Abstract

There exist many classification methods including random forest, boosting, and neural network. However, to our best knowledge, there is no existing method that can assess the goodness-of-fit of a general classification procedure. Lacking a parametric assumption makes it hard to construct a test statistic for this problem. To overcome this difficulty, we propose a methodology called BAGofT that splits the data into a training set and a test set. We first fit the classification procedure to assess and adaptively discover its potential underfitting on the training set. Then
we calculate the test statistic by the test set based on the results from the training set. The data splitting guarantees the test’s size control under the null hypothesis and consistency (power goes to one) under the alternative hypothesis. For testing parametric classification models, the BAGofT has a broader scope compared to the existing methods since it is not restricted to specific models like logistic regression. Additionally, simulation studies show that the BAGofT can be favorably better than some popular goodness-of-fit tests.

*Keywords:* goodness-of-fit test, classification procedure, adaptive partition
1 Introduction

Goodness-of-fit tests aim to assess the potential discrepancy between the observed data and the fitted model. In the field of binary classification, which is the focus of our work, the classical Pearson’s chi-squares ($\chi^2$) test and residual deviance test group the observations according to distinct covariate values. When the number of observations in each group is small, which will happen if we have a continuous covariate, the two tests are improper. Various tests have been proposed to address this situation. These include the tests based on the distribution of the Pearson’s $\chi^2$ statistic under sparse data (McCullagh, 1985; Osius and Rojek, 1992; Farrington, 1996), kernel smoothed residuals (Le Cessie and Van Houwelingen, 1991), the comparison with a generalized model (Stukel, 1988), the comparison between an estimator from the control data and an estimator from the joint data in the context of case-control studies (Bondell, 2007), the Pearson-type statistics calculated from bootstrap samples (Yin and Ma, 2013), information matrix tests (White, 1982; Orme, 1988), and grouping observations into a finite number of sets (Hosmer and Lemeshow, 1980; Pigeon and Heyse, 1999; Pulkstenis and Robinson, 2002; Xie et al., 2008; Liu et al., 2012). Lu and Yang (2019) proposed a method to assess goodness-of-fit using cross-validation to compare the parametric classification model to assess with a nonparametric model.

However, there are two weaknesses of the existing goodness-of-fit tests. Firstly, the majority of the tests only control the Type-I error under the null hypothesis without theoretical guarantees about the testing power. Secondly, existing methods are rigid. They focus on the goodness-of-fit of a specific model like logistic regression or generalized linear model without providing adaptiveness that work broadly.

To our best knowledge, there is no existing method to assess the goodness-of-fit of general classification procedures such as decision trees, neural networks, $k$-nearest neighbors,
and support vector machines. We focus on the case where the procedures provide an estimate of the conditional probability. The performance of these general procedures is usually assessed by metrics based on classification accuracy rates. However, these metrics conceptually do not characterize the deviation of a procedure to the underlying data generating process. For instance, when the conditional probabilities of two classes is almost one half, the best possible fit which is random guess gives poor classification accuracy. Moreover, the classification procedure to assess is often selected from many competing models according to their prediction accuracy on validation data. Selected from the comparison of many candidate models, the winner may be an overfitted model. The above considerations motivate the following problem. How to generalize the notion of the goodness-of-fit test to assess general classification procedures? A tool of that kind will guide data analysts to understand the selected procedure’s possible deviation away from the underlying truth.

To address the above challenges for testing the goodness-of-fit of parametric classification models and general classification procedures, we propose a new methodology named the binary adaptive goodness-of-fit test (BAGofT). For a general classification procedure, there is no workable saturated model to compare with, as is done in Pearson’s chi-squares and deviance based testings. The BAGofT uses a cross-validation type of data splitting, which is an effective and reliable way to overcome this difficulty. On the first ‘training’ set, the BAGofT applies an adaptive partition of the data space that highlights the potential underfitting of the model or procedure to assess. Then the BAGofT calculates a Pearson-type test statistic on the remaining ‘test’ part of the data based on the grouping from the adaptive partition.

For parametric classifications, the BAGofT has a theoretical guarantee for its consistency under a general range of alternative hypotheses, including those of misspecified pre-
dictors and model structures. Moreover, the adaptive nature of the BAGofT can flexibly expose different kinds of weaknesses from the parametric classification model to test.

For a general classification procedure where there is no true model or saturated model to compare with, one major challenge is to define the goodness-of-fit. Unlike parametric models, whose convergence is well understood, general classification procedures can have different rates of convergence. If we choose the splitting ratio of the BAGofT according to a specific rate, then the test size is controlled if the classification procedure to assess converges no slower than it, and the BAGofT consistently reject otherwise. In practice, if the convergence rate of the procedure to assess is unknown, we propose a method that directly applies the BAGofT with different splitting ratios. Our simulation show that this method can reveal different convergence speeds of the procedure to assess.

The outline of the paper is given below. In Section 2, we provide the background of the problem and theoretical analysis of the BAGofT for testing parametric classification models. In Section 3, we establish the theorems for the BAGofT to general classification procedures and discuss the practical implications. In Section 4, we present an efficient adaptive grouping algorithm with an excellent performance from our simulation studies. In Section 5, we discuss the use of multiple splitting to enhance the performance of BAGofT. We present simulation results in Section 6, and real-data examples in Section 7. We conclude the paper in Section 8. Proofs and additional discussions are included in the supplementary material.
2 \ BAGofT for testing parametric classification models

2.1 Settings

Let $Y$ be the binary response variable and $X$ be the vector of $p$ covariates. The support of $X$ is $S \subseteq \mathbb{R}^p$. Let

$$
\pi(x) = P(Y = 1|X = x).
$$

The data, denoted by $D$, consists of $n$ i.i.d. observations from a population distribution of the pair $(Y, X)$. We denote the fitted conditional probability obtained from a parametric classification model (or a general classification procedure to be discussed latter) by $\hat{\pi}(x)$.

A parametric classification model assumes that $\pi(\cdot) = f(\cdot; \beta)$, where $f$ is known and the unknown parameter $\beta$ is in a finite dimensional set $B$. For example, a generalized linear model assumes $f(x, \beta) = \mathbb{E}(y|\theta) = db(\theta)/d\theta = g^{-1}(x, \beta)$, where the response has a density function $f(y|\theta) = c(y) \exp(y\theta - b(\theta))$, and $g(\cdot)$ is a link function. The null and alternative hypotheses of the goodness-of-fit for testing a parametric classification model are defined by

$$
H_0 : \pi(\cdot) \in \{f(\cdot; \beta) | \beta \in B\},
$$

$$
H_1 : \pi(\cdot) \notin \{f(\cdot; \beta) | \beta \in B\}.
$$

Compared with parametric classification models, general classification procedures are not restricted to be in a parametric form. They include any modeling technique that maps the data $D$ to a fitted conditional probability function $\hat{\pi}(\cdot) : S \rightarrow [0, 1]$. For a general classification procedure, we study the convergence rate of $\hat{\pi}(\cdot)$. Let $r_n$ be the convergence
rate of the classification procedure we assess under the null hypothesis. The null and alternative hypotheses of the goodness-of-fit test for a general classification procedure are defined by

\[
H_0 : \sup_{x \in S} |\hat{\pi}_{D_n}(x) - \pi(x)| / r_n = O_p(1),
\]

\[
H_1 : \exists D \subseteq S \text{ with } P(x \in D) \text{ bounded away from 0 such that } \forall x \in D, \frac{|\hat{\pi}_{D_n}(x) - \pi(x)|}{r_n} \to_p \infty,
\]

where we write \( a_n = O_p(b_n) \) if for all \( \epsilon > 0 \), there exist \( C, N > 0 \) such that when \( n \geq N \), \( P(a_n/b_n > C) < \epsilon \) is bounded above in probability for all \( n \). So under the null hypothesis, \( \hat{\pi}_{D_n}(x) \) converges \( \pi(x) \) no slower than \( r_n \) and under the alternative hypothesis, it converges slower or does not converge to \( \pi(x) \).

Some additional notations are as follows. We use \( \to_p \) and \( \to_d \) to denote convergence in probability and in distribution, respectively. We use \( \chi^2_k \) to denote the chi-squared distribution with \( k \) degrees of freedom, Uniform\( [a, b] \) to denote the uniform distribution on the interval \( [a, b] \), and \( \mathcal{N}(\mu, \sigma^2) \) to denote the Gaussian distribution with mean \( \mu \) and variance \( \sigma^2 \).

### 2.2 Construction of the test statistic

The BAGofT is a two-stage approach where the first stage explores data-adaptive grouping and the second stage performs testing based on that grouping. To apply an adaptive grouping, we need to randomly split the data into a training set with size \( n_1 \) and a test set with size \( n_2 \). Denote the training data by \( D_{n_1} \). Let \( y_{e,i} \) be the \( i \)th response value in the test data, \( x_{e,i} \) be the \( i \)th covariates value in the test data. We first obtain \( \hat{\pi}_{D_{n_1}}(.) \) which is the fitted conditional probability function by applying the assessed classification model or
procedure to $D_{n_1}$. Then, we calculate predicted probabilities $\hat{\pi}_{D_{n_1}}(x_{e,i})$ for $i = 1, \cdots, n_2$. We then group the test data into sets based on an appropriately chosen partition. The partition is generated by applying a tree-type greedy algorithm to the training data $D_{n_1}$ (elaborated in Subsection 4). Let $\{\hat{G}_{D_{n_1,1}}, \cdots \hat{G}_{D_{n_1,K_n}}\}$ be such a partition of the support $S$, independent of the test data $D_{n_2}$. Here, the number of groups $K_n$ can be selected from the training data. For $i = 1, \cdots, n_2$, the $i$th observation in the test set is said to belong to group $k$, if $x_{e,i} \in \hat{G}_{D_{n_1,k}}$. Let

\[
R_i = y_{e,i} - \hat{\pi}_{D_{n_1}}(x_{e,i}),
\]

\[
\sigma_i^2 = \hat{\pi}_{D_{n_1}}(x_{e,i}) \{1 - \hat{\pi}_{D_{n_1}}(x_{e,i})\},
\]

\[
T = \sum_{k=1}^{K_n} \left( \frac{\sum\{i: x_{e,i} \in \hat{G}_{D_{n_1,k}}\} R_i}{\sqrt{\sum\{i: x_{e,i} \in \hat{G}_{D_{n_1,k}}\} \sigma_i^2}} \right)^2.
\]

We define the following BAGofT statistic:

\[
\text{BAG} = 1 - F_{\chi^2}(T, K_n),
\]  

where $F_{\chi^2}(c, k)$ denotes the CDF of $\chi^2_k$ evaluated at $c$. The testing statistic BAG can be treated as a $p$-value of the statistic $T$. The value of BAG tends to be small when the discrepancy between the fitted model and the underlying data generating process as quantified by $T$ is large. Compared with the Hosmer-Lemeshow test and other relevant methods, the adaptive nature of our method enables desirable features such as pre-screening candidate grouping methods to avoid Bonferroni correction, incorporating prior or practical knowledge that is potentially adversarial to the model under test (e.g., a variable not in the model to assess.), and providing interpretations of the data regions where the fitted model or procedure is likely to fail. The above flexibility enables a significantly improved statistical
power (elaborated in Section 4). It is worth noting that the BAGofT exhibits a tradeoff in data splitting. On the one hand, sufficient test data used to perform tests can enhance power. On the other hand, more training data tend to select an adversarial grouping that increases power. We will develop theoretical analysis and experimental studies to guide the use of an appropriate splitting ratio.

2.3 Theory for testing parametric classification models

We first establish that theorem that the BAGofT test statistic converges in distribution, which guarantees the size of the test under the null hypothesis. Some required conditions are as follows.

**Condition 1 (Sufficient number of observations in each group)** Let

\[ n_{\min} = \min_{k=1,\ldots,K_n} \sum_{i=1}^{n_2} I\{x_{e,i} \in \hat{G}_{Dn_1,k}\}. \]

We have \( n_{\min} \to_p \infty \) at a rate faster than \( n_2^{2/3} \) as \( n \to \infty \).

**Condition 2 (Bounded true probabilities)** There exists a positive constant \( 0 < c_2 < 1/2 \) such that \( c_2 \leq \pi(x_{e,i}) \leq 1 - c_2 \) for all \( i = 1, \ldots, n_2 \).

**Condition 3 (Parametric rate convergence under \( H_0 \))** Under the null hypothesis,

\[ \sup_{x \in \mathcal{S}} |\hat{\pi}_{Dn}(x) - \pi(x)| = O_p(1/\sqrt{n}) \] as \( n \to \infty \).

Condition 1 is a requirement for the adaptive grouping. It can be guaranteed by posing a constraint on the number of observations in each group. An example of such grouping can be found in Section 4. Condition 2 can be extended to more general cases if more
assumptions are made on the tail distributions of the covariates. Condition 3 holds for a
typical parametric model and a compact set \( S \). It will be extended to a general convergence
rate in Subsection 3.

**Theorem 1 (Convergence of BAG for parametric models under \( H_0 \))** Assume that Con-
ditions 1, 2, and 3 hold. Under the null hypothesis, if \( n_2 \to \infty \) and \( n_2^{2/3}/\sqrt{n_1} \to 0 \) as
\( n \to \infty \), we have

\[
\text{BAG} \to_d U,
\]

where \( U \sim \text{Uniform}[0, 1] \).

Accordingly, if we reject the model when the BAGofT statistic is less than 0.05, we will
have the asymptotic size 0.05. A possible test size, for instance, can be \( n_2 = 5\sqrt{n} \). Here,
the factor 5 is not essential. It is only used to guarantee a reasonable test size for a small
\( n \) in our simulation studies. The condition of \( n_1 \) and \( n_2 \) in the above theorem indicates
that the number of observations for estimating parameters and dividing groups (\( n_1 \)) needs
to be much larger than that for performing tests (\( n_2 \)). Intuitively, this is because we need
the error of the model to assess approximating the true probabilities to be much smaller
compared to the randomness of the test statistic.

Next, we establish the theorem that shows the BAGofT asymptotically rejects the
underfitted models under the alternative hypothesis.

**Condition 4 (Convergence under \( H_1 \))** There exists a function \( \pi_a : S \to [0, 1] \) such that
\[
\sup_{x \in S} |\widehat{\pi}_{D_n}(x) - \pi_a(x)| \to_p 0 \text{ as } n \to \infty \text{ under the alternative hypothesis. Moreover, there exists a positive constant } c_3 \text{ such that } c_3 \leq \pi_a(x) \leq 1 - c_3 \text{ almost surely.}
\]
Condition 5 (Identifiable difference under $H_1$) Under the alternative hypothesis, there exists a set $D_n \in S$ with $P(x \in D_n)$ bounded away from 0, such that either
\[
\text{ess inf} \left[ (\pi(x) - \pi_a(x)) \cdot I\{x \in D_n\} \right] \geq c, \quad \text{or} \quad \text{ess sup} \left[ (\pi(x) - \pi_a(x)) \cdot I\{x \in D_n\} \right] \leq -c,
\]
for some positive constant $c$.

Condition 4 requires the convergence of the model under the alternative. It can be implied by, for example, the classical asymptotic consistency of the maximum likelihood estimator under a misspecified model (White, 1982). Condition 5 guarantees that the true model and the asymptotic model under the alternative hypothesis are different.

Let $\hat{n}_{2,k} = \sum_{i=1}^{n_2} I\{x_{e,i} \in \hat{G}_{D_n, k}\}$ denote the number of observations in the $k$th group. Let $\hat{n}_{2_k}^{D_n} = \sum_{i=1}^{n_2} I\{x_{e,i} \in \hat{G}_{D_n, k} \cap D_n\}$ denote the number of observations in both the $k$th group and the set $D_n$.

Theorem 2 (Consistency of BAG for parametric models under $H_1$) Suppose that Conditions 1, 2, 4, and 5 hold. Under the alternative hypothesis, if the training and test sizes satisfy $n_1, n_2 \to \infty$, and there is at least one group $k$ with $\hat{n}_{2,k}^{D_n} / \hat{n}_{2,k} \to_p 1$ as $n \to \infty$, we have
\[
\text{BAG} \to_p 0,
\]
which implies the consistency of the test.

To obtain the consistency, we need our adaptive grouping to identify the set $D_n$ where $\pi(x)$ and $\pi_a(x)$ are different. Our algorithm in the following Subsection 4.2 adaptively searches the partitions and the number of groups $K_n$ according to the Pearson residual (which measures the deviation between $\hat{\pi}(x)$ and $\pi(x)$).
Corollary 1 (Simultaneous size control and consistency for parametric models)
Assume that Conditions 1-5 hold. Assume that the splitting ratio satisfies both $n_2 \rightarrow \infty$ and $n_2^{2/3}/\sqrt{n_1} \rightarrow 0$. Then, $\text{BAG} \rightarrow_d U$ under the null hypothesis. If there is at least one group with $\hat{n}_{2,k}^{D_n}/\hat{n}_{2,k} \rightarrow_p 1$ as $n \rightarrow \infty$, $\text{BAG} \rightarrow_p 0$ under the alternative hypothesis, as $n \rightarrow \infty$.

3 BAGofT for assessing general classification procedures
In this section, we extend our theories of the BAGofT to general classification procedures and provide guidance on its application.

3.1 Theorems for testing general classification procedures

Condition 6 (Convergence at a general rate under $H_0$) Under the null hypothesis, 
\[
\sup_{x \in S} |\hat{\pi}_{D_n}(x) - \pi(x)| = O_p(r_n) \text{ with } r_n \rightarrow 0 \text{ as } n \rightarrow \infty.
\]
Condition 6 generalizes the parametric rate required in Condition 3.

Theorem 3 (Convergence of BAG under $H_0$ for classification procedures) Under the null hypothesis, given Conditions 1, 2, and 6, if $n_1, n_2 \rightarrow \infty$ and $n_2^{2/3} r_{n_1} \rightarrow 0$ as $n \rightarrow \infty$, we have 
\[
\text{BAG} \rightarrow_d U,
\]
where $U \sim \text{Uniform}[0, 1]$.

Condition 7 (Existence of a slow converging set under $H_1$) Under the alternative hypothesis, there exists a set $D_n \in S$ with $P(x \in D_n)$ lower bounded by a constant, such
that

\[
\text{ess inf} \left( \hat{\pi}_{D_n}(x) - \pi(x) \right) \cdot I\{x \in \mathbb{D}_n\} \geq 0, \quad \text{or}
\]

\[
\text{ess sup} \left( \hat{\pi}_{D_n}(x) - \pi(x) \right) \cdot I\{x \in \mathbb{D}_n\} \leq 0,
\]

and

\[
\inf_{x \in \mathbb{S}} |\hat{\pi}_{D_n}(x) - \pi(x)| / r_n^{(a)} \rightarrow p \infty,
\]

for a sequence \( r_n^{(a)} \).

The above rate of \( r_n^{(a)} \) will pose a constraint on the splitting ratio in Theorem 4. We further let \( \hat{n}_{D_n}^{(a)} = \sum_{i=1}^{n_2} I\{x_{e,i} \in \hat{G}_{D_n, k} \cap \mathbb{D}_n\} \) denote the number of observations in both the \( k \)th group and the slow converging set \( \mathbb{D}_n \).

**Condition 8 (Bounded predicted probability)** There exists a positive constant \( c_4 \) such that \( c_4 \leq \hat{\pi}(x) \leq 1 - c_4 \) almost surely.

Condition 7 requires a region where the classification procedure to assess either converges slowly or diverges from the underlying data generating distribution. Condition 8 is for technical convenience. It can be replaced with the uniform convergence of \( \hat{\pi}_{D_n}(\cdot) \) to a function \( \pi_a(\cdot) \) in conjunction with the same boundedness constraint on that limiting function.

We write \( a_n = O(b_n) \) if there exist \( C \), for all \( n > 0 \), \( |a_n/b_n| < C \).

**Theorem 4 (Consistency of Bag under \( H_1 \) for classification procedures)** Under the alternative, assume that Conditions 2, 7, and 8 hold, the test size \( n_2 \rightarrow \infty \), the minimum number of observations in each group satisfies both \( n_2^{3/4}/n_{min}, n_{min}/n_2^{3/4} = O(1) \), and \( n_2^{5/16} r_n^{(a)} \rightarrow \infty \). Also assume that there is at least one group with \( \hat{n}_{D_n}^{(a)}/\hat{n}_{D_n} \rightarrow_p 1 \) and \( (n - \hat{n}_{D_n}^{(a)})/(\hat{n}_{D_n} r_n^{(a)}) \rightarrow_p 0 \) as \( n \rightarrow \infty \). Then, we have

\[ \text{BAG} \rightarrow_p 0. \]
Additionally, if the number of groups $K_n$ chosen by the adaptive partition is upper bounded by a fixed constant, we can relax the requirement $n_2^{5/16}(a) r_{n_1} \to \infty$ to $n_2^{3/8}(a) r_{n_1} \to \infty$. In this way, the test can be asymptotically powerful for a $r_{n_1}^{(a)}$ that converges faster to 0, which means that we are able to identify a smaller deviation.

Corollary 2 (Obtaining both size control and consistency for learning procedures)
Suppose that Conditions 2, 6 7, and 8 hold. We take the minimum number of observations in each group such that $n_{\min}/n_2^{3/4}, n_2^{3/4}/n_{\min} = O(1)$ and the splitting ratio satisfies both $n_2 \to \infty$ and $n_2^{2/3} r_{n_1} \to 0$ as $n \to \infty$. Then $\text{BAG} \to_d U$ as $n \to \infty$ under the null hypothesis, where $U \sim \text{Uniform}[0,1]$. Assume under the alternative hypothesis, $n_2^{5/16}(a) r_{n_1} \to \infty$, also, there is at least one group with $\hat{n}_{2,k}^{(a)}/\hat{n}_{2,k} \to_p 1$ and $(n - \hat{n}_{2,k}^{(a)})/(\hat{n}_{2,k} r_{n_1}^{(a)}) \to_p 0$ as $n \to \infty$. Then we have the asymptotic consistency of $\text{BAG}$.

In conclusion, if the general classification procedure to assess converges fast ($n_2^{2/3} r_{n_1} \to 0$), $\text{BAG}$ converges in distribution to $\text{Uniform}[0,1]$ as $n \to \infty$. If there exists a slow converging region ($n_2^{1/16} r_{n_1} \to \infty$), $\text{BAG} \to_p 0$ as $n \to \infty$.

3.2 Splitting ratios and interpretation for test results
In practice, the convergence rate of $r_n$ is usually unknown. Moreover, when the sample size is finite, the convergence rate $r_{n_1}$ provides limited insight in selecting a suitable splitting ratio. For practical implementations, we advocate using the training and test splitting ratios 90%, 75%, and 50%. The four typical results and our suggested conclusions are given below.

Case 1. The BAGofT fails to reject under all the three splitting ratios. The conclusion is that the general classification procedure we assess converges fast to the underlying
data generating process.

**Case 2.** The BAGofT only rejects when the training and test splitting ratio is 50%. The conclusion is that the general classification procedure we assess converges moderately fast.

**Case 3.** The BAGofT rejects when the training and test splitting ratios are 50% and 75% and fails to reject when the ratio is 90%. The conclusion is that the general classification procedure we assess converges slowly.

**Case 4.** The BAGofT rejects under all the three splitting ratios. The conclusion is that the general classification procedure we assess fails to capture at least one major aspect from the data.

A caveat of the above rule of thumb is that if the $\hat{\pi}(.)$ from the general classification procedure to assess changes abruptly when the sample size is only slightly changed, we may fail to observe the gradual change of the rejection results as listed in **Case 1 - Case 4**. In that case, we suggest the following conclusion.

**Case 5** The rejection result does not belong to **Case 1 - Case 4**. The outcome from different sample sizes is not in a sensible order.

To enhance the stability of $\hat{\pi}(.)$, we can use the same general classification procedure multiple times and obtain the averaged predictions.

### 4 Adaptive testing

The asymptotic theory of the BAGofT from the earlier sections needs us to find a grouping scheme from the training set, which asymptotically identifies the region that $\hat{\pi}_{DA}(x)$
converges slowly or does not converges to \( \pi(x) \). In this section, we introduce an adaptive grouping algorithm that can efficiently discover that region.

### 4.1 Adaptive grouping

The idea of adaptive grouping is that instead of applying one prescribed partition, we select a partition from a set of partitions based on the training data \( D_{n_1} \). According to Theorem 1 and Theorem 3, we can adaptively select a grouping rule (including the number of groups \( K_n \) and partition sets \( \hat{G}_{D_{n_1},k} \)) based on the training set and obtain weak convergence of the test statistic under the null hypothesis. Meanwhile, with adaptive grouping, the power under the alternative is expected to be adaptively high.

The adaptive grouping’s main goal is to find a partition to exploit the data regimes of model misspecification. One possible way is to fit the deviations (e.g., Pearson residuals) using a nonparametric regression model and choose a partition based on the fitted values’ contour. In this way, we group the observations with large positive deviations and those with large negative deviations into two separate groups to calculate BAG, so we avoid offsetting them. In the next section, we will introduce a random forest-based adaptive grouping scheme. We will show its excellent performance in our simulation studies. We developed this method into an R package ‘BAGofT,’ available from https://github.com/JZHANG4362/BAGofT and currently under the inspection of CRAN.

### 4.2 Random forest-based adaptive grouping scheme

We present an easy-to-implement grouping scheme that shows excellent performance in our various experimental studies. On the training set, we first fit the model or procedure to assess and calculate the Pearson residuals. Then, we fit a random forest on the training set
residuals and obtain fitted values \( \hat{q}^{(1)}_i, i = 1, \ldots, n_1 \). For different numbers of groups \( K = 1, \ldots, n_2/n_{\text{min}} \), we partition \( \{\hat{q}^{(1)}_i\}_{i=1}^{n_1} \) by its \( K \)-quantiles into intervals \( \{G^{(K)}_1, \ldots G^{(K)}_K\} \), calculate the test statistic

\[
B_K = \sum_{k=1}^{K} \left( \frac{\sum \{i: \hat{q}^{(1)}_i \in G^{(K)}_k\} \{y_{t,i} - \hat{\pi}_{D_1}(x_{t,i})\}}{\sqrt{\sum \{i: \hat{q}^{(1)}_i \in G^{(K)}_k\} \hat{\pi}_{D_1}(x_{t,i}) \left(1 - \hat{\pi}_{D_1}(x_{t,i})\right)}} \right)^2
\]

using the training set, and calculate the gaps \( B_K - B_{K-1}, i = 2, \ldots, n_2/n_{\text{min}} \). We choose the partition \( \{G^{(K_n)}_1, \ldots G^{(K_n)}_{K_n}\} \) where \( K_n = \arg \max_{K=2,\ldots,\lfloor n_2/n_{\text{min}} \rfloor} (B_K - B_{K-1}) \) \( \lfloor . \rfloor \) returns the largest integer less than the input). Next, we obtain the random forest prediction on the test set \( \hat{q}^{(2)}_i, i = 1, \ldots, n_2 \). We calculated the test statistic by

\[
T = \sum_{k=1}^{K_n} \left( \frac{\sum \{i: \hat{q}^{(2)}_i \in G^{(K_n)}_k\} R_i}{\sqrt{\sum \{i: \hat{q}^{(2)}_i \in G^{(K_n)}_k\} \sigma_i^2}} \right)^2
\]

and BAG based on the definition (2).

This algorithm is summarized in the Algorithm 1. Upon the return of the Algorithm 1, we can calculate the test statistic BAG in Equation (2) based on \( \{\hat{q}^{(2)}_i\}_{i=1}^{n_2} \) and the selected partition interval \( \{G^{(K_n)}_1, \ldots G^{(K_n)}_{K_n}\} \). Note that we can a set of covariates in the random forest different from those in the model or procedure to assess. For instance, sometimes we apply variable screening to drop some covariates before fitting the classification model or procedure to obtain a parsimonious model or make the fitting algorithm stable. In this case, our random forest-based adaptive partition can still fit on all of the available covariates to check the goodness of fit based on all available data. This algorithm also provides some insights into possible misspecifications of the model/procedure to assess by providing the variable importance. Since the random forest is fitted on the Pearson residual of the model or procedure to assess, variables with high importance are likely to be a major reason for
misspecifications. In high dimensional settings where we have many covariates, we may add a variable pre-selection step before applying the random forest partition. A possible way is to calculate the distance correlation (Székely et al., 2007) between the Pearson residual and the covariates we consider, and then select a small set of covariates with the largest distance correlation values. Some learning methods may return fitted probabilities close to 0 or 1, causing issues to the calculation of the Pearson residual. If that occurs, we may replace the Pearson residual with the raw residual. Simulations in Section 3 and Section 4 in our supplementary material show that a pre-selection of size 5 from 500 covariates may significantly improve the power of the BAGofT.

5 Combing results from multiple splitting

Recall that our test is based on a splitting of the original data into training and test parts. Due to the data splitting’s randomness, we may obtain different test results from the same data. To alleviate this randomness, we may randomly split the data multiple times and appropriately combine the test results from each splitting to generate the final result.

We propose the following procedure. First, randomly split the data into training and test sets multiple times and calculate the BAG defined in the Formula (2) based on each training-test set pair; Second, calculate the sample mean of the BAG values. Other ways to combine results from multiple splittings include taking the sample median or sample minimum of BAG values. As the theoretical distribution of these statistics from the combined results is a challenge, we evaluate the obtained statistic by empirical $p$-values.

The empirical $p$-value is obtained by a set of BAG statistics calculated from parametric bootstrap data. First, we fit the model using all the data and obtain fitted probabilities;
Algorithm 1  Random forest-based adaptive selection of partitions

1: procedure Partition($D_{n_1}, D_{n_2}, n_{min}, parVar$)  
   $\triangleright$ parVar is the set of variables to construct the partition. It can be different from the set of covariates in the model or procedure to assess (see Subsection 4.2).

2: Fit the model or procedure on the training set $D_{n_1}$ and calculate the Pearson residual.

3: Fit a random forest of the Pearson residual w.r.t. the partition variables $parVar$ and obtain the fitted value on the training set $\{q_i^{(1)}\}_{i=1}^{n_1}$ and predicted value on the test set $\{q_i^{(2)}\}_{i=1}^{n_2}$.

4: for $K$ in $1, \cdots, \lfloor n_2/n_{min} \rfloor$ do

5: Partition $\{q_i^{(1)}\}$ by its $K$-quantiles into $K$ intervals $\{G_1^{(K)}, \cdots G_K^{(K)}\}$.

6: Calculate $B_K$ in Equation (3).

7: end for

8: $K_n \leftarrow \arg \max_{K=2,\ldots,\lfloor n_2/n_{min} \rfloor} (B_K - B_{K-1})$.

9: return $\{G_1^{(K_n)}, \cdots G_{K_n}^{(K_n)}\}$.

10: end procedure
Second, we generate some bootstrap datasets from the binomial distribution with those fitted probabilities as parameters; Third, we calculate the BAG statistic on each of the bootstrap datasets, so these BAG statistics correspond to the case where the model or the procedure to assess is ‘correct’; Fourth, we compare the BAG statistic from the original data to those from bootstrapped datasets and calculate the empirical $p$-value. As for the training-test splitting for testing parametric models, we suggest taking $n_2 = 5\sqrt{n}$. For the minimum number of observations in each group, we suggest taking $n_{\text{min}} = (n_2)^{3/4}/2$, which satisfies Condition 1 and guarantees that there are at least nine observations in each group when the total sample size $n = 100$.

We will provide experimental studies in various settings in Section 6 and the supplementary material. Our results show the promising performance of the proposed method under both the null and alternative hypotheses. In practice, if the model or procedure takes a lot of time to fit or requires manual tuning, this approach with multiple splitting may be undesirable. So we recommend applying the BAGofT based on a single splitting. If the data comes with a test set (for calculating the classification accuracy), it is desirable for us to directly apply the same splitting. In this way, the test result and the classification accuracy are based on the same test set, facilitating the interpretations.

6 Experimental studies

In the following subsections, we present some simulation results to demonstrate the performance of the BAGofT in various settings.

In Subsection 6.1, we check the performance of the BAGofT in parametric settings and compare it to existing methodologies including the recently proposed Generalized Residual
Prediction (GRP) test (Janková et al., 2020). The GRP calculates a test statistic by pivoting the Pearson residuals from the model to assess. It has a different focus compared to the BAGofT. First, the GRP test works for generalized linear models (GLM) and the BAGofT tests classification models including those are not GLM, e.g., linear discriminant analysis and naive Bayes models. Secondly, the GRP test is more specific in the sense that it focuses on the cases where the link function of the generalized linear model is correctly specified while the BAGofT can have power against a general deviation between the data and the model to assess. Additionally, for the covariate pre-selected case as mentioned in Subsection 4.2, the GRP test focuses on identifying whether the model to test misses some main effects, and the BAGofT can test other misspecifications including missing quadratic effects or interactions of the dropped covariates. For comparison, we choose simulation settings that work for both the GRP and BAGofT.

In Subsection 6.2, we demonstrate the application of the BAGofT to assess general classification procedures. The outcomes from classification procedures with different convergence speeds are presented to illustrate the suggested interpretations from Subsection 3.2.

6.1 Testing parametric models

In this subsection, we choose some commonly studied parametric settings and compare the performance of the BAGofT with some other popular methods. Similar settings have been studied in (Canary et al., 2017; Yin and Ma, 2013; Pulkstenis and Robinson, 2002).

Setting 1. We generate data from

$$P(y = 1|x_1, x_2, x_3) = 1/(1 + \exp(- (\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3))),$$

where $x_1$, $x_2$, and $x_3$ are independently generated from Uniform$[-3, 3]$, $\mathcal{N}(0, 1)$, and
$\chi^2_1$, respectively. We test the correctly specified Model $A$ and the Model $B$ that misses $x_3$.

**Setting 2.** We generate data from

$$P(y = 1|x_1, x_2, x_3) = 1/(1 + \exp(- (\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2))),$$

where $x_1$ and $x_2$ are independently generated from $\text{Uniform}[-3,3]$. We test the correctly specified Model $A$ and the Model $B$ that misses the interaction term.

**Setting 3.** We generate data from

$$P(y = 1|x_1, x_2, x_3) = 1/(1 + \exp(- (\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2))),$$

where $x_1$, $x_2$, and $x_3$ are independently generated from $\text{Uniform}[-3,3]$, $\mathcal{N}(0,1)$, and $\chi^2_2$, respectively. We test the correctly specified Model $A$ and the Model $B$ that misses the quadratic term.

For Model $A$, we check the BAGofT test statistic’s null distribution. For Model $B$, we compare the power of the BAGofT with the Hosmer-Lemeshow test (Hosmer and Lemeshow, 1980), le Cessie-van Houwelingen (CH) test (Le Cessie and Van Houwelingen, 1991), and GRP test (Janková et al., 2020). These three tests are fitted by packages ResourceSelection (Lele et al., 2019), rms (Harrell Jr, 2019), and GRPtests (Jankova et al., 2019), respectively with default values. The BAGofT takes 40 splits. It takes all available covariates for the adaptive partition (namely $x_1, x_2, x_3, x_1, x_2$, and $x_1, x_2, x_3$ in the three settings, respectively).

To avoid cherry-picking, we independently generate the coefficients from zero-mean normal distributions. To reflect different degrees of deviation from the null, we take standard
deviation $\gamma = 1$ or 0.5 for $\beta_3$ in setting 1, setting 2 and $\beta_4$ in setting 3. Other coefficients have a unit standard deviation. We take sample sizes 100, 200, and 800. The testing process in each setting is repeated 100 times.

The BAGofT results from the three ways (i.e., those based on mean, median, and minimum) to combine multiple splitting results in Subsection 5 are very close, so we only present the results from the method based on the mean. For Model A, the Q-Q plots of the BAGofT test statistic versus uniform $\text{Uniform}[0,1]$ are shown in Figure 1. We observe that, in general, the BAGofT test statistic has a good approximation to $\text{Uniform}[0,1]$ under the null hypothesis. When the sample size is small, the BAGofT test statistic tends to be conservative, with large $p$-values. For Model B, the rejection rates of the BAGofT compared with other tests under the significance level 0.05 are shown in Figure 2. Due to the random generation of coefficients, a small portion of the datasets is unbalanced. It caused computation errors for the CH test GRP test. We dropped these cases when computing the rejection rates. For the results of Figure 2, the BAGofT (in circles) has the best performance in all of the cases. The GRP test (in squares) has slightly worse performance in Settings 2 and 3.

We also study the relationship between the number of splits and the variation of the BAGofT test statistic. Recall that the purpose of multiple splitting is to obtain a test statistic with fewer variations. The results show that 10 to 20 splits are usually good enough to obtain stable results. Additionally, we check the covariates with the largest variable importance in Settings 1 and 3 when the models are misspecified (Model B). Recall that our algorithm introduced in Subsection 4.2 applies the random forest to fit the Pearson residuals, so the covariates with large variable importance tend to be the major source of misspecification. For example, in our simulation, the missing variable $x_3$ in Setting
1 has the largest variable importance; The $x_1$ in Setting 2, whose quadratic effect is missing, has the largest variable importance. Additional details on the variable importance in the simulations are included in the supplementary material.
Figure 1: The Q-Q plot of the BAGoT empirical $p$-values from Model A versus Uniform[0,1] distribution. The $x$-axis and $y$-axis correspond to the theoretical quantiles and observed sample quantile, respectively. When the sample size $n$ is small, the BAGoT tends to be conservative (some dots are above the straight lines) but when the sample size is large, the approximation is desirable.
Figure 2: The rejection rates of tests at significance level 0.05 for Model B. The BAGofT is compared with the HL (Hosmer and Lemeshow, 1980), CH (Le Cessie and Van Houwelingen, 1991), and GRP (Janková et al., 2020) tests. Settings 1-3 correspond to missing a main effect, missing an interaction, and missing a quadratic effect, respectively. The larger \( \gamma \) value corresponds to a larger deviation of the model to test to the truth.
6.2 Evaluating classification learning procedures-low dimensional setting.

In this subsection, we illustrate the application of the BAGofT to measure the goodness-of-fit of several classification procedures in a low dimensional setting. We generate data from the Bernoulli distribution with conditional probability

\[ P(y = 1|x_1, x_2, x_3) = 1/(1 + \exp(\sin(x_1) + 1.8x_2x_3 + x_4)). \]

We independently generate the data of \( x_1 \) from \( \mathcal{N}(0, 2.25) \) and the data of \( x_2, x_3, \) and \( x_4 \) from \( \mathcal{N}(0, 1) \). We assess a feed-forward neural network, a random forest, and a logistic regression model. The neural network has two hidden layers, with 80 and 5 neurons, respectively. The activation function is ReLu (Nair and Hinton, 2010). For the random forest, we average over 500 trees, and each tree randomly takes two covariates. For the logistic regression model, we consider the main effects of \( x_1-x_4 \) only, so it does not converge to the data generating model.

We take 100 independent replications. In each replication, we randomly generate a sample with size 1000 and apply the BAGofT with splitting ratios 90%, 75%, and 50% to the procedures to assess. The BAGofT takes 40 splits, and its adaptive partition is based on all available covariates \( x_1-x_4 \).

Figure 3 gives an overall illustration of the \( p \)-values of the procedures to test from 100 replications. We take the significance level to be 0.05. It can be seen that the neural network is likely to be rejected except for the splitting ratio of 90%, and it corresponds to the slow convergence (Case 3). The majorities of random forest’s \( p \)-values are above 0.05, and it corresponds to the fast convergence (Case 1). Apparently, the logistic regression model with main effects only fails to capture the nonlinearity from the data generating model.
(Case 4). It is rejected in all splitting ratios and replications. For single realizations (where each single test results from 100 replications), the majority (40%) of the times the neural network falls into Case 3 and majority (59%) of the times the random forest falls into Case 1.

Figure 3: The BAGofT p-value box plots for the neural network, random forest and logistic regression with splitting ratios 90%, 75%, and 50%. The neural network falls into Case 3 (slow convergence). The random forest falls into Case 1 (fast convergence). The logistic regression falls into Case 4, where it fails to capture a major aspect (nonlinearity) from the data.
6.3 Evaluating classification learning procedures-high dimensional setting.

In this subsection, we apply the BAGofT to assess classification procedure in high dimensional settings. We consider two data-generating models.

**Setting 1:**

\[ P(y = 1|x_1, \cdots, x_{1000}) = \frac{1}{1 + \exp(-(-6 + 3 \cdot I\{-2 < x_1 < 2\} + 0.5(x_2 + x_3 + x_4 + x_5)))}. \]

**Setting 2:**

\[ P(y = 1|x_1, \cdots, x_{1000}) = \frac{1}{1 + \exp(-(0.5x_1 + 0.3x_2 + 0.1x_3 + 0.1x_4 + 0.1x_5)))}. \]

We independently generate the data of \(x_1 - x_{500}\) from Uniform\([-5, 5]\). We assess a logistic regression model with lasso penalty, a random forest, and an xgboost model (Chen and Guestrin, 2016).

We take 100 independent replications. In each replication, we randomly generate a sample with size 500 and apply the BAGofT with the three splitting ratios to the logistic regression model. The BAGofT takes 20 splits, and its adaptive partition is based on all available covariates \(x_1 - x_{1000}\). The random forest is fitted with maximum nodes 10 and xgboos is fitted with 25 iterations. The results are summarized in Figure 4.

The result of lasso logistic regression in setting 1 belongs **Case 4** since the lasso logistic fails to capture the nonlinearity in the data-generating model. The lasso logistic regression’s result of setting 2 belongs to **Case 1**, which shows the lasso logistic regression model has fast convergence speed. The lasso logistic regression fails to capture a major aspect in setting 1 (**Case 2**) in setting 1 and converges slowly (**Case 3**) in setting 2. The random forest converges moderately fast (**Case 2**) in setting 1 and converges slowly (**Case 3**) in setting 2.
**Figure 4:** The BAGofT $p$-value box plots for the lasso logistic regression model, random forest, and xgboost with splitting ratios 90%, 75%, and 50% in the high dimensional settings.
7 Real data example

In the following three subsections, we demonstrate the application of the BAGofT by real data examples. In Subsection 7.1, we test a parametric classification model and compare the BAGofT to other methodologies. In Subsection 7.2, we present a graphical illustration of the variable importance from the BAGofT’s adaptive partition. We show that it can be useful in finding out the reason for the possible underfitting. In Subsection 7.3, we apply the BAGofT to assess different classification procedures.

7.1 Testing parametric classification models: Micro-RNA data

We consider the data from the study of (Shigemizu et al., 2019), which is available from the Gene Expression Omnibus (GEO) database with accession number GSE120584. The paper studies the prediction of dementia using the logistic regression model and micro-RNA data. The study fits logistic regression models on the principal components of preselected micro-RNA data. The selection of micro-RNA and the number of principal components are determined by results from 10-fold cross-validation and area under the curve (AUC).

We study the logistic regression model for Alzheimer’s disease (AD). The data for this model is obtained by combining the data of AD and normal controls (NC). The number of observations is \( n = 1309 \). The paper selects 78 micro-RNA and computes ten principal components from the data to fit the prediction model for AD. We consider a subset model using the first 7 principal components as the covariates.

\[
\log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 \text{PC}_1 + \cdots + \beta_7 \text{PC}_7.
\]  

Since we have the data of the first 20 principal components, we test whether there is a
significant underfitting of this models as

\[ P(\text{dementia} \mid \text{PC}_1, \cdots \text{PC}_{20}) = \pi(\text{PC}_1, \cdots \text{PC}_{20}). \]

We apply the BAGofT with 40 splits and pre-selection size 5. The empirical \(p\)-value of the BAGofT is 0. We check the average variable importance of the random forest which we use to group the observations. It shows that \(\text{PC}_9\) has the largest importance and is likely to be the major reason for the underfitting. Next, we add \(\text{PC}_9\) to the model and consider:

\[
\log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 \text{PC}_1 + \cdots + \beta_7 \text{PC}_7 + \beta_9 \text{PC}_9. \tag{5}
\]

The \(p\)-value from the test statistics is 0.15. So we will not reject this model at the significance level of 0.05. As a comparison, we apply the Hosmer-Lemeshow test (‘HL’), CH test (‘CH’), and GRP test (‘GRP’) to these models. We also apply the BAGofT with the covariates in the model to test only (BAG*). The results are shown in Table 1. Interestingly, the BAGofT based on \(\text{PC}_1 - \text{PC}_7\) (the covariates in the Model 1) also rejects the Model 1, which indicates that the Model 1 fails to capture some major information from \(\text{PC}_1 - \text{PC}_7\). However, this underfitting gets less significant after the inclusion of the \(\text{PC}_9\) with BAG*’s \(p\)-value 0.47. Other tests can not reject either Model 1 or Model 2.

**Table 1:** \(P\)-values for models from Equations (4) and (5).

<table>
<thead>
<tr>
<th>Test</th>
<th>HL</th>
<th>CH</th>
<th>GRP</th>
<th>BAG</th>
<th>BAG*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>0.42</td>
<td>0.26</td>
<td>0.17</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Model 2</td>
<td>0.17</td>
<td>0.23</td>
<td>0.23</td>
<td>0.15</td>
<td>0.47</td>
</tr>
</tbody>
</table>
7.2 Testing classification procedures: Fashion MNIST data

We consider the Fashion MNIST data (Xiao et al., 2017), which has images of different clothes with a pixel size of $28 \times 28$. We take the first 500 images of trousers and the first 500 images of blouses. So the total sample size is 1000. An example snapshot of these images is shown in Figure 5. We assess a feed-forward neural network with one hidden layer and one hidden node/neuron. We apply the BAGofT with 40 splits and a pre-selection size 5. The training versus test ratios is 90%, 75%, and 50%.

The test statistic has empirical $p$-value 0 in all of the three splitting ratios, which indicates that the neural network fails to capture at least one major aspect from the data (Case 4). To interpret the testing results, we plot the variable importance of the $28 \times 28$ covariates from the BAGofT with training versus test ratio 90% in Figure 6. As we remarked in Subsection 4.2, the covariates with high variable importance are likely to be the major reason for the underfitting. In the particular image of Figure 6, it indicates that the space between the two legs of the trousers is where the model to assess is underfitting. Interestingly, this is indeed the major difference between the two kinds of clothes.

7.3 Testing classification procedures: COVID-19 CT scans

Coronavirus disease 2019 (COVID-19) has had a massive impact on the world. We consider the data in the study from (He et al., 2020), which is available at https://github.com/UCSD-AI4H/COVID-CT. The data set contains 349 positive cases and 397 negative cases.

We test classification procedures fitted on the 1000 features generated from the pre-trained MobileNetV2 (Sandler et al., 2018). The images are resized into $224 \times 224$ RGB pixels before entering MobileNetV2. We test two one-layer neural networks and a ran-
Figure 5: An example of trouser and dress images from the Fashion MNIST data (Xiao et al., 2017).

Figure 6: Variable importance of the neural network fitted to the Fashion MNIST data (Xiao et al., 2017). Covariates with higher variable importance are marked by brighter color. The neural network still has room for improvement from those highlighted covariates.
dom forest. The neural networks (with the ReLu activation) have one and five neurons, respectively.

We apply the BAGofT with 20 splits and a pre-selection size 10. The training set ratios are 90%, 75%, and 50%. We take the significance level 0.05, and the $p$-values are shown in Table 2. It can be seen that the neural network with one neuron is too restricted to capture the major information from the data (Case 4). Both the neural network with five neurons and the random forest have fast convergence (Case 1).

Table 2: $P$-values from classification procedures fitted on the COVID-19 data (He et al., 2020). NNET-1, NNET-2 and RF denote the neural network with one neuron, the neural network with five neurons, and random forest, respectively.

<table>
<thead>
<tr>
<th>Splitting ratio</th>
<th>90%</th>
<th>75%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNET-1</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>NNET-2</td>
<td>0.42</td>
<td>0.17</td>
<td>0.20</td>
</tr>
<tr>
<td>RF</td>
<td>0.12</td>
<td>0.45</td>
<td>0.20</td>
</tr>
</tbody>
</table>

8 Conclusion

We developed a new methodology of goodness-of-fit tests for parametric classification models and general classification procedures. The nature of the adaptive grouping to flexibly expose the weaknesses of the model or procedure to assess makes the developed solution highly powerful. The adaptive grouping in the BAGofT may capture the covariates that lead to underfitting. Numerical results have demonstrated the significant advantages of the BAGofT compared with some existing tests, including the state-of-the-art Hosmer-Lemeshow test.
The R package “BAGofT” is available from https://github.com/JZHANG4362/BAGofT and is currently under the inspection of CRAN.

References


