JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2022, VOL. 00, NO. 0, 1–11
https://doi.org/10.1080/10618600.2022.2026779

Taylor & Francis
Taylor & Francis Group

Nonparametric Subset Scanning for Detection of Heteroscedasticity

Charles R. Doss[a] and Edward McFowland III[b]

[a]School of Statistics, University of Minnesota, Minneapolis, MN; [b]Harvard Business School, Harvard University, Boston, MA

ABSTRACT

We propose heteroscedastic subset scan (HSS), a novel method for identifying covariates that are responsible for violations of the homoscedasticity assumption in regression settings. Viewing the problem as one of anomalous pattern detection, we use subset scanning techniques to efficiently identify the subset of covariates that are most "heteroscedastically relevant." Through simulations and a real data example, we demonstrate that HSS is capable of detecting heteroscedasticity in a wide range of settings, including in cases where existing global tests lack power. Furthermore, the global power of our method compares favorably to methods such as the Breusch–Pagan test. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2020
Revised October 2021

KEYWORDS
Anomaly detection; Model diagnostics; Regression; Scan statistics

1. Introduction

Regression techniques are used in a wide range of scientific fields to model the relationship between a set of covariates and a response. Consider a regression model $Y = m_0(X) + \epsilon$, with response $Y \in \mathbb{R}$, covariate vector $X \in \mathbb{R}^p$, error term $\epsilon \in \mathbb{R}$ and unknown regression function $m_0 : \mathbb{R}^p \mapsto \mathbb{R}$. We can take X to be fixed or random; our notation is based on X being random but with no distributional assumptions made on X so that this accomodates both cases (allowing a dirac measure for X). We observe n iid observations from this model. Assume that $E(\epsilon) = 0$. Many regression paradigms require an assumption of homoscedasticity, that is, $\text{var}(\epsilon) = \sigma^2 > 0$ is constant and does not depend on the covariate X. Problems may arise when attempting to fit a model where the constant variance assumption does not hold. Such models are said to exhibit heteroscedasticity.

The Gauss–Markov Theorem states that the minimum-variance linear unbiased estimator of the coefficient vector in a linear regression model is the ordinary least squares (OLS) estimator. However, this theorem does not hold if the error terms do not have constant variance. If heteroscedasticity is present, OLS estimates can be improved by accounting for the variance structure.

It is well-known that not accounting for heteroscedasticity when it is present can result in a "loss of efficiency" and "more importantly, the biases in estimated standard errors may lead to invalid inferences" (Breusch and Pagan 1979). And "in many cases the loss of efficiency in using procedures for homoscedastic models under heteroscedastic errors may be substantial" Dette and Munk (1998). Biased estimates of standard errors can result in higher rates of Type I errors in inference. And using inefficient procedures or statistics in place of more efficient ones can result in higher rates of Type II errors. On the other hand, in a high-dimensional setting, Li and Yao (2019) provide a simulation showing that using methods designed for heteroscedasticy when none is present can result in a very large loss of efficiency. The natural conclusion is that it is important to properly assess whether heteroscedasticity does or does not need to be accounted for.

Our work is motivated by a desire to identify covariates that are associated with violations of the homoscedasticity assumption. It is possible that ϵ is dependent on some covariates but independent of other covariates; therefore, we wish to identify the subset that is related to the residuals. We do not assume any particular model for m_0. More formally, let $\mathbb{X}_{i,j}$ denote the ith entry of the jth column of the observed covariate matrix $\mathbb{X}$. Then we call the jth covariate X_j a "heteroscedastically relevant variable" (HRV) if $\text{var}(Y_i|\mathbb{X}_{i,j})$ is nonconstant in i. Our goal is to identify HRVs given data. We resist the term "heteroscedastic variable," as heteroscedasticity is a property of a model and not of the covariates. This definition of HRVs includes only covariates that are marginally related to error variance. A set of covariates that are jointly related to the error variance could be defined as a "heteroscedastically relevant set," but such a set would not necessarily be unique. As the two concepts are closely related, we focus on HRVs and leave heteroscedastically relevant sets for future research.

In this work we propose a novel method for identifying heteroscedastically relevant variables, Heteroscedastic Subset Scan (HSS). This method adds to the body of research focused on diagnostic tests for regression and makes several new contributions:

- HSS allows for the efficient discovery of HRVs. Unlike previous methods, we view the problem as one of anomalous pattern detection. As such, we perform a search to identify the subset of covariates which exhibits the most unexpected relationship with the residuals under the null hypothesis,

CONTACT Edward McFowland III *emcfowland@hbs.edu* Harvard Business School, Harvard University, Boston, MA.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

under which the residuals are independent of the covariates. This differs from previous methods which have largely focused on global tests for heteroscedasticity. Our approach pinpoints relevant covariates in the data and gives us a better understanding of how to proceed when choosing a model that accounts for heteroscedasticity.

- HSS is computationally efficient. As the number of covariates p grows, HSS performs a linear-time scan over subsets of the covariate space. This efficiency follows from the fact that our method satisfies the linear time subset scanning (LTSS) property described by Neill (2012) and McFowland III, Speakman, and Neill (2013), as shown in Section 3. The LTSS property allows us to reduce our search over the space of all covariate subsets from $\mathcal{O}(2^p)$ to $\mathcal{O}(p)$ while still guaranteeing that the most anomalous subset is found.
- HSS can identify heteroscedasticity in many different forms. Existing methods such as the Breusch–Pagan test only test for linear relationships between the covariates and the squared residuals, giving these tests low power in situations where this relationship is nonlinear. Due to its nonparametric nature, HSS does not rely on any assumptions about the "shape" of the heteroscedasticity and is therefore, capable of detecting nonlinear violations, such as "butterfly residuals" described by Celik (2015).
- Our scoring function makes a new contribution to the subset scanning literature. Previous subset scan statistics are focused on identifying anomalies in only one direction, typically searching for anomalous patterns in the data where many values exceed their expectation. This is useful in settings such as disease surveillance, where researchers are solely interested in detecting high-risk sub-populations and can reasonably ignore low-risk subpopulations, as in McFowland III, Speakman, and Neill (2013). Our statistic is ambidirectional, meaning that it is capable of identifying anomalous patterns where values diverge from the expected value, regardless of whether this divergence is above or below expectation. While this innovation was motivated by the heteroscedasticity problem, it is potentially useful in a wide array of applications.

Once HRVs have been identified, the data analyst must decide how to proceed. One option that is often mentioned in this context is transformation of the response variable, although that approach may introduce as many problems as it solves (it may affect the mean relationship and it may cause new covariates to become HRVs). A more direct general approach is to model the heteroscedasticity in some way. In a linear regression this can mean using weighted least squares, for instance; in nonlinear models, the method for smoothing may be chosen dependent on the variance (see, e.g., Muller and Stadtmuller 1987). Many other options are possible and the choice will be very dependent on the details of the modeling situation.

We begin with a review of existing methods and gaps in the literature (Section 2). Then we present HSS, which can efficiently identify subsets of the covariate space where heteroscedasticity is present (Section 3). Next, we discuss selection of an important tuning parameter (Section 4.1) and other computational considerations (Section 4.2). Section 5 presents a simulation study that shows the efficacy of HSS in a variety of situations compared to alternative methods. Finally, Section 6 discusses our results.

2. Methods for Identifying Heteroscedasticity

Many existing methods for modeling heteroscedastic data require the user to have some knowledge of the form of the heteroscedasticity, necessitating tests for violation of the homoscedasticity assumption. We now give an overview of existing tests. All are global in that they look for heteroscedasticity but do not attempt to identify HRVs.

Goldfeld and Quandt (1965) give some of the earliest tests for heteroscedasticity. They present a parametric test in which the data are split into two groups based on the values of a single covariate X_m suspected to be heteroscedastically relevant. Regression models are fit to both partial datasets and the ratio of the sum of squared errors for the models is used as a test statistic. To consider multiple variables one must apply the Goldfeld–Quandt test separately to each one. If the error variance is a nonmonotonic function of a covariate X_m, the Goldfeld–Quandt test may have low power, as demonstrated in Section 5. Other tests from this era are also univariate, including Glejser (1969) and Harrison and McCabe (1979). Dette and Munk (1998) provide a nonparametric approach, still in the univariate case.

In the setting of multiple covariates, Breusch and Pagan (1979) present a test based on the Lagrangian multiplier statistic. Cook and Weisberg (1983) independently present a similar test. The Breusch–Pagan test is applicable to a wide range of heteroscedastic models, although it assumes that the ϵ_i are normal under the null hypothesis. Koenker (1981) proposes a studentized version of the original test wherein the test statistic can be computed by multiplying the sample size n by the R^2 of a regression of the squared residuals $\hat{\epsilon}^2$ on the covariates, testing whether the covariates and residuals show a linear relationship. Koenker and Bassett (1982) and Newey and Powell (1987) propose methods based on quantiles which can be useful if conditional variance is heavy tailed or asymmetric. White (1980) extends the Breusch–Pagan test to detect both heteroscedasticity and more general model misspecification, although this test requires p^2 to be small relative to the sample size n, so is not feasible even for moderate values of p.

Recent work by Li and Yao (2019) addresses the moderate-dimensional situation where p is large but still less than n so that a standard linear regression model can be fit. They propose a modified likelihood ratio test and a coefficient-of-variation test, both of which are suited for both low- and moderate-dimensional problems if we wish to identify heteroscedasticity present in first- or second-degree polynomials of the covariates.

In summary, existing methods for identifying heteroscedasticity have several deficiencies. Aside from Li and Yao's tests, none of the existing tests are powered to detect heteroscedasticity if a covariate has a nonmonotonic relationship with the error variance. Also, Breusch–Pagan and White assume that the residuals follow a normal distribution under the null hypothesis so that a rejection from one of these tests may be caused by nonnormality and not by nonconstant variance. Finally, all of the methods are global tests, providing no evidence of which variables are responsible for heteroscedasticity.

HSS addresses all of these deficiencies. HSS can identify non-monotonic heteroscedasticity, for example, "butterfly residuals," as described by Celik (2015) (when error variance is large at the extreme values of a covariate and small for values near the median). We make no parametric assumption about the residuals, ensuring that our test is not responding to a mis-specification of the residual distribution. Most importantly, HSS detects heteroscedastically relevant variables, allowing the user to make informed modeling decisions.

3. Heteroscedastic Subset Scan

3.1. Overview

Our method performs a scan for anomalies based on a hypothesis testing framework where the null hypothesis is that the distribution of $\epsilon_i|\mathbb{X}_{i,j}$ is constant in i. We now give a brief overview of the method with a more detailed explanation in the subsequent sections.

The model is as given in Section 1. We observe n independent observations from the model and stack the covariates into an $n \times p$ data matrix $\mathbb{X}$ with entries $\mathbb{X}_{i,j}$. We refer to the jth covariate of $\mathbb{X}$ as X_j and the column vector holding the values of X_j as $\mathbb{X}_j$. We assume that we have fit a model giving a vector of residuals $\hat{\epsilon}$ with elements $\hat{\epsilon}_i, i = 1, \ldots, n$. We omit discussion of the response variable and regression model, since we only need responses and the model to generate residuals. We fix an integer K, a user-specified tuning parameter discussed in more detail in Section 4.1. The values of $\mathbb{X}_j$ can then be broken into K disjoint quantile intervals called "K-intervals." We compare residuals whose corresponding jth covariate values are in the kth K-interval with the residuals whose jth covariate values are outside the kth K-interval. To accomplish this, we first split the data into reference and evaluation sets. Then for each $j \in \{1, \ldots, p\}$, we further split both sets into K partitions, one for each K-interval of $\mathbb{X}_j$.

Then we compute an empirical p-value for each evaluation set residual by comparing each of these *evaluation* residuals in *a given interval* to the distribution of *reference* set residuals from *the other* K-intervals. These p-values are (asymptotically) uniformly distributed under the null hypothesis. Thus, for each covariate and each $k \in \{1, \ldots, K\}$, we compare the distribution of observed p-values (for the kth K-interval of the jth covariate) to what would be expected under the uniform distribution. This yields a preliminary measure of the heteroscedastic signal present in X_j. Then we scan over the space of all possible subsets of covariates to find a "most anomalous subset." Scanning over subsets allows us to find subtle (anomalous) patterns spread across multiple covariates, that may go undetected when evaluating each covariate individually.

However, after finding this initial most anomalous subset, we still do not know if the subset we've identified is significantly anomalous or not. We discuss how we finally arrive at a (possibly null) set of HRVs in Section 4.1. In brief: we repeat the above-described procedure using different random splits of reference and evaluation sets, which allows us to compute an average "inclusion rate" for each covariate. We then estimate the distribution of these inclusion rates under the null using a bootstrap. Each variable is then included or excluded depending on whether its observed inclusion rate is larger or smaller than a corresponding bootstrap distribution quantile with multiplicity adjustment.

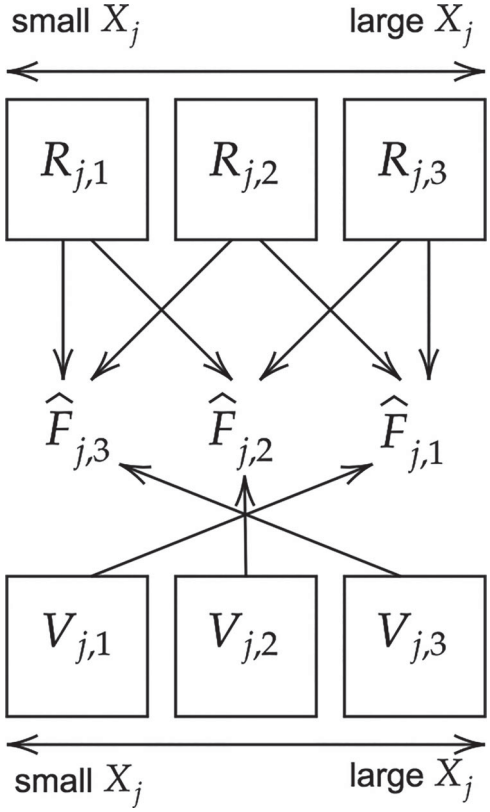

Figure 1. Illustration of use of reference and evaluation sets

3.2. Reference and Evaluation Sets

First we must obtain a reference distribution for each evaluation set residual. Consider the vector of residuals $(\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)$. Unlike previous methods, we do not assume a parametric form for the distribution of the residuals. Instead we split the residual vector into two parts, forming a reference set and an evaluation set. These sets are kept disjoint to avoid issues of dependence between the sets. For $j = 1, \ldots, p$, let Z be a random sample without replacement of size $\lfloor \rho n \rfloor$ from $\{1, \ldots, n\}$, where $\rho \in (0, 1)$ is the proportion of the data to be used in the reference set.[1] Let $Z^C = \{1, \ldots, n\} \setminus Z$. Then we define the reference set $R = \{\hat{\epsilon}_i | i \in Z\}$ and the evaluation set $V = \{\hat{\epsilon}_i | i \in Z^C\}$. The reference set is used to create several empirical estimates of the residual distribution. Residuals in the evaluation set are compared to these estimates to get empirical p-values. We do a large number of random splits into reference and evaluation sets. For simplicity of exposition we begin by describing the algorithm for a single split, and return to discuss multiple splits in Section 3.7. Figure 1 gives a pictorial representation of the process described in Sections 3.2–3.4.

[1] We have found $\rho = 1/2$ to be acceptable in practice.

3.3. Estimation of Empirical Error Distributions

In this section, we use the reference set residuals to estimate K empirical error distributions for each covariate X_j. These distributions allow us to identify portions of the data where evaluation set residuals do not follow their reference distribution.

A covariate X_j is heteroscedastically relevant if there is a relationship between the residuals and $\mathbb{X}_j$. We work under the mild assumption that the variance, as a function of each X_j, is continuous. Therefore, if X_j is an HRV, then there must be two disjoint intervals of X_j values with different residual distributions. This logic underlies the ad hoc method of inspecting residual plots to identify heteroscedasticity and informs tests such as Goldfeld–Quandt. We use this same logic to propose a more automatic procedure for identifying HRVs.

For each covariate X_j, $j = 1, \ldots, p$, we split the evaluation set residuals into K groups, based on their corresponding X_j values. We refer to the subset of V whose X_j values lie in the kth K-interval of $\{\mathbb{X}_{i,j} : i \in Z^C\}$ as the "(j,k)th evaluation set"; we denote this subset as $V_{j,k} \subset V$. Similarly, we define the "(j,k)th reference set" as $R_{j,k} \subset R$. Finally, we define the "(j,k)th comparison set" as $R^C_{j,k} := R \setminus R_{j,k}$. These are the residuals to which we compare $V_{j,k}$. If X_j is an HRV and K is sufficiently large to identify the heteroscedasticity, then for some k, the set of residuals $V_{j,k}$ will appear to have a different distribution than the residuals in $R^C_{j,k}$.

To quantify this, we first estimate empirical error distributions from the reference set. For each covariate X_j, the reference set R is used to estimate K empirical error distributions, $\hat{F}_{j,k}$, $k = 1, \ldots, K$, where

$$\hat{F}_{j,k}(t) = \frac{1}{|R^C_{j,k}|} \sum_{i \in Z} \mathbb{1}_{\left\{\hat{\epsilon}_i \in R^C_{j,k}\right\}} \mathbb{1}_{\{|\hat{\epsilon}_i| \le t\}}$$

where $|Z|$ denotes the cardinality of the set Z and $\mathbb{1}$ denotes the indicator function.

Each $\hat{F}_{j,k}$ is the empirical cdf of the magnitude of the estimated residuals in R whose X_j values fall outside of the kth K-interval of $\mathbb{X}_j$ in R. We compare residuals in V whose corresponding X_j values are in the kth K-interval of $\mathbb{X}_j$ to $\hat{F}_{j,k}$. In this way, we compare evaluation set residuals to reference set residuals with nonsimilar covariate values. If the null hypothesis is true and the distribution of residuals is the same across the domain of $\mathbb{X}_j$, then let F_ϵ denote the true distribution of the residuals. If we formed $\hat{F}_{j,k}$ using the true residuals ϵ instead of the estimated residuals $\hat{\epsilon}$, then by the Glivenko–Cantelli Theorem (Ferguson 1996), we would conclude that $\hat{F}_{j,k}$ is uniformly close to F_ϵ. If our model is correct and the sample size is large, the estimated residuals will generally be close to the true residuals and $\hat{F}_{j,k}$ will be close to F_ϵ.

We avoid parametric assumptions by fitting empirical estimates for the error distribution. This distinguishes HSS from tests like Breusch–Pagan and White, which assume the residuals follow a normal distribution with mean zero and unknown variance σ^2. A violation of the null hypothesis for these tests could be caused by nonnormally-distributed residuals, even if the distribution of residuals is constant. By abstaining from making parametric assumptions, we ensure that HSS is solely a test for heteroscedasticity.

3.4. Generation of p-values

Now we use the empirical error distributions defined above to calculate an empirical p-value for each residual in the evaluation set. For $j \in \{1, \ldots, p\}$, let $k(i,j) \in \{1, \ldots, K\}$ be the value of k such that if $i \in Z^C$, then $\mathbb{X}_{i,j}$ is in the kth K-interval of $\{\mathbb{X}_{i^*,j} : i^* \in Z^C\}$ and if $i \in Z$ then $\mathbb{X}_{i,j}$ is in the kth K-interval of $\{\mathbb{X}_{i^*,j} : i^* \in Z\}$. Then the (i,j)th empirical p-value is $p_{i,j} = 1 - \hat{F}_{j,k(i,j)}(|\hat{\epsilon}_i|)$. Under the null hypothesis, the residuals whose X_j value belongs to the kth K-interval of $\mathbb{X}_j$ and the residuals used to form $\hat{F}_{jk}$ share the same true distribution. If the null hypothesis does not hold, these groups of residuals will generally not come from the same distribution. We separate the p-values into pK sets or subgroups, one for each K-interval of each covariate. We denote these subgroups by $P_{j,k} = \{p_{i,j} : i \in Z^C, k = k(i,j)\}$.

3.5. A Scoring Function

We now present a scoring function that quantifies the amount of heteroscedasticity found in a given subset S of the covariates. Let $S \subseteq \{1, \ldots, p\}$. We begin by counting the number of α-level significant p-values in each subgroup $P_{j,k}$ for constant $\alpha \in (0,1)$:

$$N_\alpha(P_{j,k}) = \sum_{t \in P_{j,k}} \mathbb{1}(t \le \alpha), \text{ for } j = 1, \ldots, p, \text{ and } k = 1, \ldots, K.$$

We wish to compare the N_α values to their expected behavior under the null hypothesis. To do so, we must understand the null distribution of the N_α values.

We now argue that the probability mass function (pmf) of $N_\alpha(P_{j,k})$ is related to Wallenius' noncentral hypergeometric (WNH) distribution (Wallenius 1963) under the null hypothesis. Let $\text{WNH}(M, B, m, \omega)$ denote a Wallenius random variable with parameters M (population size), B (number of "successes" in the population), m (number of draws without replacement from the population), and ω (the odds ratio of drawing a "success"). A random variable $H \sim \text{WNH}(M, B, m, \omega)$ has probability mass function

$$P(H = h) = \binom{B}{h}\binom{M-B}{m-h} \int_0^1 (1 - t^{\omega/D})^h (1 - t^{1/D})^{m-h} dt$$

where $D = \omega(B - h) + (M - B) - (m - h)$. Let n_1 be the number of p-values that comprise subgroup $P_{j,k}$ and let n_2 be the size of the comparison set $R^C_{j,k}$. By way of analogy, imagine that we are drawing balls (without replacement) from an urn with n_1 white balls representing evaluation set residuals and n_2 black balls representing comparison set residuals. We let the first ball drawn from the urn be the most extreme residual, the second ball the second-most extreme, etc. For $\alpha \in (0,1)$, a p-value that reaches α-level significance must come from an evaluation set residual that is more extreme than at least $(1 - \alpha) \times 100\%$ of the comparison set residuals. If there are exactly c many α-level significant p-values in $P_{j,k}$, then the $\lfloor \alpha n_2 \rfloor + c$ most extreme residuals must have included exactly c evaluation set residuals *and* the next most extreme residual must be a comparison set residual. To complete the analogy of drawing balls from an urn, $P(N_\alpha(P_{j,k}) = c)$ is the probability of drawing exactly c white

balls in the first $\lfloor \alpha n_2 \rfloor + c$ draws and then drawing a black ball in the next draw, if white balls have weight ω capturing the propensity of evaluation set residuals to be α-level significant. We let $\omega = \frac{\alpha_0}{\alpha}$, where $\alpha_0 = \frac{1}{|Z^C|} \sum_{i \in Z^C} \mathbb{1}\left(|\hat{\epsilon}_i| > q(Z, 1-\alpha)\right)$ and $q(Z, 1-\alpha)$ is the $(1-\alpha)$-quantile of $\{|\hat{\epsilon}_i| \text{ s.t. } i \in Z\}$. Weighting the residuals accounts for overall random differences between reference and evaluation sets that might otherwise be attributed to heteroscedasticity. This process relates to the Wallenius distribution; we have

$$P(N_\alpha(P_{j,k}) = c) = P(H_1 = c)P(H_2 = 0)$$

where $H_1 \sim \text{WNH}(n_1 + n_2, n_1, \lfloor \alpha n_2 \rfloor + c, \omega)$ and $H_2 \sim \text{WNH}(n_1 + n_2 - \lfloor \alpha n_2 \rfloor - c, n_1 - c, 1, \omega)$. Within each subgroup $P_{j,k}$, we compare the observed number of significant p-values to the expected number. We can numerically compute the expectation, $E(N_\alpha(P_{j,k})) = \sum_{c=0}^{n_1} cP(N_\alpha(P_{j,k}) = c)$. We now assemble a scoring function that will give higher values to subsets of covariates that show larger differences from their expected number of α-significant p-values. We start by defining $T_\alpha(S)$, the sum of the differences between the number of expected α_0-significant p-values and the number of observed α_0-significant p-values across each of the K subgroups, in each of the covariates contained within S. We let

$$T_\alpha(S) = \sum_{j \in S} \sum_{k=1}^{K} \left| N_\alpha(P_{j,k}) - E(N_\alpha(P_{j,k})) \right|.$$

An alternative approach to defining $T_\alpha(\{X_j\})$ would be to directly aggregate all p-values for X_j (i.e., $\cup_{k=1,\ldots,K} P_{j,k}$) or equivalently, let $T_\alpha(\{X_j\})$ be $\left| \sum_{k=1}^{K} N_\alpha(P_{j,k}) - E \sum_{k=1}^{K} N_\alpha(P_{j,k}) \right|$. This does not work because if X_j is an HRV we may have in one K-interval generally lower than expected p-values and in another K-interval higher than expected p-values, so that aggregating those together (before taking absolute values) cancels out the signal.

We must adjust or normalize T_α before using it, since, under the null hypothesis, the value of T_α is dependent on α and on the size of S. Thus, we define our final scoring function $F_\alpha(S)$ to be a chi-squared-like statistic as follows. We let

$$F_\alpha(S) = \begin{cases} \frac{\left(T_\alpha(S) - E(T_\alpha(S))\right)^2}{E(T_\alpha(S))} & \text{if } T_\alpha(S) > E(T_\alpha(S)) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where

$$E(T_\alpha(S)) = \sum_{j \in S} \sum_{k=1}^{K} \sum_{x=0}^{|P_{j,k}|} \left| x - E(N_\alpha(P_{j,k})) \right| P(N_\alpha(P_{j,k}) = x).$$

Setting $F_\alpha(S)$ to zero when $T_\alpha(S) < E(T_\alpha(S))$ allows us to ignore subsets where the observed difference $T_\alpha(S)$ is actually less than its expected value and ensures that the scoring function solely rewards subsets that are more anomalous than expected.

Our scoring function $F_\alpha(S)$ represents a departure from statistics typically found in the subset scanning literature. Previously, subset scan statistics have placed greater weight on subsets of the data with unexpectedly high numbers of events and ignored subsets with unexpectedly low numbers (McFowland III, Speakman, and Neill 2013). Such a statistic would prove futile here, as heteroscedasticity can cause both unexpectedly high and unexpectedly low numbers of significant p-values. The use of the absolute value in the definition of T_α allows us to widen our search to subsets with an unexpected number of significant p-values regardless of whether this number is high or low. The flexibility brought by this change has the potential to be useful in other situations where we wish to perform an ambidirectional search for anomalies.

3.6. Maximizing the Scoring Function

Our goal is to maximize the scoring function $F_\alpha(S)$ for any α and subset S. In general, maximizing a function of S overall possible 2^p subsets of a covariate space of size p is computationally infeasible. We now introduce the linear time subset scanning (LTSS) property and show that it applies to our scoring function, reducing the problem to a search over only p possible subsets, after a single sort, which is easily feasible.

Definition 1 (Neill 2012). Fix $\alpha \in (0, 1)$. The scoring function $F_\alpha(S)$ and priority function $G(j; \mathbb{X})$ satisfy the strong LTSS property if and only if, for all $j = 1, \ldots, p$, $\max_{S:|S|=j} F_\alpha(S) = F_\alpha(\{X^{(1)}, \ldots, X^{(j)}\})$, where $X^{(j)}$ is the feature with the jth highest value of $G(\cdot; \mathbb{X})$.

The LTSS property guarantees that the subset S^* which maximizes $F_\alpha(S)$ for a given α must be the subset containing the k highest-priority covariates $\{X^{(1)}, \ldots, X^{(k)}\}$ for some k between 1 and p. Thus, to solve the global optimization problem for a fixed α, we can first sort the covariates by their priority value of $G(\cdot; \mathbb{X})$ and then compute $F_\alpha(S)$ with S taken to be one of the p subsets

$$\{X^{(1)}\}, \{X^{(1)}, X^{(2)}\}, \ldots, \{X^{(1)}, \ldots, X^{(p)}\}.$$

By inspection, the LTSS property guarantees that we have then achieved a global maximum. Neill (2012) gives a constructive theorem that produces a specific priority function $G(j; \mathbb{X})$ that follows directly from the scoring function $F_\alpha(S)$ if certain properties hold. This pair of functions is then guaranteed to satisfy the strong LTSS property.

Theorem 1 (Neill 2012). Let $F_\alpha(S) = F_\alpha(T, |S|)$ be a function of one additive statistic of subset S, $T_\alpha(S) = \sum_{j \in S} g(j; \mathbb{X})$ (where $g(j; \mathbb{X})$ depends only on feature X_j) and the cardinality of S, $|S|$. Assume that $F_\alpha(S)$ is monotonically increasing with $T_\alpha(S)$. Then $F_\alpha(S)$ satisfies the strong LTSS property with priority function $G(j; \mathbb{X}) = g(j; \mathbb{X})$.

We first note that Theorem 1 is considering a set of elements $\{X^{(1)}, \ldots, X^{(p)}\}$, that is, a fixed set of data covariates; therefore, it is concerned with $\max_{S \subseteq \{X^{(1)}, \ldots, X^{(p)}\}} F_\alpha(S)$. Moreover, the condition of monotonicity of $F_\alpha(S)$ in $T_\alpha(S)$ is considered for a fixed subset S and therefore, requires the $\frac{\partial F_\alpha(S)}{\partial T_\alpha(S)} > 0$, overall possible values of $T_\alpha(S)$ that the subset could exhibit. To see that our scan satisfies the conditions of Theorem 1, recognize that given a value of α, $F_\alpha(S)$ depends only on $T_\alpha(S)$, the cardinality of S, $E(T_\alpha(S))$, and indirectly $E(N_\alpha(S))$. $T_\alpha(S)$ is still random given S; however, the latter three values are all constant, given

S. Moreover, in the definition of $T_\alpha(S)$ we see that its outer summand ensures $T_\alpha(S)$ is an additive statistic of S, while its inner summand depends on only one feature (covariate), as required by Theorem 1. Finally, given that we set $F_\alpha(S)$ to zero when $T_\alpha(S)$ is less than its expectation, $F_\alpha(S)$ is monotonically increasing with $T_\alpha(S)$. Therefore, our scoring function $F_\alpha(S)$ satisfies the strong LTSS property with priority function

$$G(j;\mathbb{X}) = \sum_{k=1}^{K} \left| N_\alpha(P_{j,k}) - E(N_\alpha(P_{j,k})) \right|.$$

Therefore, if we prioritize covariates by the total difference between the $N_\alpha(P_{j,k})$ values and their expected values, we are guaranteed to find the subset which maximizes $F_\alpha(S)$ after searching over only p subsets.

Finally, we maximize $F_\alpha(S)$ overall α values to find the most anomalous subset, which is the subset S^* which maximizes the scoring function across all α values:

$$F(S^*) = \max_\alpha F_\alpha(S).$$

As shown by McFowland III, Speakman, and Neill (2013), we need only consider p-values found in the $P_{j,k}$ as possible values of α, as the statistic will achieve its maximum at one of these values.

3.7. Multiple Random Splits of the Reference and Evaluation Sets

In the previous section we computed a set S^* maximizing $F(\cdot)$. However, we do not know whether S^* is in fact significantly anomalous or not. Furthermore, S^* is based on a single random split of our data into reference and evaluation sets; we would potentially get a different S^* if a different split of the data were used. In our final step we deal with both of these issues, by using many different random reference and evaluation sets. We fix N_s, the number of random splits we will use. Then N_s separate times we run the procedure described in Sections 3.2–3.6, each time beginning with a newly drawn random reference and evaluation set split (as described in Section 3.2). Each iteration of the method will return its own most anomalous subset. We aggregate these subsets using a bootstrap method detailed in this section.

Specifically, to determine whether to include a covariate X_j in our final subset, we will check the proportion of random splits for which X_j was included in S^* and compare it to an estimate of what this proportion would be under the null, denoted $\widehat{\text{NIR}}_j$. We define the observed inclusion rate of X_j as $\widehat{IR}_j = N_s^{-1} \sum_{i=1}^{N_s} \mathbb{1}_{\{j \in S_i^*\}}$, where S_i^* is the most anomalous subset identified for the ith random split of the data. We include X_j in our final subset if $\widehat{IR}_j$ is significantly higher than $\widehat{NIR}_j$. We assess this via the bootstrap.

We bootstrap as follows. We leave $\mathbb{X}$ fixed. For bootstrap sample $b = 1, \ldots, B$, we draw a new vector of residuals $\hat{\epsilon}_b$ with length n randomly with replacement from the observed residual vector $\hat{\epsilon}$. Since we leave the covariates fixed but resample the residuals, our resampling scheme breaks any link between covariates and residuals. Then we perform HSS (as outlined in steps 2 through 6 of Algorithm 1) using the new residuals N_s times, with different reference and evaluation sets each time, giving $S^*_{b,l}$, for $b = 1, \ldots, B$, $l = 1, \ldots, N_s$. We then compute $\widehat{\text{IR}}_{b,j} = N_s^{-1} \sum_{i=1}^{N_s} \mathbb{1}_{\left\{j \in S^*_{b,i}\right\}}$.

Then we can find a p-value for the inclusion of each covariate by finding the quantile of the observed inclusion rate based on the estimated distribution of the null inclusion rate. After applying a Holm correction to these p-values, any covariate with a p-value below a desired significance threshold is included in the final subset.

Algorithm 1 Outline of Method

Given covariate matrix $\mathbb{X}$, estimated residual vector $\hat{\epsilon}$, parameter K, and number of random splits N_s:

1. Randomly split the residuals into reference set R and evaluation set V.
2. For $j = 1, \ldots, p$ and $k = 1, \ldots, K$, estimate empirical error distribution $\hat{F}_{j,k}(t)$
3. For each residual in V, calculate p-value $p_{i,j}$ based on the relevant $\hat{F}_{j,k}$ distribution. Form subgroups $P_{j,k}$. For a given α and subset S, $F_\alpha(S)$ can now be calculated.
4. For all α, identify the subset S which maximizes $F_\alpha(S)$. Because F_α satisfies LTSS, we use priority function $G(\cdot; X_j)$ to perform subset scanning.
5. Find the most anomalous subset S^* by maximizing $F_\alpha(S)$ overall α.

Repeat steps 1 through 5 N_s times, with different random reference and evaluation sets each time and calculate inclusion rates for each covariate.

6. Estimate the null inclusion rate distribution for all X_j using the bootstrap method.
7. Use the estimated distributions from step 6 to get a p-value for each X_j. Select covariates with significant p-values after Holm correction.

4. Further Considerations

4.1. Choosing an Appropriate K Value

The choice of K (i.e., the number of empirical distributions to estimate) is an important one. In general, it is best to use the smallest K capable of identifying the type of heteroscedasticity in the data. Choosing a K that is too large can dilute heteroscedastic signal by forcing the set of X_j values into very small partitions.

Choosing $K = 2$ is appropriate in situations where the error variance is suspected to have a monotonic relationship with a covariate. In such a scenario, because the error variance increases (or decreases) with X_j, residuals with small X_j values will have a different distribution than residuals with high X_j values. Therefore, $\hat{F}_{j,1}$ and $\hat{F}_{j,2}$ constructed from residuals with X_j values above and below the median, respectively, should be quite different from the evaluation set residuals they will be compared to.

On the other hand, nonmonotonic heteroscedasticity may not be detectable with $K = 2$. Assume, by way of example, that

the variance of the residuals increases with the distance of X_j from its median. If $K = 2$ were chosen in such a scenario, $\hat{F}_{j,1}$ and $\hat{F}_{j,2}$ would be quite similar. The evaluation set residuals will not have an anomalous signal and we are likely to conclude that X_j is not heteroscedastically relevant. This problem is solved by letting $K = 3$, where at least one of the three estimated empirical distributions will be different from the others.

Without prior knowledge of the shape of any heteroscedasticity present in the model, it is best to choose K sufficiently large to identify a wide variety of shapes. We find that in most practical situations, $K = 3$ is large enough to identify heteroscedasticity. We leave it as future work to find a data-driven method for the optimal selection of K.

4.2. Computational Considerations

The pmf of the noncentral hypergeometric distribution is challenging to compute, especially for large n. The probability of a given ball being drawn depends on the weights of all balls left in the urn, making the calculation of the pmf recursive. If computational time is a concern, then we recommend a modified version that reduces runtime considerably.

We now use the standard hypergeometric instead of Wallenius' noncentral hypergeometric. Let $\text{HG}(M, B, m)$ denote a hypergeometric random variable with parameters M (population size), B (number of "successes" in the population), and m (number of draws without replacement from the population). A random variable $H \sim \text{HG}(M, B, m)$ has pmf $P(H = h) = \frac{\binom{B}{m}\binom{M-B}{m-h}}{\binom{M}{m}}$. We rely on the same analogy as in Section 3.5, except here all balls in the urn are equally likely to be drawn. This assumption simplifies the calculations considerably but does not account for random differences between the reference and evaluation sets. We account for these differences by replacing α—the expected proportion of α-level significant p-values—with the observed proportion in the entire evaluation set, which is α_0 as defined above. Borrowing notation from Section 3.5, we have

$$P(N_\alpha(P_{j,k}) = c|\alpha_0) = P(H_1 = c)P(H_2 = 0)$$

where $H_1 \sim \text{HG}(n_1 + n_2, n_1, \lfloor \alpha_0 n_2 \rfloor + c)$ and $H_2 \sim \text{HG}(n_1 + n_2 - \lfloor \alpha_0 n_2 \rfloor - c, n_1 - c, 1)$. The rest of the method proceeds as stated above.

While this modification relies on an inexact assumption, the hypergeometric with α_0-correction is a close approximation of the WNH distribution. Intuitively, the distribution of white balls in a sample of C draws from a urn where the white balls have weight ω is similar to the distribution of white balls in a sample of ωC draws where all balls have equal weight. In the case of large n, α_0 converges in probability to α, so the weights in Wallenius' distribution converge to one and the two distributions approach equality.

5. Results

5.1. Overview

In this section, we compare the performance of HSS to several alternatives described in the following section. We consider three simulated settings as well as one real data example. The simulations and their results are described in detail in Section 5.3–5.5.

Throughout this section, we will use four metrics to assess the performance of each method: power, recall, precision, and Jaccard similarity. Let S^* be the subset of HRVs as chosen by a particular method and let S_0 be the true subset of HRVs. Then, we define the power of an individual simulation run with nonempty S_0 to be $E\mathbb{1}(|S^*| > 0)$. We define recall to be $E\frac{|S^* \cap S_0|}{|S_0|}$, the expected proportion of the HRVs identified by the method. We define precision to be $E\frac{|S^* \cap S_0|}{|S^*|}$, the expected proportion of covariates identified by the method that are truly heteroscedastically relevant. Finally, we define Jaccard similarity, a combination of precision and recall, to be $E\frac{|S^* \cap S_0|}{|S^* \cup S_0|}$. For all simulations, we use tuning parameter values $K = 3$ and $\rho = 0.5$. For each sample size setting and type of heteroscedasticity, we create 128 datasets and perform HSS with $N_s = 50$ random splits for each dataset.

5.2. Alternative Methods

Our method is the first built to identify individual HRVs. To create comparison methods, we must modify existing global tests so that they can perform the same identification task.

Koenker's modification of the Breusch–Pagan test for global heteroscedasticity tests whether there is a linear relationship between the covariates and the residuals. The test statistic is derived from the coefficient of determination of a regression of the squared residuals on the covariates. We will fit the same regression model to determine which covariates are heteroscedastically relevant. The covariates that have statistically significant coefficients after a Holm multiplicity correction are considered heteroscedastically relevant.

The Goldfeld–Quandt test is also easily modified for an identification purpose. Because this test is only specified for one covariate at a time, we simply perform the test p times and get p-values for each covariate. We again use a Holm correction and consider any covariates with significant p-values after correction to be heteroscedastically relevant.

We elect not to use the White and Li and Yao tests as comparisons due to the difficulty of modifying them for the purpose of identifying HRVs.

5.3. Linear Regression

We first construct a simple data setting where none of the p covariates have any covariance with the other covariates. We simulate $n \times p$ covariate matrix $X \sim N(\mathbf{5}, \mathbf{\Sigma})$, where $\mathbf{5} = (5, \ldots, 5)$, $\mathbf{\Sigma} = 4I_p$ and I_p is the $p \times p$ identity matrix. We let sample size n vary from 250 to 3000 and set the number of covariates p equal to 25. We let the number of true HRVs d be 5. We used $B = 499$ bootstrap resamples except when $n = 3000$ when we used $B = 249$ bootstrap resamples. Throughout this section, we will focus on three types of heteroscedasticity:

- Monotonic heteroscedasticity occurs when error variance increases (or decreases) as the value of a heteroscedastically relevant variable increases. This form of heteroscedasticity is

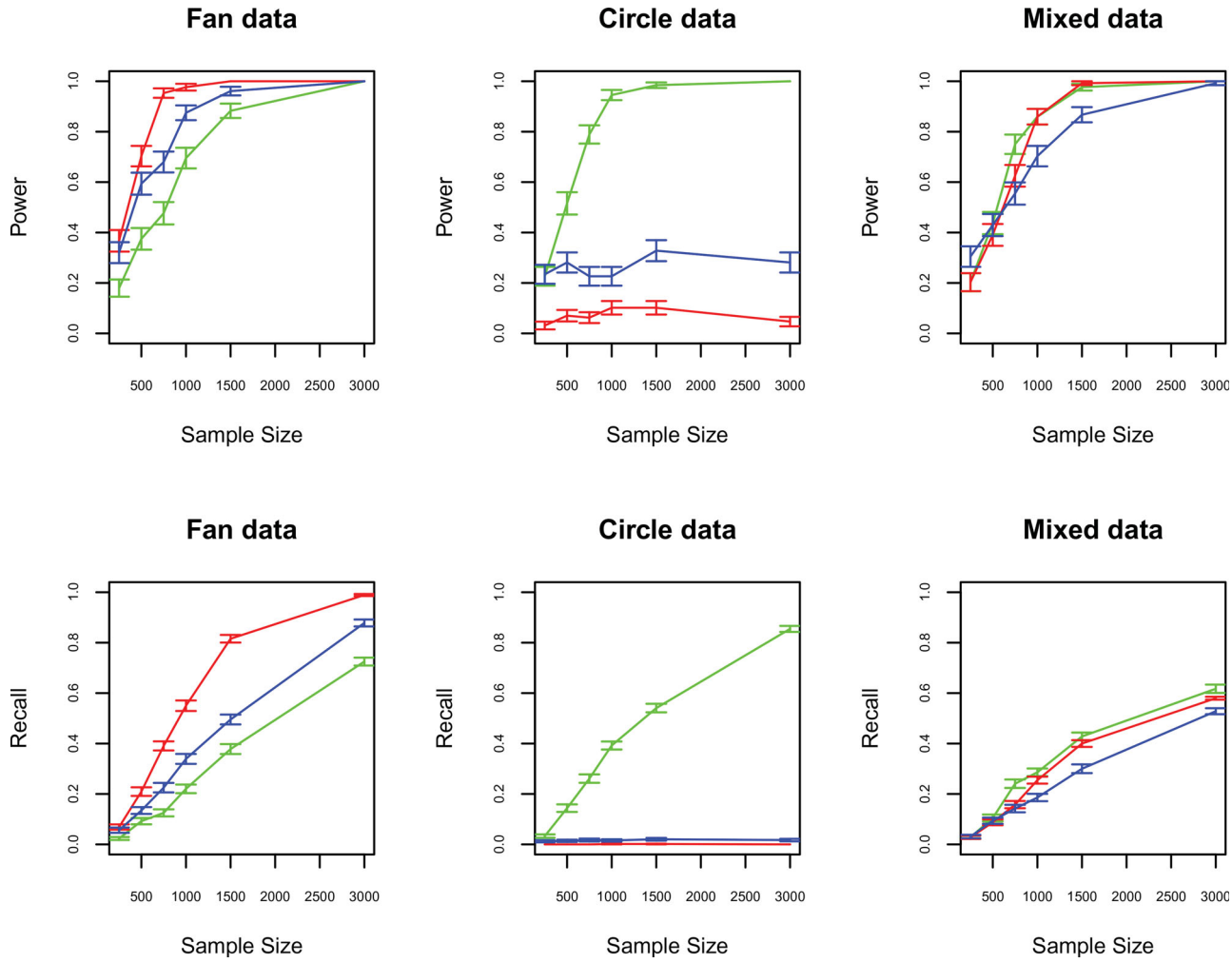

Figure 2. Model comparisons for "fan," "circle," and "mixed" data types and linear regression. Results presented for HSS (green), modified Breusch–Pagan (red), and modified Goldfeld–Quandt (blue).

often referred to as "fan" or "megaphone" error variance due to the shape of its residual plot. In our simulations, we let $\epsilon_i \sim N(0, (1 + \sum_{j=1}^{d} \mathbb{X}_{i,j})^2)$ for $i = 1, \ldots, n$.

- Nonmonotonic heteroscedasticity occurs when error variance is not a strictly increasing or strictly decreasing function of a covariate. Celik (2015) refers to a particular form of nonmonotonic heteroscedasticity as "butterfly residuals," where error variance is large at the extreme values of a covariate and small for values near the median of the covariate. For ease of simulating data, we focus on the inverse, which we call "circle" residuals. We let $\epsilon_i \sim N(0, (\sum_{j=1}^{d} \eta_{i,j})^2)$ for $i = 1, \ldots, n$, where

$$\eta_{i,j} = \begin{cases} \max(1, \mathbb{X}_{i,j}^2) & \text{if } \mathbb{X}_{i,j} < 5 \\ \max(1, (10 - \mathbb{X}_{i,j})^2) & \text{if } \mathbb{X}_{i,j} > 5 \end{cases}. \quad (2)$$

- Mixed heteroscedasticity occurs when some covariates exhibit a monotonic relationship with the error variance while others have a nonmonotonic relationship. We let half of the HRVs have fan error variance and let the other half have circle error variance. For $i = 1, \ldots, n$, and $\eta_{i,j}$ as above, we let

$$\epsilon_i \sim N\left(0, \left(2 \sum_{j=1}^{\lceil d/2 \rceil} \mathbb{X}_{i,j} + \sum_{j=\lceil d/2 \rceil + 1}^{d} \eta_{i,j}\right)^2\right).$$

After generating $\mathbb{X}$ and ϵ, we let $Y_i = \sum_{j=1}^{p} \mathbb{X}_{i,j} + \epsilon_i$ and then estimate residuals $\hat{\epsilon}_i$ by performing ordinary linear regression with all 50 covariates included as predictors.

Figure 2 compares power and recall across a range of sample sizes and types of heteroscedasticity. Precision and Jaccard similarity tend to be nearly equal to power and recall, respectively, so we omit plots for these measures. The "fan" errors are (by definition) linearly heteroscedastic, so that is where the (modified) Breusch–Pagan (and Goldfeld–Quandt) tests are built to perform optimally. Thus, in those cases they tend to outperform HSS. As expected, the modified Breusch–Pagan test outperforms the modified Goldfeld–Quandt test in all settings. The gap between HSS and existing methods is smallest for power (and precision, not presented). This indicates that much of the gap in performance between HSS and the modified Breusch–Pagan is due to HSSs ability scan over subsets of covariates—combining their individual signals—to detect the presence of heteroscedasticity.

Figure 2 also confirms what was expected to be a major advantage of our method over existing methods in that HSS has the power to detect nonmonotonic heteroscedasticity. Existing tests assume a linear relationship between error variance and HRVs and therefore, struggle in detecting nonmonotonic heteroscedasticity. The power, precision, recall, and Jaccard similarity of HSS all approach one as n grows, which is not

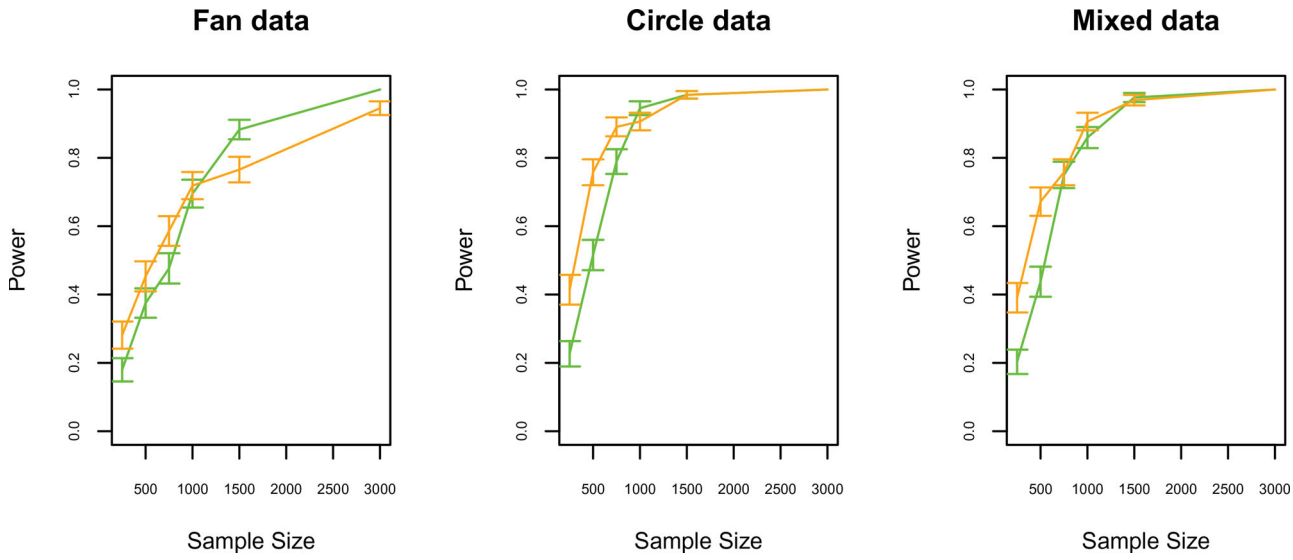

Figure 3. Power comparisons with the global method of Li and Yao. Results presented for HSS (green) based on linear regression residuals and Li and Yao (orange).

true of the modified Breusch–Pagan or Goldfeld–Quandt. Furthermore, Figure 2 shows that HSS is capable of handling multiple forms of heteroscedasticity in a single dataset. It is apparently the case that having mixed heteroscedastic "shapes" (e.g., monotonic and nonmonotonic) causes the bootstrap some difficulty (as recall is lower than for either purely "fan" or "circle"). There may be alternative methods for selecting the final subset than the bootstrap, but we leave a study of this for future work.

We compare the power of our test to the power of the global Li and Yao test (coefficient of variation version presented here), as shown in Figure 3. The Li and Yao test is powered to detect monotonic or nonmonotonic heteroscedasticity. The power of HSS is similar to the power of the Li and Yao test (with Li and Yao's tests generally having greater relative power as the ratio p/n is larger); HSS is not constructed to have optimal global power but rather to be able to detect individual HRVs, so we are pleased that HSS has good overall power.

Thus, overall, HSS performs similarly to (or sometimes better than) existing methods even when a global test for heteroscedasticity is desired. Simulations with null (homoscedastic) data indicate that Type I error rate is generally maintained near the nominal 0.1 level, with values between 0.16 (at $n = 250$) and 0.10 (when $n \geq 1500$). (Increasing B yields some [small] improvement[s] in the level at smaller sampler sizes.)

5.4. Lasso Regression

Our method is especially useful when identifying HRVs from among a large group of covariates. When p is small, manual inspection of the residual plots can help to identify heteroscedasticity. When p is large, HSS obviates the need to manually check a large number of residual plots. Lasso regression, proposed by Tibshirani (1996), is a commonly-used tool for variable selection when p is large. In this section, we investigate the performance of HSS when Lasso regression is used. We note that our methodology depends on classical large-n-fixed-p asymptotics, so that if n is not large relative to p then HSS may be less effective.

We simulate $\mathbb{X}$ as in Section 5.3. We let $p = 25$ and let n vary from 250 to 3000. We use 10-fold cross-validation to select the Lasso tuning parameter λ. We let response $Y_i = \epsilon_i$, where ϵ_i is simulated as in Section 5.3 for fan, circle, and mixed error variance. Thus, the five HRVs $X_1, \ldots, X_5$ have no direct effect on the response Y. These covariates are thus, unlikely to be included in the Lasso model after variable selection is performed. For Lasso regression, we will use residuals that arise after Lasso variable selection but will still perform our search for HRVs over the entire covariate space.

The simulation confirms that the use of Lasso regression instead of OLS does not noticeably impact the performance of HSS. Although the Lasso regression coefficients differ from the OLS regression coefficients, this does not effect the relationship between the covariates and the error terms. Even in the case where the Lasso regression coefficients corresponding to HRVs are zero (as in our simulations), the HRVs can be detected without issue. We present the global power results in Figure 4 and also include the (global) Li and Yao test (which uses the OLS residuals rather than the lasso residuals). We omit the recall plots due to their strong similarity with Figure 2. Figure 4 shows that HSS has similar power to the global Li and Yao test once we are in a large-n-fixed-p regime, with results being similar to the OLS case. Again, HSS is built to identify the individual HRVs, rather than built to be a global test, so its good overall power performance is positive.

5.5. Education Data Example

We now examine education expenditure data presented by Chatterjee and Hadi (2006). The outcome variable is per capita education expenditure in 1975, with three predictors: per capita income (X_1), number of residents per 1000 under 18 years of age (X_2), and number of residents per 1000 living in urban areas (X_3). Each row of the data corresponds to one U.S. state, so there are 50 observations with three covariates for each. The authors remove outlier Alaska from the data, so we follow suit and proceed with the other $n = 49$ states. Although manual inspection of the residual plots is feasible here, we use this

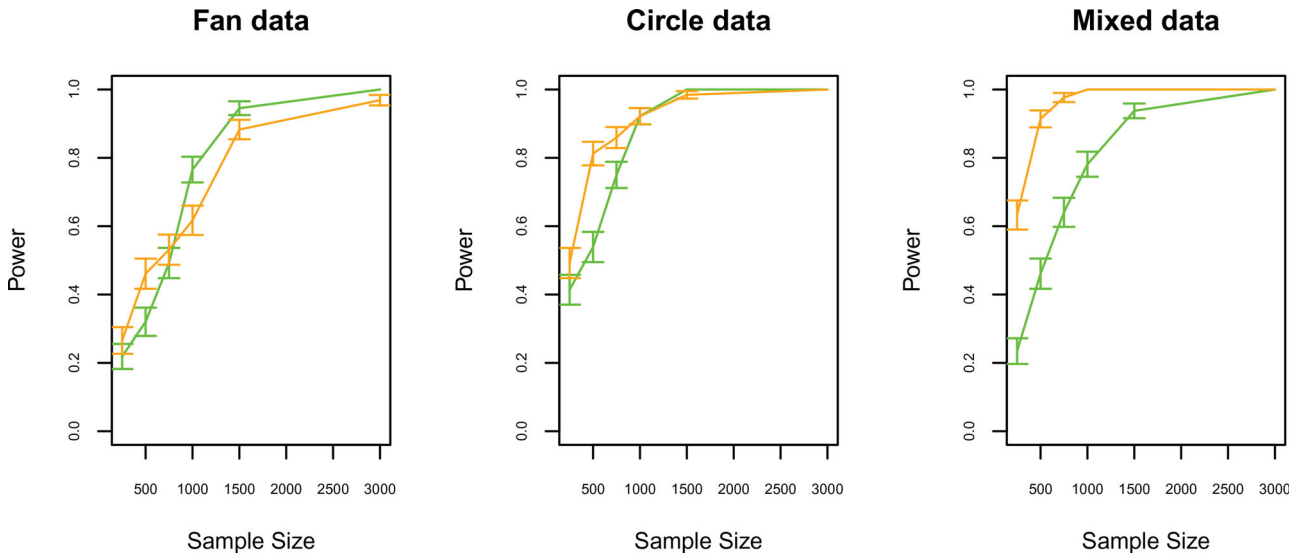

Figure 4. Power comparisons with the global method of Li and Yao, with HSS using Lasso regression residuals. Results presented for HSS (green) and Li and Yao (orange).

example to show that HSS can identifying heteroscedasticity in a real data setting where the Breusch–Pagan test lacks power. The authors note that this covariate is visibly heteroscedastic, with error variance increasing at larger per capita income levels, as seen in Figure 5.

We use ordinary linear regression with X_1, X_2, and X_3 as covariates to estimate residuals and then proceed with HSS and with the modified Breusch–Pagan method. For HSS, we use the tuning parameters of $K = 3$ and $\rho = 0.5$. We perform 100 random splits of the data and observe inclusion rates of 0.99, 0.18, and 0.08 for X_1, X_2, and X_3, respectively. Using the bootstrap with $B = 199$ to get a final subset, we get an estimated p-value of 0.01 for X_1 (and p-values near 1 for the other two variables) and conclude that X_1 is the only HRV.

The modified Breusch–Pagan test does not identify any of the covariates as HRVs. Furthermore, the standard Breusch–Pagan test is insignificant ($p = 0.143$). In this simple example, HSS has shown the ability to detect heteroscedasticity and pinpoint the covariate responsible for it. The Breusch–Pagan test cannot perform either of these tasks, including the global test that it is best suited for. This is perhaps due to the nonlinear relationship between per capita income and error variance seen in the data. While error variance appears to increase as income grows, it does not increase at a consistent rate. The Breusch–Pagan test assumes a linear relationship, while HSS looks more broadly for differences in error distribution across the range of per capita income levels, giving greater power in this setting.

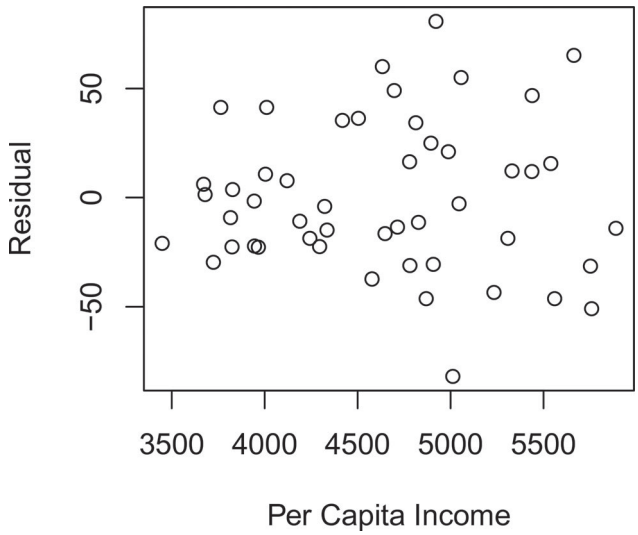

Figure 5. Residual plot for per capita income from education expenditure data.

6. Discussion

This article makes several contributions to the literature on heteroscedasticity detection. We frame the problem as one of anomalous pattern detection and present the heteroscedastic subset scan (HSS) algorithm. This novel method represents a departure from previous global tests for heteroscedasticity, as HSS seeks to identify the subset of the covariate space that is most responsible for violations of the homoscedasticity assumption. Our method uses a reference set from the data to create an estimate of the empirical error distribution which is then used to identify covariates along which the error distribution is nonconstant. The method is guaranteed to efficiently identify the most anomalous subset by calculating a scoring function for only a linear number of subsets. Furthermore, our method makes a contribution to the subset scanning literature by proposing an ambidirectional scoring function which prioritizes departures both above and below the expected value.

In simulation experiments, HSS outperformed existing methods in a wide range of scenarios. Existing methods have little power to detect nonmonotonic heteroscedasticity, but the nonparametric nature of HSS gives it the ability to detect such relationships. Furthermore, HSS can be used to identify relationships between higher moments of the error distribution and the covariates, another area where existing methods are underpowered.

Future work could investigate several outstanding areas of interest. We have recommended that the tuning parameter K be set to 3 in most settings. Future research may focus on data-driven selection of K. Another extension of the method would be to identify heteroscedastically relevant records in addition to variables. McFowland III, Speakman, and Neill (2013) present

an algorithm that iterates between scanning over records and attributes until the most anomalous subspace of the data is identified. Such an extension could help the user further narrow in on the source of homoscedasticity violations in the model. Finally, future research may choose to focus on the identification of heteroscedastically relevant sets, groups of covariates that are jointly (but not necessarily marginally) related to the error variance. While our method searches only for marginal heteroscedasticity, detection of joint heteroscedasticity could also be useful from a model diagnostic perspective.

Supplementary Materials

The online supplementary materials provide a basic R implementation of the HSS algorithm and the education expenditure data utilized in Section 5.5.

Acknowledgments

Evan Olawsky was originally an author on this paper. He has chosen to not participate in the revision, and to cede authorship of the paper.

Funding

Charles R. Doss is partially funded by NSF grant DMS-1712664 and NSF grant DMS-1712706. Edward McFowland III gratefully acknowledges funding support from the NSF Program on Fairness in Artificial Intelligence in Collaboration with Amazon, grant IIS-2040898.

ORCID

Charles R. Doss http://orcid.org/0000-0003-1364-5222
Edward McFowland III http://orcid.org/0000-0001-5249-7117

References

Breusch, T. S., and Pagan, A. R. (1979), "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, 47, 1287–1294. [1,2]

Celik, R. (2015), "Stabilizing Heteroscedasticity for Butterfly-Distributed Residuals by the Weighting Absolute Centered External Variable," *Journal of Applied Statistics*, 42, 705–721. [2,3,8]

Chatterjee, S., and Hadi, A. S. (2006), *Regression Analysis by Example* (4th ed.), Hoboken: Wiley. [9]

Cook, R. D., and Weisberg, S. (1983), "Diagnostics for Heteroscedasticity in Regression," *Biometrika*, 70, 1–10. [2]

Dette, H., and Munk, A. (1998), "Testing Heteroscedasticity in Nonparametric Regression," *Journal of the Royal Statistical Society*, Series B, 60, 693–708. [1,2]

Ferguson, T. S. (1996), *A Course in Large Sample Theory*, Texts in Statistical Science Series, Boston, MA: Chapman & Hall. [4]

Glejser, H. (1969), "A New Test For Heteroskedasticity," *Journal of the American Statistical Association*, 64, 316–323. [2]

Goldfeld, S. M., and Quandt, R. E. (1965), "Some Tests for Homoscedasticity," *Journal of the American Statistical Association*, 60, 539–547. [2]

Harrison, M. J., and McCabe, B. P. M. (1979), "A Test for Heteroscedasticity Based on Ordinary Least Squares Residuals," *Journal of the American Statistical Association*, 74, 494–499. [2]

Koenker, R. (1981), "A Note on Studentizing a Test for Heteroscedasticity," *Journal of Econometrics*, 17, 107–112. [2]

Koenker, R., and Bassett, G. (1982), "Robust Tests for Heteroscedasticity based on Regression Quantiles," *Econometrica*, 50, 43–61. [2]

Li, Z., and Yao, J. (2019), "Testing for Heteroscedasticity in High-Dimensional Regressions," *Econometrics and Statistics*, 9, 122–139. [1,2]

McFowland III, E., Speakman, S., and Neill, D. B. (2013), "Fast Generalized Subset Scan for Anomalous Pattern Detection," *Journal of Machine Learning Research*, 14, 1533–1561. [2,5,6,10]

Muller, H.-G., and Stadtmuller, U. (1987), "Estimation of Heteroscedasticity in Regression Analysis," *The Annals of Statistics*, 15, 610–625. [2]

Neill, D. B. (2012), "Fast Subset Scan for Spatial Pattern Detection," *Journal of the Royal Statistical Society*, Series B, 74, 337–360. [2,5]

Newey, Whitney K, and Powell, J. L. (1987), "Asymmetric Least Squares Estimation and Testing," *Econometrica*, 55, 819–847. [2]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, 58, 267–288. [9]

Wallenius, K. T. (1963), "Biased Sampling: The Non-central Hypergeometric Probability Distribution," Ph.D. thesis, Stanford University, Department of Statistics. [4]

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838. [2]