Statistica Sinica **32** (2022), 695-718
doi:https://doi.org/10.5705/ss.202019.0210

PERFORMANCE ASSESSMENT OF HIGH-DIMENSIONAL VARIABLE IDENTIFICATION

Yanjia Yu[1], Yi Yang[2] and Yuhong Yang[1]

[1] *University of Minnesota and* [2] *McGill University*

Abstract: Because model selection is ubiquitous in data analysis, the reproducibility of statistical results requires that we be able to evaluate the reliability of the employed model selection method, regardless of the model's apparent good properties. Instability measures have been proposed for evaluating model selection uncertainty. However, low instability does not necessarily indicate that the selected model is trustworthy, because low instability can also arise when a method tends to select an overly parsimonious model. F- and G-measures have become increasingly popular for assessing variable selection performance in theoretical studies and simulation results. However, they are not computable in practice. In this work, we propose an estimation method for F- and G-measures and prove their desirable properties of uniform consistency. This gives the data analyst a valuable tool to compare different variable selection methods based on the data at hand. Extensive simulations are conducted to show the very good finite-sample performance of our approach. Lastly, we apply our methods to several microarray gene expression data sets, with intriguing results.

Key words and phrases: F-measure, G-measure, gene expression, model averaging, reproducibility, variable selection performance.

1. Introduction

Variable selection is of interest in many fields, including bioinformatics, genomics, finance, and economics. In bioinformatics, for example, microarray gene expression data are collected to identify cancer-related biomarkers in order to differentiate affected patients from healthy individuals based on their gene expression profile. The number of variables, p, in typical microarray gene expression data is of 10^{3-5} magnitude, while the number of subjects, n, is of 10^{1-3} magnitude. For problems in which $p \gg n$, the penalized likelihood estimation provides a class of methods for selecting the variables (see, e.g., Fan and Lv (2010)). However, it is well recognized in the literature that model selection methods, including the penalization methods for high-dimensional data, often encounter instability

Corresponding author: Yanjia Yu, School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA. E-mail: yuxxx748@umn.edu.

issues (Chatfield (1995); Draper (1995); Breiman (1996a,b); Buckland, Burnham and Augustin (1997); Yuan and Yang (2005); Lim and Yu (2016)). For example, removing a few observations or adding small perturbations to the data may result in dramatically different sets of variables being selected (Meinshausen and Bühlmann (2006); Nan and Yang (2014); Lim and Yu (2016)). This uncertainty in variable selection, as is well known, may have severe practical consequences. On a larger scale, reproducibility is a major problem in the science community (McNutt (2014); Stodden (2015)).

Variable selection uncertainty is mainly evaluated using instability measures, which test how sensitive a variable selection method is to small changes in the data because of subsampling (Chen, Giannakouros and Yang (2007)), resampling (Breiman (1996b); Buckland, Burnham and Augustin (1997)), or perturbations (Breiman (1996b)). However, a low instability measure does not necessarily indicate that a variable selection result is reliable, because low instability can also arise when a method tends to select an overly parsimonious model (e.g., the intercept-only model, in the extreme case).

There is therefore a great need for measures that can fully evaluate the uncertainty of variable selection beyond instability. In variable selection, researchers focus on two types of errors: including unnecessary variables, and excluding important variables. F- and G-measures are popular in the field of information retrieval (Billsus and Pazzani (1998)) for assessing overall variable selection performance (see, e.g., Lim (2011); Lim and Yu (2016)). Specifically, the F-measure is the harmonic mean of *precision* and *recall*, where precision (or positive predictive value) is defined as the fraction of the selected variables that are true variables, and recall (also known as sensitivity) is defined as the fraction of the true variables that are selected. The G-measure is the geometric mean of precision and recall. By combining precision and recall into one measure, one can evaluate the overall accuracy of a given variable selection method. Clearly, a higher F (or G) value indicates better selection performance, in an overall sense. However, existing approaches calculate the F- (or G-) measure of a given selection method for simulated data only (where the true model is known), and do not work for real data.

In this paper, we propose a method for the **p**erformance **a**ssessment of (high-dimensional) **v**ariable **i**dentification (PAVI), in which we estimate the F- or G-measure based on a combination of multiple candidate models under a proper weighting scheme. Our proposal works for both regression and classification, and applies to both synthetic and real data. Under sensible conditions, we show that our estimates are uniformly consistent in estimating the true F- and G-

measures for any set of models to be checked. The candidate models can be very flexible. For example, they can be obtained by penalization using the Lasso (Tibshirani (1996)), smoothly clipped absolute deviations (SCAD) penalty (Fan and Li (2001)), adaptive Lasso (Zou (2006)), minimax concave penalty (MCP) (Zhang (2010)) or other variable selection techniques. Two weighting schemes are considered in this work: adaptive regression by mixing (Yang (2001)), and weighting via information criteria (see, e.g., Nan and Yang (2014)). In the simulation section, we show the reliable estimation performance of our method for both classification and regression. We further demonstrate our methods by analyzing several microarray gene expression data from real applications. The results of the real data analysis suggest that the PAVI method is very useful for evaluating the variable selection performance of high-dimensional linear-based models. It provides useful information on the reliability and reproducibility of a given model when the true model is unknown. For example, one may justifiably doubt the reproducibility of a model that has very small estimated F- and G- values.

The remainder of the paper is organized as follows. In Section 2, we define the F- and G-measures and introduce our estimation methods. Section 3 provides the theoretical justification for the PAVI estimators of the F- and G-measures. Section 4 shows how to implement the PAVI method for both regression and classification, including how to obtain the candidate models and assign weights. Simulation results are presented in Section 5. We demonstrate our methods by analyzing three well-studied gene expression data sets in Section 6. Section 7 concludes the paper. All technical proofs are relegated to the Supplementary Material along with additional numerical results.

2. Methodology

Let us consider the generalized linear model framework. Denote $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\intercal$ as the $n \times p$ design matrix with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal$, for $i = 1, \ldots, n$. Let $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$ be the n-dimensional response vector. For a regression with a continuous response, we consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is the vector of n independent errors, and $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^\intercal$ is a p-dimensional coefficient vector of the true underlying model that generates the data. For classification, we consider the binary logistic regression model, for ease of presentation. Let $Y \in \{0, 1\}$ be a binary response variable, and $X \in \mathbb{R}^p$ be a p-dimensional predictor vector. We assume that Y follows the Bernoulli

distribution given $X = \mathbf{x}$, with conditional probability

$$\Pr(Y = 1|X = \mathbf{x}) = 1 - \Pr(Y = 0|X = \mathbf{x}) = \frac{e^{\mathbf{x}^\intercal \boldsymbol{\beta}^*}}{1 + e^{\mathbf{x}^\intercal \boldsymbol{\beta}^*}}. \tag{2.1}$$

Let $\mathcal{A}^* = \text{supp}(\boldsymbol{\beta}^*) \equiv \{j : \beta_j^* \neq 0\}$ be the index set of the variables in the true model with size $|\mathcal{A}^*|$, where $|\cdot|$ denotes the cardinality of a set. For both regression and classification, we assume that the true model is sparse. In other words, most coefficients in $\boldsymbol{\beta}^*$ are exactly zero, such that $|\mathcal{A}^*|$ is small.

Let $\mathcal{A}^0 = \{j : \beta_j^0 \neq 0\}$ be an index set of all nonzero coefficients from any given variable selection result $\boldsymbol{\beta}^0$. One can use F- and G- measures to evaluate the performance of $\mathcal{A}^0$. F- and G-measures take values between zero and one, where a higher value indicates better performance of the variable selection method. The definitions of F- and G- measures are based on *precision* and *recall*. The precision pr for $\mathcal{A}^0$ is the fraction of true variables in the given model $\mathcal{A}^0$; that is, $pr(\mathcal{A}^0) \equiv pr(\mathcal{A}^0; \mathcal{A}^*) = |\mathcal{A}^0 \cap \mathcal{A}^*|/|\mathcal{A}^0|$. The recall re for $\mathcal{A}^0$ is the fraction of variables in the true model $\mathcal{A}^*$ that are selected; that is, $re(\mathcal{A}^0) \equiv re(\mathcal{A}^0; \mathcal{A}^*) = |\mathcal{A}^0 \cap \mathcal{A}^*|/|\mathcal{A}^*|$. The F-measure for a given model $\mathcal{A}^0$ is defined as the harmonic mean of the precision and recall, and the G-measure is defined as the geometric mean of the two. Specifically,

$$F(\mathcal{A}^0) = F(\mathcal{A}^0; \mathcal{A}^*) \equiv \frac{2 \times pr(\mathcal{A}^0) \times re(\mathcal{A}^0)}{pr(\mathcal{A}^0) + re(\mathcal{A}^0)} = \frac{2|\mathcal{A}^0 \cap \mathcal{A}^*|}{|\mathcal{A}^0| + |\mathcal{A}^*|},$$

and

$$G(\mathcal{A}^0) = G(\mathcal{A}^0; \mathcal{A}^*) \equiv \sqrt{pr(\mathcal{A}^0) \times re(\mathcal{A}^0)} = \frac{|\mathcal{A}^0 \cap \mathcal{A}^*|}{\sqrt{|\mathcal{A}^0| \cdot |\mathcal{A}^*|}}.$$

In a penalized regression, it is well known that when the penalty level is increased, fewer active variables are selected. Therefore, false positives are less likely to happen, whereas false negatives become more likely. By taking the harmonic (or geometric) mean of the precision and recall, the F-measure (or G-measure) integrates both false-positive and false-negative aspects into a single characterization. Given $\mathcal{A}^0$, a high F- or G-measure indicates that the false-positive and false-negative rates are both low. For example, if $\mathcal{A}^* = (1, 1, 1, 0, 0, 0, 0)$ and $\mathcal{A}_1^0 = (1, 1, 1, 0, 0, 0, 1)$, then $pr(\mathcal{A}_1^0) = 3/4$, $re(\mathcal{A}_1^0) = 1$, $F(\mathcal{A}_1^0) = 6/7$, and $G(\mathcal{A}_1^0) = \sqrt{3}/2$. For the same $\mathcal{A}^*$, if we consider a worse case $\mathcal{A}_2^0 = (1, 1, 0, 0, 0, 0, 1)$, then $pr(\mathcal{A}_2^0) = 2/3$, $re(\mathcal{A}_2^0) = 2/3$, $F(\mathcal{A}_2^0) = 2/3$, and $G(\mathcal{A}_2^0) = 2/3$. The F- and G-measures are smaller than those in the first case owing to the existence of both under-selection and over-selection. In general, F- and G-measures are conservative, in the sense that both are more sensitive to under-

selection than they are to over-selection. Specifically, suppose $|\mathcal{A}^*| = m$. If $\mathcal{A}_3^0$ over-selects one variable, then $|\mathcal{A}_3^0| = m+1$, $F(\mathcal{A}_3^0) = 2m/(2m+1)$, and $G(\mathcal{A}_3^0) = \sqrt{m/(m+1)}$. However, if $\mathcal{A}_4^0$ under-selects one variable, then $|\mathcal{A}_4^0| = m-1$, $F(\mathcal{A}_4^0) = (2m-2)/(2m-1)$, and $G(\mathcal{A}_4^0) = \sqrt{(m-1)/m}$. One can easily see that $F(\mathcal{A}_3^0) > F(\mathcal{A}_4^0)$ and $G(\mathcal{A}_3^0) > G(\mathcal{A}_4^0)$.

In real applications, the true model $\mathcal{A}^*$ is usually unknown, and thus we cannot directly know $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$ for any given model $\mathcal{A}^0$. However, by borrowing information from a group of given models, we can estimate $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$ from the data. Suppose that we have a set of candidate models $\mathbb{S} = \{\mathcal{A}^1, \ldots, \mathcal{A}^K\}$, which can be obtained from a preliminary analysis. When the model size p is small, we can use a full collection of all-subset models $\mathbb{S} = \mathbb{C}$, where

$$\mathbb{C} = \{\varnothing, \{1\}, \ldots, \{p\}, \{1,2\}, \{1,3\}, \ldots, \{1, \ldots, p\}\},$$

where $1, \ldots, p$ represents the indices of the p variables. If p is too large, we can choose $\mathbb{S}$ as a group of models obtained from penalized methods, such as the Lasso, adaptive Lasso, SCAD, and MCP. Define $\mathbf{w} = \{w_1, \ldots, w_K\}$ as the corresponding data-driven weights for $\mathbb{S} = \{\mathcal{A}^1, \ldots, \mathcal{A}^K\}$, where $w_k \geq 0$, for $k = 1, \ldots, K$, and $\sum_{k=1}^K w_k = 1$. In Section 4.1, we further describe how we acquire $\mathbb{S}$ and $\mathbf{w}$. For now, we assume these are already properly acquired. For each $\mathcal{A}^k$, we define the estimated precision and recall for $\mathcal{A}^0$ (relative to $\mathcal{A}^k$) as $pr(\mathcal{A}^0; \mathcal{A}^k) = |\mathcal{A}^0 \cap \mathcal{A}^k|/|\mathcal{A}^0|$ and $re(\mathcal{A}^0; \mathcal{A}^k) = |\mathcal{A}^0 \cap \mathcal{A}^k|/|\mathcal{A}^k|$, and propose the following $\widehat{F}(\mathcal{A}^0)$ using PAVI to estimate $F(\mathcal{A}^0)$:

$$\widehat{F}(\mathcal{A}^0) = \sum_{k=1}^K w_k F(\mathcal{A}^0; \mathcal{A}^k) = 2\sum_{k=1}^K w_k \frac{|\mathcal{A}^0 \cap \mathcal{A}^k|}{|\mathcal{A}^0| + |\mathcal{A}^k|}. \tag{2.2}$$

Similarly, we propose $\widehat{G}(\mathcal{A}^0)$ using PAVI to estimate $G(\mathcal{A}^0)$:

$$\widehat{G}(\mathcal{A}^0) = \sum_{k=1}^K w_k G(\mathcal{A}^0; \mathcal{A}^k) = 2\sum_{k=1}^K w_k \frac{|\mathcal{A}^0 \cap \mathcal{A}^k|}{\sqrt{|\mathcal{A}^0| \cdot |\mathcal{A}^k|}}. \tag{2.3}$$

We define the (sample) standard deviation of $\widehat{F}(\mathcal{A}^0)$ as

$$\mathrm{sd}\big(\widehat{F}(\mathcal{A}^0) = \sqrt{\sum_{k=1}^K w_k \big(F(\mathcal{A}^0; \mathcal{A}^k) - \widehat{F}(\mathcal{A}^0)\big)^2}. \tag{2.4}$$

Similarly, the (sample) standard deviation of $\widehat{G}(\mathcal{A}^0)$ is

$$\mathrm{sd}\big(\widehat{G}(\mathcal{A}^0)\big) = \sqrt{\sum_{k=1}^{K} w_k \big(G(\mathcal{A}^0;\mathcal{A}^k) - \widehat{G}(\mathcal{A}^0)\big)^2}. \tag{2.5}$$

In (2.2) and (2.3), $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ are estimated using the candidate models $\mathcal{A}^k \in \mathbb{S}$ and weights $w_k \in \mathbf{w}$, for $k = 1, \ldots, K$. Intuitively, if higher weights w_k are assigned to those $\mathcal{A}^k$ that are closer to the true model $\mathcal{A}^*$, then $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ should better approximate the true values of $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$, respectively. In Section 4.2, we discuss the methods for computing the weights $\mathbf{w}$ from the data.

3. Theory

In this section, we show that the proposed estimators $\widehat{F}$ and $\widehat{G}$ are uniformly consistent for the true F and G, respectively, over the set of all models to be checked. The theory relies on the *weak consistency* (see Definition 1 and Nan and Yang (2014)) of the data-dependent model weights $\mathbf{w} = \{w_1, \ldots, w_K\}$, and the *weak inclusion property*, which indicates whether a model screening process is applied to reduce the model list (Definition 2).

Definition 1. (Weak consistency). The weighting vector $\mathbf{w} = (w_1, \ldots, w_K)^\intercal$ is weakly consistent if

$$\frac{\sum_{k=1}^{K} w_k \cdot |\mathcal{A}^k \nabla \mathcal{A}^*|}{|\mathcal{A}^*|} \xrightarrow{p} 0 \text{ as } n \to \infty,$$

where ∇ denotes the symmetric difference between two sets.

Remark 1. The definition basically says that $\mathbf{w}$ is sufficiently concentrated around the true model $\mathcal{A}^*$, such that the weighted deviation $|\mathcal{A}^k \nabla \mathcal{A}^*|$ eventually diminishes relative to the size of the true model. When the true model is allowed to increase in dimension as n increases, including the denominator $|\mathcal{A}^*|$ in the definition makes the condition more likely to be satisfied.

The following theorem shows that under the weak consistency condition, the estimators $\widehat{F}$ and $\widehat{G}$ are uniformly consistent (the proof is provided in the Supplementary Material).

Theorem 1. *(Uniform consistency of $\widehat{F}$ and $\widehat{G}$). Suppose the model weighting $\mathbf{w}$ is weakly consistent. Then, $\widehat{F}$ and $\widehat{G}$ based on PAVI are uniformly consistent,*

in the sense that

$$\sup_{\mathcal{A}^0\in\mathbb{C}} |\widehat{F}(\mathcal{A}^0) - F(\mathcal{A}^0)| \xrightarrow{p} 0 \quad as \;\; n \to \infty;$$
$$\sup_{\mathcal{A}^0\in\mathbb{C}} |\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)| \xrightarrow{p} 0 \quad as \;\; n \to \infty.$$

From this theorem, we see that if the model weighting focuses mostly on models that are sensibly close to the true model, then our estimated $\widehat{F}$ and $\widehat{G}$ will be close to their respective true values. Clearly, we also have $E|\widehat{F}(\mathcal{A}^0) - F(\mathcal{A}^0)| \to 0$ and $E|\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)| \to 0$, uniformly.

Theorem 2. *(Uniform convergence of* $\mathrm{sd}\big(\widehat{F}\big)$ *and* $\mathrm{sd}\big(\widehat{G}\big)$*). Suppose the model weighting* $\mathbf{w}$ *is weakly consistent. Then* $\mathrm{sd}\big(\widehat{F}\big)$ *and* $\mathrm{sd}\big(\widehat{G}\big)$ *based on PAVI converge to zero in probability uniformly, in the sense that*

$$\sup_{\mathcal{A}^0\in\mathbb{C}} |\mathrm{sd}\big(\widehat{F}(\mathcal{A}^0)\big)| \xrightarrow{p} 0 \quad as \;\; n \to \infty;$$
$$\sup_{\mathcal{A}^0\in\mathbb{C}} |\mathrm{sd}\big(\widehat{G}(\mathcal{A}^0)\big)| \xrightarrow{p} 0 \quad as \;\; n \to \infty.$$

From this theorem, we see that if the model weighting is sensible, then $\mathrm{sd}\big(\widehat{F}\big)$ and $\mathrm{sd}\big(\widehat{G}\big)$ will be close to zero. The results also support the reliability of our PAVI method.

Theorems 1 and 2 rely on the weak consistency of $\mathbf{w}$. Clearly, when the candidate models in $\mathbb{S}$ are all poor, weak consistency may not be plausible. One can choose all-subset models $\mathbb{C}$ as $\mathbb{S}$ when p is small, because it always contains $\mathcal{A}^*$. However, in the high-dimensional case, it would be computationally infeasible to use $\mathbb{C}$, and a model screening process may be applied (e.g., considering solution paths of model selection methods).

Definition 2. (Weak inclusion property). A set of candidate models $\mathbb{S}$ obtained by a model screening process is called weakly inclusive with respect to $\mathbf{w}$ on $\mathbb{C}$ if $\sum_{k\in\mathbb{S}} w_k$ is bounded away from zero in probability.

Theorem 3. *Under the assumption that the weighting vector* $\mathbf{w}$ *on the all-subset models* $\mathbb{C}$ *is weakly consistent, as long as* $\mathbb{S}$ *is weakly inclusive, the conclusions of Theorems* 1 *and* 2 *still hold.*

Remarks on this result are given in the Supplementary Material.

4. Implementation

4.1. Candidate models

We discuss how to choose the candidate models for computing $\widehat{F}$ and $\widehat{G}$. To obtain the candidate models, we can use a complete collection of all-subset models; that is, we can choose $\mathbb{S} = \mathbb{C}$. However, in the high-dimensional case, where $p \gg n$, it is almost impossible to use all subsets owing to the high computational cost.

Here, we show how to choose the candidate models for linear and logistic regression models in the high-dimensional setting. Similar procedures apply to other likelihood-based models. Given n independent observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ for the pair (X, Y), we can fit the linear or logistic regression model by minimizing the penalized negative log-likelihood

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} -\ell(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(\beta_j), \tag{4.1}$$

where $-\ell(\boldsymbol{\beta}) = (2n)^{-1}\sum_{i=1}^n (y_i - \mathbf{x}_i^\intercal \boldsymbol{\beta})^2$ for the linear regression, and

$$-\ell(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} \{-y_i \log \pi_i - (1-y_i)\log(1-\pi_i)\}$$

for the logistic regression, where $\pi_i = \Pr(Y_i = 1 | X_i = \mathbf{x}_i)$ is the probability in (2.1) for observation i. The nonnegative penalty function $p_\lambda(\cdot)$, with $\lambda \in [0, \infty)$, can be the Lasso (Tibshirani (1996)), SCAD (Fan and Li (2001)), MCP (Zhang (2010)), or some other regularizer.

We compute the models $\mathbb{S} = \{\mathcal{A}^{\lambda_1}, \ldots, \mathcal{A}^{\lambda_L}\}$ for the Lasso, SCAD, and MCP on the solution paths $\{\widehat{\boldsymbol{\beta}}^{\lambda_1}, \ldots, \widehat{\boldsymbol{\beta}}^{\lambda_L}\}$ for decreasing sequences of tuning parameters $\{\lambda_1, \ldots, \lambda_L\}$. These models are then combined as a set of candidate models $\mathbb{S} = \{\mathbb{S}_{\text{Lasso}}, \mathbb{S}_{\text{SCAD}}, \mathbb{S}_{\text{MCP}}\}$. One can efficiently compute all solution paths of the Lasso using **glmnet** (Friedman, Hastie and Tibshirani (2010)), and those of the SCAD and MCP using **ncvreg** (Breheny and Huang (2011)).

4.2. Weighting methods

There are several different methods in the literature for determining the weights $\mathbf{w} = \{w_1, \ldots, w_K\}$. For example, Buckland, Burnham and Augustin (1997) and Leung and Barron (2006) proposed information-criterion-based methods for weighting, such as those using the AIC (Akaike (1973)) and BIC (Schwarz (1978)). Hoeting et al. (1999) proposed the Bayesian model averaging (BMA)

method for weighting, and Yang (2001) studied a weighting strategy called the adaptive regression by mixing (ARM), which computes the weights using data splitting and cross-assessment. It is proven in Yang (2001) that the weighting by the ARM delivers the best rate of convergence for regression estimation. Yang (2000) also extend the ARM weighting method to the classification setting. When the number of models in the candidate-model set is fixed, the BMA weighting is consistent (and thus weakly consistent). From Yang (2007), when one properly chooses the data-splitting ratio, the ARM weighting can be consistent. More recently, Lai, Hannig and Lee (2015) proposed Fisher's fiducial-based methods for deriving probability density functions as weights on the set of candidate models. They showed that, under certain conditions, their method is consistent when p is diverging and the size of the true model is either fixed or diverging. Here, we consider only the ARM weighting and a weighting based on an information criterion.

4.2.1. Weighting using ARM for linear regression

To get the ARM weights, we randomly split the data $\mathbf{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ into a training set $\mathbf{D}_1$ and a test set $\mathbf{D}_2$ of (approximately) equal size. We train the linear regression model on $\mathbf{D}_1$ and evaluate its prediction performance on $\mathbf{D}_2$, based on which the weights $\mathbf{w} = \{w_1, \ldots, w_K\}$ can be computed. Let $\boldsymbol{\beta}_s^{(k)}$ be the sub-vector of $\boldsymbol{\beta}^{(k)}$ representing the nonzero coefficients of model $\mathcal{A}^k$, and let $\mathbf{x}_s^{(k)} \in \mathbb{R}^{|\mathcal{A}^k|}$ be the corresponding subset of selected predictors. When p is large, the ARM weighting performs poorly in terms of measuring the model deviation. One way to fix this is to add a non-uniform prior $e^{-\psi C_k}$ to the weighting computation, with

$$C_k = s_k \log \frac{ep}{s_k} + 2\log(s_k + 2), \tag{4.2}$$

where s_k is the number of non-constant predictors for model k. The first term $s_k \log ep/s_k$ is an upper bound of $\log \binom{p}{s_k}$, which characterizes which model it is among the $\log \binom{p}{s_k}$ possibilities. This is followed by

$$\binom{p}{s_k} = \frac{\prod_{j=0}^{s_k-1}(p-j)}{s_k!} \leq \frac{p^{s_k}}{s_k!} \leq \left(\frac{pe}{s_k}\right)^{s_k}, \tag{4.3}$$

using Stirling's approximation. The second term in (4.2) represents the number of variables to be estimated. From an information-theoretic perspective, C_k can be regarded as an upper bound on the descriptive complexity of model $\mathcal{A}^k$. This concept plays a crucial role in model selection theory (Yang (1999); Wang et al. (2014); Ye and Yang (2019)). In addition to this interpretation, one can also treat

$e^{-\psi C_k}$ as the prior probability assigned to the models, from a Bayesian viewpoint. The constant $\psi > 0$ controls the relative importance of the prior weight on the final weights, which can be specified by the user. From a theoretical point of view, when ψ is bigger than 5.1, the complexity term is big enough to control the selection bias, and results in minimax optimal estimations (Yang (1999)). However, the bound 5.1 is more due to technical reasons. In practice, a smaller choice often works very well. Based on previous works (Nan and Yang (2014); Ye, Yang and Yang (2018); Ye and Yang (2019)) and our own numerical studies (see Section 6 of the Supplementary Material), we found that $\psi = 1$ or 2 often delivers the best numerical results.

The ARM weighting method for linear regression models is summarized in Algorithm 1.

Algorithm 1: ARM weighting procedure for linear regression.

1 Randomly split $\mathbf{D}$ into a training set $\mathbf{D}_1$ and a test set $\mathbf{D}_2$ of equal size.

2 For each $\mathcal{A}^k \in \mathbb{S}$, fit a standard linear regression of y on $\mathbf{x}_s^{(k)}$ using the training set $\mathbf{D}_1$ and get the estimated regression coefficient $\widehat{\boldsymbol{\beta}}_s^{(k)}$ and the estimated standard deviation $\widehat{\boldsymbol{\sigma}}_s^{(k)}$.

3 For each $\mathcal{A}^k$, compute the prediction $\mathbf{x}_s^{(k)\intercal}\widehat{\boldsymbol{\beta}}_s^{(k)}$ on the test set $\mathbf{D}_2$.

4 Compute the weight w_k for each candidate model $\mathcal{A}^k$:

$$w_k = \frac{e^{-\psi C_k}(\widehat{\boldsymbol{\sigma}}_s^{(k)})^{-n/2}\prod_{(\mathbf{x}_{si}^{(k)},y_i)\in\mathbf{D}_2}\exp(-(\widehat{\boldsymbol{\sigma}}_s^{(k)})^{-2}(y_i-\mathbf{x}_s^{(k)\intercal}\widehat{\boldsymbol{\beta}}_s^{(k)})^2/2)}{\sum_{l=1}^{K} e^{-\psi C_l}(\widehat{\boldsymbol{\sigma}}_s^{(l)})^{-n/2}\prod_{(\mathbf{x}_{si}^{(l)},y_i)\in\mathbf{D}_2}\exp(-(\widehat{\boldsymbol{\sigma}}_s^{(l)})^{-2}(y_i-\mathbf{x}_s^{(l)\intercal}\widehat{\boldsymbol{\beta}}_s^{(k)})^2/2)},$$

for $k = 1, \ldots, K$, where C_k, for $k = 1, \ldots, K$ is defined in (4.2).

5 Repeat the steps above (with random data splitting) L times to get $w_k^{(l)}$, for $l = 1, \ldots, L$, and get $w_k = (1/L)\sum_{l=1}^{L} w_k^{(l)}$.

4.2.2. Weighting using ARM for logistic regression

The ARM weighting method for logistic regression models is similar. We summarize it in Algorithm 2.

4.2.3. Weighting using modified BIC for linear and logistic regression

Information criteria such as the BIC can be used as alternative ways for computing the weights. Let ℓ_k be the maximized likelihood for model k. Recall that the BIC is $I_k^{\mathrm{BIC}} = -2\log\ell_k + s_k \log n$. To accommodate the huge number of models, an extra term was added by Yang and Barron (1998) to reflect the additional price one needs to pay for searching through all the models. Including

Algorithm 2: ARM weighting procedure for logistics regression.

1 Randomly split $\mathbf{D}$ into a training set $\mathbf{D}_1$ and a test set $\mathbf{D}_2$ of equal size.

2 For each $\mathcal{A}^k \in \mathbb{S}$, fit a standard logistic regression of y on $\mathbf{x}_s^{(k)}$ using the data in $\mathbf{D}_1$ and get the estimated conditional probability function $\hat{p}^{(k)}(\mathbf{x}_s^{(k)})$, for $k = 1, \ldots, K$,

$$\hat{p}^{(k)}(\mathbf{x}_s^{(k)}) \equiv \Pr(Y = 1 | X_s^{(k)} = \mathbf{x}_s^{(k)}) = \frac{\exp(\mathbf{x}_s^{(k)\intercal} \widehat{\boldsymbol{\beta}}_s^{(k)})}{1 + \exp(\mathbf{x}_s^{(k)\intercal} \widehat{\boldsymbol{\beta}}_s^{(k)})}.$$

3 For each $\mathcal{A}^k$, evaluate $\hat{p}^{(k)}(\mathbf{x}_s^{(k)})$ on the test set $\mathbf{D}_2$.

4 Compute the weight w_k for each model $\mathcal{A}^k$ in the candidate models:

$$w_k = \frac{e^{-\psi C_k} \prod_{(\mathbf{x}_{s,i}^{(k)}, y_i) \in \mathbf{D}_2} \hat{p}^{(k)}(\mathbf{x}_{s,i}^{(k)})^{y_i} \left(1 - \hat{p}^{(k)}(\mathbf{x}_{s,i}^{(k)})\right)^{1-y_i}}{\sum_{l=1}^{K} e^{-\psi C_l} \prod_{(\mathbf{x}_{s,i}^{(l)}, y_i) \in \mathbf{D}_2} \hat{p}^{(l)}(\mathbf{x}_{s,i}^{(l)})^{y_i} \left(1 - \hat{p}^{(l)}(\mathbf{x}_{s,i}^{(l)})\right)^{1-y_i}}, \ k = 1, \ldots, K.$$

5 Repeat the steps above (with random data splitting) L times to get $w_k^{(l)}$, for $l = 1, \ldots, L$, and get $w_k = (1/L) \sum_{l=1}^{L} w_k^{(l)}$.

this extra term, we calculate the weights using a modified BIC (BIC-p) informationcriterion:

$$w_k = \frac{\exp(-I_k/2 - \psi C_k)}{\sum_{l=1}^{K} \exp(-I_l/2 - \psi C_l)}, \ k = 1, \ldots, K. \tag{4.4}$$

5. Simulation

In order to study the performance of the estimated F- and G-measures, we conduct simulations for several well-known variable selection methods (for both regression and classification) under various settings. We consider numerical experiments for both the $n < p$ and the $n \geq p$ cases, with specified structural feature correlation (i.e., independent/correlated). We also consider special settings of the true coefficients, such as decaying coefficients.

5.1. Setting I: regression models

For the regression case, the response Y is generated from the following model:

$$Y = X\boldsymbol{\beta} + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$. To study how the estimation performance varies with the noise level σ^2, we choose nine σ-values, evenly spaced between 0.01 and 5. The

predictors $\mathbf{x}_i$ and the coefficient vector $\boldsymbol{\beta}$ are generated according to the following settings:

Example 1. $n = 200$, $p = 8$, $\boldsymbol{\beta} = (3, 1.5, 2, 0, 0, 0, 0, 0)^\intercal$. Predictors $\mathbf{x}_i$, for $i = 1, \ldots, n$ are generated as n independent and identically distributed (i.i.d.) observations from $N(0, \boldsymbol{I}_p)$.

Example 2. Same as Example 1, except $n = 1{,}000$.

Example 3. $n = 200$, $p = 2{,}000$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$, where $(\beta_1, \beta_2, \beta_3) = (3, 1.5, 2)$ and $(\beta_4, \ldots, \beta_{2000})$ are zero. Predictors $\mathbf{x}_i$, for $i = 1, \ldots, n$, are sampled as n i.i.d. observations from $N(0, \boldsymbol{I}_p)$.

Example 4. $n = 200$, $p = 30$, components 1–5 of $\boldsymbol{\beta}$ are 10.5, components 6–10 are 5.5, components 11–15 are 0.5, and the rest are zero. Therefore, there are 15 nonzero predictors, including five large ones, five moderate ones, and five small ones. Predictors $\mathbf{x}_i$, for $i = 1, \ldots, n$, are generated from $X \sim N_p(0, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = (0.4^{|j-k|})_{p \times p}$; thus, the pairwise correlation between X_j and X_k is $0.4^{|j-k|}$.

Example 5. $n = 200$, $p = 200$, components 1–5 of β are 10.5, components 6–10 are 5.5, components 11–15 are 0.5, and the rest are zero. Predictors $\mathbf{x}_i$, for $i = 1, \ldots, n$, are generated from $X \sim N_p(0, \boldsymbol{\Sigma})$. The covariance structure $\boldsymbol{\Sigma}$ is set as follows: the first 15 predictors $(X_1, \ldots, X_{15})$ and the remaining 185 predictors $(X_{16}, \ldots, X_{200})$ are independent. The pairwise correlation between X_j and X_k in $(X_1, \ldots, X_{15})$ is $0.4^{|j-k|}$, with $j, k = 1, \ldots, 15$. The pairwise correlation between X_j and X_k in $(X_{16}, \ldots, X_{200})$ is $0.4^{|j-k|}$, with $j, k = 16, \ldots, 200$.

We apply four penalized methods, namely, the Lasso, adaptive Lasso, MCP, and SCAD to the data from Examples 1–5, and denote the resulting models as $\mathcal{A}^{\text{Lasso}}$, $\mathcal{A}^{\text{AdLasso}}$, $\mathcal{A}^{\text{MCP}}$, and $\mathcal{A}^{\text{SCAD}}$, respectively. We use **glmnet** to compute $\mathcal{A}^{\text{Lasso}}$ and $\mathcal{A}^{\text{AdLasso}}$, and **ncvreg** for computing $\mathcal{A}^{\text{MCP}}$ and $\mathcal{A}^{\text{SCAD}}$. Five-fold cross-validation is used for penalty parameter tuning in all these procedures. Because we know the true model $\mathcal{A}^* = \{j : \beta_j \neq 0\}$ in the simulation, we can report the true $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$ measures for each model $\mathcal{A}^0 \in \{\mathcal{A}^{\text{Lasso}}, \mathcal{A}^{\text{AdLasso}}, \mathcal{A}^{\text{MCP}}, \mathcal{A}^{\text{SCAD}}\}$. For comparison, we also compute the estimated $\widehat{F}$ and $\widehat{G}$ using two different weighting methods, the ARM and the BIC-p (modified BIC), with prior adjustment $\psi = 1$. The number of observations in the training set used to compute the ARM weight is half of the sample size $\lfloor n/2 \rfloor$, and the corresponding repetition time is 100.

All simulation cases are repeated 100 times, and the corresponding values are computed and averaged. We compare $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ with the true $F(\mathcal{A}^0)$ and

$G(\mathcal{A}^0)$ in Figure 1 for Example 1, and in Figures A1–A4 of the Supplementary Material for Examples 2–5. Overall, $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ using the ARM and the BIC-p weightings well reflect the trends of $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$, in the sense that both the true curves and the estimated curves trend down as σ^2 increases. Furthermore, the estimation accuracy drops as σ^2 increases. The estimated $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ properly reflect the true performance of a given $\mathcal{A}^0$. For example, in Figures A2, A3, and A4, we see that the performance of the Lasso deteriorates significantly as σ^2 increases, because it tends to over-select variables under higher noise levels. In contrast, the adaptive Lasso, MCP, and SCAD have more robust performance against high noise levels. $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ correctly reflect these aforementioned facts. From the results, we find that the MCP performs best, with the highest true/estimated F- and G-measures in Examples 2–5, while the adaptive Lasso performs best in Example 1.

Comparing Figures 1 and A1, we see that the sample size influences the estimation performance: large samples produce more accurate $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$. The gain in the estimation accuracy from an increased sample size is because more information results in better assigned weights on the candidate models.

In Figure A4, the over-estimation in the adaptive Lasso, SCAD, and MCP when σ is large occurs because highly weighted candidate models miss several small coefficients variables, which is caused by the decaying coefficients, and worsened by correlation between the variables. For the Lasso, when σ is small, PAVI identifies good candidate models on which to put high weights. Thus, the estimation is good. When σ is larger, the candidate models with high weights miss several true variables. At the same time, the Lasso chooses more redundant variables as σ becomes larger. Therefore, the precision is under-estimated, as is the F-measure.

5.2. Setting II: classification models

For the classification case, we randomly generate n i.i.d observations $\{y_i, \mathbf{x}_i\}_{i=1}^n$. Each binary response $y_i \in \{0, 1\}$ is generated from a Bernoulli distribution with conditional probability $\Pr(Y = 1|X = \mathbf{x}_i) = \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta})/(1 + \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}))$. The explanatory variables X and the coefficient vector $\boldsymbol{\beta}$ are set under the same configurations as in Examples 1–5.

The absolute differences between the true and estimated measures,

$$d_F = |\widehat{F}(\mathcal{A}^0) - F(\mathcal{A}^0)| \quad \text{and} \quad d_G = |\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)|,$$

are used to evaluate the estimation performance, where a smaller d_F and d_G

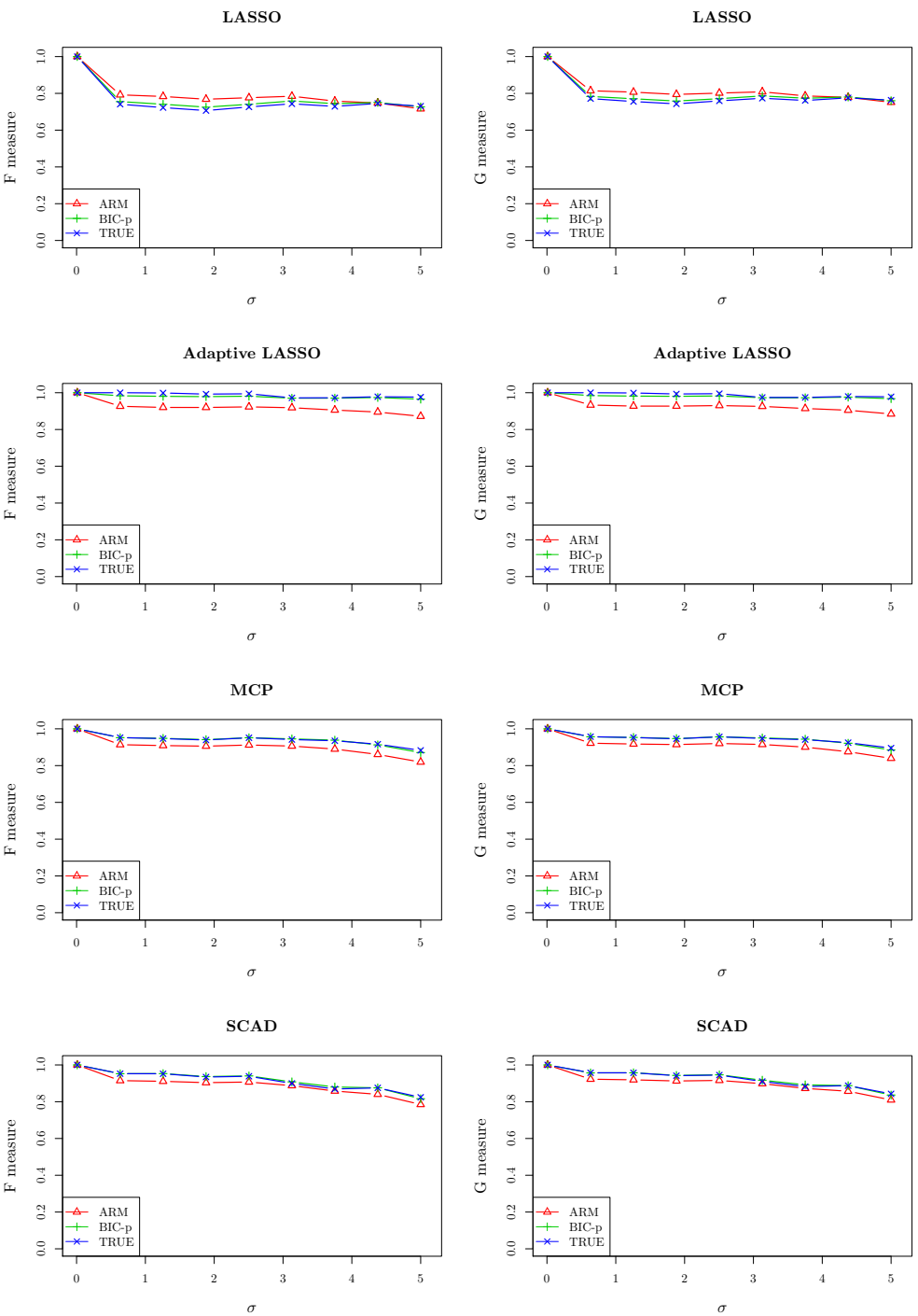

Figure 1. Regression case (Example 1).

indicate better estimation performance.

All simulation cases are repeated 100 times, and the corresponding $F(\mathcal{A}^0)$, $G(\mathcal{A}^0)$, $\widehat{F}(\mathcal{A}^0)$, $\widehat{G}(\mathcal{A}^0)$, d_F, and d_G values are computed and averaged. The results are summarized in Table 1 for Example 1, and in Tables A1–A4 of the Supplementary Material for Examples 2–5. The standard errors are also shown (in parentheses). As shown in the tables, d_F and d_G are generally small, which indicates that the estimated $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ are good approximations of the true $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$, respectively. The estimated $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ reflect the true advantage of a given variable selection method. For example, in Table 1, and in Tables A1–A4, the adaptive Lasso, MCP, and SCAD have better variable selection performance than that of the Lasso, according to their larger true values of $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$. The estimated $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ correctly reflect these differences in performance.

Our estimation method still performs very well in the high-dimensional setting, as can be seen from the small d_F and d_G in Table A2. However, the results from Tables 4 and 5 show that the decaying coefficients and feature correlation make the estimation of $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ more difficult. In these two cases, the BIC-p methods tend to overestimate $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$ for the MCP and SCAD, whereas the ARM tends to underestimate $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$ for the Lasso and adaptive Lasso.

The overestimation problem of the BIC-p method mainly comes from that of the recall part. The final model selected by the SCAD misses several true variables; thus, the true recall is very small. However, if we were to use the heavily weighted candidate models that miss several true variables in the PAVI calculation, the recall would be overestimated.

For the SCAD and ARM combination, using the heavily weighted models that miss several true variables in PAVI will overestimate of the recall and underestimate the precision, although these two effects cancel each other to some degree.

The underestimation by the ARM methods mainly comes from that of the precision part, while the estimated recall is close to (slightly overestimates) the true recall. The Lasso tends to miss true variables and over-select redundant variables in the examples. Thus, the true precision of the Lasso is small.

For the Lasso and BIC combination, using the heavily weighted models that miss several true variables with small coefficients in PAVI overestimates the recall and underestimates the precision, although these two effects cancel each other to some degree.

Both issues are mainly caused by the fact that the candidate models with

Table 1. Classification case (Example 1).

	F	G	d_F	d_G
		Lasso		
True	0.670 (0.010)	0.712 (0.009)		
ARM	0.711 (0.009)	0.747 (0.007)	0.046 (0.003)	0.039 (0.002)
BIC-p	0.687 (0.010)	0.726 (0.008)	0.017 (0.002)	0.014 (0.001)
		AdLasso		
True	0.944 (0.009)	0.949 (0.008)		
ARM	0.899 (0.004)	0.908 (0.004)	0.066 (0.003)	0.060 (0.003)
BIC-p	0.946 (0.007)	0.950 (0.007)	0.018 (0.002)	0.016 (0.001)
		MCP		
True	0.968 (0.009)	0.971 (0.008)		
ARM	0.903 (0.005)	0.913 (0.004)	0.079 (0.003)	0.072 (0.002)
BIC-p	0.961 (0.007)	0.965 (0.006)	0.019 (0.002)	0.017 (0.001)
		SCAD		
True	0.902 (0.012)	0.911 (0.010)		
ARM	0.881 (0.006)	0.892 (0.006)	0.054 (0.003)	0.050 (0.003)
BIC-p	0.911 (0.010)	0.919 (0.009)	0.018 (0.002)	0.016 (0.001)

large weights cannot recover all the variables with small true coefficients. Then, the problem is worsened by the high correlation between the features.

6. Real Data

In this section, we apply PAVI using candidate models from several model selection methods to gene expression data for cancer-related biomarker identification. The biomarker selection process is usually under a high-dimensional, small-sample, and high-noise setting involving highly correlated genes (Golub et al. (1999); West et al. (2001)). As such, the sets of genes identified may be subject to substantial changes, owing to small perturbations in the data (Baggerly, Morris and Coombes (2004); Henry and Hayes (2012)). Here, we use $\widehat{F}$ and $\widehat{G}$ to evaluate such selection uncertainty.

Our goal is to provide a serious and careful analysis of the outcomes of several variable selection methods from multiple angles to understand the key issues of interest. We hope our analysis provides strong enough evidence that the estimated F and G values yield valuable information.

6.1. Data description

We consider three well-studied benchmark cancer data sets: `Colon` (Alon et al. (1999)), `Leukemia` (Golub et al. (1999)), and `Prostate` (Singh et al. (2002)).

Table 2. Summary of `Colon`, `Leukemia`, `Prostate`.

Data	n	n_1 $(y=1)$	n_2 $(y=0)$	p (number of genes)	Data source
`Colon`	62	40	22	2,000	Alon et al. (1999)
`Leukemia`	72	25	47	7,129	Golub et al. (1999)
`Prostate`	102	52	50	12,600	Singh et al. (2002)

Table 2 provides a brief summary.

6.2. Methods/Models to be examined

Using the three datasets, we compare the variable selection performance of four commonly used penalization methods: the Lasso, adaptive Lasso, MCP, and SCAD. We first obtain the final model $\mathcal{A}^0$ for each method, where the tuning parameter λ is selected using five-fold cross-validation. Then, we use PAVI to estimate $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ with two weighting schemes, ARM and BIC-p. The procedure is repeated 100 times to average out randomness in the tuning parameter selection, and the averages of $\widehat{F}(\mathcal{A}^0)$, $\mathrm{sd}\big(\widehat{F}(\mathcal{A}^0)\big)$ and $\widehat{G}(\mathcal{A}^0)$, $\mathrm{sd}\big(\widehat{G}(\mathcal{A}^0)\big)$ are summarized in Tables 3, 4, and A5. For comparison, we also include several other models studied in the existing literature. Specifically, we consider Leung and Hung, 2010 (L10), Yang and Song, 2010 (Y10), Chandra and Gupta, 2011 (C11), and Lee and Leu, 2011 (L11) for `Colon`, Leung and Hung, 2010 (L10), Yang and Song, 2010 (Y10), and Ji, Yang and You, 2011 (J11; two kinds of models are provided using different importance criteria in this work, denoted by $\mathrm{J11}^1$ and $\mathrm{J11}^2$, respectively) for `Leukemia`, and Leung and Hung, 2010 (L10) and Sharma, Imoto and Miyano, 2012 (S12) for `Prostate`.

Y10, J11, and S12 use linear-based variable selection techniques without initial variable screening. Specifically, Y10 uses a probit regression model, J11 uses a linear kernel support vector classifier (SVC), and S12 uses the linear discriminant analysis (LDA) technique with nearest centroid classifier (NCC). In contrast, L10, C11, and L11 use nonparametric variable selection techniques: L10 uses a support vector machine (SVM), C11 uses a naïve Bayes classifier (NBC), and SVM, and L11 uses an SVM. In addition, we consider the importance screening method (ImpS) of Ye, Yang and Yang (2018), which uses sparsity-oriented importance learning for variable screening.

6.3. Results

The estimated $\widehat{F}$ and $\widehat{G}$ of each model on `Colon`, `Leukemia`, and `Prostate` are reported in Tables 3, 4, and A5 (in the Supplementary Material), respectively.

Table 3. Estimated F- and G-measures and standard deviations for `Colon`. L10 has numerically zero $\widehat{F}$ and $\widehat{G}$ values (shown in bold).

	ARM				BIC-p			
	F	$sd.F$	G	$sd.G$	F	$sd.F$	G	$sd.G$
Lasso	0.147	0.024	0.280	0.022	0.205	0.066	0.332	0.058
AdLasso	0.194	0.165	0.255	0.211	0.309	0.191	0.361	0.209
MCP	0.349	0.045	0.459	0.035	0.460	0.130	0.544	0.093
SCAD	0.149	0.032	0.274	0.039	0.211	0.074	0.331	0.071
ImpS	0.524	0.081	0.596	0.065	0.656	0.176	0.698	0.118
L11	0.111	0.110	0.175	0.175	0.112	0.105	0.157	0.151
Y10	0.103	0.017	0.233	0.018	0.146	0.048	0.276	0.047
C11	0.184	0.020	0.317	0.022	0.223	0.076	0.333	0.082
L10	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**

Table 4. Estimated F- and G-measures and standard deviations for `Leukemia`. J11[1] and J11[2] have numerically zero $\widehat{F}$ and $\widehat{G}$ values (shown in bold).

	ARM				BIC-p			
	F	$sd.F$	G	$sd.G$	F	$sd.F$	G	$sd.G$
Lasso	0.083	0.025	0.206	0.026	0.079	0.012	0.203	0.014
AdLasso	0.323	0.044	0.432	0.031	0.322	0.039	0.434	0.033
MCP	0.168	0.170	0.221	0.210	0.061	0.089	0.078	0.108
SCAD	0.094	0.028	0.220	0.028	0.090	0.013	0.216	0.015
ImpS	0.525	0.065	0.591	0.042	0.573	0.129	0.636	0.102
J11[1]	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
J11[2]	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Y10	0.108	0.014	0.236	0.009	0.105	0.002	0.233	0.012
L10	0.212	0.180	0.265	0.224	0.336	0.089	0.419	0.110

We find that ImpS achieves almost the largest estimated $\widehat{F}$ and $\widehat{G}$ on all three data sets. L10 has basically zero $\widehat{F}$ and $\widehat{G}$ for `Colon` and `Prostate`. J11[1] and J11[2] have basically zero $\widehat{F}$ and $\widehat{G}$ for `Leukemia`. (These cases are shown in bold in Tables 3, 4, and A5.) This suggests that, from a logistic regression modeling perspective, they may have chosen "wrong" variables and have very low recall or precision.

6.4. Are the zero $\widehat{F}$ and $\widehat{G}$ values too harsh for the methods?

It is striking that the $\widehat{F}$ and $\widehat{G}$ values for some selections are numerically zero, which seems rather extreme. Does this mean those models are truly poor, or does it mean our performance assessment methodology fails? We examine the matter from three perspectives.

6.4.1. First perspective: the labels of the selected genes

First, let us examine the labels of the selected genes. We obtain the selected genes in the literature. We use five-fold cross-validation for the penalty parameter tuning to obtain selected genes for the penalized regression models. In Tables A6, A7, and A8, the results show that the genes selected by L10 (`Colon` and `Prostate`), $J11^1$, and $J11^2$ (`Leukemia`) are mostly not supported by other models. More specifically, the choices of variables by L10, $J11^1$, and $J11^2$ in those cases share zero, one, or at most two genes with the other methods, respectively. (These cases are underlined in Tables A6, A7, and A8.)

6.4.2. Second perspective: predictive accuracy

Here, we examine the issue from a predictive accuracy perspective. We randomly split the data set into 4/5 observations for training, and 1/5 observations for testing. We fit the SVM models using the selected genes on the training data using **kernlab** (Karatzoglou et al. (2004)), and evaluate the predictive accuracy on the testing data. The procedure is repeated 100 times, and the averaged classification accuracy and "standard errors" (w.r.t. the permutations) are recorded in Table 5. Alternatively, we may consider parametric models. We fit the logistic regression using the selected genes (in Table 5). We find that L10, $J11^1$, and $J11^2$ have worse predictive accuracy (shown in bold in Table 5) than that of the simpler model selected by ImpS, supporting the validity of their low $\widehat{F}$ and $\widehat{G}$ values.

6.4.3. Third perspective: traditional model fitting

For the third perspective, we investigate the AIC, BIC, and deviance measures. When comparing models fitted using the maximum likelihood to the same data, the smaller the AIC or BIC value, the better is the model.

From Table 6, the model for `Colon` with zero $\widehat{F}$ and $\widehat{G}$ values also has relatively large AIC, BIC, and deviance values (shown in bold in the table) compared with those of the models with large $\widehat{F}$ and $\widehat{G}$ values. The results are similar for the other two data sets, except that the deviance values for `Leukemia` are extremely small, owing to the easy classification nature of the data.

In summary, we see that the low (near zero) $\widehat{F}$ and $\widehat{G}$ values for the investigated sets of selected genes are supported from the three perspectives. Our PAVI approach provides a valid tool for checking the reliability and reproducibility of a given set of selected variables when the true model is not known. To be fair, we want to emphasize that the poor $\widehat{F}$ and $\widehat{G}$ values of some of the selection methods are based on the logistic regression perspective, although Table 5 seems

Table 5. Comparisons of classification accuracy on `Colon`, `Leukemia`, and `Prostate` using a logistic regression and an SVM.

Logistic Model					
`Colon`		`Leukemia`		`Prostate`	
ImpS	86.3 (0.8)	ImpS	97.1 (0.3)	ImpS	94.0 (0.4)
Lasso	80.0 (1.0)	Lasso	99.8 (0.1)	Lasso	97.0 (0.4)
AdLasso	85.5 (0.8)	AdLasso	93.9 (0.5)	AdLasso	99.8 (0.1)
MCP	85.1 (0.8)	MCP	99.5 (0.1)	MCP	98.7 (0.2)
SCAD	84.3 (0.8)	SCAD	97.9 (0.3)	SCAD	97.1 (0.2)
L11	80.4 (0.8)	J11[1]	**89.4 (0.8)**	S12	96.5 (0.4)
Y10	90.9 (0.9)	J11[2]	**89.8 (0.7)**	L10	**59.0 (0.8)**
C11	79.6 (1.0)	Y10	91.2 (0.7)		
L10	**83.0 (0.9)**	L10	95.5 (0.4)		
SVM Model					
`Colon`		`Leukemia`		`Prostate`	
ImpS	84.0 (0.9)	ImpS	97.6 (0.3)	ImpS	95.3 (0.4)
Lasso	75.8 (0.9)	Lasso	99.1 (0.2)	Lasso	96.3 (0.4)
AdLasso	79.0 (0.9)	AdLasso	95.8 (0.4)	AdLasso	96.6 (0.3)
MCP	83.1 (1.1)	MCP	99.0 (0.2)	MCP	97.1 (0.3)
SCAD	86.0 (0.9)	SCAD	99.1 (0.2)	SCAD	96.4 (0.3)
L11	79.0 (1.1)	J11[1]	**88.6 (0.8)**	S12	95.5 (0.4)
Y10	78.3 (1.0)	J11[2]	**87.4 (0.9)**	L10	**59.3 (0.9)**
C11	77.1 (0.9)	Y10	90.2 (0.6)		
L10	**72.4 (0.9)**	L10	92.2 (0.6)		

Table 6. Estimated AIC, BIC, and deviance for `Colon`, `Leukemia`, and `Prostate`.

`Colon`				`Leukemia`				`Prostate`			
	AIC	BIC	Dev.		AIC	BIC	Dev.		AIC	BIC	Dev.
Lasso	26.0	53.6	0.0		56.0	119.7	0.0		62.0	143.3	0.0
AdLasso	34.9	49.8	20.9		12.0	25.6	0.0		22.0	50.8	0.0
MCP	32.1	44.9	20.1		16.0	34.2	0.0		16.0	36.9	0.0
SCAD	26.0	53.6	0.0		48.0	102.6	0.0		38.0	87.8	0.0
ImpS	35.5	44.1	27.5		8.0	17.1	0.0		12.0	27.7	9.4
L11	51.4	70.5	33.4	J11[1]	**20.0**	**42.7**	**0.0**	S12	36.1	49.2	26.1
Y10	40.0	82.5	0.0	J11[2]	**18.0**	**38.4**	**0.0**	L10	**140.1**	**158.5**	**126.1**
C11	45.2	68.6	23.2	Y10	38.0	81.2	0.0				
L10	**48.6**	**63.5**	**34.6**	L10	10.0	21.3	0.0				

to suggest that a logistic regression works at least as well as an SVM.

7. Conclusion

Despite there being many variable selection methods, most investigations of their behaviors are limited to theoretical studies and somewhat scattered simulation results, which may have little to do with a specific data set. There is a severe lack of valid performance measures that are computable based on data alone. This leads to the pessimistic view that for real data, nothing can be said strongly about which method is better for describing the data generation mechanism since no one knows the truth. Sound implementable variable selection diagnostic tools can provide insight into the matter.

Nan and Yang (2014) proposed an approach to investigate how many variables are likely missed, and how many are not quite justifiable for an outcome of a variable selection process. In real applications, it is often of interest and important to summarize the two types of selection errors into a single measure to characterize the behavior of a variable selection method. As a result, F- and G-measures are gaining in popularity in the model selection literature. If we are given a data set for which several model selection methods are considered, prior to this work, the available model diagnostic tools could only tell us (a) which methods were more unstable, and (b) how many terms are likely missed or unsupported. This information, unlike the F- and G-measures, may not be enough to give one a good sense of the overall model selection performance. In this study, we have advanced this line of research on model selection diagnostics by providing a valid estimation of F- and G-measures.

We have proved that the estimated F- and G-measures are uniformly consistent, as long as the weighting is weakly consistent. The simulation results clearly show that the $\widehat{F}$ and $\widehat{G}$ values based on our PAVI approach nicely characterize the overall performance of the model selection outcomes. This information can be used to compare different methods for the data at hand.

We used three real-data examples to demonstrate the utility of our PAVI methodology. Many variable selection results have been reported in the literature based on these data sets. A careful study with multiple perspectives has provided strong evidence to suggest that some of the variable selection outcomes may be far removed from the best set of variables to use for a logistic regression or an SVM with the given information.

Supplementary Material

All proofs are relegated to the online Supplementary Material, along with additional numerical results.

Acknowledgments

The authors thank the editor, an associate editor, and two anonymous referees for their helpful comments and suggestions. The work of Yi Yang was partially supported by NSERC RGPIN-2016-05174 and FRQNT NC-205972.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, 267–281. Akadémiai Kiadó, Budapest.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences* **96**, 6745–6750.

Baggerly, K. A., Morris, J. S. and Coombes, K. R. (2004). Reproducibility of SELDI-TOF protein patterns in serum: Comparing datasets from different experiments. *Bioinformatics* **20**, 777–785.

Billsus, D. and Pazzani, M. J. (1998). Learning collaborative information filters. In *The Proceedings of the 15th International Conference on Machine Learning* (Edited by J. W. Shavlik), 46–54. Morgan Kaufmann Publishers, San Francisco.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics* **5**, 232–253.

Breiman, L. (1996a). Bagging predictors. *Machine Learning* **24**, 123–140.

Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24**, 2350–2383.

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

Chandra, B. and Gupta, M. (2011). An efficient statistical feature selection approach for classification of gene expression data. *Journal of Biomedical Informatics* **44**, 529–535.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **158**, 419–466.

Chen, L., Giannakouros, P. and Yang, Y. (2007). Model combining in factorial data analysis. *Journal of Statistical Planning and Inference* **137**, 2920–2934.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **57**, 45–97.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101–148.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P. et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

Henry, N. L. and Hayes, D. F. (2012). Cancer biomarkers. *Molecular Oncology* **6**, 140–146.

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

Ji, G., Yang, Z. and You, W. (2011). PLS-based gene selection and identification of tumor-specific genes. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **41**, 830–841.

Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A. (2004). kernlab - An S4 package for kernel methods in R. *Journal of Statistical Software* **11**, 1–20.

Lai, R. C. S., Hannig, J. and Lee, T. C. M. (2015). Generalized fiducial inference for ultrahigh-dimensional regression. *Journal of the American Statistical Association* **110**, 760–772.

Lee, C.-P. and Leu, Y. (2011). A novel hybrid feature selection method for microarray data analysis. *Applied Soft Computing* **11**, 208–213.

Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* **52**, 3396–3410.

Leung, Y. and Hung, Y. (2010). A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **7**, 108–117.

Lim, C. (2011). *Modeling High Dimensional Data: Prediction, Sparsity, and Robustness*. Ph.D. thesis, University of California, Berkeley.

Lim, C. and Yu, B. (2016). Estimation stability with cross-validation (ESCV). *Journal of Computational and Graphical Statistics* **25**, 464–492.

McNutt, M. (2014). Raising the bar. *Science* **345**, 9.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**, 1436–1462.

Nan, Y. and Yang, Y. (2014). Variable selection diagnostics measures for high-dimensional regression. *Journal of Computational and Graphical Statistics* **23**, 636–656.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Sharma, A., Imoto, S. and Miyano, S. (2012). A top-r feature selection algorithm for microarray gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9**, 754–764.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C. et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203–209.

Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Application* **2**, 1–19.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **58**, 267–288.

Wang, Z., Paterlini, S., Gao, F. and Yang, Y. (2014). Adaptive minimax regression estimation over sparse lq-hulls. *The Journal of Machine Learning Research* **15**, 1675–1711.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R. et al. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences* **98**, 11462–11467.

Yang, A.-J. and Song, X.-Y. (2010). Bayesian variable selection for disease classification using gene expression data. *Bioinformatics* **26**, 215–222.

Yang, Y. (1999). Model selection for nonparametric regression. *Statistica Sinica* **9**, 475–499.

Yang, Y. (2000). Adaptive estimation in pattern recognition by combining different procedures.

Statistica Sinica **10**, 1069–1090.

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

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics* **35**, 2450–2473.

Yang, Y. and Barron, A. R. (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory* **44**, 95–116.

Ye, C. and Yang, Y. (2019). High-dimensional adaptive minimax sparse estimation with interactions. *IEEE Transactions on Information Theory* **65**, 5367–5379.

Ye, C., Yang, Y. and Yang, Y. (2018). Sparsity oriented importance learning for high-dimensional linear regression. *Journal of the American Statistical Association* **113**, 1797–1812.

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

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Yanjia Yu

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.

E-mail: yuxxx748@umn.edu

Yi Yang

Department of Mathematics and Statistics, McGill University, Montreal, Quebec H3A 0B9, Canada.

E-mail: yi.yang6@mcgill.ca

Yuhong Yang

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.

E-mail: yangx374@umn.edu

(Received July 2019; accepted August 2020)