MULTIPLE DATA SPLITTING FOR TESTING

By

Joseph P. Romano
Cyrus DiCiccio

Technical Report No. 2019-03
April 2019

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

MULTIPLE DATA SPLITTING FOR TESTING

By

Joseph P. Romano
Stanford University

Cyrus DiCiccio
LinkedIn

Technical Report No. 2019-03
April 2019

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

Multiple Data Splitting for Testing

Joseph P. Romano
Departments of Statistics and Economics
Stanford University
romano@stanford.edu

Cyrus DiCiccio
Stanford University
and LinkedIn
cyrusd@stanford.edu

April 1, 2019

Abstract

Data splitting is a ubiquitous tool based on the following simple idea. Partition the data into disjoint "splits" so that one portion of the data can be allocated to simplifying the statistical problem by: learning about the underlying model; estimating, selecting, or eliminating nuisance parameters; reducing the dimension of the problem; choosing the form of an estimator or test statistic, etc. Then, formal inference can be based on the independent (complement) second portion of the data. Typically, the problem of constructing tests that control a measure of Type 1 error is made much easier by data splitting. Moreover, in single or multiple testing problems that include a large number of parameters, there can be a dramatic increase of power by reducing the number of parameters tested, particularly if the underlying parameter is relatively sparse. A criticism of single splitting is a loss of power or efficiency due to testing on only a fraction of the data, though carefully selection of a test statistic may in turn improve power. In addition, a single split method can be viewed as unsatisfactory since it would rely on an arbitrary split or ordering of the data (and hence violate the sufficiency principle). To combat the possible loss of power with single splitting, we propose repeated splitting by computing tests over many splits of the data. Assuming each split results in a p-value, the problem becomes how to combine these now very dependent p-values to construct one overall procedure. In this paper, we present some general methods that exhibit rigorous error control, both in finite samples and asymptotically. The power of these methods are compared with each other, as well as the power of full data tests and tests using only a single split of the data.

KEY WORDS: Data Splitting, Hypothesis Testing, P-values, Subsampling, U-statistics.

1 Introduction

Data splitting, a technique which involves partitioning a data set into disjoint "splits" or subsamples which can then be used for various statistical tasks, has widespread application in the statistical literature. Typically, one portion of the data is used for some form of selection (such as model fitting, dimension reduction, or choice of tuning parameters), and then the second, independent portion of the data is used for some further purpose such as estimation and model fitting. In addition, data splitting can be used in prediction to assess the performance of models (where a portion of the data has been used to select and/or fit a model and the remainder is used to assess the performance of the selected model) or in inference to perform tests of significance after hypotheses or test statistics have been selected. Data splitting has become a useful remedy for data-snooping (giving valid inference after selection of a hypothesis), estimating nuisance parameters, and avoiding over-fitting in prediction problems. Some examples of problems that benefit from the use of data splitting are given in Section 2.

Early works employing data splitting, such as Larson (1931) and Horst (1941), focused mainly on prediction error in regression models. They demonstrated the decrease in predictive power that a linear model has when used for out of sample predictions when compared with in sample predictions. The ideas pioneered in these works have since evolved into the now ubiquitous cross-validation method for assessing predictive models. A detailed history of data splitting as applied to assessment of statistical predictions is given in Stone (1974).

Another, often neglected, use of data splitting is as a remedy for "data snooping." Researchers who do not specify hypotheses to test prior to collecting data often formulate hypotheses based on the data that they observe. After computing p-values for many possible hypotheses, only those that are most significant are reported. It becomes difficult to distinguish results that are truly present in the population from artifacts of the collected data. Ioannidis (2005) suggests that data snooping is a primary reason why it is more likely for a research claim to be false than true. While multiple testing procedures controlling the familywise error rate or the false discovery rate can address the problem of data snooping, these methods are often not done in practice. More subtle forms of data snooping, which are often not taken into account, involve goodness of fit testing prior to hypothesis testing, and using several models for inference while only reporting the most significant results. A simpler and more interpretable solution is to use a portion of the data to suggest hypotheses, and a second, independent portion to test the selected hypotheses.

The downside to this data splitting method is a potential loss of power, although it is somewhat unclear how the loss of power resulting from only testing on a split of the data compares with the loss of power resulting from using a correction for multiplicity such as

the Bonferroni correction. Unlike estimation problems, where simply averaging estimates taken over several splits of the data can be expected to improve accuracy, it is not obvious how to combine information across several splits of the data in inference problems, such as controlling error rates in testing.

Data splitting is often not used by researchers who fear that their results will appear less significant if only tested on a fraction of the available data. To avoid the inherent loss of power associated with data splitting, it may be beneficial to perform selection and inference on the full data. In certain problems, valid inference after model selection using the entirety of the data for both the model selection and inference steps can be performed by conditioning on the selected model. This approach is studied in Potscher (1991) as well as Fithian et al. (2015). Once again, it remains unclear whether the loss of power from conditioning on a selected model is better than the loss of power from data splitting.

A somewhat more obscure use of a data splitting technique is to use a portion of the data to select an efficient test statistic, and perform the test on the remainder of the data. While the applications of data splitting to improve prediction (and assessment of predictors) are well studied, there is very little written on the efficacy of data splitting for improving inference.

Works in this vein include Cox (1975) and Moran (1973), although it is worth noting that Cox mentions he was "unable to trace its history." The hope for data splitting in this context is that the loss of power associated with performing a test on a smaller subsample than the full data could be offset by selecting a suitably efficient test statistic, leading to a more powerful test than naively using the full data. Even in testing problems where an "optimal" (e.g. UMP invariant or maximin) testing procedure exists, such tests may have low power against particular alternatives. Using a portion of the data to estimate how the parameter departs from the null can suggest a more efficient statistic based on the second split of the data. Having some knowledge about the direction of the alternative may compensate for the loss of power due to splitting the data.

An important example is testing that a vector of parameters is zero against a sparse alternative. Examining the first part of the data can suggest which parameters are likely to be non-zero, leading perhaps to a more efficient test statistic based on the second part of the data. However, it is unclear whether the improvement in power attained from reducing the dimension of the problem is sufficient to justify the loss of power from splitting the data, not to mention the challenge of deriving valid conditional procedures.

More recent works exploring the use of data splitting to improve power pertain to multiple testing problems. Rubin et al. (2006) employs an initial split of the data to select thresholds used for testing on the second split. Ignatiadis et al. (2016) gives a weighted testing procedure where the first split of the data is used to choose weights of individual hypotheses, and the second split is used for testing. Interestingly, in these examples, use of the initial split of the

data is not used to select efficient statistics; however, these examples are similar in spirit to selecting efficient test statistics in that both methods use an initial split of the data to gain some knowledge about the particular alternative hypothesis, and demonstrate that this can indeed improve power.

A major criticism of data splitting is that it introduces randomization into data analysis. Two researchers nominally applying the same methodology to the same data set could reach different conclusions if they choose different subsamples for each stage. In i.i.d. problems, such a method violates the sufficiency principle, which demands that any inference should be invariant with respect to the ordering of the data. But, combining p-values across a large number of splits could help lessen the sensitivity of inference after selection to the random allocation of data to each split. There have been several recent papers addressing the issue of combining p-values obtained over multiple splits of the data. van de Wiel et al. (2009) addresses combining p-values over multiple splits of a data testing the hypothesis that two models give the same prediction error. When assessing prediction error, the two splits are used as training and test sets, respectively, resulting in a p-value based on the residuals. Then, p-values are aggregated using the median.

A similar, albeit less general, method of combining a fixed number of splits is proposed by Meinshausen et al. (2009). In high dimensional linear models, data splitting is often used to reduce the dimension of the problem. A portion of the data is used to choose predictors (using an algorithm that gives asymptotic guarantees of selecting predictors including the true model such as SURE selection, given in Fan and Lv (2008)), while the remainder is used for inference about the selected predictors. This approach is taken in Wasserman and Roeder (2009) and Barber and Candes (2016). Meinshausen et al. (2009) propose a method of combining p-values in these regression problems while giving asymptotic control of the FWER and FDR, and they provide numerical evidence that multi-split methods can provide an improvement over using only a single split. These papers consider only a fixed number of splits of the data. It would be desirable to give methods that allow use of an arbitrary number of splits of the data, and to study the effect of the number of splits used for testing.

The main goal of this paper is to develop testing procedures that allow an arbitrary number of splits of the data by combining results across splits. The main focus is primarily on rigorous Type 1 error control, though power considerations are discussed as well. We introduce four general classes of methods of combining p-values over an arbitrary number of splits of the data. The first is similar in spirit to the approach taken in Meinshausen et al. (2009), but has *finite* sample control of error rates in a very *general* setup. While easy to apply, the proposed method is conservative in that the resulting error rate is strictly below the nominal level. This leads to the following questions. Are conservative methods for combining splits an improvement over using a single split? How conservative are the

conservative methods, and can they be improved upon? How do the multi-split methods depend on the number of splits used, and is there a best number of splits?

Section 2 presents some examples to fix ideas. Section 3 gives a formal description of data splitting for inference, as well as several very general methods of combining test decisions or p-values computed over multiple splits of the data. The methods in this section give exact finite sample control of the null rejection probabilities, but can be conservative in the sense that the null rejection probabilities can be below the nominal level (even asymptotically). We also offer an alternative approach (based on concentration inequalities) which can be more powerful when the split sizes are small.

Section 4 gives approximate methods of combining splits which are generally applicable under mild assumptions, but have asymptotic level equal to the nominal level. Rather than bounding tail probabilities for the average (or median) p-value over all splits, one can accurately estimate the limiting distribution as a basis for inference, resulting in more powerful procedures. For example, for a small number of splits (or arbitrary size), the limiting distribution of the average p-value (averaged over the splits) can be consistently approximated via simulation or resampling. For a larger number of splits (of smaller size), we can approximate the distribution by exploiting the U-statistics structure of the average (or median) p-value. Section 5 provides a conclusion, as well as some important questions that remain for future work. The Appendix (Sections 6, 7, 8) concern some auxiliary results and proofs. Some general asymptotic theory for U-statistics (as well as corresponding M-statistics) is given in Section 6 which allows for the kernel order to increase with sample size. Section 7 concerns verification of a non degeneracy condition in Corollary 4.1. Section 8 gives useful results and proofs of the results stated in the main text.

Numerical and theoretical comparisons of these methods, as well as recommendations for choice of tuning parameters are provided in the Supplement. First, the power of the methods combining splits compares with the UMP test in the case of testing for a single normal mean. While there is no selection in this example, it is helpful to understand how various methods of combining splits compares with the UMP test. Theoretical results and simulations are reported. Surprisingly, split sampling based on small splits attain the optimal limiting local power function in this example.

Then, we examine single (but multivariate) testing for the null hypothesis that many means are zero as well as for the moment inequality problem. Special attention is payed to sparse alternatives, and whether power can be improved by selecting means which are suspected to be non-zero. In these examples, we compare the various methods of splitting the data, consider several split sizes, and multiple methods of selection.

2 Some Motivating Examples

In this section, we mention some examples that will help fix ideas and provide motivation for data splitting.

Example 2.1. (Multivariate Mean) Suppose $X_i = (X_{i,1}, ..., X_{i,p})$, $i = 1, ..., n$ are i.i.d. multivariate normal with mean $(\mu_1, ..., \mu_p)$ and identity covariance. We are interested in testing $H_0 : \mu_1 = \cdots = \mu_p = 0$. We could use the classical Chi-squared test, but since the critical value increases with the degrees of freedom p, it may be helpful to first "weed out" those indices for which the mean appears close to zero. Instead, we use an initial split of the data to select which means to include. Suppose that S is a randomly chosen subset of size b from the set $\{1, ..., n\}$, which defines a "split" of the data. We could "select" means using the data in $S^c_{n,i}$ (the data not corresponding to S) according to whether

$$\sqrt{n-b}\left|\bar{X}_j(S^c)\right| := \frac{1}{\sqrt{n-b}}\left|\sum_{i\in S^c} X_{i,j}\right|$$

exceeds $z_{1-\beta}$ for some $\beta < 1/2$, and then perform a test using the Chi-squared statistic

$$\sum_{j \text{ selected}} b\bar{X}_j^2(S)$$

to get a p-value using the quantiles of the Chi-squared distribution with degrees of freedom equal to the number of selected means. Alternatively, one may test using other test statistics, such as the maximum absolute sample mean or Tukey's higher criticism; see Arias-Castro et al. (2011).

Cox (1975) studied a single split approach in the case where the alternatives are restricted so that only of the populations may have a nonzero mean. But rather than an arbitrary split of the data, we may wish to combine p-values over many splits of the data. Moran (1973) investigated the usefulness of using a portion of the data to determine the direction in which a parameter departs from a null hypothesized value and using the remainder of the data to construct a powerful test against alternatives in the selected direction. Both these authors considered one split of the data. Based on one split, the p-value is $U(0, 1)$ under H_0. Here, we compute many (dependent) p-values across multiple splits of the data. The problem we wish to address is how to combine these p-values in a valid way so that rigorous error control is maintained.

Example 2.2. (Regression Problems) Consider the linear regression model

$$Y = X\beta + \epsilon$$

where $Y \in \mathbb{R}^n$ is a vector of response variables, X is an $n \times p$ design matrix and $\epsilon \in \mathbb{R}^n$ is a vector of error terms with ϵ_i i.i.d. $N(0, \sigma^2)$.

In model selection problems, the majority of selection algorithms may include many erroneous noise variables and it can be challenging to report models that control an appropriate error rate. This difficulty is particularly apparent in the "high-dimensional" setting where $p > n$. These high dimensional regression problems can be greatly simplified through the use of data splitting. Several proposals for inference in these problems, including those by Wasserman and Roeder (2009) and Barber and Candes (2016) involve using a portion of the data to reduce the dimension of the problem to a tractable size and then using the remainder of the data to perform inference. Wasserman and Roeder (2009) consider multi-stage regression problems. They use a portion of the data to select a model among some selected candidate models by cross validation. The remainder of the data is then used for hypothesis testing on the selected variables. As mentioned in the introduction, Meinshausen et al. (2009) propose a method of combining p-values in these regression problems, and they provide numerical evidence that multi-split methods can provide an improvement over using only a single split. Our first approach in the next section is similar to that of Meinshausen et al. (2009), but we prove a very general result that applies to any i.i.d. testing problem (with fixed or arbitrary number of splits) and it has finite sample validity.

Example 2.3. (Moment Inequality Problem) The moment inequality problem, which is useful in many econometric applications, relates to testing a number of moments are non-positive against the alternative that at least one is positive. A survey of the moment inequality problem is given by Canay and Shaikh (2017). In this problem, data splitting may be useful to reduce the number of moments under consideration. Romano et al. (2014) and Andrews and Soares (2010) develop tests that make use of selecting components for testing and also provide an overview of commonly used test statistics.

Let $X_1, ..., X_n$ be iid random variables with distribution $P \in \mathbf{P}$ on $\mathbb{R}^p$. For convenience, we will assume $\mathbf{P}$ is a family of distribution having finite second moments. Consider the problem of testing the null hypothesis

$$H_0 : P \in \mathbf{P}_0$$

against the null hypothesis

$$H_a : P \in \mathbf{P}_1$$

where $\mathbf{P}_0 = \{P \in \mathbf{P} : E_P(X_1) \leq 0\}$ (with the inequality is component-wise) and $\mathbf{P}_1 = \mathbf{P} \backslash \mathbf{P}_0$. That is, we are interested in testing the null hypothesis that the component-wise means are all non-positive against the alternative that the mean of at least one component is non-negative. Throughout, we will assume that the X_i have covariance matrix Σ, and we will denote the sample covariance matrix by $\hat{\Sigma}$. Furthermore, denote by s_j^2 the sample variance of the $X_{i,j}$'s.

The classical approach to testing in this problem is to choose a test statistic such as a maximum or Chi-squared statistic and estimate the distribution under $E(X) = 0$. Some test statistics include

$$M_n := \max_{1 \le j \le p} \sqrt{n} \frac{\bar{X}_j}{s_j} ,$$

$$Q_n := \inf_{t=(t_1,...,t_p)<0} Z_n'(t) \hat{\Sigma}^{-1} Z_n(t)$$

where

$$Z_n(t) := \left(\sqrt{n} \frac{\bar{X}_1 - t_1}{s_1}, ..., \sqrt{n} \frac{\bar{X}_p - t_p}{s_p} \right)$$

and

$$W_n := \sum_{j=1}^{k} \left(\sqrt{n} \frac{\bar{X}_j}{s_j} \right)^2 I \left\{ \bar{X}_j > 0 \right\}$$

Similar to Example 2.1, the null distribution (based on a "least favorable" configuration) assuming that all means are 0 has a critical value which increases with the dimension p, making it difficult to achieve good power. More sophisticated methods of testing use test statistics which adapt to the negative mean components and estimate the distribution of the test statistic under the true null distribution rather than assuming all components have mean zero. In the interest of constructing powerful tests, it would be useful to identify which components have negative means and exclude those from a test statistic. But, the technical hurdle (without data splitting) is the construction of critical values, which has proven to be challenging. Rather than using full data methods, this task can be greatly simplified by using data splitting methods. Allocating a portion of the data to identify the negative components of the mean and ultimately testing on an independent portion of the data eliminates the burden of accounting for the selection procedure.

Example 2.4. (Goodness of fit) In the quintessential goodness of fit testing problem, assume $U_1, \ldots, U_n$ are i.i.d. with c.d.f. F on $(0, 1)$ and H_0 specifies the underlying distribution is $U(0, 1)$ versus the nonparametric alternatives that F is not $U(0, 1)$. Let $h_1, h_2, \ldots$ be an orthonormal basis in $L_2(0, 1)$. Define the normalized averages $Z_j = n^{-1/2} \sum_i h_j(U_i)$. Then, tests can be based on $Z_1, Z_2, \ldots$ (such as the Neyman smooth tests). In large samples, the problem is approximately the same as Example 2.1 and similar considerations apply. Thus, one might hope to use a small number of the Z_j in which to direct power and data splitting offers an approach.

3 Conservative Finite Sample Methods of Combining Splits

Suppose $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) according to a probability distribution P on some sample space S. The basic goal is to construct a test of some null hypothesis $H_0 : P \in \omega$. Sometimes, it is useful to have two independent samples, so that the first sample may be used to make the testing problem easier in some sense, say by estimating or even eliminating nuisance parameters, dimension-reduction, etc. Then, a test may more easily be constructed using the second sample. For any independently sampled data, one can arbitrarily create two independent samples by splitting the data.

Fix an integer b between 1 and n. Consider the $N = \binom{n}{b}$ subsets of $\{1, ..., n\}$ of size b and let $S_{n,i}$ denote the ith subset of size b, ordered in any fashion. For each $S_{n,i}$, $i = 1, \ldots, N$, construct a level α test of H_0 using the data with indices falling in $S_{n,i}$. For example, suppose a p-value $\hat{p}_{n,i}$ is available based on the test using $S_{n,i}$. Note, however, we specifically do allow for selection of the test statistic to also depend on the data with indices falling in $S^c_{n,i}$ in the sense that the choice of test performed on $S_{n,i}$ can depend on $S^c_{n,i}$, the complement of $S_{n,i}$, i.e., the $n - b$ observations not in $S_{n,i}$. However, we do require that, conditional on the data indexed by $S^c_{n,i}$, the p-value $\hat{p}_{n,i}$ is still valid. That is, for any $P \in \omega$,

$$P\{\hat{p}_{n,i} \leq u | D^c_{n,i}\} \leq u \quad 0 < u < 1 \ . \tag{3.1}$$

where $D^c_{n,i}$ is the data indexed by $S_{n,i}$. Then, unconditionally, we still have that $\hat{p}_{n,i}$ is indeed a genuine p-value in that, for any $P \in \omega$, it satisfies

$$P\{\hat{p}_{n,i} \leq u\} \leq u \quad 0 < u < 1 \ . \tag{3.2}$$

Of course, in the i.i.d. case, the data indexed by $S_{n,i}$ and by its complement are independent so that (3.1) equals (3.2). The difficulty lies in the fact that, even in the i.i.d. setting, all these N p-values are quite dependent, and so the problem becomes how to combine them in order to construct an overall test that controls the Type 1 error rate (and is powerful).

3.1 Quantiles of p-values

Fix a positive integer k and $\beta \in (0, 1)$. Consider the overall test that rejects H_0 if at least k out of the N p-values are less than or equal to β. What can we say about the overall level of this test? Let B denote the number of p-values less than or equal to β. By Markov's inequality,

$$P\{\text{reject}\} = P\{B \geq k\} \leq E(B)/k \leq \beta N/k \ . \tag{3.3}$$

Therefore, if β, N and k are such that $\beta N/k = \alpha$, then the Type 1 error is controlled at level α. For a given k and N, simply choose $\beta = k\alpha/N$.

For example, suppose N is even and $k = N/2$. Then, taking $\beta = \alpha/2$, the test that rejects if at least half of the p-values are less than $\alpha/2$ controls the probability of a Type 1 error. In other words, most of the p-values should be significant at level $\alpha/2$. To put it another way, the median of the N p-values must be less than $\alpha/2$ (where we define the median of an even number N of p-values to be the $N/2$ ordered value from smallest to largest).

For N odd, the median p-value is less than or equal to β if and only if $(N+1)/2$ of the p-values are significant at level β. Thus, applying (3.3) with $k = (N+1)/2$ yields

$$P\{\text{reject}\} = P\{B \geq \frac{N+1}{2}\} \leq \beta 2N/(N+1) \ ,$$

and so taking $\beta = \frac{N+1}{N} \cdot \frac{\alpha}{2}$ controls the Type 1 error.

In general, if $k = N/2$ (whether or not k is an integer), the overall test that rejects if $N/2$ or more of the p-values are significant at level $\alpha/2$ controls the Type 1 error at level α. Equivalently, the test rejects if the median p-value is less than or equal to $\alpha/2$. Thus, it seems at this point there is a small but not unnoticeable price to pay for using multiple splits rather than just one split.

Notice that the above argument holds whether or not all splits of size b are used. This is important because $\binom{n}{b}$ may be large, though finite sample control still holds. In fact, one could in principle compute an arbitrary number, say M, of p-values based on subsamples of even possibly different sizes. The only requirement is p-values must be valid in the sense of (3.2). In order to ensure this, we always assume that subsamples or splits of the data are chosen by picking M sets of indices from $\{1, \ldots, n\}$ either deterministically or at random, but without regard to the values of $X_1, \ldots, X_n$. Furthermore, a set of of b_1 indices could be chosen to compute a p-value while using an independent subsample of b_2 indices that would dictate which test statistic one might use (for example, after dimension reduction), with $b_1 + b_2 \leq n$ and possible $< n$. To put it bluntly, as long as we are not first looking at the data to see which splits lead to significant p-values, then the method applies. Indeed, if test statistics are computed on multiple splits of the data, and only the splits producing significant results are chosen, then the conservative method may no longer control the Type 1 error rate.

Rather than the median p-value, one can look at the rth quantile. So, if k is the smallest integer $\geq rM$, then, we could take $\beta = \alpha k/M \approx \alpha r$ and control the Type 1 error. We state the procedure as follows.

Procedure 3.1. *(Conservative Method of Combining p-values)*

- *Choose M splits of the data independently of the observed data values.*

- *For each split, $S_{n,i}$, compute a p-value satisfying equation (3.2).*
- *Fix a value $r \in (0, 1]$.*
- *Reject H_0 if the proportion of p-values less than or equal to αr is at least r.*

The fact that Procedure 3.1 controls the overall rejection rate of H_0 at level α is recorded next in Theorem 3.1. All proofs are given in the appendix.

Theorem 3.1. *For each of M splits of the data, a test of a null hypothesis H results in p-values that satisfy (3.2). Let B be the number of such tests which reject H at level β, i.e. with p-values $\leq \beta$. For any $0 < r < 1$, consider the overall procedure which rejects H if the proportion of rejections, B/M, is greater than or equal to r. Then,*

$$P\{\text{Type 1 error}\} \leq \beta/r \ .$$

Hence, if each test based on a split $S_{n,i}$ is tested at level $\beta = r\alpha$, then the probability of a Type 1 error is bounded above by α.

This method requires a choice of r and M and the power of the overall test can be quite sensitive to the values chosen. Recommendations based on theoretical and numerical results are given in the Supplement.

3.1.1 Interesting Special Cases

A particularly interesting case of Theorem 3.1 is basing an overall test on the median p-value.

Corollary 3.1. *Take M splits of the data, with indices chosen independently of the data. If p-values satisfy (3.2), then the overall test that rejects the null hypothesis if the median p-value is less than $\alpha/2$ is of level α.*

Rather than basing an overall test on the quantiles of the p-values, a conservative test can be based on the extremes, rejecting for small values of either the minimum or the maximum of the p-values.

Corollary 3.2. *Take M splits of the data, with indices chosen independently of the data. If p-values satisfy (3.2), then the overall test that rejects the null hypothesis if the smallest p-value is less than α/M is of level α. Similarly, the overall test that rejects if the maximum p-value is smaller than α is of level α.*

While it is not immediately obvious how the tests based on the median or minimum of the p-values compare to simply using a single split of the data, it is clear that using the maximum

will result in a loss of power over only using a single split. Typically, the loss of power when using the maximum is relatively small when the p-values are heavily correlated (when a large fraction of the data is used for testing) and fairly large when the p-values are not heavily correlated. Conversely, the test based on the minimum tends to be less conservative when the p-values are independent (when a small fraction of the data is used for testing).

3.2 Conservative Method Based on Average p-values

The following procedure replaces the median p-value of Corollary 3.1 with the sample mean of the p-values, and the same cutoff of $\alpha/2$.

Procedure 3.2.

- *Choose a fixed number, say M, splits of the data independently of the observed data values.*
- *For each split, $S_{n,i}$, compute a p-value satisfying equation (3.2).*
- *Reject H_0 if $\frac{1}{M}\sum_{m=1}^{M}\hat{p}_{n,m}$ is smaller than $\alpha/2$.*

The validity of this procedure relies on the fact that twice the average p-value is again a p-value, which is established by Ruschendorf (1982) and discussed further in Vovk and Wang (2012).

Theorem 3.2. *(Ruschendorf, 1982) Suppose that p-values $\hat{p}_{n,1}, ..., \hat{p}_{n,M}$ satisfying equation (3.2) are computed over $M \leq N$ splits of the data. Then, for any α, b, and M,*

$$P\left(\frac{1}{M}\sum_{m=1}^{M}\hat{p}_{n,m} \leq \frac{\alpha}{2}\right) \leq \alpha \ .$$

Thus, either the median p-value or the average p-value can be compared with $\alpha/2$ in order to control the Type 1 error at level α. Or to put it another way, twice the median p-value or twice the average p-value serves as a valid p-value for the overall test.

3.3 Conservative Methods Based on Concentration Inequalities

We derive alternative conservative methods of combining p-values by exploiting a U-statistic structure when p-values are computed over smaller subsamples of the data. As will be seen, the resulting methods sometimes provide a striking improvement. If p-values are computed based on subsamples of size b (where $b = b_1 + b_2$ and b_1 of the observations are used for testing and b_2 for "selection"), then the average p-value is a U-statistic of degree b. So, rather than using Markov's inequality (3.3), we can start by exploiting an exponential inequality of Hoeffding (1963) who provides the following bound on U-statistics. (Note that we are

specifically not using $b_2 = n - b_1$ since b will typically be much less than n in order for the approach to provide an improvement, as will be seen.)

Lemma 3.1. *(Hoeffding, 1963) Let $X_1, ..., X_n$ be independent random variables. For $n \geq b$, define the random variable*

$$U := \frac{1}{n^{(b)}} \sum g(X_{i_1}, ..., X_{i_b}) \ , \tag{3.4}$$

where the sum is taken over all b-tuples of distinct positive integers not exceeding n and $n^{(b)}$ is the number of such tuples. Then, if g is bounded below by some constant c and above by some constant d,

$$P\left(U - EU \geq t\right) \leq \exp\left[-2kt^2/(d-c)^2\right] \tag{3.5}$$

where $k = \lfloor n/b \rfloor$ is the largest integer smaller than n/b.

Remark 3.1. Two points are notable in order to apply this inequality to our problem. First, we do not need to assume that g is symmetric (as it need not be if part of the subsample is used for testing and part of it is used for selection). Second, it is not crucial that the statistic U in (3.4) is computed by averaging over all $n^{(b)} = n(n-1)\cdots(n-b+1)$-tuples of size b. Consider the balanced U-statistic, say U', which is constructed as follows. First, for some permutation $(i_1, \ldots, i_n)$ of $(1, \ldots, n)$, take an average of g computed on the $k = \lfloor n/b \rfloor$ subsamples of size b of distinct indices

$$\frac{1}{k}\left[g(X_{i_1}, \ldots, X_{i_b}) + g(X_{i_{b+1}}, \ldots, X_{i_{2b}}) + \cdots + g(X_{i_{(k-1)b+1}}, \ldots, X_{i_{kb}})\right] \ . \tag{3.6}$$

The choice of permutation of indices can be at random or deterministic, but independent of the observations. Then, U' can be obtained by averaging (3.6) over additional permutations of indices (chosen independent of the data). So, the sampling scheme can just be taking just one identity permutation, all permutations (to yield the complete U), or averaging over random choices of permutations, as long as there includes an "inner" average over k independent components. Note that, if $k = n/b$, then each observation is used in an equal number of splits.

Lemma 3.2. *Lemma 3.1 holds verbatim with U replaced by U', if the incomplete U-statistic U' is computed as described in Remark 3.1, even if g is not symmetric.*

The proof of Lemma 3.2 follows Hoeffding's proof of the inequality stated in Lemma 3.1.

We now apply these lemmas when the function g is a p-value computed on a subsample of size b. Suppose that b_1 observations are used for testing and b_2 observations are used for selection. Using the bound of Hoeffding (1963), stated in Lemma 3.1 as well as that in Lemma 3.2, yields the following procedure.

Procedure 3.3. • *Choose permutations and average p-values over subsamples as described in Remark 3.1.*

- *Reject H if the average of p-values computed over these splits is smaller than $1/2 - \sqrt{-\log(\alpha)/2k}$, where $k = \lfloor n/b \rfloor$.*

Error control of Procedure 3.3 is provided in the next result.

Theorem 3.3. *Let $\bar{p}_{n,b}$ be the average p-value computed over all $n^{(b)}$-tuples of size b, and let $\bar{p}'_{n,b}$ be the average p-value obtained by sampling some number of permutations as described in Remark 3.1. Then, if H_0 is true,*

$$P\left(\bar{p}_{n,b} - \frac{1}{2} \leq -\sqrt{-\log(\alpha)/2k}\right) \leq \alpha \ , \tag{3.7}$$

where $k = \lfloor n/b \rfloor$. Consequently, Procedure 3.3 is level α. The same result holds with $\bar{p}_{n,b}$ replaced by $\bar{p}'_{n,b}$.

When $\alpha = .05$, it is easily seen that if $k = \lfloor n/b \rfloor$ is strictly larger than 6, the threshold for rejection given by Procedure 3.3 is larger than the threshold given by Procedure 3.2 and therefore Procedure 3.3 is preferable since it makes it easier to reject H_0 while still controlling Type 1 error. Otherwise, Procedure 3.2 is preferable. For instance, if $k = 6$ (i.e a sixth of the data is used for testing and selection), the threshold given by Procedure 3.3 is 0.00036, but if $k = 7$, the threshold is 0.03742. In contrast, the cutoff of $\alpha/2 = 0.025$ is required for the average p-value if using Procedure 3.2. With $k = 10$, the difference becomes more dramatic with the cutoff of Procedure 3.3 being 0.113.

Next, we would like to develop a similar result for the median p-value computed over M splits of the data. Let p_b denote the p-value based on b observations and define

$$\tilde{p}_{n,b} = \text{median}_{i_1,\ldots,i_b} p_b(X_{i_1}, \ldots, X_{i_b}) \ , \tag{3.8}$$

where the median is taken over all $n^{(b)}$ b-tuples of distinct positive integers not exceeding n. Analogously, if not all $n^{(b)}$ b-tuples are used, but instead M b-tuples are taken by sampling permutations as described in Remark 3.1, then the median p-value over the restricted set is denoted $\tilde{p}'_{n,b}$.

Procedure 3.4. • *Choose permutations and average p-values over subsamples as described in Remark 3.1.*

- *Reject H if the median of p-values computed over these splits is smaller than $1/2 - \sqrt{-\log(\alpha)/2k}$.*

Theorem 3.4. *Theorem 3.3 holds verbatim when Procedure 3.3 is replaced by Procedure 3.4, i.e., if the average p-values $\bar{p}_{n,b}$ (or $\bar{p}'_{n,b}$) are replaced by median p-values $\tilde{p}_{n,b}$ (or $\tilde{p}'_{n,b}$).*

3.4 Combining Independent p-values

In the proofs of Theorems 3.1 and 3.4, the approach was to bound $P\{B \geq k\}$, where B is the number of p-values $\leq \beta$ (as in (3.3)). If M splits of the data are based on disjoint subsamples of the data, then the p-values $\hat{p}_1, \ldots, \hat{p}_M$ are independent. Under H_0 when each $\hat{p}_i$ is distributed as $U(0,1)$, B has the binomial distribution with parameters M and β. Hence, for a given k, let $c_{\alpha,k}$ satisfy

$$\sum_{m=k}^{M} \binom{M}{m} c_{\alpha,k}^{m} (1 - c_{\alpha,k})^{M-m} = \alpha \ .$$

Procedure 3.5. *Reject H_0 if $B \geq k$, where B is the number of rejections at level $c_{\alpha,k}$.*

Procedure 3.5 is exact level α under independence, for any k. Alternatively, $\frac{1}{\sqrt{M}} \sum_{i=1}^{M} \Phi^{-1}(\hat{p}_i)$ is $N(0,1)$ under H_0, which leads to the following level α procedure.

Procedure 3.6. *Reject H_0, if $\frac{1}{\sqrt{M}} \sum_{i=1}^{M} \Phi^{-1}(\hat{p}_i) < z_\alpha$.*

Such methods of combining p-values are common in meta-analysis based on independent studies; see Wilkinson (1951) and Hedges and Olkin (1985). For completeness, we include the following modified procedure, which is level α under general dependence.

Procedure 3.7. *Reject H_0, if $\frac{1}{M} \sum_{i=1}^{M} \Phi^{-1}(\hat{p}_i) < z_\alpha$.*

Remark 3.2. The conservative procedure for dependent p-values given by Procedure 3.1 rejects if the kth order statistic of the p-values is smaller than $\alpha k/M$. Especially for large values of k, $c_{\alpha,k}$ can be substantially larger than $\alpha k/M$. For example, when $\alpha = .05$ and $M = 10$, $c_{\alpha,k} \approx .2224$ whereas $\alpha k/M = .025$ when $k = 5$, and $c_{\alpha,k} \approx .4931$ whereas $\alpha k/M = .04$ when $k = 8$.

4 Asymptotically Level α Methods of Combining Splits

Each of the methods in Section 3 provide valid tests combining p-values over (possibly) arbitrarily many splits under very minimal assumptions. However, these methods are conservative in the sense that their null rejection probabilities can be dramatically under the nominal level, even asymptotically, and can lead to a loss of power.

Methods of combining p-values over many splits which are exact, or at least asymptotically level α, may provide an improvement in power. In this section, we provide asymptotically valid procedures under more restrictive assumptions than the methods presented in the previous section.

4.1 Combining p-values Over a Fixed Number of Splits

Suppose we are interested in testing a null hypothesis of the form

$$H_0 : \theta_1 = \cdots = \theta_p = 0 \ .$$

for some parameters $\theta_1, ..., \theta_p$. In many situations, each of the θ_i can be estimated by some asymptotically normal statistic, say $\hat{\theta}_i$, and a test statistic for H_0 is a function of these estimators. Of course, data splitting can be used in this setting to first identify a subset of the parameters under consideration that may be non-zero to inform a choice of test statistic. If only a fixed number of splits of the data are used to find p-values, the asymptotic joint distribution of the test statistics can be estimated using asymptotic or resampling methods. For certain methods of combining p-values over splits, such as taking the average or median p-values, the limiting joint distribution of the test statistics can be used to simulate a critical value for the overall test. We will justify the approach for a fixed number of splits. If the number of splits grows with n but not too quickly, it may be possible to justify the approach, though this requires further investigation.

Example 4.1. *(Continuation of Example 2.1)* For testing whether a mean vector is zero, one can first select which means to include based on a subsample of size $n - b$ and then apply the Chi-squared test based on the remaining data of size b. If several p-values are computed in this way, it may be possible to approximate the joint distribution. Suppose $S = \{1, ..., n/2\}$. We could calculate two p-values: one from using S for selection and S^c for testing, and the other using S^c for selection and S for testing. While there are many possibilities for combining these two p-values, a natural choice might be to reject for small values of the average of the p-values. The joint distribution is a function of $\sqrt{n/2}\left(\bar{X}_1(S), ..., \bar{X}_p(S), \bar{X}_1(S^c), ..., \bar{X}_p(S^c)\right)$, which has an asymptotically normal limiting distribution. In this example, the continuous mapping theorem ensures that simulating the distribution of the sum of the two p-values computed on the limiting normal distribution of the test statistics gives an asymptotically valid rejection region.

More generally, for a split $S_{n,k}$ of the data, assume the resulting p-value, $\hat{p}_{n,k}$, can be written as a function, $p(\cdot)$, of the statistics $\hat{\theta}_{n,j}(S_{n,k})$, and $\hat{\theta}_{n,j}(S^c_{n,k})$, for $j = 1, ..., p$; that is,

$$\hat{p}_{n,k} = p\left(\sqrt{n_k}\hat{\theta}_{n,1}(S_{n,k}), \sqrt{n-n_k}\hat{\theta}_{n,1}(S^c_{n,k}), ..., \sqrt{n_k}\hat{\theta}_{n,p}(S_{n,k}), \sqrt{n-n_k}\hat{\theta}_{n,p}(S^c_{n,k})\right) \ ,$$

where $n_k = |S_{n,k}|$. For compactness of notation, write

$$\Theta_{n,k} = \left(\sqrt{n_k}\hat{\theta}_{n,1}(S_{n,k}), \sqrt{n-n_k}\hat{\theta}_{n,1}(S^c_{n,k}), ..., \sqrt{n_k}\hat{\theta}_{n,p}(S_{n,k}), \sqrt{n-n_k}\hat{\theta}_{n,p}(S^c_{n,k})\right) \ .$$

Suppose M splits of the data, $S_{n,k}$, $k = 1, ..., M$, are chosen such that $(\Theta_{n,1}, ..., \Theta_{n,M})$ has (under H_0) an asymptotically normal distribution with a covariance that can be found exactly

or consistently estimated. Then, the asymptotic distribution of the average p-value

$$\frac{1}{M}\sum_{k=1}^{M}\hat{p}_{n,k}$$

can be approximated using the continuous mapping theorem provided the computed p-values are an almost surely continuous function of the statistics. While here we use the average p-value, the p-values can be combined using some other function $f(\cdot)$ of the p-values, such as the median.

Appropriate cutoffs for combining p-values computed on a fixed number of splits can be obtained by simulating from a multivariate normal distribution in cases where the asymptotic variance Σ is known, or can be consistently estimated by some $\hat{\Sigma}$. The method of combining p-values over a finite number of splits of the data can be summarized as follows.

Procedure 4.1.

- *Choose M splits of the data either randomly or deterministically, but independently of the observed data values.*
- *For each split, $S_{n,i}$, compute a p-value $\hat{p}_{n,i}$ satisfying equation (3.2) that is a function of an asymptotically normal estimator.*
- *Choose some function f, such as the average or median, of the p-values.*
- *Approximate the distribution of this function of the p-values by simulating from the asymptotic null distribution of the test statistics, either $N(0,\Sigma)$ or $N(0,\hat{\Sigma})$ depending on whether the variance is known or can be estimated.*
- *Reject H_0 if $f(\hat{p}_{n,1},...,\hat{p}_{n,M})$ is sufficiently extreme relative to the appropriate simulated quantiles.*

The asymptotic validity of this procedure is given by the following theorem.

Theorem 4.1. *Suppose that M p-values satisfying (3.2) are computed as functions of $\Theta_{n,1},...,\Theta_{n,M}$. If*

$$(\Theta_{n,1},...,\Theta_{n,M}) \xrightarrow{d} (\Theta_1,...,\Theta_M)$$

where $(\Theta_1,...,\Theta_M)$ has a multivariate normal distribution with mean 0 and covariance Σ (under H_0), then for any function $f(\cdot)$ of the p-values,

$$f\left(\hat{p}_{n,1},...,\hat{p}_{n,M}\right) \xrightarrow{d} f\left(p\left(\Theta_1\right),...,p\left(\Theta_M\right)\right)$$

provided the set of discontinuity points of $f\left(p\left(\Theta_1\right),...,p\left(\Theta_M\right)\right)$ has measure zero. Therefore, an asymptotically level α test rejects if

$$f\left(\hat{p}_{n,1},...,\hat{p}_{n,M}\right) > f_\alpha \ .$$

where f_α is the α quantile of $f\left(p\left(\Theta_1\right),...,p\left(\Theta_K\right)\right)$.

Remark 4.1. *A commonly used choice of fixed splits is K-fold cross validation, where the data is split into K folds, say $D_1, ..., D_K$, and each of $M = K$ splits is given by the union of $K-1$ of the folds, i.e.*

$$S_{n,k} = \cup_{i \neq k} D_i \quad .$$

Example 4.2. *Suppose that the $\hat{\theta}_{n,j}$, $j = 1, ..., p$ are asymptotically linear in the sense that under H_0,*

$$\hat{\theta}_{n,j} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_j(X_i) + o_P(1) \ .$$

for some known functions ψ_j satisfying $E(\psi_j(X_i)) = 0$ (under H_0) and $E\left(\psi_j^2(X_i)\right) < \infty$. Suppose K splits of the data are chosen according to the K-fold cross validation method described in Remark 4.1 and the statistics computed on each split $S_{n,k}$ are $\hat{\theta}_{n_K,j}(S_{n,k})$ and $\hat{\theta}_{n-n_K,j}(S_{n,k}^c)$, where $n_K = n(K-1)/K$, which satisfy

$$\hat{\theta}_{n_K,j} = \frac{1}{\sqrt{n_K}} \sum_{i \in S_{n,k}} \psi_j(X_i) + o_P(1)$$

and

$$\hat{\theta}_{n-n_K,j} = \frac{1}{\sqrt{n-n_K}} \sum_{i \in S_{n,k}^c} \psi_j(X_i) + o_P(1) \ .$$

Then, in the above notation,

$$(\Theta_{n,1}, ..., \Theta_{n,K}) \xrightarrow{d} N(0, \Sigma)$$

where the entries of the covariance matrix Σ is given by the appropriate limiting variances and covariances:

$$\lim Var\left(\hat{\theta}_{n_K,j}(S_{n,k})\right) = \lim Var\left(\hat{\theta}_{n-n_K,j}(S_{n,k}^c)\right) = Var\left(\psi_j(X_1)\right) \ ,$$

$$\lim Cov\left(\hat{\theta}_{n_K,j}(S_{n,k}), \hat{\theta}_{n-n_K,l}(S_{n,k}^c)\right) = \lim Cov\left(\hat{\theta}_{n-n_K,j}(S_{n,k_1}^c), \hat{\theta}_{n-n_K,l}(S_{n,k_2}^c)\right) = 0 \ ,$$

$$\lim Cov\left(\hat{\theta}_{n_K,j}(S_{n,k_1}), \hat{\theta}_{n_K,l}(S_{n,k_2})\right) = \frac{K-2}{K-1} Cov\left(\psi_j(X_1), \psi_l(X_1)\right) \ ,$$

and

$$\lim Cov\left(\hat{\theta}_{n_K,j}(S_{n,k_1}), \hat{\theta}_{n-n_K,l}(S_{n,k_2}^c)\right) = \frac{1}{\sqrt{K-1}} Cov\left(\psi_j(X_1), \psi_l(X_1)\right) \ ,$$

for any k, j, l, and $k_1 \neq k_2$.

For each split of the data, means to be included in the test statistic can be chosen according to which $\hat{\theta}_j(S_{n,k}^c)$ exceed some threshold t. Then a p-value can be obtained by the Chi-squared approximation to

$$\sum_{j \ selected} \left(\hat{\theta}_j(S_{n,k})\right)^2 \ .$$

If the p-values are aggregated by taking the average of the p-values over each of the K splits, then Theorem 4.1 ensures that the distribution of this overall test statistic can be approximated by simulating from the appropriate normal distribution, and averaging the resulting p-values. In this example, for any j and l, $cov(\psi_j(X_1), \psi_l(X_1))$ can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\left(\psi_j(X_i) - n^{-1/2}\hat{\theta}_{n,j}\right)\left(\psi_l(X_i) - n^{-1/2}\hat{\theta}_{n,l}\right) \ .$$

Under typical moment assumptions, these estimates, and the above formulas are enough to give a consistent estimator of Σ.

Remark 4.2. *In this setting, rather than simulating the asymptotic normal distribution, the distribution can also be approximated using the bootstrap.*

4.2 Combining Results Across a Growing Number of Splits

4.2.1 Average p-value Over All Splits

Ideally, using all splits of the data should improve power over using a fixed number of splits. It was seen in Section 3.1 that the quantiles of the p-values computed over all splits can be combined to give a conservative test. Results on U-statistics can be leveraged to give tests which have asymptotic level α, i.e. not asymptotically conservative with a value strictly less than α.

For a subsample of size b, consider using the first b_1 observations in the split to select a test statistic, and the remaining b_2 observations to perform a test, so $b = b_1 + b_2$. Let $S_{n,1}, ..., S_{N,n}$ be an enumeration of all such subsamples of size b_1 and b_2, so that $N = \binom{n}{b}\binom{b}{b_1}$. Let $p_b(\cdot, \ldots, \cdot)$ be a p-value computed on b observations. (Note that we are tacitly assuming that p_b is symmetric in its first b_1 arguments as well as symmetric in its last b_2 arguments.) For each split $S_{n,i}$, compute a p-value $\hat{p}_{n,i} = p_b(S_{n,i})$. Assume that the p-value satisfies $E(\hat{p}_{n,i}) = 1/2$ under H_0 (although the procedure is also valid whenever $E(\hat{p}_{n,i}) \leq 1/2$ which is guaranteed if 3.1 holds). Of course, we can consider the case where there is no selection, so $b_1 = 0$, but in general due to selection, p_b is not a symmetric function of its b arguments. However, we can symmetrize p_b in the obvious manner by defining

$$p_b^S(x_1, \ldots, x_b) = \frac{1}{b!}\sum_{i_1,\ldots,i_b} p_b(X_{i_1}, ..., X_{i_b}) \ ,$$

where the sum is taken over all permutations of $\{1, \ldots, b\}$.

Then, the average p-value taken over all N combinations of b_1 and b_2 choices can be expressed as

$$\bar{p}_n = \bar{p}_n(X_1, \ldots, X_n) = \frac{1}{N}\sum_{i_1,\ldots,i_b} p_b^S(X_{i_1}, \ldots, X_{i_b}) \ , \tag{4.1}$$

where the sum is taken over all combinations of b_1 and b_2 subsets of indices. In order to use $\bar{p}_n$ for inference purposes, we need to know its (approximate) distribution under the null hypothesis. But importantly, $\bar{p}_n$ as defined is a U-statistic of degree b. Therefore, we can use Hoeffding's methods in order to obtain the limiting distribution of $\bar{p}_n$, at least in the classical case with fixed degree b. We will also consider the case where $b \to \infty$.

Define

$$p_{1,b}(x_1) = E[p_b^S(X_1, \ldots, X_b)|X_1 = x_1] \ ,$$

and

$$\tilde{\sigma}_b^2 = \mathrm{Var}[p_{1,b}(X_1)]$$

A general CLT for U-statistics of growing degree is given in Theorem 6.1. The following theorem specializes the result in Theorem 6.1 to the context of p-values. The corollary shows how the assumptions of the first part simplify in our context because the Lindeberg Condition always holds (since p-values are uniformly bounded by one). Note that, in the case b_1 and b_2 are fixed (which means we are implicitly assuming the choice of p_b is the same for all n as well), then $\tilde{\sigma} = \tilde{\sigma}_b$ does not depend on b (though it is easy to generalize if this is not the case).

Theorem 4.2. *Assume $b = b_1 + b_2$ satisfies $b^2/n \to 0$. If for all $\delta > 0$,*

$$\lim_{n\to\infty} \frac{1}{\tilde{\sigma}_b^2} \int_{|p_{1,b}(x)|>\delta\sqrt{n\tilde{\sigma}_b^2}} p_{1,b}^2(x) dP(x) = 0 \tag{4.2}$$

then

$$\frac{\sqrt{n}\left(\bar{p}_n - \frac{1}{2}\right)}{\sqrt{b^2\tilde{\sigma}_b^2}} \xrightarrow{d} N(0,1). \tag{4.3}$$

This result also holds if the p-values are computed over M_n uniformly randomly chosen splits of the data provided $n/M_n \to 0$.

Corollary 4.1. *Suppose that b_1 and b_2 are fixed based on a fixed kernel (not depending on n and possibly asymmetric) of degree $b = b_1 + b_2$ with $n \to \infty$, and $\tilde{\sigma} > 0$. Then, (4.3) holds.*

Corollary 4.2. *If $b^2/n \to 0$ and $n\tilde{\sigma}_b \to \infty$, then (4.3) holds.*

First, suppose the kernel p_b and hence b are fixed, so that $\tilde{\sigma}_b = \tilde{\sigma}$ does not depend on n (or b). In order to use Corollary 4.1 for inference purposes, we first need to verify that $\tilde{\sigma}$ is not zero. This condition holds quite generally for the problems we consider; see Section 7.

Remark 4.3. Using this method requires some knowledge of the variance $\tilde{\sigma}_b^2$. In some applications, this variance can be computed exactly (for example in the single mean example in the Supplement); however, it is typically easy to obtain a consistent estimator. This

variance can be consistently estimated by the estimator in Equation 6.10 proposed in Section 6.2,

$$\hat{\sigma}_n^2 = \frac{1}{n}\binom{n-1}{b-1}^{-1}\binom{n-b-1}{b-1}^{-1}\sum_{i,j}\frac{1}{(2b-1)!}\sum\left(\hat{p}_{n,i}-\frac{1}{2}\right)\left(\hat{p}_{n,j}-\frac{1}{2}\right) \tag{4.4}$$

where the sum is taken over all pairs of splits S_i and S_j such that $|S_i \cap S_j| = 1$ and the second sum is taken over all permutations of the observations in these two splits. This estimator is also a U-statistic of degree $2b-1$. So, for example, when b is fixed, $\hat{\sigma}_n$ is consistent in the sense $\hat{\sigma}_n/\tilde{\sigma} \xrightarrow{P} 1$. This estimator is also consistent whenever the sum is taken over M_n randomly chosen splits satisfying $n/M_n \to 0$. ■

Using either the exact variance, or the estimated variance, the following procedure provides an asymptotically level α overall test.

Procedure 4.2.

- *Randomly choose $M_n \leq N$ splits of the data where $n/M_n \to 0$.*
- *For each split, S_i, compute a p-value $\hat{p}_{n,i}$ satisfying equation (3.2).*
- *Reject H_0 if the average of the p-values is smaller than $1/2 - z_{1-\alpha}\sqrt{b^2\tilde{\sigma}_n/n}$ if $\tilde{\sigma}_b$ is known, or smaller than $1/2 - z_{1-\alpha}\sqrt{b^2\hat{\sigma}_n/n}$ where $\hat{\sigma}_n^2$ is defined by (4.4).*

Consequently, an overall asymptotically level α test is obtained by rejecting the null hypothesis whenever the average p-value is smaller than $1/2 - z_{1-\alpha}\sqrt{b^2\tilde{\sigma}_b}$.

4.2.2 Average Test Decisions Over All Splits

Instead of averaging p-values, an overall level α test can be performed based on the average of test decisions computed over splits of the data. For each split $S_{n,i}$ of size $b = b_1 + b_2$ suppose that $\phi_{n,i} = \phi_b(S_{n,i})$ is the test decision from a level β (not necessarily equal to α) test which used the first b_1 observations in the split to select a test statistic, and the remaining b_2 observations to perform the test. Assume that the test decisions satisfy $E(\phi_{n,i}) = \beta$ under H_0 (that is, the tests are of level β), but note that if the tests on subsamples are conservative, the procedure based on the average test decision will be (asymptotically) conservative.

Define

$$\tilde{\phi}_{1,b}(x_1) = E\left(\frac{1}{b!}\sum_{i_1,\ldots,i_b}\phi_b\left(X_{i_1},\ldots,X_{i_b}\right)|X_1 = x_1\right)$$

where the sum is taken over all permutations of $\{1,\ldots,b\}$ and

$$\tilde{\zeta}_b^2 = \text{var}\left(\tilde{\phi}_{1,b}\left(X\right)\right) \ .$$

An overall testing procedure which rejects if the proportion of rejections at level β is sufficiently large can be performed as follows.

Procedure 4.3.

- *Randomly choose $M_n \leq N$ splits of the data where $n/M_n \to 0$.*
- *For each split, S_i, compute a test decision $\phi_{n,i}$ at level β.*
- *Reject H_0 if the average of the test decisions is larger than $\beta + z_{1-\alpha}\sqrt{b^2\tilde{\zeta}_b^2/n}$ if $\tilde{\zeta}_b$ is known, or smaller than $\beta + z_{1-\alpha}\sqrt{b^2\hat{\zeta}_n^2/n}$ if the variance is estimated as outlined in Remark 4.4.*

Once again, the validity of this procedure relies on the asymptotic normality of the average of the test decisions which is given in the following theorem.

Theorem 4.3. *Suppose that b_1 and b_2 are either fixed, or growing in such a way that $b^2/n \to 0$. If for all $\delta > 0$,*

$$\lim_{n\to\infty} \frac{1}{\tilde{\zeta}_b^2} \int_{|\tilde{\phi}_{1,b}(x)|>\delta\sqrt{n\tilde{\zeta}_b^2}} \tilde{\phi}_{1,b}^2(x) dP(x) = 0 \tag{4.5}$$

then

$$\frac{\sqrt{n}\left(\frac{1}{N}\sum_i \phi_{n,i} - \beta\right)}{\sqrt{b^2\tilde{\zeta}_b^2}} \xrightarrow{p} N(0,1).$$

Consequently, an overall level α test can be performed based on the number of rejections of tests at some possibly different level β on each subsample of the data. Note that, since $\tilde{\phi}_{1,b}$ is uniformly bounded, a sufficient condition for (4.5) is $n\tilde{\zeta}_b \to \infty$.

Remark 4.4. *This variance can be consistently estimated by the estimator in Equation 6.10 proposed in section 6.2 which in this situation becomes*

$$\hat{\zeta}_n^2 = \frac{1}{n}\binom{n-1}{b-1}^{-1}\binom{n-b-1}{b-1}^{-1}\sum_{i,j}\frac{1}{(2b-1)!}\sum(\phi_{n,i}-\beta)(\phi_{n,j}-\beta)$$

where the sum is taken over all pairs of splits S_i and S_j such that $|S_i \cap S_j| = 1$ and the second sum is taken over all permutations of the observations in these two splits. This estimator is also consistent whenever the sum is taken over M_n randomly chosen splits satisfying $n/M_n \to 0$.

Note that this suggests the number of rejections at some level β can be much smaller than required by the conservative method. For example, the conservative method requires that the proportion of tests that reject at level $\beta = \alpha/2$ be at least $1/2$. The U-statistic method, on the other hand, can reject if the proportion of rejections is larger than $\alpha/2 + z_{1-\alpha}\sqrt{b^2\tilde{\zeta}_n^2/n} \to \alpha/2$. Hence, the cutoff for rejection can be substantially smaller than $1/2$, leading to much greater power. Some simulations of this method are reported in the Supplement.

4.2.3 The Median p-value Over All Splits

The conservative method derived in Section 3.1 rejects if at least half of the tests reject at level $\alpha/2$. Instead of using a portion of the data for selection and the remainder of the data for testing, suppose that only $b_1 < n$ observations are used for selection, and $b_2 < n - b_1$ observations are used for testing. For each subsample, $S_{n,i}$ of size $b := b_1 + b_2$, the resulting p-value $\hat{p}_{n,i}$ is assumed to be uniform (which is needed for an asymptotically exact test, but sub-uniformity still yields an asymptotically valid test). The conservative method of combining p-values rejects if the median p-value is smaller than $\alpha/2$, however, in this setting, there exists an asymptotically level α test which may provide some improvement in the cutoff for rejecting (under some technical conditions discussed below).

Let $\hat{p}_{n,i}$ denote the p-value computed on the ith split $S_{n,i}$ of size $b = b_1 + b_2$. Define $\tilde{p}_n$ to be the median of the p-values computed over all pairs of splits of size b_1 and b_2. Define

$$\tilde{p}_b(x_1, ..., x_b; t) := \frac{1}{b!} \sum I \left\{ p_b(x_{i_1}, ..., x_{i_b}) > 1/2 + t \right\} \ ,$$

where the sum is extended over all permutations of $1, ..., b$. Also define

$$\tilde{\zeta}_{1,b}(t) = \mathrm{Var}(\tilde{p}_{1,b}(X; t)) \ ,$$

where

$$\tilde{p}_{1,b}(x; t) = E(\tilde{p}_b(x, X_2, ..., X_b; t)) \ .$$

An overall test based on the median p-value can be performed as follows.

Procedure 4.4.

- *Randomly choose $M_n \leq N$ splits of the data where $M_n/N \to 0$.*
- *For each split, S_i, compute a p-value $\hat{p}_{n,i}$ satisfying equation (3.2).*
- *Reject H_0 if the median p-value is smaller than $1/2 - z_{1-\alpha}\sqrt{b^2\tilde{\zeta}_{1,b}(0)/n}$.*

Under some regularity conditions, this procedure asymptotically controls the Type I error rate at level α. The following is a special case of a more general result for M-statistics as given in Theorem 6.2.

Theorem 4.4. *Suppose that b_1 and b_2 are either fixed, or growing in such a way that $b^2/n \to 0$ and that for any fixed t and $\delta > 0$,*

$$\lim_{n\to\infty} \frac{1}{\tilde{\zeta}_{1,b}} \int_{|\tilde{p}_{1,b}(x;t)| > \delta\sqrt{n\tilde{\zeta}_{1,b}}} \tilde{p}^2_{1,b}(x; t) dP(x) = 0 \ .$$

and

$$\lim_{n\to\infty} \tilde{\zeta}_{1,b}\left(\sqrt{\frac{\tilde{\zeta}_{1,b}(0)b^2}{n}}t\right)/\tilde{\zeta}_{1,b}(0) \to 0$$

Then, under H_0*,*

$$\sqrt{\frac{n}{b^2\tilde{\zeta}_{1,b}}}(\tilde{p}_n - 1/2) \xrightarrow{d} N(0,1).$$

Remark 4.5. *This variance can be consistently estimated by the estimator in Equation 6.10 proposed in section 6.2 which in this case becomes*

$$\hat{\zeta}_n^2 = \frac{1}{n}\binom{n-1}{b-1}^{-1}\binom{n-b-1}{b-1}^{-1}\sum_{i,j}\frac{1}{(2b-1)!}\sum(I\{\hat{p}_{n,i} > 1/2\} - 1/2) \cdot (I\{\hat{p}_{n,j} > 1/2\} - 1/2)$$

where the sum is taken over all pairs of splits S_i *and* S_j *such that* $|S_i \cap S_j| = 1$ *and the second sum is taken over all permutations of the observations in these two splits. This estimator is also consistent whenever the sum is taken over* M_n *randomly chosen splits satisfying* $n/M_n \to 0$*.*

5 Conclusion and Further Questions

Although using a single split of a dataset can improve power in some situations, typically data splitting reduces power. We have proposed several conservative methods of combining p-values over multiple splits of a dataset which are broadly applicable and give finite-sample exact inference (when exact p-values are available). While these methods can be sensitive to the fraction of data used for testing as well as the level of the individual tests, using half of the data for testing, and using two splits of the data tends to outperform a single split. These methods can be extremely conservative (even asymptotically), especially when many splits of the data are used, leading to suboptimal performance. Asymptotically exact methods, which are valid under more restrictive assumptions, can provide further improvements over the conservative methods. One such approach exploits a U-statistic structure, and we have provided some general theory for U-statistics with growing kernel size. These methods, when applicable, can give comparable, if not better, performance than full data tests.

The supplement is devoted to applying the methods in this paper to several examples. We treat the toy example of testing a single normal mean as a test case. Here, we can calculate the local limiting power of the various methods because the analysis is more tractable. Perhaps surprisingly because there is no selection or dimension reducing applied, the limiting power attains the optimal limiting power of the usual UMP test. Cases with selection, such as

the multivariate mean are less tractable but simulations are given to compare the various procedures.

We have provided the groundwork for multiple split testing methods with rigorous Type 1 error control in finite samples or asymptotically. Yet, many questions remain for future work, such as:

- Procedures 3.1 are 3.3 are generally applicable, but they are a family of methods, depending on the choice of quantile as well as choices for the number and size of the splits. We have provided some recommendations based on simulations on the supplement, but further work is needed to specify these parameters.

- How to choose among the now various methods? We have been able to calculate limiting power in some situations, but further technology is needed to advance further.

- Combining multiple splits of the data can be helpful for improving power in testing. Can multi-split methods be used to construct shorter confidence intervals than using a single split?

- The U-statistic results suggest that using a larger number of splits is preferable. In the case of testing for a single mean, the complete U-statistic base on the average p-value is asymptotically as powerful as the UMP test. When an insufficient number of splits is used, the Incomplete U-statistic test can have worse power than the UMP test, suggesting that it is better to use more splits. However, the conservative tests tend to become more conservative when using many splits and may have worse power than using fewer splits. It would be useful to refine the rejection regions for the conservative method to perform better when many splits are used.

- The U-statistic methodology is fairly restrictive regarding the amount of data used for selecting and performing a test. Can these methods be refined to apply to using the full data for each p-value, especially in the case where a large portion of the data is used for selection.

- The results presented here have focused mainly on selection parameters to reduce dimensionality of the testing problem. Does using a portion of the data to select a test statistic (e.g. between a test based on a Chi-squared statistic or the max statistic) perform well? That is, can we use a first split of the data to suggest an optimal or adaptive test statistic to be used for the test data?

- In cases where finding the optimal portion of data used for testing is intractable, does using splits of varying size work well?

- Especially in cases such as high dimensional regression problems where variable screening via data splitting is often necessary to achieve valid inference, it would be useful to understand when using multiple splits of the data can improve power over a single split, or perhaps even give improvements in performance relative to full data methods.

- The methods of this paper can be applied to multiple testing problems. Splitting has a clear potential benefit. By using a portion of a data to informally distinguish true null hypotheses from false ones, one many reduce the number of hypotheses tested in the second portion of the data, thereby reducing the effect of multiplicity (or more generally reweight hypotheses based on the first portion of a split). Many multiple testing methods take as input individual p-values for the individual tests. So, as an example, twice the median or average p-value computed over splits is a valid p-value for the individual tests. Less conservative construction of p-values based on the results in Section 4 can also be used. Indeed, the marginal p-values for each test can be used as input into some well=known multiple testing procedures. Usually such methods assume a least favorable configuration of p-values (such as the Holm procedure) or some kind of dependence structure (such as the Benjamini-Hochberg procedure). Can we refine these methods by approximating the joint distribution of the marginal p-values, especially when these p-values use split samples for selection?

- In multiple testing problems, allocating a first split of the data may bring some new advantages. We mention two. In the challenging directional errors problem when testing many two-sided parameters, the first portion of the data can be used to not only reduce the number of parameters tested but the direction in which the parameters are tested (so that in the second phase one is tested a reduced number of one-sided hypotheses). Another avenue of new methodology is offered by the following. When hypotheses are ordered, it is well-known that one can test them sequentially, each at level α, without having to adjust for multiplicity. Therefore, one might use the first split of the data to suggest an ordering by significance in the first split. In both of these problems, using one split of the data is straightforward to control multiple testing error rates, but combining over many splits of the data offers the possibility of improved power.

6 Appendix A

In this section, some general results are developed for U-statistics with growing kernel order, as well as the corresponding M-statistic.

6.1 A General U-statistic CLT Under Growing Kernel Order

Suppose $X_1, \ldots, X_n$ are i.i.d. P. Consider the U-statistic

$$U_n(X_1, ..., X_n) = \binom{n}{k}^{-1} \sum h_k(X_{i_1}, ..., X_{i_{k_n}})$$

where h_k is a symmetric kernel of order $k = k_n$, and the sum is taken over all $\binom{n}{k}$ combinations of k observations taken from the sample. We specifically allow the order $k = k_n$ of the kernel h_{k_n} to depend on n, as does the kernel itself. For cleaner notation, we may just write k and h_k rather than k_n and h_{k_n}, but we will allow k to be fixed as well as $k \to \infty$ as $n \to \infty$. (Note that, if h_k were not symmetric in its arguments, it can always be symmetrized by further averaging. So, for the purposes of the CLT, we will assume h_n is symmetric.) Define $\theta_k = E(h_k(X_1, ..., X_k))$, and

$$\zeta_{1,k} = \mathrm{Var}(h_{1,k}(X)) \ ,$$

where

$$h_{1,k}(x) = E(h_k(x, X_2, ..., X_k)) \ .$$

Sufficient conditions for asymptotic normality of such U-statistics are given in Mentch and Hooker (2016), but their conditions are not general enough to apply to our setting. In particular, they assume $\zeta_{1,k} \nrightarrow 0$, which as we will see fails for our applications. (Indeed, if h is a p-value based on a subsample of size k_n and the null hypothesis holds, then it must be the case that $\zeta_{1,k}$ is of order $1/k_n \to 0$, and typically this is the exact order; thus, $\zeta_{1,k} \to 0$.)

First, define for $1 \leq c \leq k$,

$$h_{c,k}(X_1, \ldots, X_c) = E[h_k(X_1, \ldots, X_k)|X_1, \ldots, X_c]$$

and

$$\zeta_{c,k} = Var(h_{c,k}(X_1, \ldots, X_c)) \ , \tag{6.1}$$

so that $\zeta_{k,k}$ is the variance of the kernel based on a sample of size k equal to the order of the kernel. In our applications, the kernel will typically be a p-value and hence uniformly bounded, so that the $\zeta_{c,k}$ are also uniformly bounded as c, k, and n vary.

Remark 6.1 (Simple Consistency)**.** Under weak conditions, U_n is consistent in the sense $U_n - \theta_n \xrightarrow{P} 0$. It suffices to show $Var(U_n) \to 0$. But, as is well-known, $Var(U_n) \leq k\zeta_{k,k}/n$. So if the $\zeta_{k,k}$ are uniformly bounded (which follows if the kernels are uniformly bounded), and $k/n \to 0$, then consistency follows.

The theorem below applies in a triangular array setup, where n i.i.d. observations are i.i.d. P_n. Then, quantities like $\zeta_{c,k}$ in (6.1) are computed under P_n.

Theorem 6.1. *Assume the order $k = k_n$ of the kernel h_k satisfies $k^2/n \to 0$. Further assume that $\zeta_{k,k}/k\zeta_{1,k}$ is bounded.*

(i) Then,

$$\frac{nVar(U_n)}{k^2\zeta_{1,k}} \to 1 \quad . \tag{6.2}$$

(ii) If, in addition, for all $\delta > 0$,

$$\lim_{n\to\infty} \frac{1}{\zeta_{1,k}} \int_{|h_{1,k}(x)|>\delta\sqrt{n\zeta_{1,k}}} h_{1,k}^2(x) dP_n(x) = 0 \tag{6.3}$$

then

$$\frac{\sqrt{n}\,(U_n(X_1, ..., X_n) - \theta_n)}{\sqrt{k^2\zeta_{1,k}}} \xrightarrow{d} N(0, 1). \tag{6.4}$$

This result also holds for the "incomplete" U-statistic which is the average of the kernels computed over M_n randomly and uniformly chosen subsamples of the data provided $n/M_n \to 0$.

Proof of Theorem 6.1. Following, for example, the argument in van der Vaart (1998), it suffices to show $Var(U_n)/Var(\hat{U}_n) \to 1$, where

$$\hat{U}_n = \frac{k_n}{n} \sum_{i=1}^{n} \phi_{1,k}(X_i) \; . \tag{6.5}$$

Indeed, Theorem 11.2 of van der Vaart (1998) applies not only for fixed k but when $k = k_n \to \infty$. As is well-known (and argued in the proof of Theorem 12.3 of van der Vaart (1998)),

$$Var(U_n) = \sum_{c=1}^{k} \binom{n}{k}^{-1} \binom{k}{c} \binom{n-k}{k-c} \zeta_{c,k} \; , \tag{6.6}$$

where

$$\zeta_{c,k} = Cov\left[h_k(X_1, \ldots, X_c, X_{c+1}, \ldots, X_k), h_k(X_1, \ldots, X_c, X_{k+1}, \ldots, X_{2k-c})\right] \tag{6.7}$$

is the covariance between the kernel based on two data sets with exactly c variables in common; by conditioning on $X_1, \ldots, X_c$, it is readily seen that (6.1) and (6.7) agree. First note that the $c = 1$ term in (6.6) divided by $Var(\hat{U}_n) = k^2\zeta_{1,k}/n$ tends to one, i.e.

$$\frac{\frac{k}{\binom{n}{k}}\binom{n-k}{k-1}\zeta_{1,k}}{\frac{k^2}{n}\zeta_{1,k}} = \frac{(n-k)!(n-k)!}{(n-1)!(n-2k+1)!} \to 1 \; .$$

The last limit uses $k^2/n \to 0$ and can be seen by applying Stirling's formula, taking logs and using a Taylor's expansion. What remains is to show that the sum from $c = 2$ to $c = k$ in (6.6) divided by $k^2\zeta_{1,k}/n$ tends to 0. But,

$$\frac{\sum_{c=2}^{k} \binom{n}{k}^{-1} \binom{k}{c} \binom{n-k}{k-c} \zeta_{c,k}}{\frac{k^2}{n}\zeta_{1,k}} \leq \frac{\sum_{c=2}^{k} \frac{1}{c!} \left[\frac{k!}{(k-c)!}\right]^2 \frac{(n-k)!}{n!} \frac{(n-k)!}{(n-2k+c)!} \zeta_{c,k}}{\frac{k^2}{n}\zeta_{1,n}}$$

$$\leq \frac{\sum_{c=2}^{k} \frac{k^{2c}}{c!} \frac{1}{(n-k+1)^c} \zeta_{c,k}}{\frac{k^2}{n}\zeta_{1,k}} \leq \sum_{c=2}^{k} \frac{1}{c!} \epsilon_n^{c-1} \zeta_{c,k}/\zeta_{1,k} \ , \tag{6.8}$$

where

$$\epsilon_n = \frac{k^2}{n-k+1} \ .$$

Using the inequality $\zeta_{c,k} \leq c\zeta_{k,k}/k$ (see Hoeffding (1948)) gives that (6.8) is bounded above by

$$\frac{\zeta_{k,k}}{k\zeta_{1,k}} \sum_{c=2}^{k} \frac{1}{(c-1)!} \epsilon_n^{c-1} \leq \frac{\zeta_{k,k}}{k\zeta_{1,k}} \sum_{j=1}^{k-1} \epsilon_n^j = \frac{\zeta_{k,k}}{k\zeta_{1,k}} \cdot \frac{\epsilon_n - \epsilon_n^k}{1-\epsilon_n} \ . \tag{6.9}$$

The second factor in the last expression for (6.9) tends to zero since $\epsilon_n \to 0$. Thus, as long as $\zeta_{k,k}/k\zeta_{1,k}$ stays bounded, the result follows. ■

Corollary 6.1. *Under the above notation, if $k^2/n \to 0$, the kernel h_k is uniformly bounded (both as k and the data vary), and $k\zeta_{1,k} \nrightarrow 0$, then asymptotic normality (6.4) holds.*

Proof of Corollary 6.1. Since the h_k are uniformly bounded, so are the $\zeta_{k,k}$. Hence, the condition in Theorem 6.1 $\zeta_{k,k}/k\zeta_{1,k}$ is bounded, since $k\zeta_{1,k} \nrightarrow 0$. Moreover, the Lindeberg condition (6.3) necessarily holds because $n\zeta_{1,k} = (n/k) \cdot k\zeta_{1,k} \to \infty$, so that the region of integration in the integral is empty for large n. ■

Remark 6.2. *In our applications, the condition that $k\zeta_{1,k} \nrightarrow 0$ holds because $k\zeta_{1,k}$ is of strict order one.*

6.2 Estimating the Variance

In order to use Theorem 6.1, we may require a consistent estimator of $\zeta_{1,k}$. Note that $\zeta_{1,k}$ can be equivalently expressed as

$$\zeta_{1,k} = \mathrm{Cov}(h_k(X_1, ..., X_k), h_k(X_1, X_{k+1}, ..., X_{2k-1})) =$$

$$E[h_k^*(X_1, \ldots, X_{2k_n-1})] - \theta_k^2 \ ,$$

and h_k^* is a symmetric kernel of degree $2k-1$ given by

$$h_k^*(X_1, ..., X_{2k-1}) = \frac{1}{(2k-1)!} \sum h_k(X_{i_1}, ..., X_{i_k}) h_k(X_{i_1}, X_{i_{k+1}}, ..., X_{i_{2k-1}}) \tag{6.10}$$

where the sum is extended over all permutations of $1, ..., 2k-1$. Then, the U-statistic of degree $2k-1$ with symmetric kernel $h_n^* - \theta_k^2$,

$$\hat{\zeta}_{1,n} = \binom{n}{2k-1}^{-1} \sum h_k^*(X_{i_1}, ..., X_{i_{2k-1}}) - \theta_k^2$$

where the sum is taken over all subsamples of $\{1, ..., n\}$ of size $2k-1$, is an unbiased estimator of $\zeta_{1,k}$. (Note that, in order to compute this estimator, one must know θ_k^2, which in our applications is $1/4$ under the null hypothesis. Otherwise, one can estimate θ_k^2 by a U-statistic of degree $2k$ in a similar fashion.) Define

$$\zeta_{m,k}^* = \text{Cov}(h_k^*(X_1, ..., X_m, ..., X_{2k-1}), h_k^*(X_1, ..., X_m, X_{2k}, ..., X_{4k-m-1})) \ .$$

Using Theorem 6.1 under the conditions based on the kernel h_k^* rather than h_k, one can conclude

$$\frac{Var(\hat{\zeta}_{1_n})}{\frac{(2k-1)^2}{n}\zeta_{1,k}^*} \to 1 \ . \tag{6.11}$$

By Remark 6.1,

$$\hat{\zeta}_{1,n} - \zeta_{1,k} \xrightarrow{P} 0 \tag{6.12}$$

if $k/n \to 0$ and the kernels are uniformly bounded. In order to show the stronger result

$$\frac{\hat{\zeta}_{1,n}}{\zeta_{1,k}} \xrightarrow{P} 1 \ , \tag{6.13}$$

it suffices to show

$$Var(\hat{\zeta}_{1,n}/\zeta_{1,k}) \to 0,$$

or using (6.11), it suffices to show

$$\frac{\frac{(2k-1)^2}{n}\zeta_{1,k}^*}{\zeta_{1,k}^2} \to 0 \ .$$

But if the h_k are bounded, so are the h_k^* and $\zeta_{m,k}^*$, and so it is sufficient to show $k/(n\zeta_{1,k}^2) \to 0$ to finally conclude (6.13). (Note that if $\zeta_{1,k}$ is of exact order $1/k$, then this condition becomes $k^3/n \to 0$.) It then follows that (6.4) holds if $\zeta_{1,k}$ is replaced by $\hat{\zeta}_{1,n}$.

Remark 6.3. *Wang and Lindsay (2014) propose a similar but alternative estimator of the variance of a general U-statistic. They show consistency in the sense (6.12). The above argument yields the stronger consistency result (6.13) even with $k \to \infty$.*

6.3 Asymptotic Normality of the M-statistic

Suppose instead of using U_n as an estimator, where the kernel is averaged over all subsamples of size k of the data, we are interested in using the median of the values of the kernel computed on all subsamples of size k, i.e.

$$\tilde{U}_n := \text{median}\left\{h_k(X_{i_1}, ..., X_{i_k})\right\} \ ,$$

which we refer to as an M-statistic. In this section, we do not assume h_k is symmetric, and so the median is taken over all $n!/(n-k)!$ ordered indices $i_1, \ldots, i_k$ taken without replacement from $1, \ldots, n$. We would like to prove a triangular array CLT for $\tilde{U}_n$ when $k = k_n$ varies with n.

Suppose that h_k has a c.d.f. F_k and that $\tilde{\theta}_k$ satisfies $F_k(\tilde{\theta}_k) = 1/2$.

Define

$$\tilde{h}_k(x_1, ..., x_k; t) := \frac{1}{k!} \sum I\left\{h_k(x_{i_1}, ..., x_{i_k}) > \tilde{\theta}_k + t\right\} \ , \tag{6.14}$$

where the average is taken over all permutations of $1, ..., k$. Also define

$$\tilde{\zeta}_{1,k}(t) = Var[\tilde{\phi}_{1,k}(X; t)]$$

with

$$\tilde{\phi}_{1,k}(x; t) = E[\tilde{h}_k(x, X_2, ..., X_k; t)] \ .$$

We will assume that the sequence $\{F_k\}$ is asymptotically (as $k = k_n \to \infty$) equidifferentiable relative to the sequence $\tilde{\theta}_k$ if, for any $\epsilon_k \to 0$,

$$F_k(\tilde{\theta}_k + \epsilon_k) - F_k(\tilde{\theta}_k) = \epsilon_k F'(\tilde{\theta}_k) + o(\epsilon_k) \ . \tag{6.15}$$

We will apply (6.15) with the particular choice $\epsilon_k = \delta_k$ defined by

$$\delta_k = \sqrt{\frac{\tilde{\zeta}_{1,k}(0)k^2}{n}} \ .$$

Note that $\tilde{\zeta}_{1,k}$ is bounded in k, so that if we assume that $k^2/n \to 0$, then $\delta_k \to 0$. Then,

$$E\left(h_n\left(X_1, ..., X_k; \delta_k\right)\right) = 1/2 - F_k'(\tilde{\theta}_k)\delta_k + o(\delta_k) \ . \tag{6.16}$$

Finally, assume that $F_k'(\tilde{\theta}_k) \to f(\tilde{\theta})$, which is just some positive constant. (Note, f and $\tilde{\theta}$ separately need not have meaning, but typically F_k' tends to some f and $\tilde{\theta}_k \to \tilde{\theta}$.)

Theorem 6.2. *In the above notation, assume that, $k^2/n \to 0$, $k\zeta_{1,k}(0) \nrightarrow 0$ and for any fixed t*

$$\tilde{\zeta}_{1,k}(\delta_n t)/\tilde{\zeta}_{1,k}(0) \to 1 \tag{6.17}$$

as $n \to \infty$. Then,

$$\sqrt{\frac{n}{\tilde{\zeta}_{1,k}(0)k^2}}\left(\tilde{U}_n - \tilde{\theta}_k\right) \xrightarrow{d} N(0, 1/f^2(\tilde{\theta})) \ .$$

Proof of Theorem 6.2: For any fixed t,

$$P\left\{\sqrt{\frac{n}{\tilde{\zeta}_{1,k}(0)k^2}}\left(\tilde{U}_n - \tilde{\theta}_k\right) \le t\right\} = P\left\{\tilde{U}_n \le \tilde{\theta}_k + \delta_k t\right\}$$

$$= P\left\{ \binom{n}{k}^{-1} \sum \tilde{h}_k \left(X_{i_1}, ..., X_{i_k}; \delta_k t\right) \leq 1/2 \right\}$$

$$= P\left\{ \sqrt{\frac{n}{\tilde{\zeta}_{1,k}(0) k_n^2}} \binom{n}{k}^{-1} \sum \left(\tilde{h}_k(X_{i_1}, ..., X_{i_k}; \delta_k t) - [1/2 F_k'(\tilde{\theta}_k)\delta_k + o(\delta_k)\right) \leq t F_k'(\tilde{\theta}_k) \right\}$$

Hence, this last expression has the same limit (if any) as

$$P\left\{ \sqrt{\frac{n}{\tilde{\zeta}_{1,k}(0) k_n^2}} [U_n(t) - E(U_n(t)] \leq t F_k'(\tilde{\theta}_k) \right\} \quad , \tag{6.18}$$

where $U_n = U_n(t)$ is a U-statistic with symmetric kernel $\tilde{h}_k(\cdot; t)$ defined by

$$U_n(t) = \binom{n}{k}^{-1} \sum \tilde{h}_k(X_{i_1}, \ldots, X_{i_k}; \delta_k t) \ .$$

But, by Corollary 6.1,

$$\sqrt{\frac{n}{k^2 \tilde{\zeta}_{1,k}(\delta_k t)}} [U_n(t) - E(U_n(t)] \xrightarrow{d} N(0,1) \ .$$

Using the assumption $\tilde{\zeta}_{1,k}(\delta_k t)/\tilde{\zeta}_k(0) \to 1$ and Slutsky's theorem gives that the limiting value of (6.18) is $\to \Phi\left(f(\tilde{\theta})t\right)$. ■

7 Appendix B: Verification of Conditions in Corollary 4.1

In Corollary 4.1, we need to verify the condition $\tilde{\sigma}_b = \tilde{\sigma} > 0$. Here, we explain why this condition holds quite generally in the problems we consider. Note that, in order for the condition $\tilde{\sigma} > 0$ to hold, we equivalently need $p_{1,b}(X_1)$ to not be a constant with probability one.

First consider the case where there is no selection. The easiest situation occurs when the p-value is based on a function $p_b(x_1, \ldots, x_b)$ (assumed symmetric) such that, for any fixed $x_j, j \geq 2$, $p_b(x_1, \ldots, x_b)$ is a strictly monotone decreasing function of x_1. (Of course, it can be strictly monotone decreasing as well; the point is that it can't be increasing for some choice of the x_j, $j \geq 2$, but decreasing for another choice.)

In this situation, the variance condition $\tilde{\sigma} > 0$ holds. To see why, for any $x < x'$,

$$p_b(x, X_2, \ldots, X_b) > p_b(x', X_2, \ldots, X_b)$$

and so

$$p_{1,b}(x) = E[p_b(x, X_2, \ldots, X_b) > p_b(x', X_2, \ldots, X_b) = p_{1,b}(x') \ .$$

Therefore, $p_{1,b}(X_1)$ cannot be constant with probability one unless X_1 is constant with probability one. As a function of x_1, it can be weakly monotone, as long as it is not constant with probability one.

For example, if $p_b(x_1, \ldots, x_b) = 1 - \Phi(T_b)$, where $\Phi(\cdot)$ is the standard normal c.d.f. and $T_b = b^{-1/2} \sum_{i=1}^n x_i$, then the monotonicity condition holds. More generally, assume $p_b = 1 - G(T_b)$, where $T_b(x_1, \ldots, x_b)$ is monotone (and not constant) and $G(\cdot)$ is some c.d.f. (corresponding to the null distribution of T_b. Assume G is strictly increasing on its support with density g. Then,

$$\frac{d}{dx} p_{1,b}(x) = -E\left[g(T_b(x, X_2, \ldots, X_b)] \frac{d}{dx} T_b(x, X_2, \ldots, X_b) \right] < 0$$

since $g > 0$ with probability one and $\frac{d}{dx} T_b(x, X_2, \ldots, X_b) > 0$ with positive probability.

Sometimes, the monotonicity condition fails. For example, let $T_b = b^{-1/2} |\sum_i x_i|$. In order to verify $\tilde{\sigma} > 0$, note that, for any fixed $x_2, \ldots, x_b$, $T_b \to \infty$ as $|x_1| \to \infty$ and so $p_b(x_1, \ldots, x_b) \to 0$ as $|x_1| \to \infty$. By dominated (or bounded) convergence, it follows that

$$p_{1,b}(x_1) = E[p_b(x_1, X_2, \ldots, X_b)] \to 0$$

as $|x_1| \to \infty$. But, for any finite x_1, $p_{1,b}(x_1) > 0$, so that $p_{1,n}(X_1)$ cannot be constant with probability one.

For a slightly more complicated example, let T_b be the classical t-statistic and G be the t-distribution with $b - 1$ degrees of freedom (or any other c.d.f. that is not constant in $[-1, 1]$). Then, T_b doesn't satisfy the monotonicity assumption. But, note that, for any fixed $x_2, \ldots, x_b$, $T_b(x_1, x_2, \ldots, x_b) \to 1$ as $x_1 \to \infty$ and $T_b(x_1, \ldots, x_b) \to -1$ as $x_1 \to -\infty$. Again by dominated convergence,

$$p_{1,b}(x_1) = E[p_b(x_1, X_2, \ldots, X_b)] \to 1 - G(\pm 1)$$

as $x_1 \to \pm\infty$, from which $\tilde{\sigma} > 0$ follows easily.

Next, we consider a multivariate situation, but still with no selection. Let $x_i = (x_{i,1}, \ldots x_{i,d})$ be a vector in d-dimensions. Consider the Chi-squared statistic

$$T_b(x_1, \ldots, x_b) = b \sum_{j=1}^d \bar{x}_j^2 \ ,$$

where $\bar{x}_j = b^{-1} \sum_{i=1}^b x_{i,j}$. As before, let $p_b = 1 - G(T_b)$, where G is the Chi-squared distribution with d degrees of freedom. if, for any j with $|x_{1,j}| \to \infty$, then $T_b \to \infty$ and so $p_b(x_1, x_2, \ldots, x_b) \to 0$. By the same dominated convergence argument as above, it follows that $p_{1,b}(x_1)$ is not constant with probability one, and so $\tilde{\sigma} > 0$ follows.

Finally, we consider the situation where p_b is not symmetric because we allow selection. Here $b = b_1 + b_2$ and $p_b(x_1, \ldots, x_b)$ is symmetric in the first b_1 variables (used for selection) and also symmetric in the last b_2 variables (used for testing). Now, the nonzero variance condition corresponds to the projection of the symmetrized version of p_b; that is, $\bar{p}_b$. Consider a generic term in the average of $\bar{p}_b$ given by $p_b(X_{i_1}, \ldots, X_{i_b})$ for some permutation $(i_1, \ldots, i_b)$ of $(1, \ldots, b)$. If the index i_j corresponding to 1 satisfies $i_j \leq b_1$, so that x_1 is used for selection, then fixing $X_{i_j} = X_1$ at x_1 gives (by symmetry)

$$Ep_b(x_1, X_2, \ldots, X_{b_1}, X_{b_1+1}, \ldots X_{b_1+b_2}) = E\left[Ep_b(x_1, X_2, \ldots, X_b)|X_2, \ldots X_{b_1}\right] . \tag{7.1}$$

But, the inner expectation is, under the null hypothesis, equal to 1/2, because regardless of the selection mechanism and choice of test statistic, we assume that we have constructed a valid p-value conditional on selection. On the other hand, if the index corresponding to 1, say i_j, exceeds b_1, so that X_1 is used only for testing, then

$$Ep_b(X_1, \ldots, X_{b_1}, \ldots, X_{b-1}, x) = E\left[Ep(X_1, \ldots, X_{b_1}, \ldots, X_{b-1}, x)|X_1, \ldots, X_{b_1})\right]$$

Now, we can apply any of the above methods already dealing with testing alone. As an example, consider again the Chi-squared statistic, but where now selection can determine which components are tested. (For the sake of argument, if it is possible that none are selected, we will nonetheless test based on either the most significant component, or simply test all of them.) Then, since x is a vector of length d, if all of its components tend to ∞, then $p_b(X_1, \ldots, X_{b-1}, x) \to 0$, and so does its expectation. Thus, the projection based on the U-statistic $\bar{p}_b$ satisfies

$$p_{1,b}(x) = \frac{b_1}{b} \cdot \frac{1}{2} + \frac{b_2}{b} Ep_b(X_1, \ldots, X_{b-1}, x) \to \frac{b_1}{2b}$$

as the components of x diverge, but clearly this is not the case for fixed components of x, because the factor with the expectation will not be exactly zero. Hence, $\tilde{\sigma} > 0$. Other examples can be treated similarly. ■

8 Appendix C: Proofs of results in text

Proof of Theorem 3.1. Note that the proportion of rejections at level $r\alpha$ is

$$\frac{1}{N} \sum_{i=1}^{N} I\left\{\hat{p}_{n,i} \leq r\alpha\right\} .$$

Using Markov's inequality, the probability that the proportion of rejections is $\geq r$ is bounded above by

$$\begin{aligned} P\left(\frac{1}{N}\sum_{i=1}^{N} I\left\{\hat{p}_{n,i} \leq r\alpha\right\} \geq r\right) &\leq \frac{1}{Nr} E\left(\sum_{i=1}^{N} I\left\{\hat{p}_{n,i} \leq r\alpha\right\}\right) \\ &= \frac{1}{Nr} Nr\alpha \\ &= \alpha \ . \end{aligned}$$

Consequently, the level of the overall test is bounded by α. ■

Proof of Theorem 3.4: For fixed β let B be the number of p-values (out of the M) that are $\leq \beta$. Then,

$$P\{\tilde{p}_{n,b} \leq \beta\} = P\{B \geq \frac{M+1}{2}\} = P\{\frac{B - E(B)}{M} \geq \frac{\frac{M+1}{2} - E(B)}{M}\}$$

$$\leq \exp\left[-2k\left(\frac{\frac{M+1}{2} - E(B)}{M}\right)^2\right] = \exp\left[-2k\left(\frac{\frac{M+1}{2} - MP\{\hat{p}_b \leq \beta\}}{M}\right)^2\right]$$

$$= \exp\left[-2k\left(\frac{1}{2} + \frac{1}{2M} - P\{\hat{p}_b \leq \beta\}\right)^2\right] \ .$$

Since $P\{\hat{p}_b \leq \beta\} \leq \beta$, then for $\beta < 1/2$, we can further bound the above by

$$\exp\left[-2k\left(\frac{1}{2} + \frac{1}{2M} - \beta\right)^2\right] \ .$$

Therefore, we can choose β so that the last expression is α. The solution β^* is

$$\beta^* = \frac{1}{2} + \frac{1}{2M} - \sqrt{\frac{\log(\alpha)}{-2k}} \ ,$$

which is actually slight better than the bound in the theorem, since

$$\alpha \geq P\{\tilde{p}_{n,b} \leq \beta^*\} \geq P\{\tilde{p}_{n,b} \leq \frac{1}{2} - \sqrt{\frac{\log(\alpha)}{-2k}}\} \ .$$

(Note that it is easy to see that the solution β^* is always less than $1/2$ as long as $\alpha \leq \exp(-1/2) \approx .6$, by using the inequality $M \geq k$.) The same argument works for the "incomplete" median $\tilde{p}'_{n,b}$, since the inequality (3.5) holds for both the complete and incomplete U-statistic (as long as subsamples are chosen by splitting permutations as described in Remark 3.1). ■

PROOF OF THEOREM 4.1. This result follows immediately from the asymptotic normality of the test statistics as well as the continuous mapping theorem. ■

PROOF OF THEOREM 4.2. This result follows from the U-statistic CLT Theorem 6.1 presented in Section 6.1. ■

PROOF OF THEOREM 4.3. This result follows from the U-statistic CLT Theorem 6.1 presented in Section 6.1. ■

PROOF OF THEOREM 4.4. This follows immediately from the M-statistic CLT Theorem 6.2 noting that the derivative of the distribution function of a uniform p-value is one. ■

References

Andrews, D. W. K. and Soares, G. (2010). Inference for parameters defined by momoment inequalities using generalized moment selection. *Econometrica*, 78(1):119–157.

Arias-Castro, E., Cands, E. J., and Plan, Y. (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556.

Barber, R. and Candes, E. (2016). A knockoff filter for high-dimensional selective inference. *arXiv:1602.03574*.

Canay, I. A. and Shaikh, A. M. (2017). *Practical and Theoretical Advances in Inference for Partially Identified Models*, volume 2 of *Econometric Society Monographs*, pages 271–306. Cambridge University Press.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fithian, W., Sun, D., and Taylor, J. (2015). Optimal inference after model selection. *arXiv:1410.2597v2*.

Hedges, L. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, 19(3):293–325.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Horst, P. (1941). Prediction of personal adjustment. *New York: Social Science Research Council*, page (Bulletin 48).

Ignatiadis, N., Klaus, B., Zaugg, J., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13:577–580.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, 2(8).

Larson, S. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, 22:45–55.

Meinshausen, N., Meier, L., and Buhlmann, P. (2009). p-values for high-dimensional regression. *Annals of Statistics*, 104:1671–1681.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(26):1–41.

Moran, P. A. P. (1973). Dividing a sample into two parts a statistical dilemma. *Sankhy: The Indian Journal of Statistics, Series A (1961-2002)*, 35(3):329–333.

Potscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, 7(2):163–185.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2014). A practical two-step method for testing moment inequalities. *Econometrica*, 82(5):1979–2002.

Rubin, D., Dudoit, S., and van der Laan, M. (2006). A method to increase the power of multiple testing procedures through sample splitting. *Statistical Applications in Genetics and Molecular Biology*, 5(1):Article 19.

Ruschendorf, L. (1982). Random variables with maximum sums. *Advances in Applied Probability*, 14(3):623–632.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Ser.B*, 36:111–147.

van de Wiel, M., Berkhof, J., and van Wieringen, W. (2009). Testing the prediction error difference between 2 predictors. *Biostatistics*, 10:550–560.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Vovk, V. and Wang, R. (2012). Combining p-values via averaging. *ArXiv e-prints*.

Wang, Q. and Lindsay, B. (2014). Variance estimation of a general u-statistic with application to cross-validation. *Statistica Sinica*, 24(3):1117–1141.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Annals of Statistics*, 37:2178–2201.

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin*, 48(2):156–158.