functions of λ in Fig. 5b and 5c under the same setting. The solid black, dash purple and dash green curves represent skewness, MSD^- and MSD^+, respectively. The bottom x-axis displays λ changing from 0.5 to 3, while the top x-axis represents the mode model complexity changing from 5 to 1. The vertical lines separate the figures into regions within which the mode model complexity is the same. Within each complexity, the skewness decreases from positive to negative, MSD^- increases, and MSD^+ decreases. There also exits periodicity in skewness, MSD^-, and MSD^+ across different complexities. This is because that, within each complexity, the distribution of the selected model changes from skew to the large complexity to skew to the small complexity as λ increases. If λ continues to increase, the mode model complexity decreases by 1, causing the distribution to become skew to large complexity again. Similar patterns can be seen in the G-plots in the supplementary materials.

From now on,we consider the mode model and its complexity, the model selection deviations, and the skewness as functions of λ, and denote them as $m^*(\lambda)$, $|m^*(\lambda)|$, $\mathrm{MSD}^-(\lambda)$, $\mathrm{MSD}^+(\lambda)$, and $\mathrm{Skewness}(\lambda)$, respectively. We use the functions of MSDs as measures of overall model selection uncertainty.

We further define the weighted model selection deviation WMSD^- and WMSD^+ to take the coefficient magnitude and variable importance into consideration

$$\mathrm{WMSD}^- = \mathbb{E}\sum_{j\in\{m^*\setminus\widehat{m}\}} |\beta_j^0| = \mathbb{E}H_w^-, \quad \text{and} \quad \mathrm{WMSD}^+ = \mathbb{E}\sum_{j\in\{\widehat{m}\setminus m^*\}} |\widehat{\beta}_j^m| = \mathbb{E}H_w^+.$$

They use the true coefficients of the missing variables and the estimated coefficients of the redundant variables as the weights, respectively. WMSD^- and WMSD^+ measure different uncertainty because of the subtle distinction between H_w^- and H_w^+. WMSD^- measures the model selection uncertainty, whereas WMSD^+ measures both the model selection uncertainty and a little bit of the parameter estimation uncertainty and is an integrated measure. This also makes WMSD^+ more complex and harder to estimate.



(a) Mode model (b) Skewness (c) MSD^{-} and MSD^{+}

Fig. 5. Illustration of how the mode model, skewness, MSD^{-} and MSD^{+} change as λ increases. We use Lasso to select the model under the setting $p = 10, n = 300, \boldsymbol{\beta} = (3, 2.5, 2, 1.5, 1, 0, 0, 0, 0, 0)$.

Algorithm 1: Bootstrap estimation of MSD^{-} as a function of tuning parameter λ.

```
Input: $\mathbf{X}$, $\mathbf{Y}$.
Output: $\widehat{\text{MSD}}^{-}(\lambda)$
foreach $b \in \{1, \dots, B\}$ do
    Obtain bootstrap sample $(\mathbf{X}^{(b)}, \mathbf{Y}^{(b)})$ via pairwise bootstrap.
    Apply a model selection method on the bootstrap sample to obtain the bootstrap model as a function of the tuning parameter, $\widehat{m}^{(b)}(\lambda)$, for $\lambda \in (0, \lambda_{\max}]$.
foreach $\lambda$ in a grid on $[0, \lambda_{\max}]$ do
    Find the bootstrap mode model $\widehat{m}^{*}(\lambda) = \arg\max_{m \in \mathcal{M}} \sum_{b=1}^{B} \mathbf{1}\{m = \widehat{m}^{(b)}(\lambda)\}$.
    Obtain $\widehat{\text{MSD}}^{-}(\lambda) = \sum_{b=1}^{B} |\widehat{m}^{*}(\lambda) \setminus \widehat{m}^{(b)}(\lambda)|/B$.
```

4.2. Model selection deviation estimation

Although the model selection deviation offers us a measure of uncertainty, it is unknown in practice and needs to be estimated. In this section, we propose to use bootstrap to estimate the model selection deviation. Given a data set, we first generate B bootstrap samples and apply a model selection method (e.g., Lasso) to each bootstrap sample to obtain bootstrap models $\widehat{m}^{(b)}$. Conceptually, we propose to estimate model selection deviation by

$$\widehat{\text{MSD}}^{-} = \frac{1}{B}\sum_{b=1}^{B} |\widehat{m}^{*} \setminus \widehat{m}^{(b)}|, \quad \widehat{\text{MSD}}^{+} = \frac{1}{B}\sum_{b=1}^{B} |\widehat{m}^{(b)} \setminus \widehat{m}^{*}|, \quad \text{and} \quad \widehat{\text{MSD}} = \widehat{\text{MSD}}^{-} + \widehat{\text{MSD}}^{+},$$

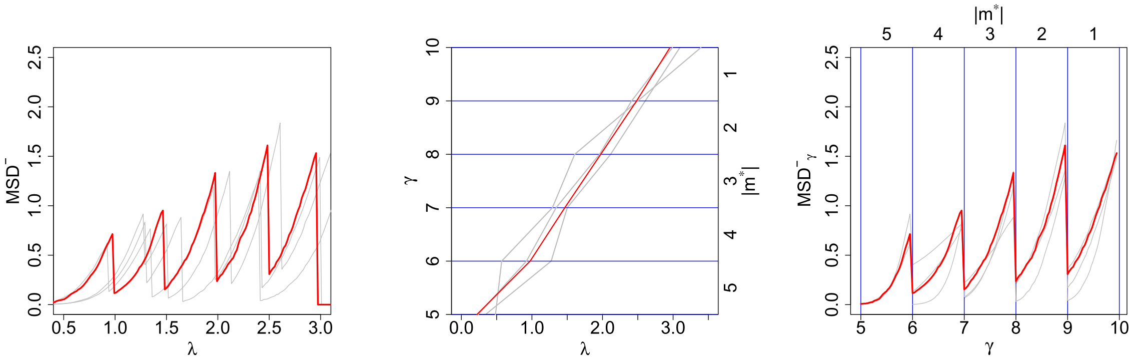
where $\widehat{m}^{*} = \arg\max_{m \in \mathcal{M}} \sum_{b=1}^{B} \mathbf{1}(m = \widehat{m}^{(b)})$ is the bootstrap mode model. Weighted model selection deviation estimation $\widehat{\text{WMSD}}^{-}$ and $\widehat{\text{WMSD}}^{+}$ can also be obtained with estimated coefficients of bootstrap mode model and bootstrap models. Denote

$$\widehat{\text{WMSD}}^{-} = \frac{1}{B}\sum_{b=1}^{B} \sum_{j \in \{\widehat{m}^{*} \setminus \widehat{m}^{(b)}\}} |\widehat{\beta}_j^{*}| \quad \text{and} \quad \widehat{\text{WMSD}}^{+} = \frac{1}{B}\sum_{b=1}^{B} \sum_{j \in \{\widehat{m}^{(b)} \setminus \widehat{m}^{*}\}} |\widehat{\beta}_j^{(b)}|,$$

where $\widehat{\beta}_j^{*}$ is the average estimate of the j-th coefficient of the bootstrap mode model and $\widehat{\beta}_j^{(b)}$ is the j-th coefficient estimate from the b-th bootstrap model $\widehat{m}^{(b)}$. Such an estimation procedure can be conducted at different levels of the tuning parameter λ. The estimation procedure for $\widehat{\text{MSD}}^{-}$ is outlined in Algorithm 1 and $\widehat{\text{MSD}}^{+}$, $\widehat{\text{WMSD}}^{-}$ as well as $\widehat{\text{WMSD}}^{+}$ can be obtained similarly.

Following Knight and Fu (2000), we have adopted pairwise bootstrap throughout the article. Another common choice is residual bootstrap (Chatterjee and Lahiri, 2011). The two bootstrap strategies perform similarly in practice, especially when the model selection uncertainty is low. The reason for using pairwise bootstrap is that we would like to avoid negative impact of the originally selected model on the bootstrap samples. Algorithm 1 can be applied to many model selection methods, such as Lasso, SCAD, and others, making it an ideal tool to evaluate and compare model selection uncertainty.

As an illustration, we simulate $M = 3$ data sets under the setting in Section 3.1, display their corresponding $\widehat{\text{MSD}}^{-}(\lambda)$ in gray using $B = 3000$, and compare them with the true $\text{MSD}^{-}(\lambda)$ in red in Fig. 6a. Obviously, $\text{MSD}^{-}(\lambda)$ and $\widehat{\text{MSD}}^{-}(\lambda)$ show



(a) $\mathrm{MSD}^-(\lambda)$ versus $\widehat{\mathrm{MSD}}^-(\lambda)$ (b) Transformation from λ to γ. (c) MSD_γ^- versus $\widehat{\mathrm{MSD}}_\gamma^-$

Fig. 6. Comparison of the true model selection deviation and estimated model selection deviation. The true is in red while the bootstrap estimate is in gray. We use Lasso under the setting $p = 10, n = 300, \boldsymbol{\beta} = (3, 2.5, 2, 1.5, 1, 0, 0, 0, 0, 0)$.

the same trend, but the estimation error seems to be significant. These curves seem to follow different periodicity, leading to the unsatisfactory estimation performance. Therefore, it seems difficult to directly use $\widehat{\mathrm{MSD}}^-(\lambda)$ to estimate $\mathrm{MSD}^-(\lambda)$. On the other hand, the trend of $\widehat{\mathrm{MSD}}^-(\lambda)$ is quite similar to that of the true $\mathrm{MSD}^-(\lambda)$ within each mode model complexity. Therefore, to improve the estimation accuracy, we propose to transform $\widehat{\mathrm{MSD}}^-(\lambda)$ and $\mathrm{MSD}^-(\lambda)$.

Note that in penalized regressions, as the tuning parameter λ increases from 0 to C (a sufficiently large positive constant), the mode model complexity $|m^*(\lambda)|$ decreases from p to $p-1$, and eventually to 0. Suppose there exists a sequence of tuning parameters $0 = \lambda^*_{p+1,p} \le \lambda^*_{p,p-1} \le \cdots \le \lambda^*_{1,0} \le \lambda^*_{0,-1} = C$ such that, if $\lambda \in [\lambda^*_{k+1,k}, \lambda^*_{k,k-1})$, then we have $|m^*(\lambda)| = k$. We refer to these intervals as λ-intervals. The widths of the λ-intervals are often drastically different, making the estimation of MSD challenging. Then, we define the transformed tuning parameter γ as a function of λ as

$$\gamma = h(\lambda) = p - k + (\lambda - \lambda^*_{k+1,k})/(\lambda^*_{k,k-1} - \lambda^*_{k+1,k}), \quad \text{if } \lambda \in [\lambda^*_{k+1,k}, \lambda^*_{k,k-1}) \text{ for } 0 \le k \le p.$$

Here $\gamma \in [0, p+1]$ can be regarded as the standardized version of λ. When $|m^*(\lambda)| = k$, we have $\gamma \in [p-k, p-k+1)$. We refer to these intervals as γ-intervals. Clearly, the widths of γ-intervals are always 1. Furthermore, the transformed $\mathrm{MSD}^-(\lambda)$ can be obtained as

$$\mathrm{MSD}_\gamma^- = \mathrm{MSD}^-(h^{-1}(\gamma)).$$

Similarly, we transform $\widehat{\mathrm{MSD}}^-(\lambda)$ using the bootstrap version of h, denoted as $\widehat{h}$. Specifically, we obtain $\widehat{h}$ by replacing the $\lambda^*_{k,k-1}$ in h with its bootstrap version $\widehat{\lambda}^*_{k,k-1}$. Note that $\widehat{\lambda}^*_{k,k-1}$ is essentially the tuning parameter at which the complexity of the bootstrap mode model changes from k to $k-1$. Finally, $\widehat{\mathrm{MSD}}^-(\lambda)$ can be transformed as

$$\widehat{\mathrm{MSD}}_\gamma^- = \widehat{\mathrm{MSD}}^-(\widehat{h}^{-1}(\gamma)).$$

The same transformation can be applied to obtain MSD_γ^+ and $\widehat{\mathrm{MSD}}_\gamma^+$. Note that, in practice, the complexity of either the true mode model or the bootstrap mode model may skip certain values as the tuning parameter changes. In that case, we can just skip the model selection deviation under that particular mode model complexity.

For the examples in Fig. 6a, we present their transformed MSDs in Fig. 6c, where the bottom x-axis represents the transformed tuning parameter γ and the top x-axis represents the mode model complexity. As we can see, the transformed bootstrap estimate is very close to the true value, indicating the potential of such an estimation procedure. We also plot the transformation of $\widehat{h}$ and h in Fig. 6b.

The motivation of such a transformation is as follows. The original data set and the bootstrap data sets often require different tuning parameters λ to achieve the same level of sparsity, which makes the estimate difficult because the model selection deviation highly depends on the mode model. Such a phenomenon makes the estimation of model selection deviation much harder. The proposed transformation map the tuning parameter in the original and bootstrap data sets to the same scale so that estimate is easier. Therefore, we shift our focus to the model selection deviation as a function of γ instead of λ. Such a shift make the comparison and estimation of different model selection methods possible.

Take MSD_γ^- as an example. Although it is similar to MSD^-, it offers a few advantages. For example, MSD_γ^- always have the γ-intervals of width 1, which makes the comparison of MSD_γ^- as a measure of selection uncertainty across different selection methods easier. In addition, because of the fixed interval width, the estimate of MSD_γ^- is also more accurate than to that of MSD^-.

5. Simulations

In this section, we demonstrate the advantages of the proposed methods using simulation. All simulated covariates are standardized with zero mean and unit variance. We use R packages `glmnet` (Friedman et al., 2010) and `ncvreg` (Breheny and Huang, 2011) to perform Lasso, SCAD, adaptive Lasso, MCP, and elastice net.

5.1. Visualization

We first show the effectiveness of the proposed graphical tools in visualizing the distribution of the selected model. We start by comparing the distributions of the model selected by Lasso under different tuning parameters and sample sizes. Under the same setting in Section 3.1, we simulate $M = 100000$ data sets using sample size $n = 100, 300$ and apply Lasso with tuning parameter $\lambda = 0.4, 0.7$. We present the G- and H-plots in Fig. 7. The mode model is the true model.

As n increases from 100 to 300, we observe: The true model is selected more frequently. The distribution of the selected model becomes less dispersed, meaning the selected model tends to be closer to the true model. The Hamming distance histogram shrinks towards zero. The model complexity histogram shrinks towards the true model complexity. The histograms of H^+ and H^- shrink to zero.

As λ decreases from 0.7 to 0.4, we observe: The selected models tend to have larger complexities. Most of the selected models contain the true model, i.e., groups #2 and #4. The Hamming distance histogram becomes wider, indicating the selected model tends to be further from the true model. The model complexity histogram becomes more skewed to the large complexity. The histogram of H^+ becomes more dispersed, meaning the selected model tends to include more redundant variables. The histogram of H^- shrinks, meaning the selected model rarely misses important variables. These G- and H-plots allow us to see many interesting patterns in the distribution of the selected model.

We now visualize the distributions of the models selected by Lasso with $\lambda = 0.8$, SCAD with $\lambda = 0.6$, $a_{\text{SCAD}} = 3$ and MCP with $\lambda = 0.55$, $a_{\text{MCP}} = 3.7$. We simulate data sets from $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2)$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$. Let $p = 10$, $n = 300$, $\sigma = 5$, $\boldsymbol{\beta}^0 = (3, 2.5, 2, 1.5, 1, 0, 0, 0, 0, 0)$, and $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{ij} = 0.5^{|i-j|}$. The G- and H-plots of these methods are presented in Fig. 8. The mode model is the true model.

As we can see, Lasso selects the true model most frequently. When Lasso misses the true model, it includes one redundant variable or miss one important variable. It rarely misses two or more variables. The histograms confirm these observations. In comparison, SCAD makes more mistakes. It selects the true model with only 50% probability. Its selected models are more spread out than Lasso. There is a non-negligible probability that SCAD will miss two variables or more. Lastly, MCP performs similarly to SCAD with slightly higher probabilities to make mistakes, but it tends not to select too large models.

More visualization results can be found in the supplementary materials.

5.2. Estimation of visualization

When the true data generating process is known, we can apply the proposed visualization. When the true data generating process is unknown and we are given only one data set, we need to estimate the distribution of the selected model and its visualization. In this section, we use a naive bootstrap procedure for estimating G-plot, the H-plot can be estimated similarly, and conduct a simulation to test the bootstrap performance.

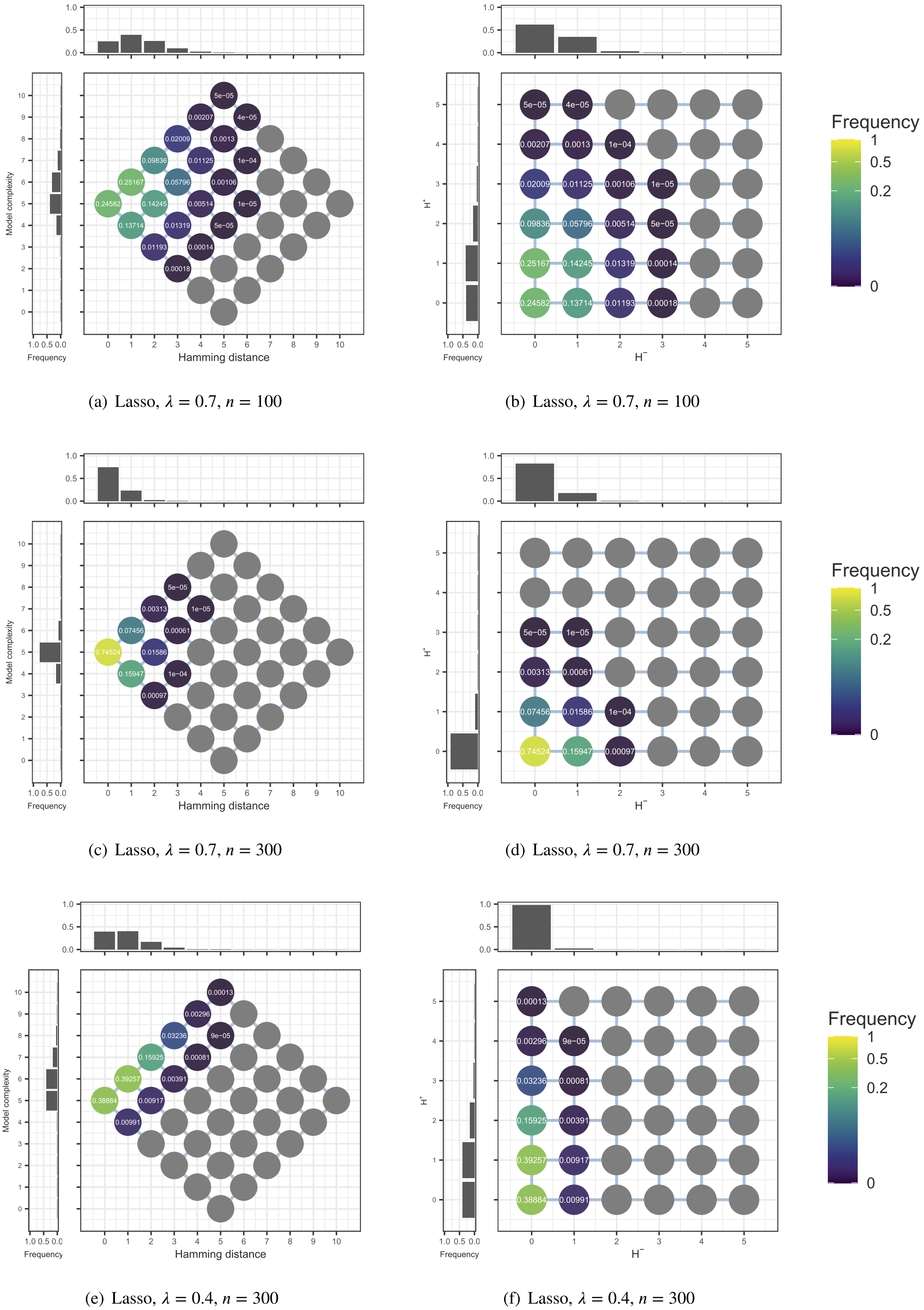
Given a data set, we first obtain bootstrap models $\widehat{m}^{(1)}, \ldots, \widehat{m}^{(B)}$, and then find the bootstrap mode model $\widehat{m}^* = \arg\max_m \{\sum_{b=1}^{B} \mathbb{1}\{\widehat{m}^{(b)} = m\}/B\}$. We use it as the estimate of group #1, i.e., $\widehat{\mathcal{G}}_1 = \{\widehat{m}^*\}$. The rest of groups #2, #3, #4, $\ldots$, can be estimated accordingly. For instance, $\widehat{\mathcal{G}}_2 = \{m : H(\widehat{m}^* \| m) = 1, |m| = |\widehat{m}^*| + 1\}$. Based on these estimated groups, we can obtain the bootstrap group frequency estimate for group #i as $\widehat{P}_{\#i} = \sum_{b=1}^{B} \mathbb{1}(\widehat{m}^{(b)} \in \widehat{\mathcal{G}}_i)/B$.

We generate $M = 1000$ data sets under the same setting in Section 3.1, and select the models by Lasso with $\lambda = 0.7$. We compare the true group frequencies with the bootstrap estimate frequencies of G-plot ($B = 1000$) in Table 2. The first row reports the true frequencies and the second row reports the mean bootstrap estimate. The estimation performance is not perfect but shows some promise. The overall pattern in group frequencies is preserved in the estimate, but the individual estimate sometimes shows large bias.

We further divide these $M = 1000$ data sets into four cases. Case 1: the bootstrap mode model is the same as the true mode model, i.e., $\widehat{m}^* = m^*$. Case 2: the bootstrap mode model is not the true mode model, but shares the same model complexity, i.e., $\widehat{m}^* \neq m^*$ and $|\widehat{m}^*| = |m^*|$. Cases 3 (and Case 4): the bootstrap mode model has a larger (and smaller) model complexity than the true mode model, i.e., $|\widehat{m}^*| > |m^*|$ (and $|\widehat{m}^*| < |m^*|$). In Table 2, we also report the mean bootstrap estimate within each case. Case 1 performs better than the other cases because it identifies the true mode model with a higher frequency. Case 2 overestimates group #5's frequency because the true mode model belongs to group #5. Similarly, Case 3 overestimates groups #3 and #6 and Case 4 overestimates groups #2 and #4. Finally, we show an example of bootstrap estimates of G-plot under Cases 1 through 4, and the true G-plot in Fig. 9.

The estimation procedure for H-plot can follow the same logic. For the scatterplot and heatmap, we need to estimate the coefficients $\boldsymbol{\beta}^0$ using our initial estimate from the original sample, which is even more challenging.

Based on these results, we conclude that it could be hopeful to estimate G-plot and H-plot through bootstrapping. However, straightforward bootstrap implementation would not be ideal. This is because the bootstrapped model selection



(a) Lasso, $\lambda = 0.7$, $n = 100$

(b) Lasso, $\lambda = 0.7$, $n = 100$

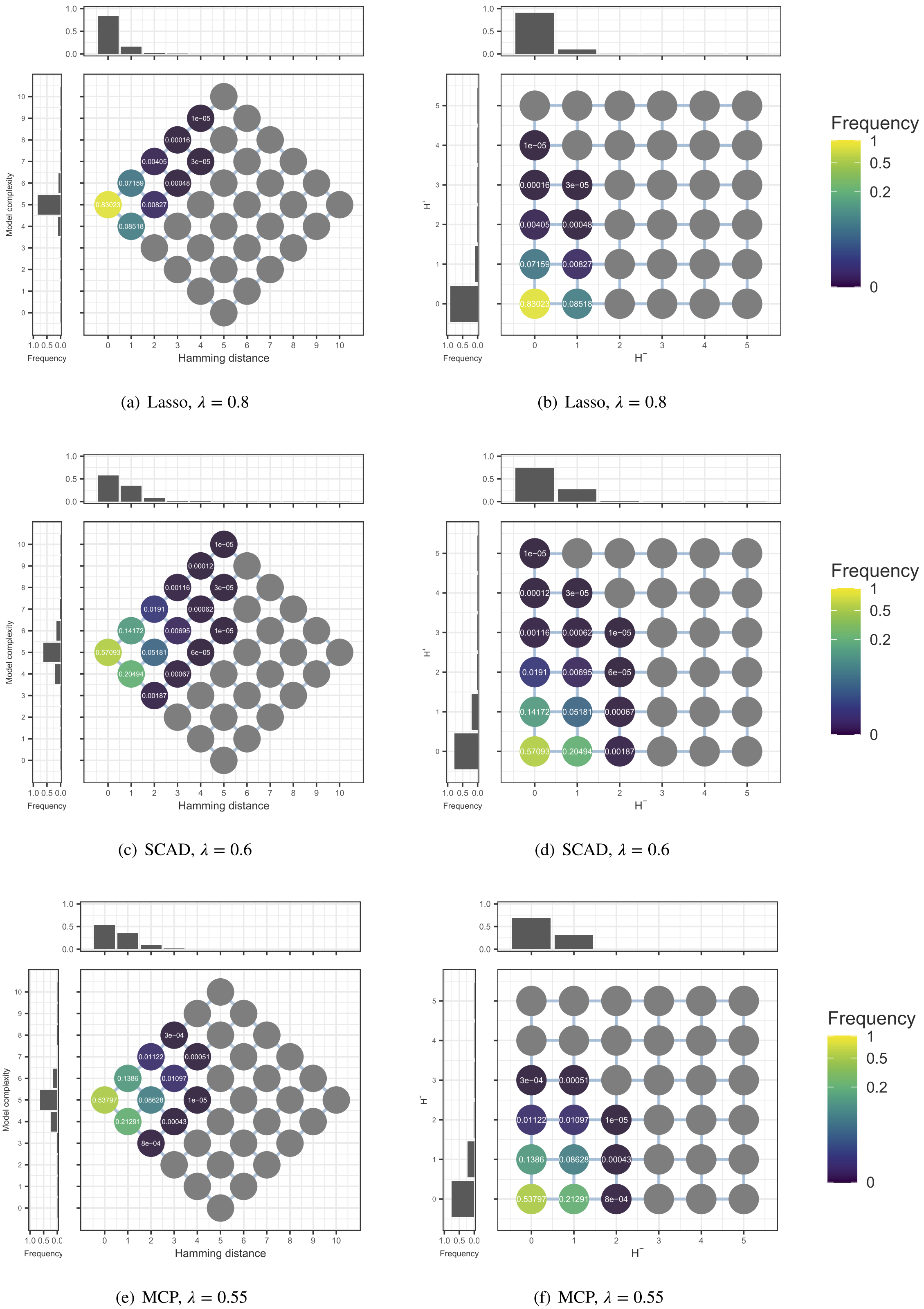
(c) Lasso, $\lambda = 0.7$, $n = 300$

(d) Lasso, $\lambda = 0.7$, $n = 300$

(e) Lasso, $\lambda = 0.4$, $n = 300$

(f) Lasso, $\lambda = 0.4$, $n = 300$

Fig. 7. Comparison of the distribution of selected model for Lasso under different tuning parameter values and different sample sizes using G-plots and H-plots. The first column is for G-plot while the second column is for H-plot.



(a) Lasso, $\lambda = 0.8$

(b) Lasso, $\lambda = 0.8$

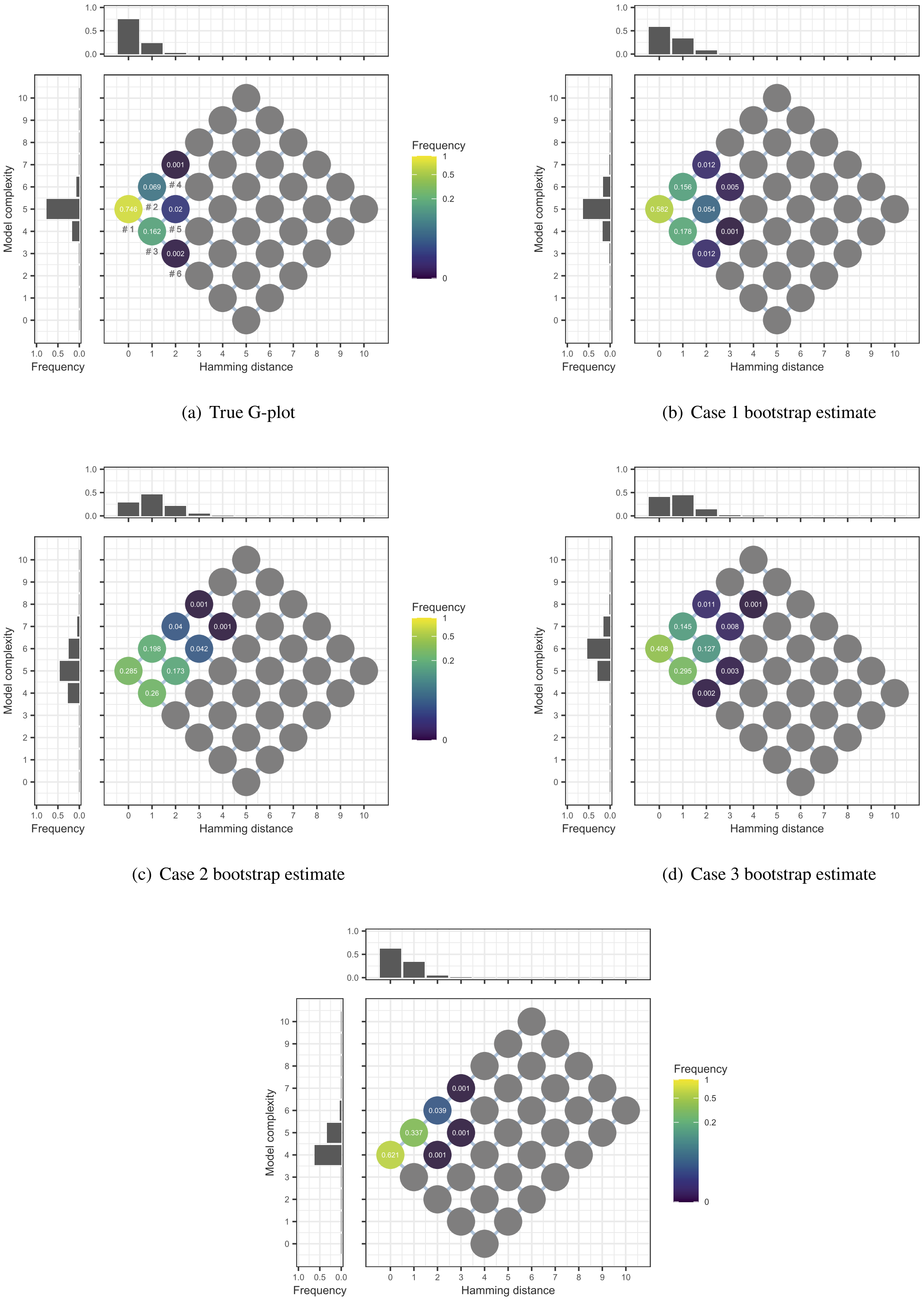
(c) SCAD, $\lambda = 0.6$

(d) SCAD, $\lambda = 0.6$

(e) MCP, $\lambda = 0.55$

(f) MCP, $\lambda = 0.55$

Fig. 8. Comparison of the distribution of selected model by Lasso, SCAD and MCP using G-plots and H-plots. The first column is for G-plots and the second column is for H-plots.



(a) True G-plot

(b) Case 1 bootstrap estimate

(c) Case 2 bootstrap estimate

(d) Case 3 bootstrap estimate

(e) Case 4 bootstrap estimate

Fig. 9. Examples of bootstrap G-plot estimates under Cases 1 to 4.

Table 2
Comparison of bootstrap group frequencies estimates (averages). $p = 10, M = 1000, B = 1000, n = 300, \lambda_{\text{Lasso}} = 0.7$.

GMSD	Group						MC Iterations
	#1	#2	#3	#4	#5	#6	
True group freq.	0.7460	0.0690	0.1620	0.0010	0.0200	0.0020	
Mean boot. est.	0.4912	0.2520	0.1167	0.0559	0.0542	0.0073	1000 out of 1000
Case 1	0.5253	0.2315	0.1236	0.0419	0.0557	0.0042	741 out of 1000
Case 2	0.2971	0.2592	0.1648	0.0693	0.1493	0.0062	20 out of 1000
Case 3	0.3292	0.1225	0.3001	0.0166	0.1137	0.0667	60 out of 1000
Case 4	0.4260	0.3792	0.0214	0.1254	0.0176	0.0001	179 out of 1000

distribution varies significantly under different observed data sets, and its behavior is more complicated than we have understood so far. When the initial selected model based on the original sample is correct, the bootstrapped distribution tend to be close to the true distribution. However, if the initial selected model is incorrect, the bootstrapped distribution would not be a good approximation. Therefore, most reliable estimation is still needed to fill this gap.

5.3. Bootstrap estimation for model selection deviation

We test the proposed estimation procedure under four different settings. The responses are generated from $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2)$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$. Here are the details of the four settings.

S1: Let $p = 10, n = 1000, \sigma = 5$, $\boldsymbol{\beta}^0 = (3, 2.5, 2, 1.5, 1, 0, 0, 0, 0, 0)$, and $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{I}_{p\times p})$.

S2: Let $p = 10, n = 1000, \sigma = 5$, $\boldsymbol{\beta}^0 = (3, 2.5, 2, 1.5, 1, 0, 0, 0, 0, 0)$, and $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{ij} = 0.5^{|i-j|}$.

S3: Let $p = 100, n = 1000, \sigma = 20$, $\boldsymbol{\beta}^0 = (10, 9.5, \ldots, 1, 0.5, 0, \ldots, 0)$, i.e., the first twenty elements in $\boldsymbol{\beta}^0$ are non-zeroes and the rests are zeroes. $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{I}_{p\times p})$.

S4: Let $p = 100, n = 1000, \sigma = 20$, $\boldsymbol{\beta}^0$ is the same as S3, and $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{ij} = 0.5^{|i-j|}$.

Under each setting, we compare Lasso, adaptive Lasso, SCAD, MCP and elastic net for uncertainty. The weights $\boldsymbol{\omega}$ in adaptive Lasso are obtained by linear regression. We also consider another alternative, adaptive Lasso with fixed weights (ALFW). We set the weights as $\boldsymbol{\omega} = (3.5, 3, 2.5, 2, 1.5, 1, \ldots, 1)$ for S1-S2, and $\boldsymbol{\omega} = (20, 19, \ldots, 2, 1, \ldots, 1)$ for S3-S4, i.e., the first twenty elements are decreasing from 20 to 1.

Under S1, we generate $M = 300$ data sets and compare the bootstrap estimates, the averages of bootstrap estimates, and the true model selection deviations. Pairwise bootstrap with $B = 3000$ are used. We present the results for Lasso and SCAD in Fig. 10. The x-axis is the transformed tuning parameter γ ranging from 5 to 10, which means the mode model complexity ranges from 5 to 1. The average of estimates and true value are close to each other in most cases, indicating that the bootstrap estimates shows little bias in this setting. The individual estimates randomly distribute around its average, indicating the estimation variance is not negligible. The additional results for other selection methods are provided in the supplementary materials.

We compare the true and the average estimated MSDs of six selection methods under S1-S4 in Fig. 11.

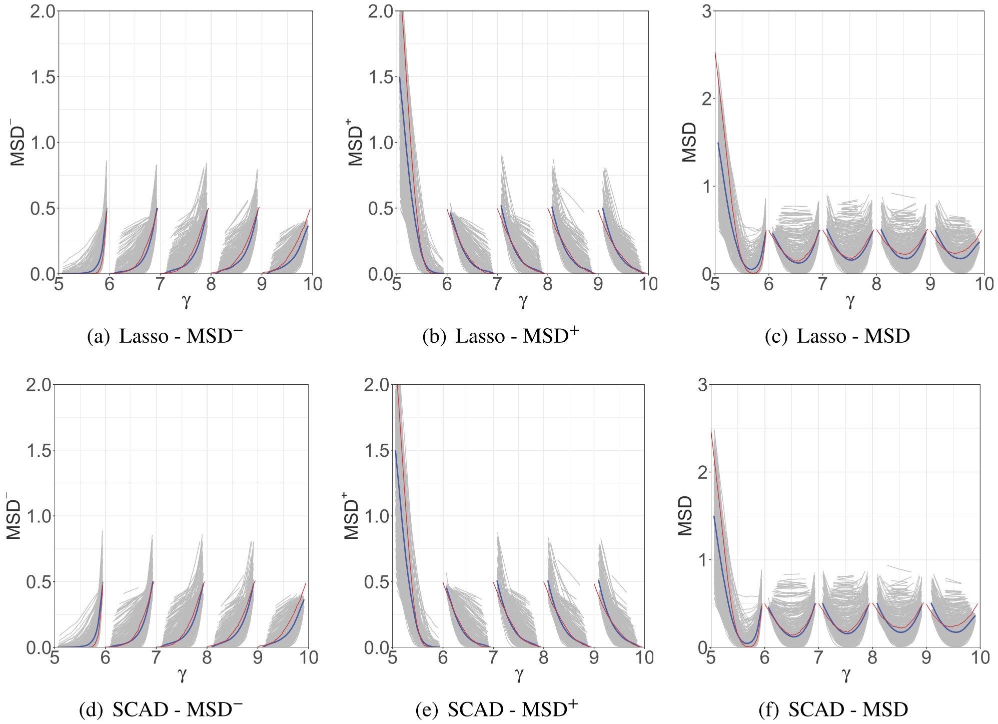
Under S1, the uncertainty of these methods follows the order of adaptive Lasso $>$ Lasso $\approx$ elastic net $\approx$ SCAD $\approx$ MCP $>$ ALFW. The same order can be recovered using the bootstrap estimates. The reason why most selection methods perform similarly is because S1 is relatively simple. These MSDs appear synchronised because we plot them against the standardized tuning parameter γ as opposed to λ. As long as γ is in the same region, all methods' mode model complexities are the same. This is true for both the true MSD and the bootstrap estimate. Therefore, plotting against γ makes the comparison across different selection easier. Note that, some lines may be missing in the true distribution since the complexity of the true mode model when using some model selection method may skip certain values as the tuning parameter changes. The same phenomenon happens in the case of S2.

Under S2, a few selection methods become slightly more unstable as collinearity is introduced, evidenced by the elevated model selection deviation. The stability of MCP worsens the most, indicating it is the most fragile one against collinearity. The overall uncertainty order is MCP $>$ adaptive Lasso $>$ SCAD $\approx$ Lasso $\approx$ ElasticNet $>$ ALFW. The bootstrap estimates are able to capture this order as well.

Under S3, the model selection uncertainty of all methods are inevitably higher because of the large number of covariates. Note that γ ranges from 83 to 100, meaning the mode model complexity ranges from 17 to 1. The MSD$^+$ at the mode model complexity around 17 is much higher than that at the complexity below 10. This is because there are a few small coefficients that are very hard to select. The differences in selection uncertainty among six model selection methods are small, but the bootstrap estimates are able to mimic these patterns.

Under S4 where both collinearity and high-dimension are introduced, some methods start to become more unstable. In particular, MCP shows significantly higher selection uncertainty than others. The order of selection uncertainty is MCP $>$ adaptive Lasso $>$ SCAD $\approx$ Lasso $>$ elastic net $>$ ALFW. The bootstrap estimates are able to reflect this change in selection uncertainty, indicating MCP has high uncertainty under collinearity.

There are some limitations of our proposed method. The bootstrap estimation of MSD works well when dimension p is not too large. In the high-dimensional cases S3 and S4, the bootstrap estimation may not be as accurate as the low



(a) Lasso - MSD^-

(b) Lasso - MSD^+

(c) Lasso - MSD

(d) SCAD - MSD^-

(e) SCAD - MSD^+

(f) SCAD - MSD

Fig. 10. Comparison of the true and estimated model selection deviations for Lasso and SCAD. The true value is in red, the bootstrap estimate is in gray, and the average of the bootstrap estimate is in blue. The first, second, and third columns are for MSD^-, MSD^+, and MSD, respectively. The first and second rows are for Lasso and SCAD, respectively.

dimensional case. Another limitation is the collinearity case S4. If there exists collinearity among the variables, the model selection uncertainty usually increases, but its bootstrap estimate quality worsens. Although these limitations degrade the estimation quality, the estimation still preserves the ranking of the uncertainty of the model selection methods. We present the deviation of the bootstrap estimate to the true value in the supplementary materials.

5.4. Bootstrap estimation for weighted model selection deviation

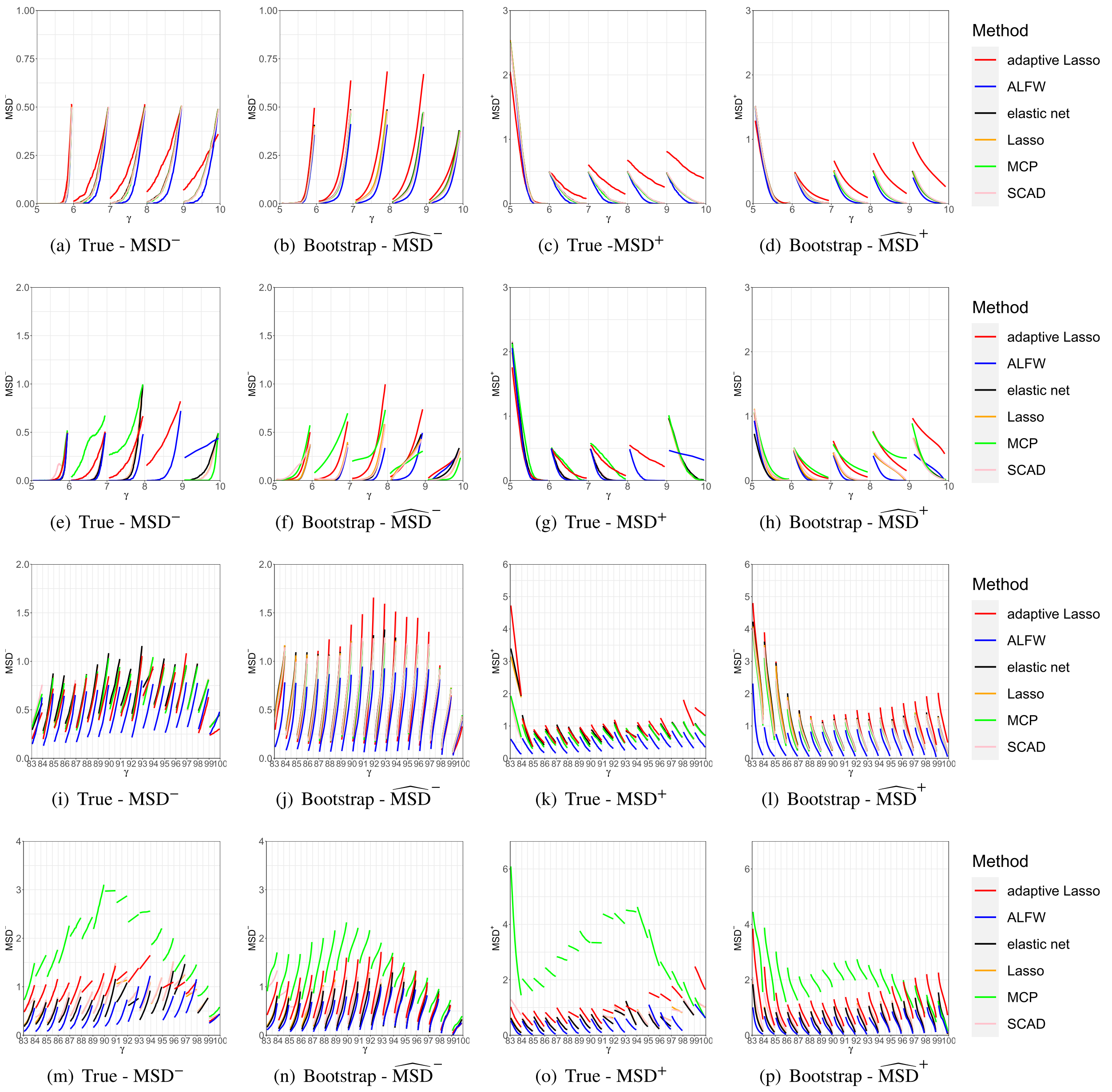
In this section, we show the performance of the bootstrap estimation for weighted model selection deviation using Lasso as an example. We simulate $M = 1000$ data sets with $n = 300$ and $p = 10$ according to $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$ is an $n \times p$ design matrix of p predictors with $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{I}_{p\times p})$, $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T \sim N_n(\mathbf{0}, 5^2\mathbf{I}_{n\times n})$, and $\boldsymbol{\beta} = (3, 2.5, 2, 1.5, 1, 0, 0, 0, 0, 0)$. Bootstrap with $B = 3000$ and Lasso are applied to each data set. We plot WMSD^-, WMSD^+, and their bootstrap estimates in Fig. 12. The true value is in red, the bootstrap estimate is in gray, and the mean bootstrap estimate is in blue.

In Fig. 12, the mean bootstrap estimate is close to the true value, which indicates the potential in estimating WMSD^- and WMSD^+. If the model setup becomes more complicated, the performance of these estimates usually deteriorates. It is because that WMSD^+ highly depends on the parameter values which are sensitive. We have found that the weighted versions of the model selection deviation are generally more sensitive and harder to estimate.

5.5. Model selection deviation with cross-validation

We test the estimation of MSD in the case of model selection with the cross-validated tuning parameter in this section. The proposed MSD is the function of λ for the reason that the selection uncertainty here is induced by different penalized loss function without influenced by the randomness of tuning parameter. In other words, if MSD is defined as a value based on a certain λ, such as a selected one by cross-validation, the additional uncertainty of selecting tuning parameter will be uncontrolled. We explain this with the following simulation example.

MSD in the case of model selection with cross-validation is referred to as MSD_{CV}. We simulate $M = 1000$ data sets under the same setting in Section 3.1. We select models using 10-fold cross-validated Lasso, and compare the true MSD_{CV}, the average of bootstrap estimates $\widehat{\text{MSD}}_{\text{CV}}$ in Table 3. As we can see, the mean bootstrap estimate deviates from the true value, especially for MSD^+_{CV}. Different from MSD, MSD_{CV} measures the total selection uncertainty due to the penalized loss function and tuning parameter, while the tuning parameter selection uncertainty is often overestimated in the bootstrap



(a) True - MSD$^-$ (b) Bootstrap - $\widehat{\text{MSD}}^-$ (c) True -MSD$^+$ (d) Bootstrap - $\widehat{\text{MSD}}^+$

(e) True - MSD$^-$ (f) Bootstrap - $\widehat{\text{MSD}}^-$ (g) True - MSD$^+$ (h) Bootstrap - $\widehat{\text{MSD}}^+$

(i) True - MSD$^-$ (j) Bootstrap - $\widehat{\text{MSD}}^-$ (k) True - MSD$^+$ (l) Bootstrap - $\widehat{\text{MSD}}^+$

(m) True - MSD$^-$ (n) Bootstrap - $\widehat{\text{MSD}}^-$ (o) True - MSD$^+$ (p) Bootstrap - $\widehat{\text{MSD}}^+$

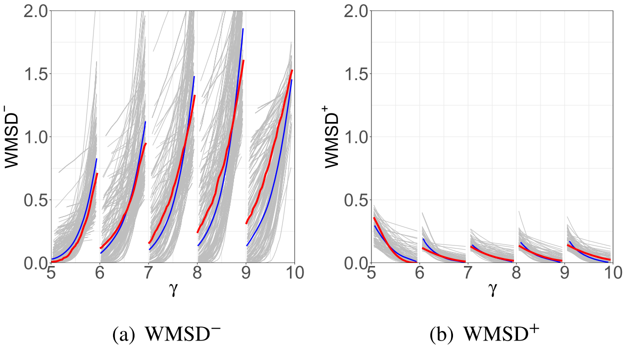
Fig. 11. Comparison of true model selection deviation and average estimated model selection deviation for various model selection methods under S1 - S4. Row 1 through row 4 are for S1 through S4, respectively. Columns 1 and 2 are for true and estimated MSD$^-$. Columns 3 and 4 are for true and estimated MSD$^+$. Lasso is orange, SCAD is pink, MCP is green, elastic net is black, adaptive Lasso is red, and ALFW is blue.

samples because the selected tuning parameters in each bootstrap sample are different. Therefore, the MSD$_{\text{CV}}$ usually cannot be estimated accurately.

6. Real data example

In this section, we illustrate the proposed method using two real data sets with different dimensionalities.

We first analyze a yeast cell-cycle gene expression data set collected in the experiment of Spellman et al. (1998). To understand the cell-cycle process, biologists are interested in identifying transcription factors (TFs) that regulate the expression levels of cell cycle-regulated genes. Therefore, we analyzed a data set with $n = 1132$ gene expression levels of yeast as response variable and the standardized binding probabilities of a total of $p = 96$ transcription factors as covariates. These binding probabilities are obtained from a mixture model approach of Wang et al. (2007) based on the ChIP data of Lee et al. (2002). The data set is publicly available in the R package PGEE. Previous studies in this area have focused on identifying either the individual transcription factor effects (Wang et al., 2007; Cheng and Li, 2008; Wang et al., 2012) or the synergistic

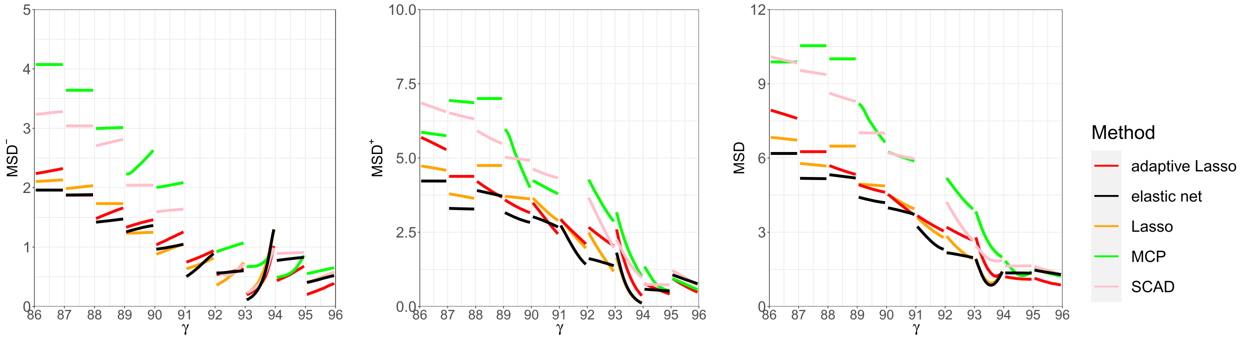


(a) WMSD$^-$ (b) WMSD$^+$

Fig. 12. Comparison of the true and estimated WMSD$^-$ and WMSD$^+$ using Lasso. The true value is in red, the bootstrap estimate is in gray, and the mean bootstrap estimate is in blue.

Table 3
Comparison of the true model selection deviation, the bootstrap estimates and their average when applying model selection with the cross-validated tuning parameter. $p = 10, M = 1000, B = 1000, n = 300$.

Comparison	Model Selection Deviation		
	MSD^-_{CV}	MSD^+_{CV}	MSD_{CV}
True MSD_{CV}	0.140	0.238	0.378
Average bootstrap estimate	0.199	0.822	1.021



(a) Estimated MSD$^-$ (b) Estimated MSD$^+$ (c) Estimated MSD

Fig. 13. Estimated model selection deviations based on the genetic data. Lasso is orange, SCAD is pink, MCP is green, elastic net is black and adaptive Lasso is red.

effects where multiple transcription factors may cooperate to regulate transcription in the cycle process (Das et al., 2004; Wang et al., 2007; Cheng and Li, 2008). None of them investigate on the stability of these results. In contrast, we focus on the uncertainty of the model selection methods on this data set.

To evaluate different methods' uncertainty on this data set, we estimate their model selection deviations with $B = 5000$ in Fig. 13. Note that x-axis represents γ which ranges from 86 to 96, and the mode model complexity ranges from 10 to 1. As we can see, the model selection deviation by MCP and SCAD are on top of the other methods at various level of γ for estimated MSD$^-$, MSD$^+$, and MSD. Meanwhile, elastic net is always below most of the other methods. The Lasso and adaptive Lasso are close to each other with adaptive Lasso being slightly higher. Therefore, we conclude that the model selection uncertainty follows MCP > SCAD > adaptive Lasso > Lasso > elastic net. Note that, as the mode model complexity increases, the differences in model selection deviation become more obvious.

We apply the proposed method on another low dimensional data set, which consists of the measurements on $n = 442$ diabetic patients (Efron et al., 2004). We use the disease progression one year after baseline as the response variable. The covariates include $p = 10$ baseline variables, such as age, sex, body mass index (BMI), average blood pressure (ABP) and six blood serum measurements (S1,...,S6). The data set is publicly obtained in the R package `care`.

We estimate the model selection deviation of different model selection methods using bootstrap $B = 5000$, and plot the estimated MSD in Fig. 14. The x-axis represents γ ranging from 0 to 10, which corresponds to the bootstrap mode model complexity changing from 10 to 1. When the bootstrap mode model complexity is at 9, SCAD and MCP's model selection deviations are missing because there are no mode models with such complexities in the bootstrap distributions for these

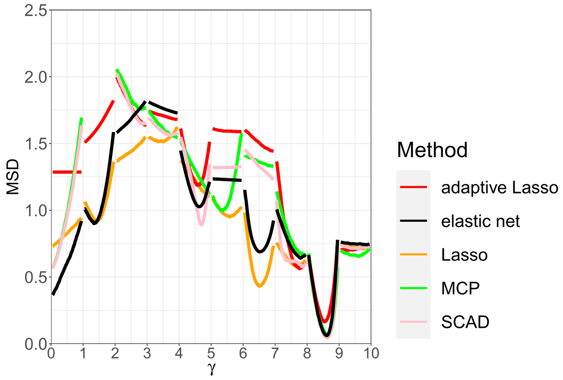


Fig. 14. Estimated model selection deviation based on the diabetes data. Lasso is orange, SCAD is pink, MCP is green, elastic net is black, and adaptive Lasso is red.

methods. Since the differences of uncertainty among various methods are not obvious when $\gamma \in (8, 10)$, we mainly focus on the comparisons when $\gamma \in (0, 8)$. We find that Lasso and elastic net have lower uncertainty than the rest. Adaptive Lasso always has the highest uncertainty. SCAD is mostly close to MCP and they also have high uncertainty. Although the ranking of these methods within each γ-interval is slightly different, overall, we conclude that the model selection uncertainty has the following order: adaptive Lasso $>$ MCP $>$ SCAD $>$ elastic net $>$ Lasso.

7. Discussion

In this article, we have proposed several new graphical tools to visualize the distributions of the selected model under various model selection procedures. The visualization helps us to understand the behavior of the model selection procedure. To the best of our knowledge, there is the first attempt in visualizing such a complex distribution. We further propose a few numerical attributes on the distribution to quantify its central tendency, dispersion, and skewness. Among them, the model selection deviation allows quantitative comparison of the model selection uncertainty of various model selection procedures. The proposed visualization tools and uncertainty measures have potential uses for various data structures, such as time series data (Behrendt and Schweikert, 2021), functional data analysis (Mousavi and Sørensen, 2017; Fan and Reimherr, 2017), and interaction analysis (Chai et al., 2017; Li et al., 2019a).

In this article, we mainly focus on the computational aspects of the proposed method and do not provide comparison theoretically. The corresponding theory would rely on the analytical form or approximation of the model selection distribution under the finite sample size, which remains challenging due to the complex nature of such a distribution. To the best of our knowledge, the theory of the bootstrap estimation of model selection distribution under the finite sample size is quite limited. We hope that our results can be one small step toward this important goal.

There are many directions for future research. For example, the current uncertainty measure depends heavily on the mode model. However, as a central tendency measure, the mode sometimes fails when the data is highly skewed or suffers from the uniqueness issue. How to better capture the center of the distribution is a key to a successful measure of the model selection uncertainty. Lastly, our model selection deviation focuses on the "variance" of the model selection procedure. A natural question is how to quantify the "bias" in model selection. We believe these topics are important for future investigation.

Acknowledgements

Y Li's work is supported by the National Natural Science Foundation of China (72271237) and the Platform of Public Health & Disease Control and Prevention, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2022.107598.

References

Bayer, S., 2018. Combining value-at-risk forecasts using penalized quantile regressions. Econom. Stat. 8, 56–77.
Behrendt, S., Schweikert, K., 2021. A note on adaptive group lasso for structural break time series. Econom. Stat. 17, 156–172.
Breheny, P., Huang, J., 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann. Appl. Stat. 5 (1), 232–253.

Chai, H., Zhang, Q., Jiang, Y., Wang, G., Zhang, S., Ahmed, S.E., Ma, S., 2017. Identifying gene-environment interactions for prognosis using a robust approach. Econom. Stat. 4, 105–120.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. J. R. Stat. Soc., Ser. A, Stat. Soc. 158 (3), 419–444.

Chatterjee, A., Lahiri, S.N., 2011. Bootstrapping lasso estimators. J. Am. Stat. Assoc. 106 (494), 608–625.

Cheng, C., Li, L.M., 2008. Systematic identification of cell cycle regulated transcription factors from microarray time series data. BMC Genomics 9 (1), 116.

Das, D., Banerjee, N., Zhang, M.Q., 2004. Interacting models of cooperative gene regulation. Proc. Natl. Acad. Sci. USA 101 (46), 16234–16239.

Ding, J., Tarokh, V., Yang, Y., 2018. Model selection techniques: an overview. IEEE Signal Process. Mag. 35 (6), 16–34.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al., 2004. Least angle regression. Ann. Stat. 32 (2), 407–499.

Ewald, K., Schneider, U., et al., 2020. On the distribution, model selection properties and uniqueness of the lasso estimator in low and high dimensions. Electron. J. Stat. 14 (1), 944–969.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96 (456), 1348–1360.

Fan, Z., Reimherr, M., 2017. High-dimensional adaptive function-on-scalar regression. Econom. Stat. 1, 167–183.

Ferrari, D., Yang, Y., 2015. Confidence sets for model selection by F-testing. Stat. Sin. 25 (4), 1637–1658.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–22.

Hansen, P.R., Lunde, A., Nason, J.M., 2011. The model confidence set. Econometrica 79 (2), 453–497.

Hennig, C., Sauerbrei, W., 2019. Exploration of the variability of variable selection based on distances between bootstrap sample results. Adv. Data Anal. Classif. 13 (4), 933–963.

Knight, K., Fu, W., 2000. Asymptotics for lasso-type estimators. Ann. Stat. 28 (5), 1356–1378.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al., 2002. Transcriptional regulatory networks in saccharomyces cerevisiae. Science 298 (5594), 799–804.

Li, Y., Li, R., Qin, Y., Wu, M., Ma, S., 2019a. Integrative interaction analysis using threshold gradient directed regularization. Appl. Stoch. Models Bus. Ind. 35 (2), 354–375.

Li, Y., Luo, Y., Ferrari, D., Hu, X., Qin, Y., 2019b. Model confidence bounds for variable selection. Biometrics 75 (2), 392–403.

Lim, C., Yu, B., 2016. Estimation stability with cross-validation (ESCV). J. Comput. Graph. Stat. 25 (2), 464–492.

Liu, X., Li, Y., Jiang, J., 2020. Simple measures of uncertainty for model selection. Test, 1–20.

Meinshausen, N., Bühlmann, P., 2010. Stability selection. J. R. Stat. Soc., Ser. B, Stat. Methodol. 72 (4), 417–473.

Mousavi, S.N., Sørensen, H., 2017. Multinomial functional regression with wavelets and lasso penalization. Econom. Stat. 1, 150–166.

Nan, Y., Yang, Y., 2014. Variable selection diagnostics measures for high-dimensional regression. J. Comput. Graph. Stat. 23 (3), 636–656.

Pötscher, B.M., Leeb, H., 2009. On the distribution of penalized maximum likelihood estimators: the lasso, scad, and thresholding. J. Multivar. Anal. 100 (9), 2065–2082.

Sauerbrei, W., Buchholz, A., Boulesteix, A.-L., Binder, H., 2015. On stability issues in deriving multivariable regression models. Biom. J. 57 (4), 531–555.

Seri, R., Martinoli, M., Secchi, D., Centorrino, S., 2020. Model calibration and validation via confidence sets. Econom. Stat. 20, 62–86.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell 9 (12), 3273–3297.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B, Methodol. 58 (1), 267–288.

Wang, L., Chen, G., Li, H., 2007. Group scad regression analysis for microarray time course gene expression data. Bioinformatics 23 (12), 1486–1494.

Wang, L., Qin, Y., Li, Y., 2021. Confidence graphs for graphical model selection. Stat. Comput. 31 (52).

Wang, L., Zhou, J., Qu, A., 2012. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. Biometrics 68 (2), 353–360.

Yang, W., Yang, Y., 2017. Toward an objective and reproducible model choice via variable selection deviation. Biometrics 73 (1), 20–30.

Ye, C., Yang, Y., Yang, Y., 2018. Sparsity oriented importance learning for high-dimensional linear regression. J. Am. Stat. Assoc. 113 (524), 1797–1812.

Yu, Y., Yang, Y., Yang, Y., 2022. Performance assessment of high-dimensional variable identification. Stat. Sin. 32, 1–24.

Zhang, C.-H., et al., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. 38 (2), 894–942.

Zheng, C., Ferrari, D., Yang, Y., 2019a. Model selection confidence sets by likelihood ratio testing. Stat. Sin. 29 (2), 827–851.

Zheng, C., Ferrari, D., Zhang, M., Baird, P., 2019b. Ranking the importance of genetic factors by variable-selection confidence sets. J. R. Stat. Soc., Ser. C, Appl. Stat. 68 (3), 727–749.

Zhou, Q., 2014. Monte Carlo simulation for lasso-type problems by estimator augmentation. J. Am. Stat. Assoc. 109 (508), 1495–1516.

Zou, H., 2006. The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. 101 (476), 1418–1429.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc., Ser. B, Stat. Methodol. 67 (2), 301–320.