Supplement to
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

In this supplement, we provide some further theory and simulations for the methods introduced in the main paper.

S.1 Comparison of Methods in Single Mean Example

In this section, the power of the methods presented in the main text is studied in the context of testing for a single mean. Obviously, this is a toy example as these methods are not needed here. But this simple model admits simple expressions of asymptotic power, which facilitates comparisons of the methods. Some numerical evidence will be provided as well.

Let $X_1, \ldots, X_n$ be i.i.d. real-valued with unknown mean μ. The problem is to test the null hypothesis H_0 that the mean is 0 versus greater than 0. For the purposes here of studying the power of tests combining splits of the data, further assume the underlying distribution is $N(\mu, 1)$. The limiting power of the UMP test against contiguous alternatives $h/\sqrt{n}$ is

$$1 - \Phi(z_{1-\alpha} - h)$$

when using the full data, and

$$1 - \Phi(z_{1-\alpha} - \sqrt{\tau}h)$$

when using a single split of size b satisfying $b/n = \tau$.

S.1.1 Power of Conservative Quantile Method

Assume $b/n \to \tau \in (0,1)$, the fraction in the sample used for testing. Let $\bar{X}_{n,b,i}$ be the sample mean of the ith subset of size b. The combined split sampling test rejects H_0 if at least $100r\%$ of the values $\sqrt{b}\bar{X}_{n,b,i}$ exceed $z_{1-r\alpha}$, i.e. if the $(1-r)$th quantile of $\sqrt{b}\bar{X}_{n,b,i}$ exceeds $z_{1-r\alpha}$.

Theorem S.1.1. *Suppose that $X_1, ..., X_n$ are i.i.d. according to a normal distribution with mean μ and variance one. The limiting power of the one sided test using Procedure 3.1 with r fixed $(0 < r \leq 1)$ of $H_0 : \mu = 0$ against contiguous alternatives $h/\sqrt{n}$ is given by*

$$1 - \Phi\left[\frac{1}{\sqrt{\tau}}(z_{1-r\alpha} - z_{1-r}\sqrt{1-\tau}) - h\right] . \tag{S.1}$$

Note this shows that, even asymptotically, the approach is conservative, i.e. when $h = 0$, the limiting rejection probability is below α. It further implies that the limiting power for small positive h can be less than α and loss of power results. By comparison, the limiting power against $h/\sqrt{n}$ of a single split sample test by taking one sample of size b is given by

$$1 - \phi(z_{1-\alpha} - h\sqrt{\tau}) . \tag{S.2}$$

Even with $\tau < 1$, the single test has better limiting power for small h. On the other hand, for large enough h, (S.1) will be larger than (S.2). In this case, the many split sample test is an improvement over the single sample test, even though it conservatively controls the Type 1 error.

Remark S.1.1. *Note that the conservative method compares the quantiles of the test statistics with $z_{1-\alpha r}$. Using the above power formula, the threshold for the overall test can be chosen as the largest constant c satisfying*

$$1 - \Phi\left[\frac{1}{\sqrt{\tau}}(c - z_{1-r}\sqrt{1-\tau})\right] = \alpha .$$

Therefore, the cutoff can be improved to $\sqrt{\tau}z_{1-\alpha} + \sqrt{1-\tau}z_{1-r}$. If this cutoff is used, the resulting test has the same power as the UMP test, regardless of the split size or the quantile used.

One can also consider the minimum quantile and use the procedure which rejects if the minimum p-value is $\leq \alpha/M$. Assume the M splits of size b are disjoint. Then, the exact power is given by

$$P\left(\max \sqrt{b}\bar{X}_{n,b,i} \geq z_{1-\alpha/M}\right) = 1 - \Phi\left(z_{1-\alpha/M} - \sqrt{\tau}h\right)^M . \tag{S.3}$$

If $b/n \to \tau \in (0,1)$ and $M \to \infty$, (S.3) tends to $1 - \exp(-\alpha)$, for any h. (A proof follows by Theorem 1.5.3 in Leadbetter, et. al. (1983).) Therefore, the power is very poor (even though the test is quite close to the nominal level and Bonferroni type procedures usually perform reasonably under independence). A similar result holds for overlapping random splits.

S.1.2 Power of Averaging p-values using Procedure 4.1

Suppose there are M splits, each of size b. The p-value based on the ith split is $1-\Phi(\sqrt{b}\bar{X}_{n,b,i})$. First, consider the case where the splits are disjoint, so there are M splits with $b = \lfloor n/M \rfloor$. The test based on the average p-value equivalently rejects when

$$M^{-1/2}\sum_{i=1}^{M}[\Phi(\sqrt{b}\bar{X}_{n,b,i}) - \frac{1}{2}] > c_M(1-\alpha) \ ,$$

where $c_M(1-\alpha)$ is the $1-\alpha$ quantile of the distribution of $M^{-1/2}\sum_{i=1}^{n}(U_i - \frac{1}{2})$ when the U_i are i.i.d. $U(0,1)$. Note $c_M(1-\alpha) \to z_{1-\alpha}/\sqrt{12}$ as $M \to \infty$.

Theorem S.1.2. *Consider Procedure 4.1 with M independent splits of size $b = \lfloor n/M \rfloor$.*
(i) If b is fixed and $M \to \infty$, the limiting power against alternatives $h/\sqrt{n}$ is (for any fixed b) given by

$$1 - \Phi\left(z_{1-\alpha} - \sqrt{\frac{3}{\pi}}h\right) \ . \tag{S.4}$$

(ii) If M is fixed but $b \to \infty$ with $b/n \to 1/M$, then the limiting power against $h/\sqrt{n}$ can be expressed as

$$P\{M^{-1/2}\sum_{i=1}^{M}\left[\Phi(Z_i + \frac{h}{\sqrt{M}}) - \frac{1}{2}\right] > c_M(1-\alpha)\} \ ,$$

where the Z_i are i.i.d. $N(0,1)$. As $M \to \infty$, this expression also tends to (S.4).

Since $\sqrt{3/\pi} \approx 0.977 \approx 1$, (S.4) is very nearly the limiting power of the UMP test. For example, with $b = 1$ and $M = n$, a test based on the statistic $\sum_i \Phi(X_i)$ has surprisingly near the performance of the test based on $\sum_i X_i$.

The next result considers overlapping splits using K-fold cross validation (though there is no selection here).

Theorem S.1.3. *If splits are chosen according to K-fold cross validation (so subsets of size $(K-1)n/K$ are used for testing), then the limiting power of Procedure 4.1 against contiguous alternatives $h/\sqrt{n}$ can be expressed as f*

$$P\left(\sum_{k=1}^{K} 1 - \Phi(Z_i + h) \leq d_K\right)$$

where $(Z_1, ..., Z_K)$ follows the multivariate normal distribution with means zero, variances one, and $cov(Z_i, Z_j) = (K-2)/(K-1)$ and d_K is the α quantile of the distribution of $\sum_{k=1}^{K} 1 - \Phi(Z_i)$. As $K \to \infty$, the probability tends to

$$P\left(1 - \Phi\left(Z_1 + h\right) > \alpha\right) = 1 - \Phi\left(z_{1-\alpha} - h\right) \ .$$

Consequently, if a relatively large amount of data is used for testing, the power of the procedure that averages p-values is nearly that of the UMP test. When the fraction of data used for testing is large, this method of combining splits gives better power than using a single split.

Procedure 3.6 is level α (under independence) and is equivalent to the UMP test when splits are chosen so that each observation appears in an equal number of splits. Next, we consider the conservative Procedure 3.7.

Theorem S.1.4. *Assume $b/n \to \tau$. The limiting powerusing Procedure 3.7 against contiguous alternatives $h/\sqrt{n}$ which rejects for small values of*

$$\frac{1}{M}\sum_{m=1}^{M}\Phi^{-1}(\hat{p}_{n,m})$$

is

$$1-\Phi\left(\frac{1}{\sqrt{\tau}}z_{1-\alpha}-h\right)$$

if the splits are each of size b and are chosen such that each X_i appears in an equal number of splits.

Interestingly, the power of the conservative method for combining p-values depends only on the size of the split (and not the number of splits used) when each observation appears in an equal number of splits, but in this situation, the power is decreasing as the size of the splits decreases.

S.1.3 Power of U-statistic Based Methods

S.1.3.1 Power of Procedure 4.3 using average p-value

The results of the previous section indicate that although the conservative method based on the median p-value (given by Procedure 3.1) can have better power than simply using one split, the power is only larger for values of the local parameter where the power is already near one. Moreover, the conservative method is often noticeably worse than the full data method and the conservative method can be very conservative, especially when a large number of splits are used. By comparison, using the asymptotically level α test based on the average p-value (given by Procedure 4.3) gives an improvement in limiting local power.

Define the average p-value taken over all splits of the data to be

$$U_n(X_1,...,X_n)=\frac{1}{N}\sum_{i=1}^{N}\hat{p}_{n,b,i}=\frac{1}{N}\sum_{i=1}^{N}1-\Phi(\sqrt{b}\bar{X}_{n,b,i})\ ,$$

with $N=\binom{n}{b}$.

Theorem S.1.5. *Let $X_1, ..., X_n$ be i.i.d according to a normal distribution with mean μ and variance one. If $b \to \infty$ and $b/\sqrt{n} \to 0$, then under $H_0 : \mu = 0$,*

$$\sqrt{n}\frac{U_n - 1/2}{\sqrt{b/(4\pi)}} \xrightarrow{d} N(0,1) \ .$$

The limiting power of the one sided test of $H_0 : \mu = 0$ using Procedure 4.3 against contiguous alternatives $h/\sqrt{n}$ is

$$P(N(h,1) > z_\alpha) = 1 - \Phi(z_{1-\alpha} - h)$$

Remark S.1.2. *If b is fixed, the average of the p-values computed over all splits of the data remains asymptotically normal; however, the overall test is less powerful asymptotically than the UMP test against local alternatives. A justification of this is implicit in the proof of Theorem S.1.5.*

Despite testing on small portions of the data, using the average p-value has the same limiting local power as the UMP test. Using the asymptotic normality of the p-value, the test rejects for an average p-value below $1/2 + z_\alpha\sqrt{b/(4\pi n)}$. By contrast, the conservative method rejects when the average p-value is below $\alpha/2$, which can be quite substantially lower than this threshold.

S.1.3.2 Power of Procedure 4.3 based on the median p-value

An asymptotically level α test can also be performed using the median of the p-values (given by Procedure 4.4). The power of this method is as follows.

Theorem S.1.6. *Suppose that $X_1, ..., X_n$ are i.i.d. according to a normal distribution with mean μ and variance one. Suppose $b \to \infty$ in such a way that $b/\sqrt{n} \to 0$. Then, under a sequence of local alternatives $h/\sqrt{n}$,*

$$\sqrt{\frac{2\pi n}{b}}\,(\tilde{p}_n - 1/2) \xrightarrow{d} N(h,1) \ ,$$

where $\tilde{p}_n$ is the median p-value computed over all splits. Thus, the limiting power of the one sided test of $H_0 : \mu = 0$ using Procedure 4.3 against $h/\sqrt{n}$ is

$$1 - \Phi(z_{1-\alpha} - h) \ .$$

Note that the asymptotically level α test rejects if the median is less than $1/2 + z_\alpha\sqrt{b/n}$, which can be substantially larger than $\alpha/2$ (for example, if $\alpha = .1$, $n = 100$, and $b = 10$, $1/2 + z_\alpha\sqrt{b/n} = .0947$ whereas $\alpha/2 = .05$). The power of this test is the same as the UMP test.

S.1.4 Conclusions

S.1.4.1 Theoretical results

The conservative methods, based on the quantiles or average of the p-values always perform worse than the full data tests. In some circumstances, these methods can outperform using a single split, suggesting that using multiple splits is beneficial when data splitting is necessary.

However, if thresholds for rejection based on the quantiles of the p-values are found that make the overall test exact, the power is the same as the UMP test. Consequently, there is hope that finding refined thresholds for rejection can lead to power that not only improves upon a single split, but is comparable to full data methods.

When using the minimum p-value, it is best to use a small number of splits. For a large number of splits, the power is poor and, unlike the conservative method based on fixed quantiles of the p-values, an improved threshold cannot improve the power.

The U-statistic method based on either the average or median p-value is asymptotically equivalent to the UMP test.

S.1.4.2 Numerical Results for the Conservative Methods

In this section, we study the effect of the split size and number of splits used in the context of testing a single mean. Using multiple splits of the data, we reject $H : \mu = 0$ if the proportion of p-values smaller than $r\alpha$ exceeds r. To see the effect of using multiple splits, and the split size on the power of the test, we give simulated power curves (at nominal level $\alpha = .05$ based on 5,000 iterations) in the setting of $n = 100$ with data distributed as $N(\mu/\sqrt{n}, 1)$ for values of μ specified in the plots.

Figures S.1.1, S.1.2 and S.1.3 give the power of the conservative quantile method using $\tau = .25$ with $r = .2$, $r = .5$ and $r = .8$, respectively.

Figures S.1.4, S.1.5 and S.1.6 give the power of the conservative quantile method using $\tau = .5$ with $r = .2$, $r = .5$ and $r = .8$, respectively.

Figures S.1.7, S.1.8 and S.1.9 give the power of the conservative quantile method using $\tau = .75$ with $r = .2$, $r = .5$ and $r = .8$, respectively.

Rather than using the quantiles of the p-values, rejecting if twice the average p-value is smaller than α is a conservative test. Figures S.1.10, S.1.11 and S.1.12 give the power of this method using $\tau = .25$, $\tau = .5$, and $\tau = .75$, respectively.

Findings for the conservative quantile method:

- For large values of r ($r = .8$), combining splits performed worse than using a single split, regardless of the sample size. For large values of r, the method is extremely conservative regardless of the number of splits used.

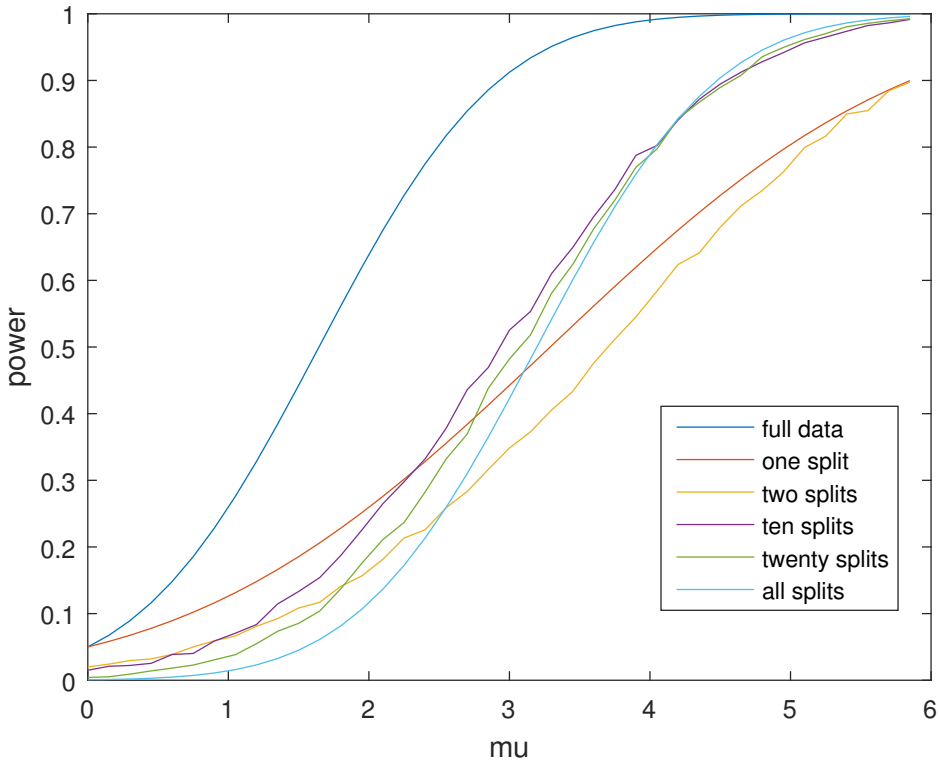

Figure S.1.1: Power of the conservative method with $r = .2$, $\tau = .25$

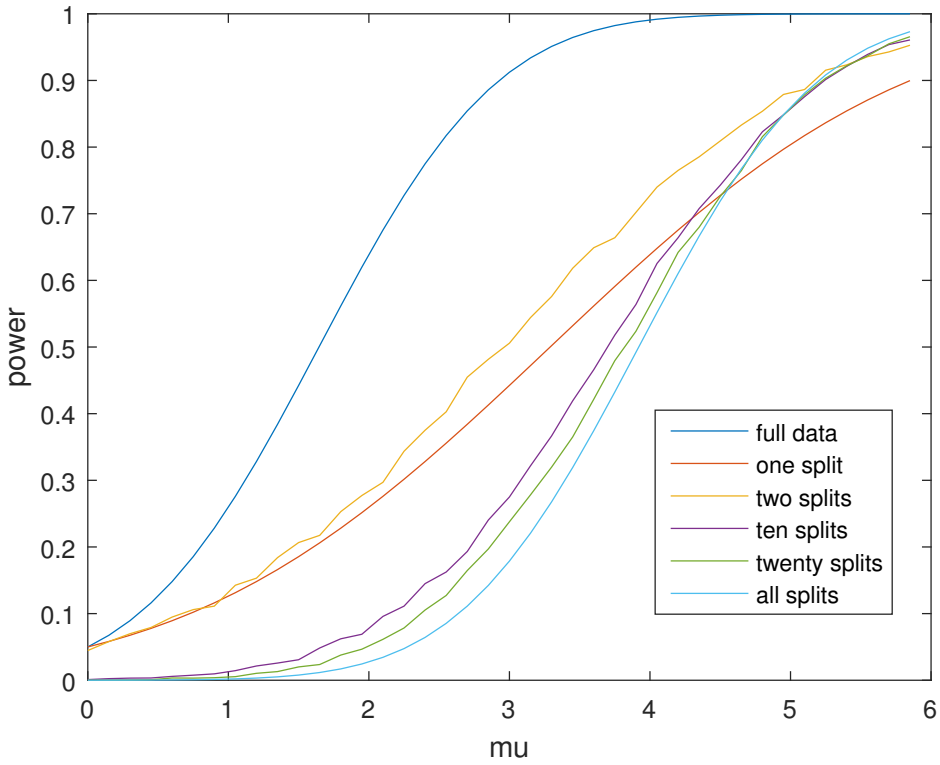

Figure S.1.2: Power of the conservative method with $r = .5$, $\tau = .25$

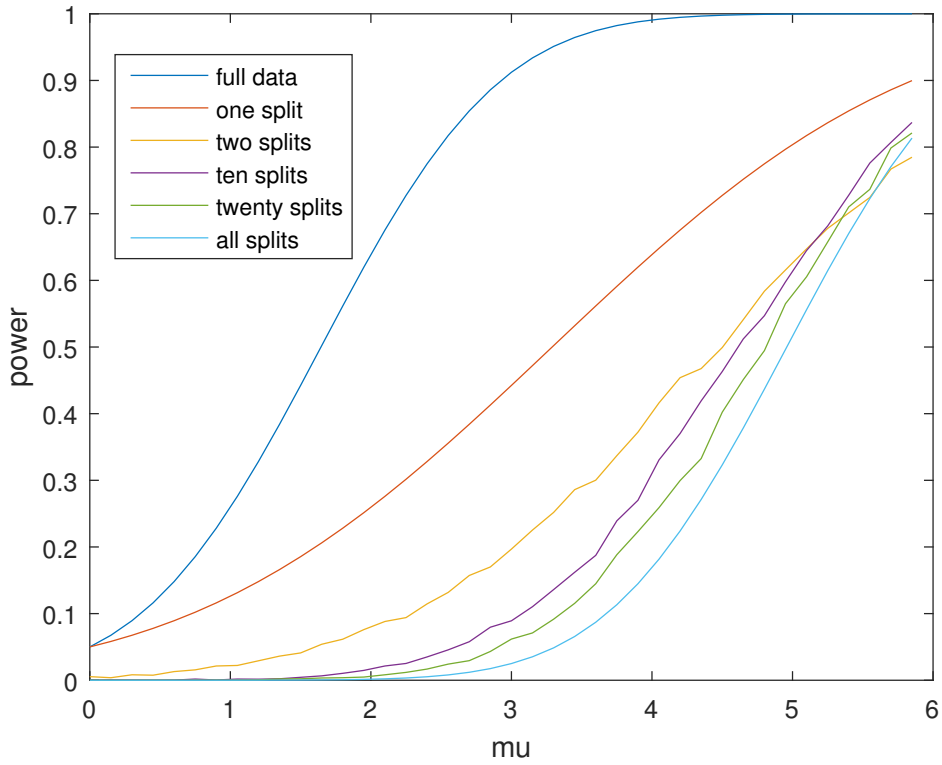

Figure S.1.3: Power of the conservative method with $r = .8$, $\tau = .25$

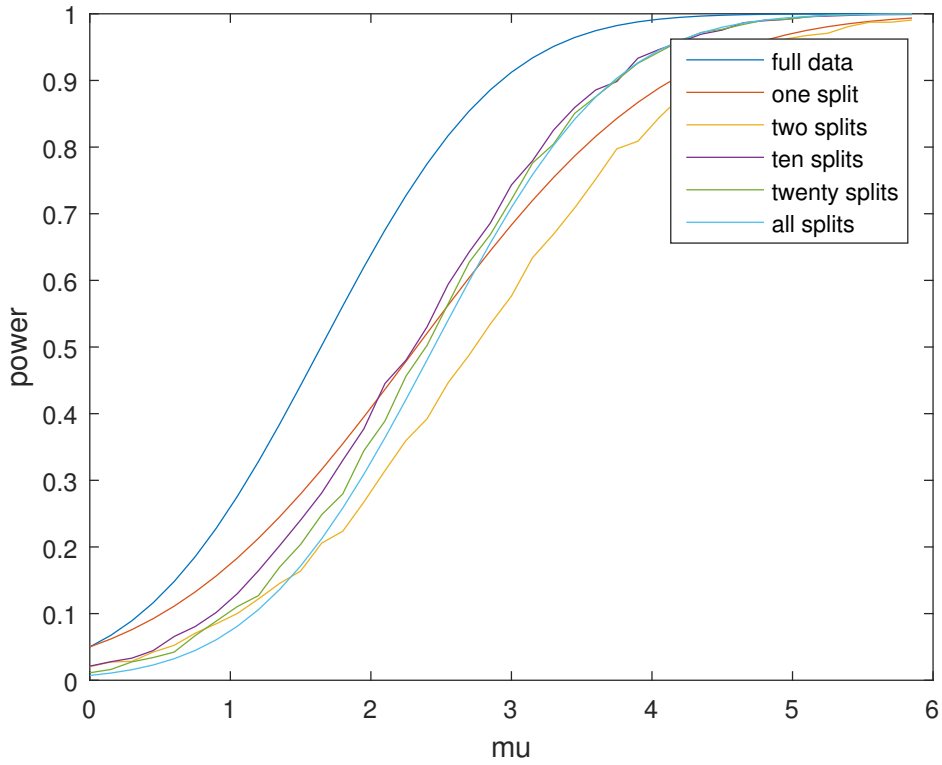

Figure S.1.4: Power of the conservative method with $r = .2$, $\tau = .5$

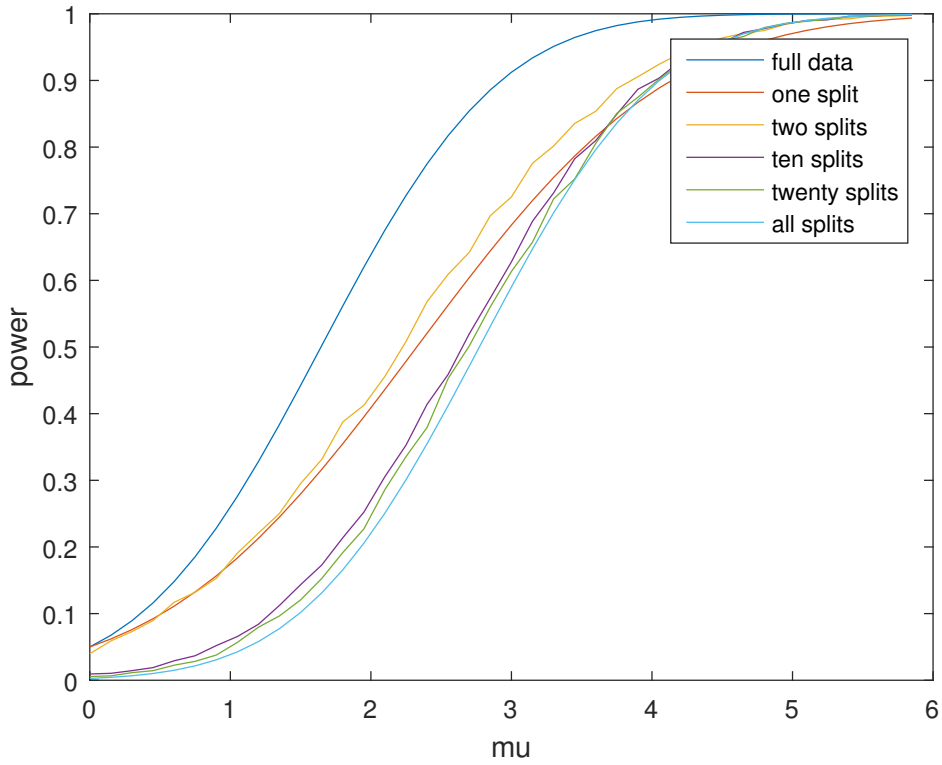

Figure S.1.5: Power of the conservative method with $r = .5$, $\tau = .5$

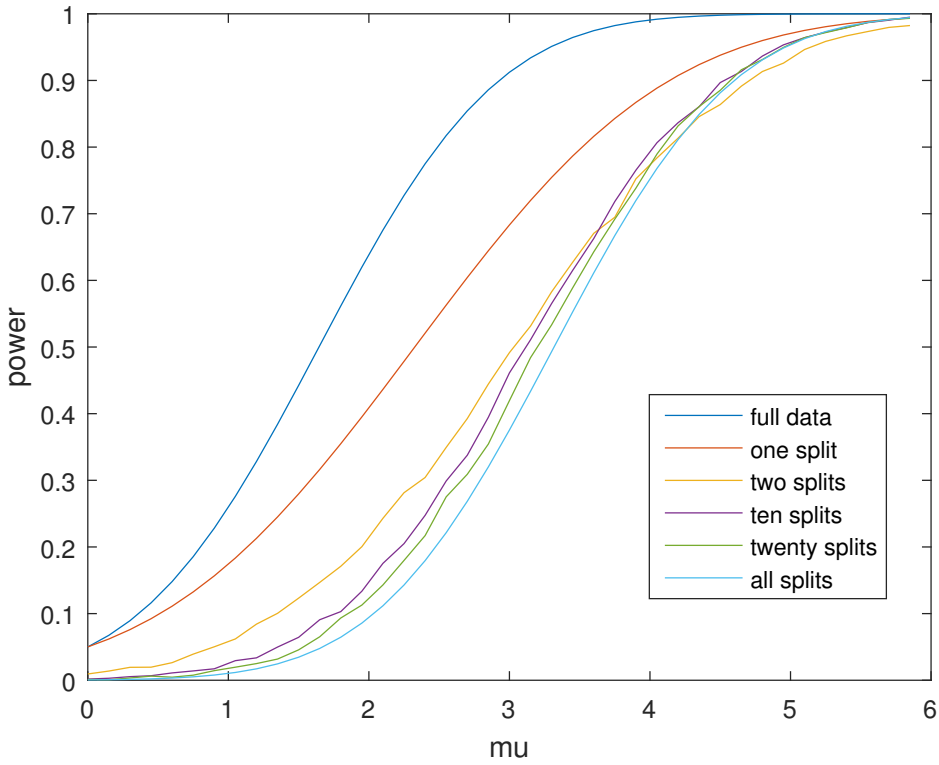

Figure S.1.6: Power of the conservative method with $r = .8$, $\tau = .5$

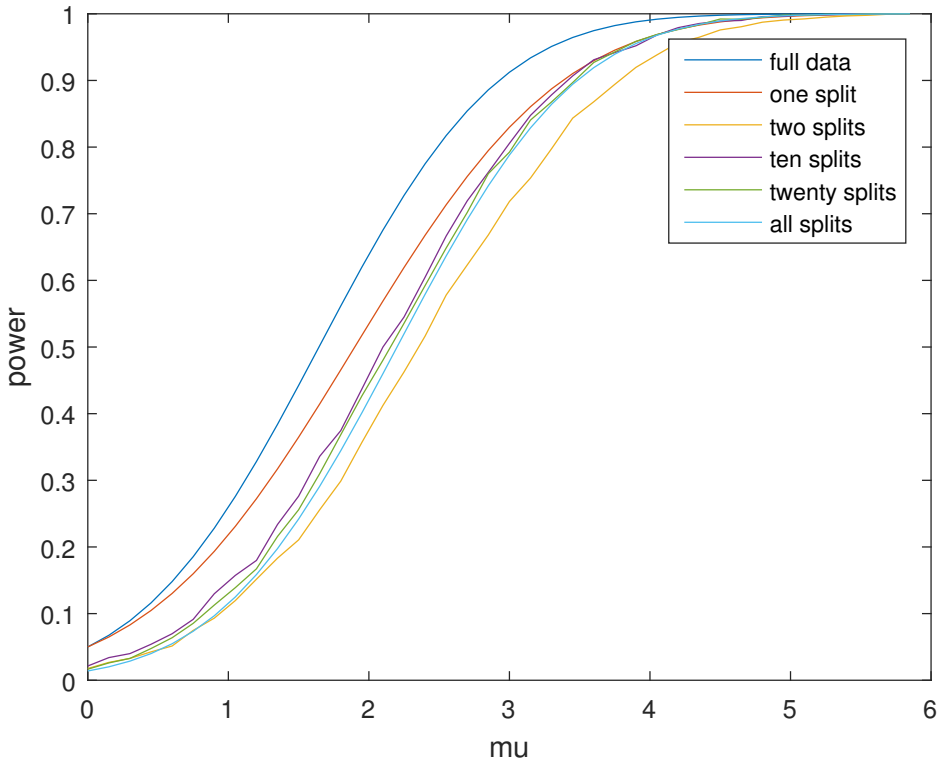

Figure S.1.7: Power of the conservative method with $r = .2$, $\tau = .75$

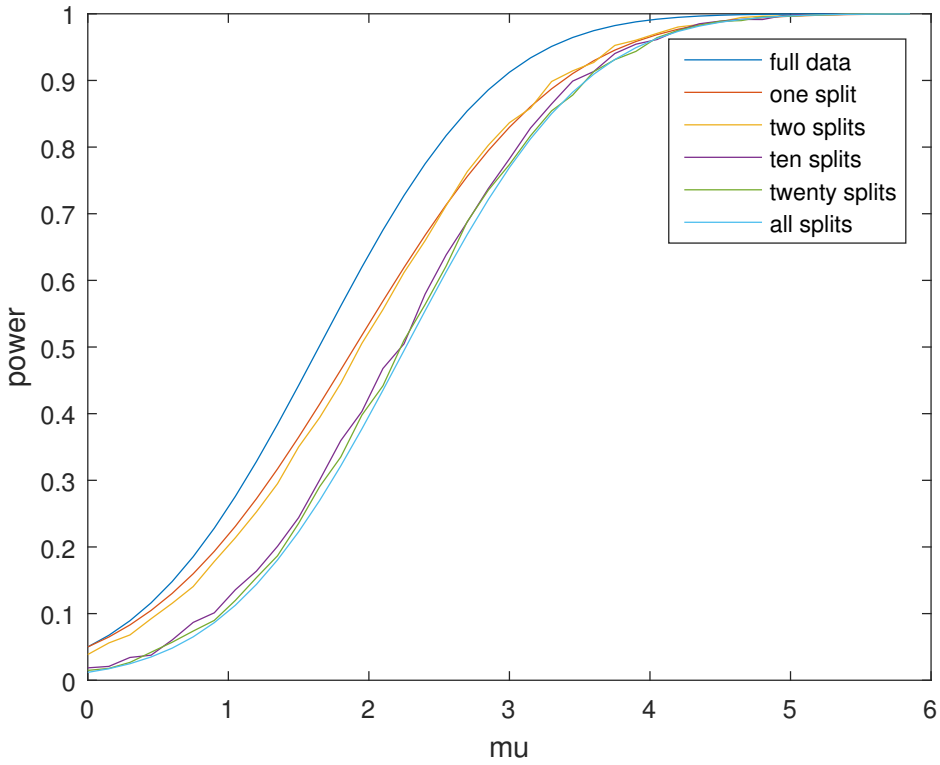

Figure S.1.8: Power of the conservative method with $r = .5$, $\tau = .75$

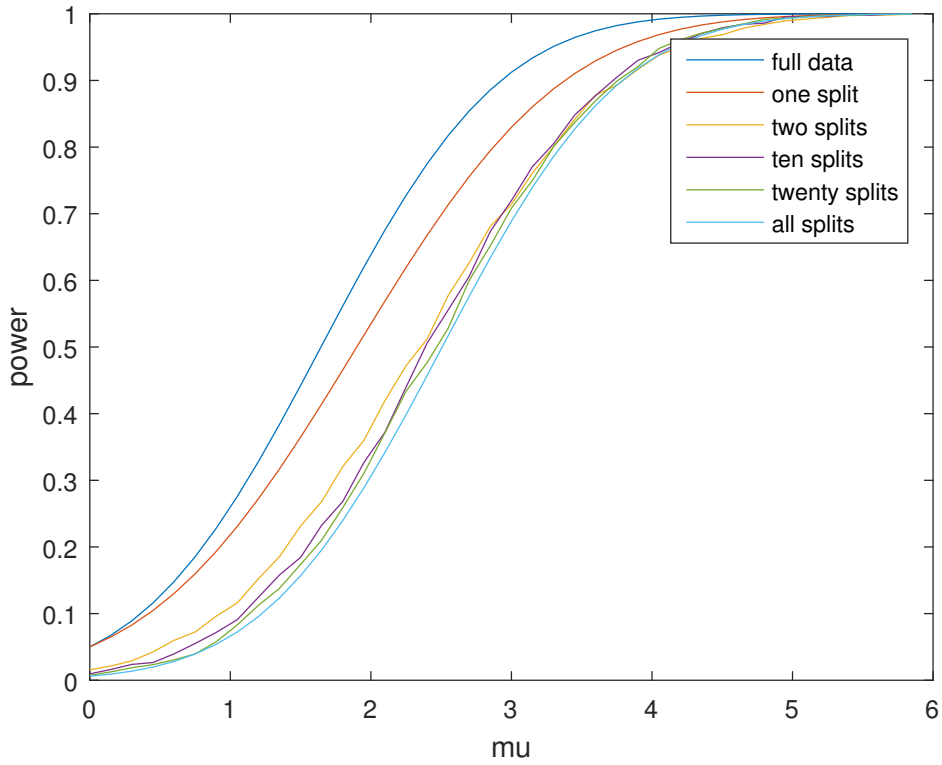

Figure S.1.9: Power of the conservative method with $r = .8$, $\tau = .75$

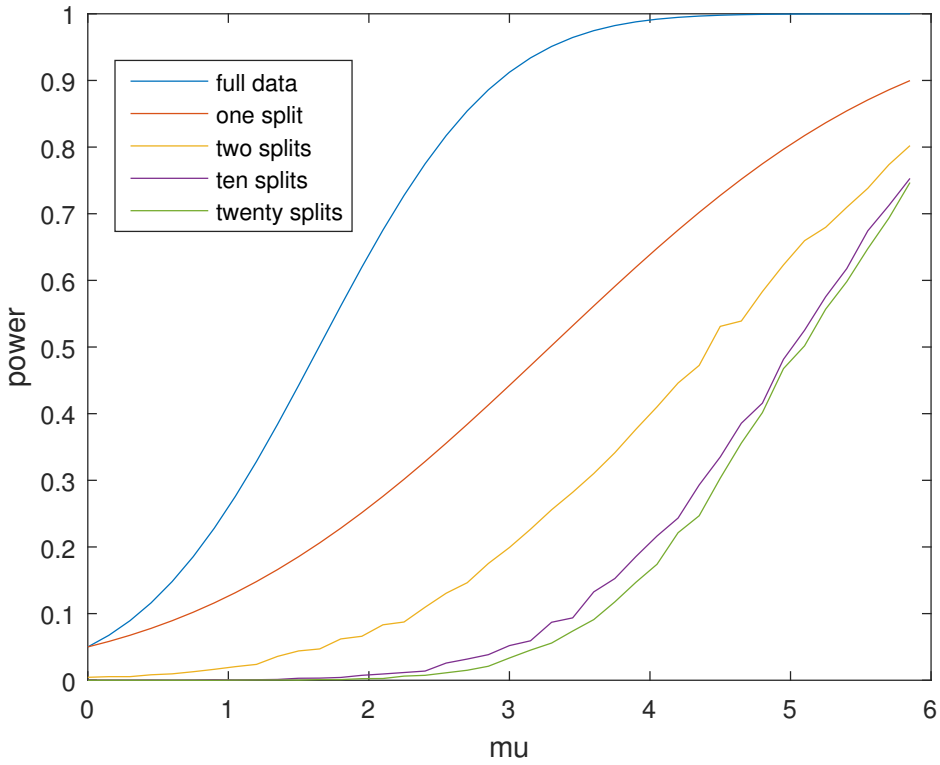

Figure S.1.10: Power of the conservative average method with $\tau = .25$

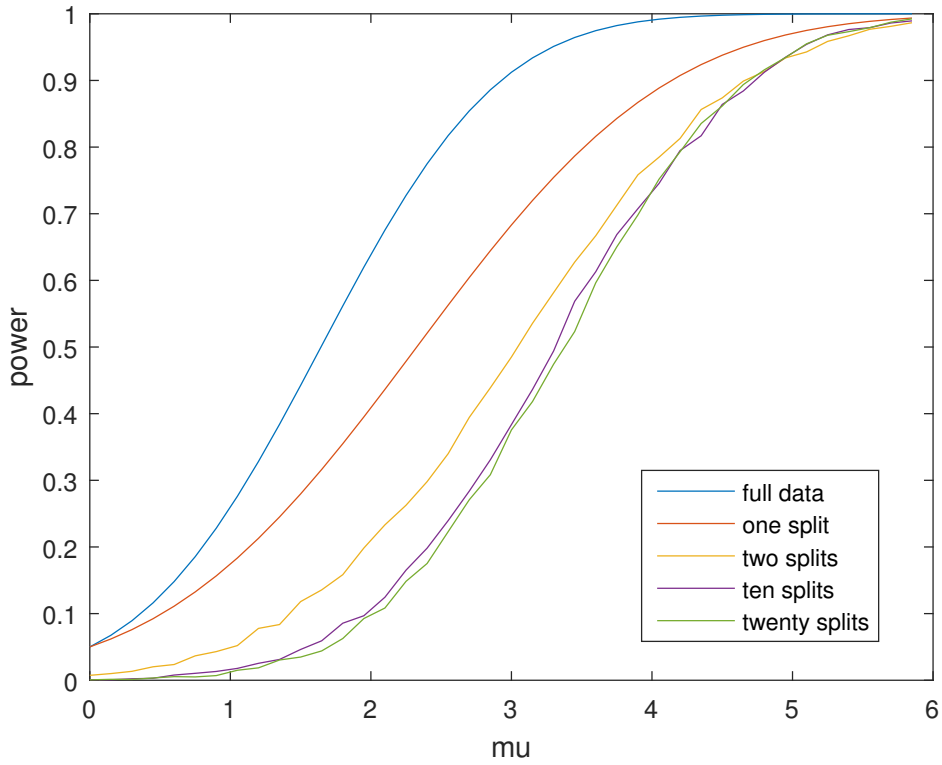

Figure S.1.11: Power of the conservative average method with $\tau = .5$

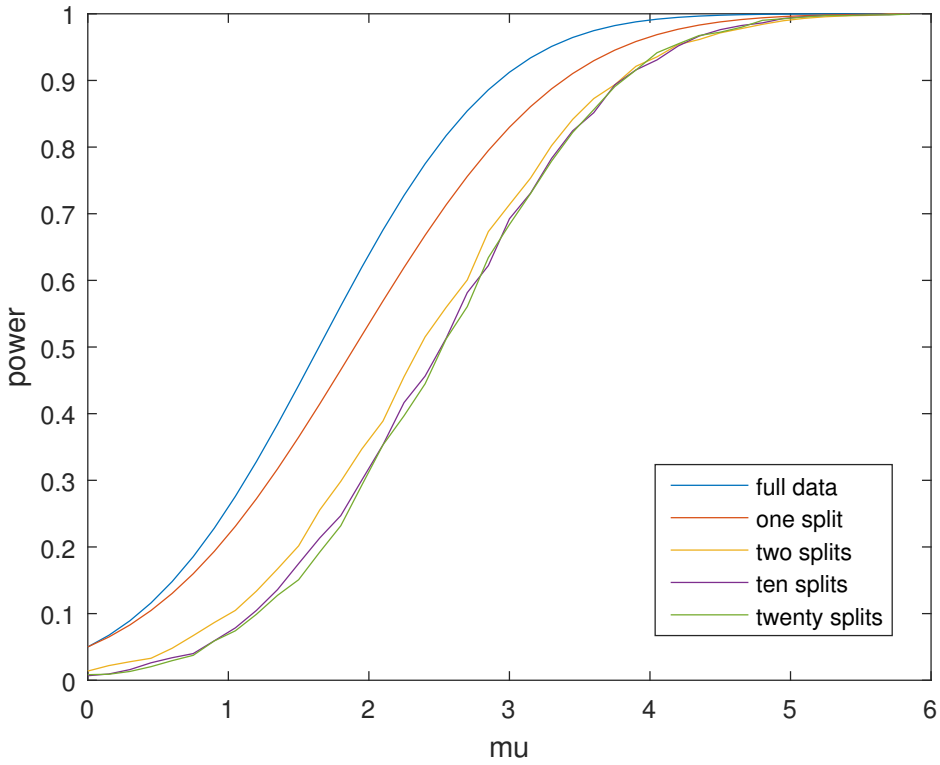

Figure S.1.12: Power of the conservative average method with $\tau = .75$

- For small r ($r = .2$), using many splits of the data can improve power over using a single splits. When the fraction of data used for testing is small ($\tau = .25$), the more splits performs substantially better than using a single split for moderate and large values of μ.
- For moderate r ($r = .5$), using two splits improves power over a single split, except when the fraction of data used for testing is large, in which case two splits and one splits performs nearly identically.
- For any τ and r, the power using 10, 20 and all splits are nearly identical, suggesting that using many more splits than 10 is unnecessary.
- Generally, when using a smaller portion of data for testing, smaller values of r are preferable.
- Using two splits and $r = .5$ works reasonably well in all scenarios.

Findings for the conservative average method:

- For each of the simulation settings, using multiple splits performs worse than using a single split.
- The power is decreasing in the number of splits used.

S.2 Simulations for Testing a Multivariate Hypothesis

In this section, we apply the methods presented in Sections 3 and 4 to the problems presented in Section 2. The simulations are performed at nominal level $\alpha = .05$ and are based on 5,000 iterations.

S.2.1 Cox's Example

Suppose we are interested in testing the null hypothesis

$$H_0 : \mu_1 = \cdots = \mu_p = 0 \tag{S.1}$$

that specifies the means $\mu_1, ..., \mu_p$ of p normal populations are zero against the alternative that exactly one of the populations has non-zero mean. In the simulations presented here, $n = 50$ samples are taken from $p = 20$ populations. For data splitting, half of the data is used to select a population to test according to which sample mean is largest. The remainder of the data is used to test that the mean of the selected population is zero. Figure S.2.1 plots

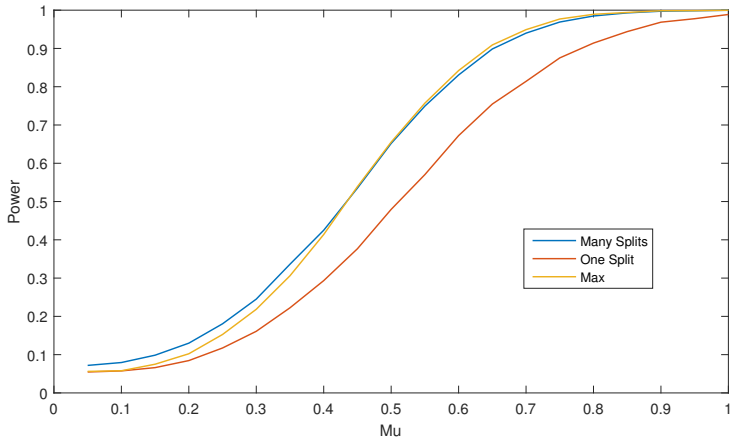

Figure S.2.1: Power in Cox's example

the power of the full data test based on the max statistic, the data splitting test using a single split, and the data splitting test averaging p-values over 200 randomly chosen subsamples of size $b = 10$. The threshold for rejection is found by simulation.

Findings:

- As Cox found, using a single split of the data reduces power.
- The method averaging all p-values gives nearly identical power to the full data test.

S.2.2 Estimated Neyman-Pearson Test

Once again consider (S.1) but the alternative vector now is unrestricted (so many could be nonzero). Data splitting can be used to estimate the parameters giving some idea of how the particular alternative departs from the null hypothesis.

Suppose again that $(X_{i,1}, ..., X_{i,p})$ have means $\mu_1, ..., \mu_p$. The Neyman-Pearson test for testing

$$H_0 : \mu_1 = \cdots = \mu_p = 0$$

against the alternative

$$H_a : \mu_1 = \mu_1^0, ..., \mu_p = \mu_p^0$$

rejects for large values of

$$T_{NP} = \mu_1^0 \sqrt{n} \bar{X}_1 + \cdots + \mu_p^0 \sqrt{n} \bar{X}_p$$

Of course, without knowing the alternative hypothesis, we cannot use this test. However, we may be able to use data splitting to try and estimate the true mean vector using the first portion of the data, and to test the hypothesis using the remaining data. For a split S, write

$\bar{X}_k(S)$ for the sample average computed on the first split, and $\bar{X}_k(S^c)$ for the sample average computed on the remainder of the data. For each split, we can use the statistic

$$T^S_{n,p} = \frac{\sum_k \sqrt{b}\bar{X}_k(S)\bar{X}_k(S^c)}{\sqrt{\sum_k \left(\bar{X}_k(S^c)\right)^2}}$$

and reject if this statistic is larger than $z_{1-\alpha}$. Following the theoretical results presented earlier, half of the data will be used for selection and half will be used for testing.

Under alternatives where many of the populations may have mean zero, it is unnecessary to include these populations in the test statistic. This data driven Neyman-Pearson test can be improved in this setting by incorporating selection of components. Each population could be tested at some preliminary level β, i.e. we could test the individual hypotheses $H_j : \mu_k = 0$, and only include populations for which the test rejects. For the purposes of simulation, we test these hypotheses at level $p^{-1/2}$ (which was seen in Section **??** to be a reasonable choice of threshold when the fraction of data used for testing is 1/2) and include the rejected populations.

To compare the performance of these procedures, we present simulations for a sample of size $n = 40$ from $p = 20$ populations. q of the populations have non-zero mean taking value $\mu/\sqrt{n}$. The power plots compare the performance of the full data Chi-squared test, the Neyman-Pearson test using one split with half the data used for testing, two splits, the U-statistic method with $b = 10$ and the U-statistic method with selection with $b = 10$ (where half of these are used for selection and the other half for testing). Figure S.2.2 gives the power for $q = 2$, and Figure S.2.3 gives the power for $q = 10$. For the methods averaged over many splits, the threshold for rejection is found using Monte Carlo simulation.

Findings:

- One split using either Neyman-Pearson statistic gives fairly severe loss of power over the Chi-squared test.
- Two splits performs comparably to one split.
- Taking the average p-value over all splits using the U-statistic method gives power that is very close to the Chi-squared test.
- Taking the average p-value over all splits using the U-statistic method with selection gives power that is noticeably better than the Chi-squared test.

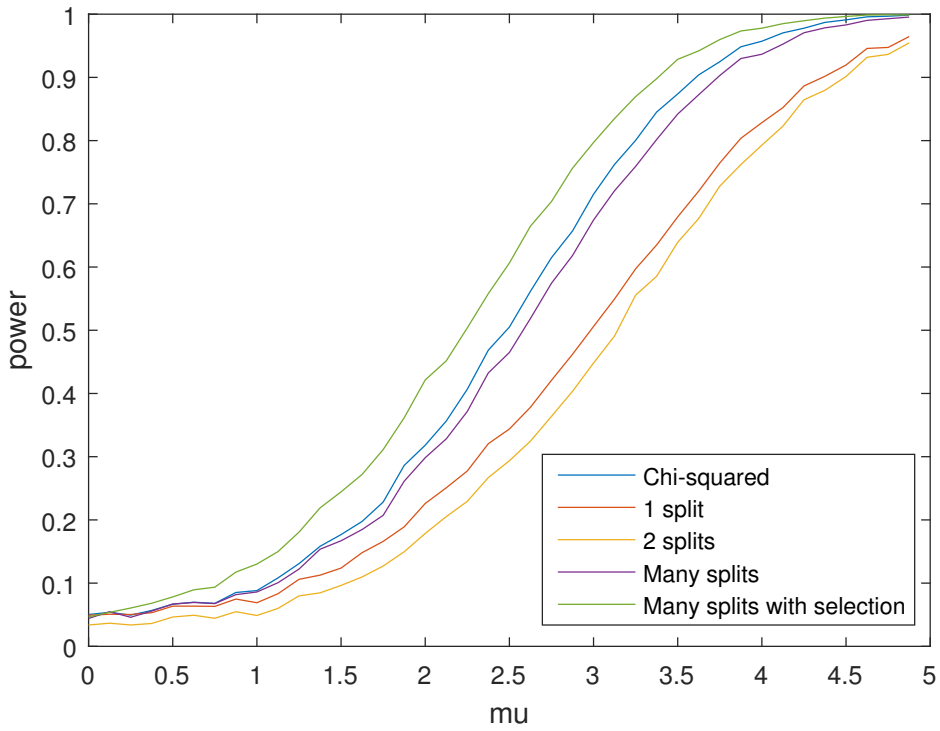

Figure S.2.2: Power of the Neyman-Pearson tests with $q = 2$

S.2.3 Moment Inequality Problem

In the spirit of K-fold cross validation, suppose we split the data into K equally sized parts, say $S_1, ..., S_k$ and compute K p-values by using the data in S_k to choose a test statistic, and using S_k^c to compute a p-value based on the chosen test statistic, for $k = 1, ..., K$.

For example, if $X_i = (X_{i,1}, ..., X_{i,p}) \sim N(\mu, \Sigma)$, $i = 1, ..., n$, and we are interested in testing the null hypothesis

$$H : \mu_i \leq 0, i = 1, ..., p \ ,$$

one might use the test statistic

$$T_{n,D}(\bar{X}_1, ..., \bar{X}_p) = \max_{i \in D} \sqrt{n} \bar{X}_i$$

for some subset D of $\{1, ..., p\}$.

If we split the data, we can use part of the data, say those with indices in S_k to choose a test statistic, and the remaining portion of the data to perform the test. Here, we could choose $D = \left\{ i : \sqrt{|S_k|} \bar{X}_{i,S_k} > z_\beta \right\}$ for some $\beta > 0$, where $\bar{X}_{i,S_k} = \frac{1}{|S_k|} \sum_{j \in S_k} X_{j,i}$.

Each split S_k chooses some subset D of predictors, and then uses the observations in S_k^c to find a p-value based on the test statistic. If the selection based on S_k is computed using

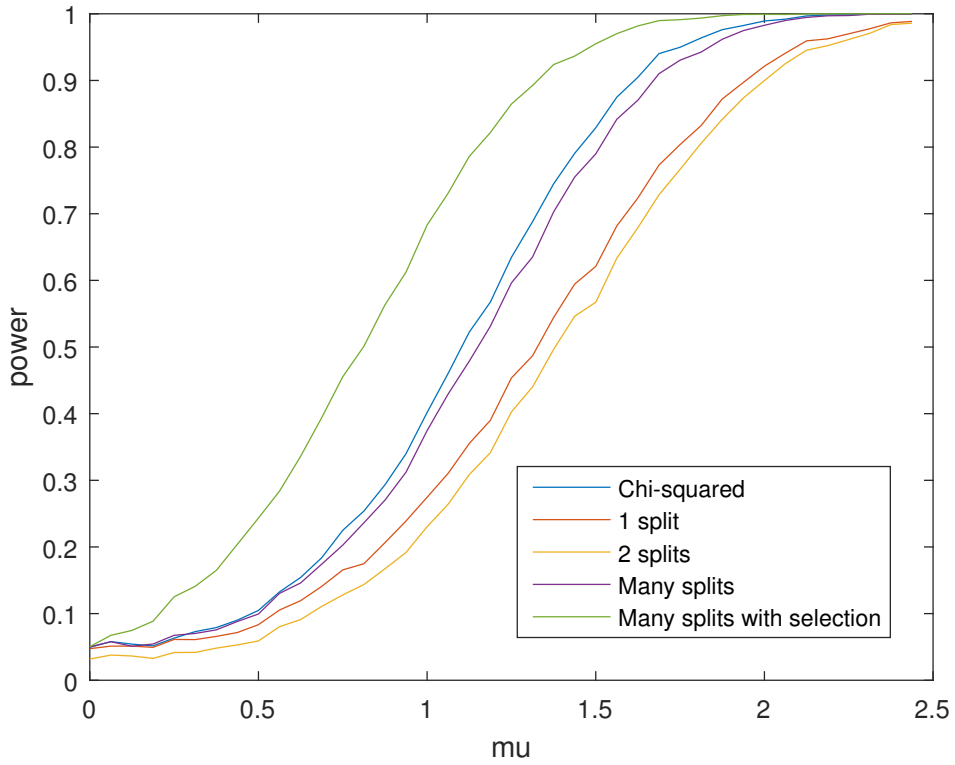

Figure S.2.3: Power of the Neyman-Pearson tests with $q = 10$

the sample means computed on this split, then we can use the bootstrap to approximate the joint distribution of the p-values under $\mu_1 = \cdots = \mu_p = 0$.

To test this hypothesis, we can use the average of the p-values computed on each of the K splits of the data as a test statistic. This statistic is a smooth function of the means computed on the splits, which have an asymptotically normal distribution:

$$\sqrt{n/K}\left(\bar{X}_{S_1} - \mu, ..., \bar{X}_{S_K} - \mu\right) \to N(0, I_K \otimes \Sigma)$$

Under $\mu = 0$, the distribution of the means computed on the splits of the data can be approximated by the distribution of

$$\sqrt{n/K}\left(\bar{X}^*_{S_1} - \bar{X}, ..., \bar{X}^*_{S_K} - \bar{X}\right) \to N(0, I_K \otimes \Sigma)$$

and we can use the bootstrap to approximate the distribution of the average p-value.

Consider the case $p = 2$, where the test statistic used on each split S is $\max\left\{\bar{X}_{1,S^c}, \bar{X}_{2,S^c}\right\}$ if $\bar{X}_{1,S}, \bar{X}_{2,S} > -z_\alpha$, $\bar{X}_{1,S^c}$ if $\bar{X}_{1,S^c} = \min\left\{\bar{X}_{1,S}, \bar{X}_{2,S}\right\} < -z_\alpha$, or $\bar{X}_{2,S^c}$ if $\bar{X}_{2,S^c} = \min\left\{\bar{X}_{1,S}, \bar{X}_{2,S}\right\} < -z_\alpha$.

If the test statistic is of the form

$$\sum_k 1 - \Phi\left(\sum_j I\left\{\left|\bar{X}_j(S_k)\right| \geq z_{1-\beta}\right\}\bar{X}_j(S_k)\right)$$

then the distribution of the test statistic can be approximated by the subsampling distribution of

$$\sum_k 1 - \Phi\left(\sum_j I\left\{\left|\bar{X}^*_j(S_k)\right| \geq z_{1-\beta}\right\}\left(\bar{X}^*_j(S_k) - \bar{X}_j I\left\{\bar{X}_j > 0\right\}\right)\right) \ .$$

In particular, when approximating the distribution of the test statistic, it is helpful that the selection be done using an un-centered statistic. Subsampling gives a better approximation to the null distribution than the bootstrap, which simulates the distribution under $\mu_1 = \cdots = \mu_p = 0$ (rather than the true null which may have some negative means). Here we give simulation results for a sample of size $n = 50$ of a $p = 2$ dimensional independent bivariate normal random variables. The full data test based on the maximum is compared with the data splitting method with one-fifth of the data used for selection according to one sided tests at the five percent level that each mean is negative, and the five-fold cross validation method averaging this data splitting procedure over all five folds. For the five-fold test, the distribution of the test statistic is found using subsampling with subsample size 20. Figure S.2.4 gives the power when the first mean is $-.5$ and the second mean is μ specified in the plot. Figure S.2.5 gives the power when the first mean is 0 and the second mean is μ specified in the plot.

Findings:

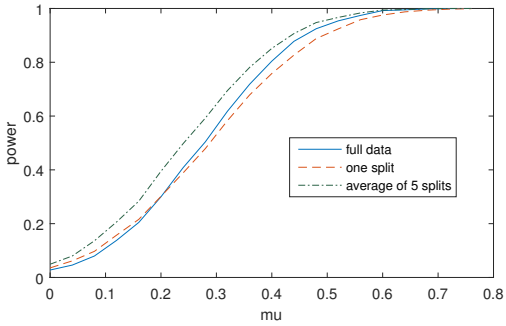

Figure S.2.4: Power for the moment inequality problem with $p = 2$, one component having mean $-.5$ and the other having mean μ.

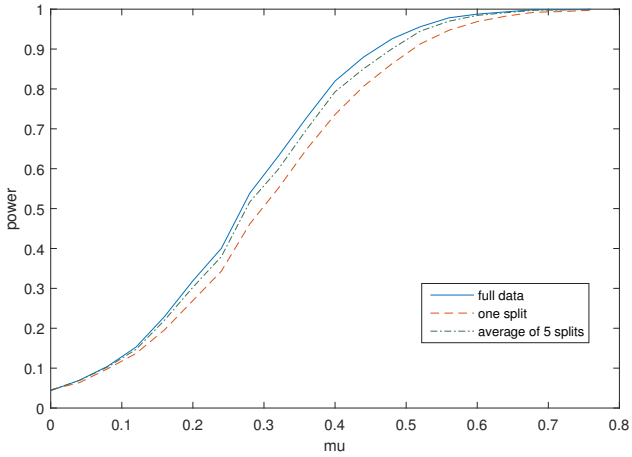

Figure S.2.5: Power for the moment inequality problem with $p = 2$, one component having mean 0 and the other having mean μ.

- Using one split decreases power over the full data test.
- The data splitting test averaged over five splits outperforms the full data test when one mean is negative and reducing the components under consideration is beneficial.
- When only one mean is positive, the many split procedure does not dramatically reduce power over the full data test.

S.2.4 Testing for Many Means

In this section, we consider testing that the means of p populations are zero against the unrestricted alternative. More precisely, assume that $X_1, ..., X_p$ are iid according to a normal distribution with mean μ and variance one for $j = 1, ..., p$. Once again, we are interested in testing

$$H : \mu_1 = \cdots = \mu_p = 0 \ .$$

S.2.4.1 Testing with the Max or Chi-squared Statistics Using the Conservative Quantile Method

In this section, we compare the performance of the conservative quantile method with the full data tests based on the max and Chi-squared statistics. For the purposes of the simulations done here, the alternative is that q of the populations have non-zero means, all taking value $\sqrt{2r\log(p)/n}$. In the simulations given here, we use $p = 5000$.

Figure S.2.6 gives power when the fraction of data used for testing is $\tau = 1/2$ and the tests reject if either p-value computed over two splits is smaller than $\alpha/2$ when $r = .8$. Figure S.2.7 repeats the simulations in Figure S.2.6 but with $r = .2$.

To study the effect of the number of splits used, Figure S.2.8 repeats the simulations in Figure S.2.7 for only the max statistic with the conservative test that rejects if the median p-values is smaller than $\alpha/2$.

To study the effect of the size of splits used, Figure S.2.9 repeats the simulations in Figure S.2.6 for only the Chi-squared statistic with the conservative test using K disjoint splits of the data that rejects if the smallest p-value is smaller than α/K.

Findings:

- When using the max statistic, using a single split decreases power when compared with the full data test.
- When using the Chi-squared statistic, a single split increases power when compared with the full data test.

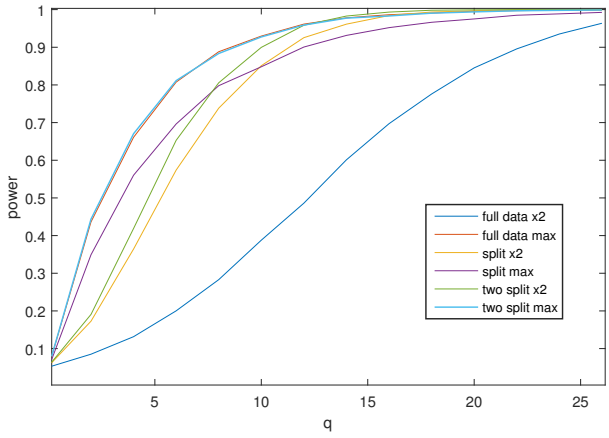

Figure S.2.6: Power when q of the means take value $\sqrt{2 \cdot .8 \cdot \log(p)/n}$

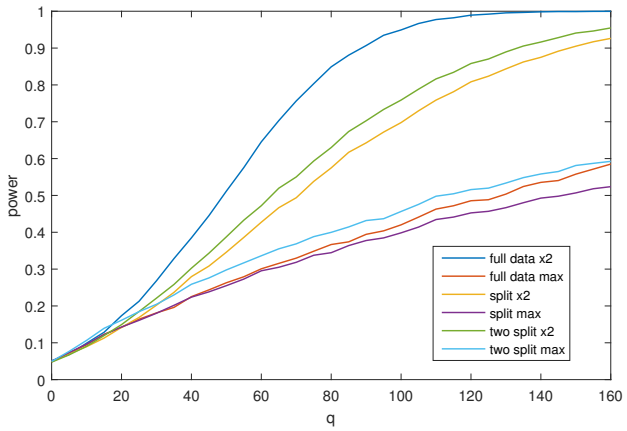

Figure S.2.7: Power when q of the means take value $\sqrt{2 \cdot .2 \cdot \log(p)/n}$

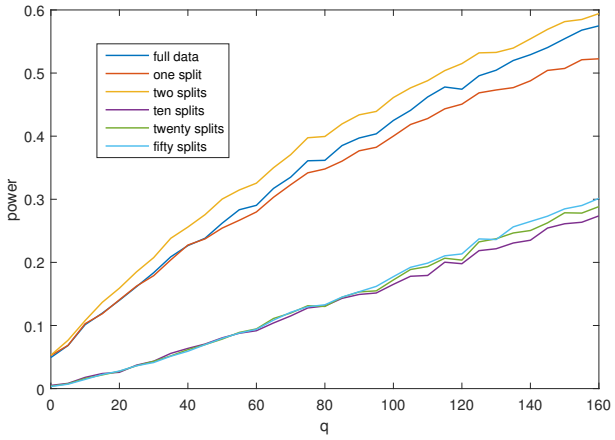

Figure S.2.8: Power when q of the means take value $\sqrt{2 \cdot .2 \cdot \log(p)/n}$

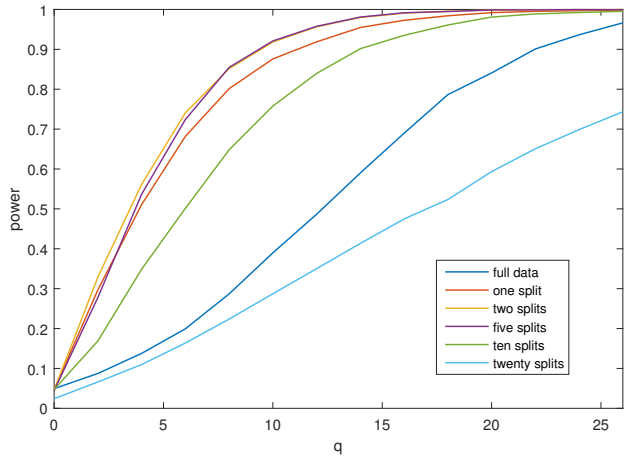

Figure S.2.9: Power when q of the means take value $\sqrt{2 \cdot .8 \cdot \log(p)/n}$

- Using two splits improves power over using a single split.
- Using two splits for the max statistic has similar or better power than the full data test.
- For the Chi-squared statistic, using two splits is substantially better than the full data test when the magnitude of the non-zero means is large ($r = .8$), but is not as efficient when the magnitude is small ($r = .2$).
- Increasing the number of splits used with the max statistic decreases power when the overall test is based on the median p-values.
- Using two or five splits with the Chi-squared statistic with a small fraction of data used for testing has comparable power but the power decreases when using more splits.

The results here are comparable to the results seen with testing a single mean. Using two splits of the data performs fairly well in all situations.

S.2.4.2 Using the U-statistic Method for Testing With the Max or Chi-squared Statistics

In this section, we compare the performance of the full data max and Chi-squared tests with the data-splitting U-statistic methods averaging the p-values for the problem of testing that the means of $p = 100$ independent populations are all zero. For simulations, we simulate $n = 40$ observations from each population. Under the alternative, the means of q (specified later) of the populations are $\mu/\sqrt{n}$. Half of a subsample of size $b = 10$ is used to select means

to test. The cutoff for selection is chosen according to the γ maximizing

$$\frac{\sum_i \left(I\left\{p_i \leq \gamma\right\} - \gamma\right)}{\sqrt{\gamma(1-\gamma)}},$$

i.e., once we find the maximizing γ, say $\hat{\gamma}$, we include populations in the test statistic if the sample average computed on the first portion of the data exceeds $\sqrt{2\log(p^{\hat{\gamma}})/n}$. The remaining data in the subsample is used to test the global null using either the maximum or Chi-squared statistics. The bootstrap (performed by computing the average p-value on a sample of size n taken with replacement from the centered data) is used to find a threshold for rejection.

Figures S.2.10, S.2.11, and S.2.12 plot the power for a range of μ values when the number of non-zero means is $q = 2$, $q = 25$, and $q = 50$, respectively. Figure S.2.13 repeats the simulations in S.2.11 but instead uses the conservative threshold given in Procedure 3.3 and $n = 100$.

Findings:

- In the sparse case ($q = 2$), the test based on the maximum has best power. The split Chi-squared and maximum tests have similar power, and both provide improvements in power over the full data Chi-squared test.
- In the less sparse cases ($q = 25$ or 50), the split Chi-squared test has nearly identical power to the full data Chi-squared test which has the best power. The split maximum test has much better power than the full data maximum, but is not quite as powerful as either of the Chi-squared tests.
- For the Chi-squared statistic, data splitting provides an improvement in power against sparse alternatives with only a slight loss of power against non-sparse alternatives.
- For the max statistic, the data splitting procedure provides a noticeable loss of power against sparse alternatives, but a dramatic improvement in power against non-sparse alternatives.
- Using the conservative threshold severely reduces power.

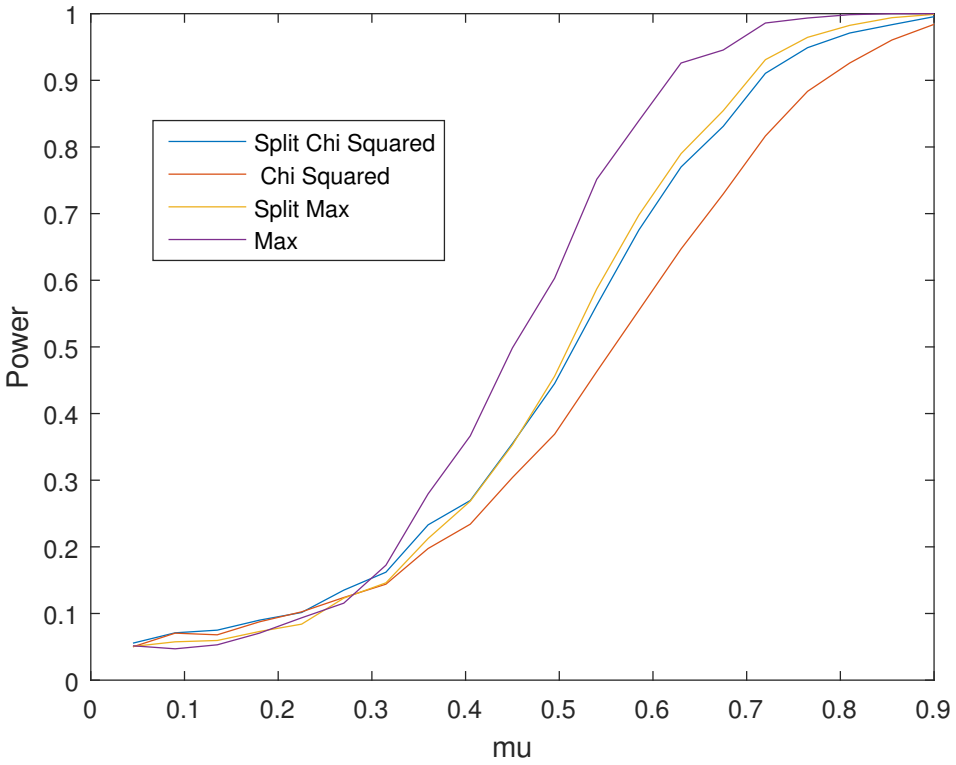

Figure S.2.10: Power with $p = 100$ normal populations with $q = 2$ having non-zero mean $\mu/\sqrt{n}$

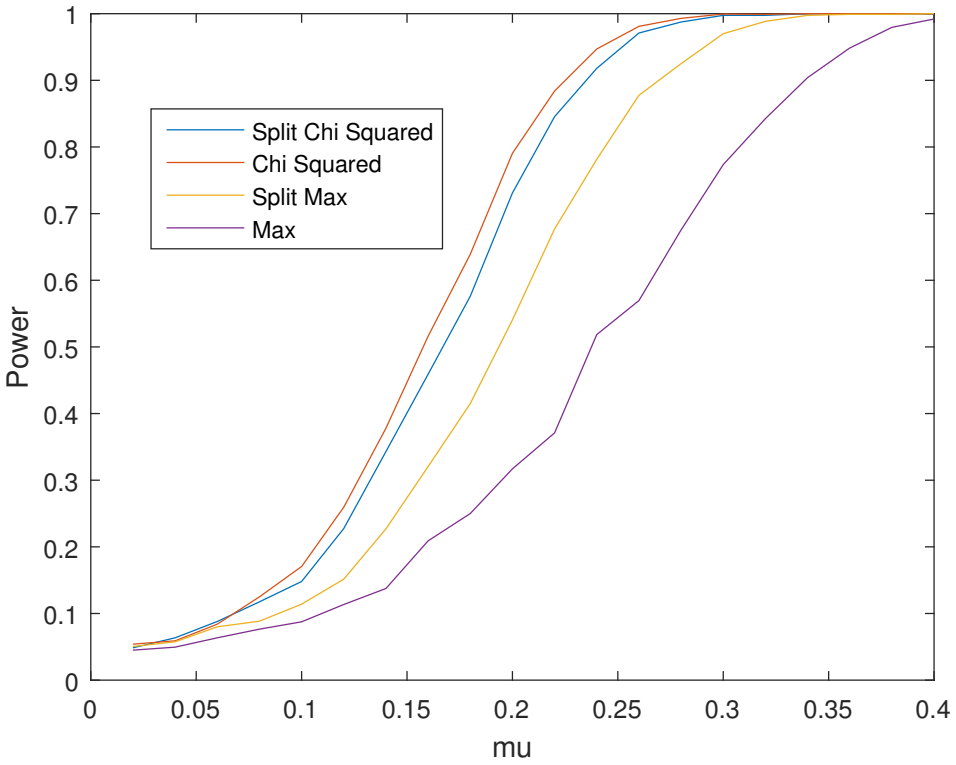

Figure S.2.11: Power with $p = 100$ normal populations with $q = 25$ having non-zero mean $\mu/\sqrt{n}$

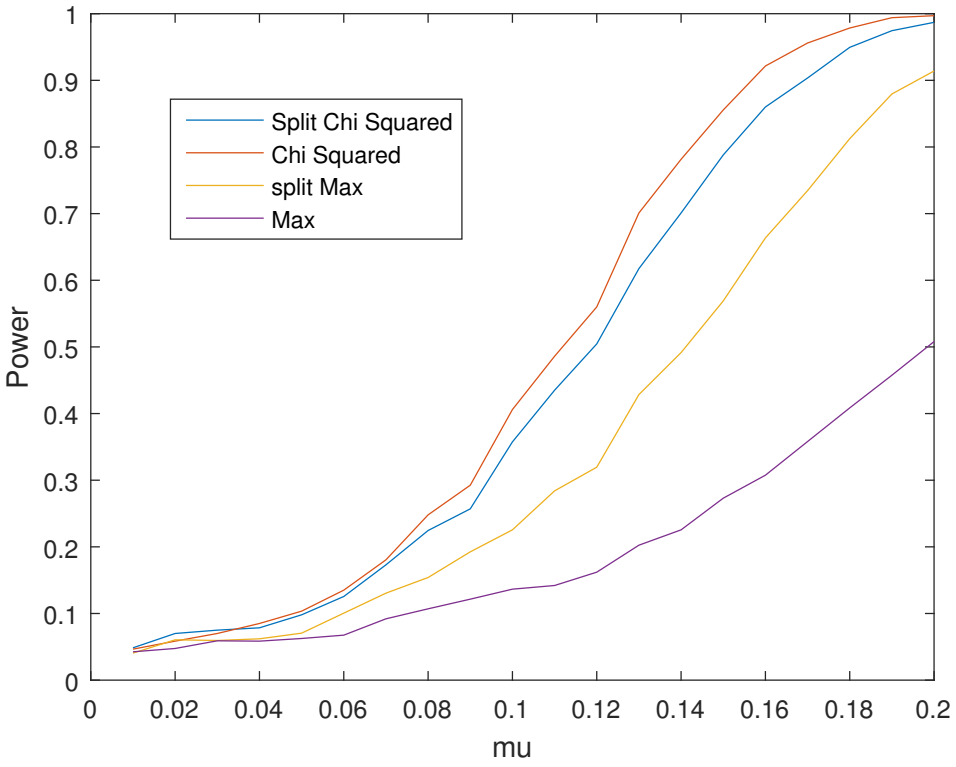

Figure S.2.12: Power with $p = 100$ normal populations with $q = 50$ having non-zero mean $\mu/\sqrt{n}$

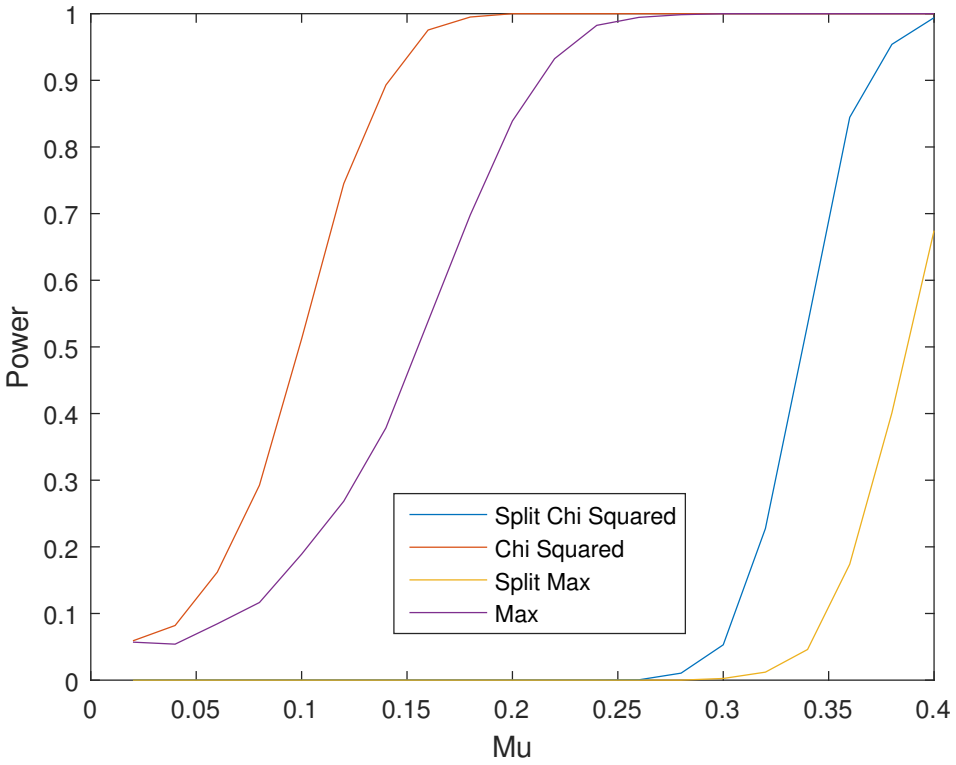

Figure S.2.13: Power with $p = 100$ normal populations with $q = 25$ having non-zero mean $\mu/\sqrt{n}$ using the conservative threshold

S.3 Proofs of Results in Supplement:

PROOF OF THEOREM S.1.1. The overall test rejects if the $(1-r)$th quantile of the $\sqrt{b}\bar{X}_{n,b,i}$ exceeds $z_{1-r\alpha}$. The test statistics based on each subsample can be written as

$$\sqrt{b}\bar{X}_{n,b,i} = (1-\tau)\frac{1}{\sqrt{b}}\left(\sum_{j\in S_{n,i}} X_j - \frac{b}{n-b}\sum_{j\in S_{n,i}^c} X_j\right) + \sqrt{b}\bar{X} \ .$$

Appealing to the results of **?**, the $(1-r)$th quantile of

$$\frac{1}{\sqrt{b}}\left(\sum_{j\in S_{n,i}} X_j - \frac{b}{n-b}\sum_{j\in S_{n,i}^c} X_j\right)$$

converges in probability to $z_{1-r}/\sqrt{1-\tau}$. Therefore, the test is asymptotically equivalent to rejecting if

$$\sqrt{1-\tau}z_{1-r} + \sqrt{b}\bar{X}$$

exceeds $z_{1-r\alpha}$. Consequently, the power satisfies

$$P\left(\sqrt{n}\bar{X} > \frac{1}{\sqrt{\tau}}\left(z_{1-r\alpha} - \sqrt{1-\tau}z_{1-r}\right)\right) \to 1 - \Phi\left(\frac{1}{\sqrt{\tau}}\left(z_{1-r\alpha} - \sqrt{1-\tau}z_{1-r}\right)\right) \ .$$

■

PROOF OF THEOREM S.1.2. The proof is fairly routine, using a Taylor approximation and noting

$$\frac{1}{M}\sum_{i=1}^{M}\phi\left(\sqrt{b}(\bar{X}_{n,b,i} - \frac{h}{\sqrt{n}})\right) \xrightarrow{P} E[\phi(Z)] = 1/\sqrt{4\pi} \ ,$$

where $Z \sim N(0,1)$. ■

PROOF OF THEOREM S.1.3. Note that

$$\Phi\left(\sqrt{b}\bar{X}(S_i)\right) = \Phi\left(\sqrt{n}\bar{X}\right) + \frac{\sqrt{b}-\sqrt{n}}{\sqrt{b}}\sqrt{b}\bar{X}(S_i) + \sum_{i\in S_i^c} + o_p\left(1/b\right)$$

Averaging these over all K splits gives the convergence result as $K \to \infty$.

■

PROOF OF THEOREM S.1.4. Since $\Phi^{-1}(\hat{p}_{n,m}) = \Phi^{-1}(1-\Phi(\sqrt{b}\bar{X}_{n,m})) = \Phi^{-1}(\Phi(-\sqrt{b}\bar{X}_{n,m})) = -\sqrt{b}\bar{X}_{n,m}$, the test rejects for large values of

$$\frac{1}{M}\sum_{i=1}^{M}\sqrt{b}\bar{X}_{n,i} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{\sqrt{b}}\sum_{i\in S_{n,m}} X_i$$

If the splits are chosen so that each of the X_i in an equal number of splits, then each X_i appears in the sum exactly Nb/n times. Therefore, the test rejects if

$$\frac{1}{M}\sum_{m=1}^{M}\Phi^{-1}(\hat{p}_{n,m}) = \sqrt{\frac{b}{n}}\sqrt{n}\bar{X}$$

exceeds $z_{1-\alpha}$. The power of the test is

$$1-\Phi\left(\frac{1}{\sqrt{\tau}}z_{1-\alpha}-h\right) \ .$$

If the splits are chosen so that each of the X_i in an equal number of splits, then each X_i appears in the sum exactly Nb/n times. Therefore, the test rejects if

$$\frac{1}{M}\sum_{m=1}^{M}\Phi^{-1}(\hat{p}_{n,m}) = \sqrt{\frac{b}{n}}\sqrt{n}\bar{X}$$

exceeds $z_{1-\alpha}$. The power of the test is

$$1-\Phi\left(\frac{1}{\sqrt{\tau}}z_{1-\alpha}-h\right) \ .$$

■

Proof of Theorem S.1.5. We apply Theorem 6.1, first in the case where the order of the kernel b is fixed (so $b=k$). Define the kernel

$$h_n(X_1,...,X_b) = 1-\Phi(\sqrt{b}\bar{X}_b) \ ,$$

which is the p-value of a test of H_0 computed on a subsample of size b. For this choice of kernel,

$$h_{1,b}(x) = 1-E\left(\Phi(\sqrt{b}\bar{X}_b)|X_1=x\right) = 1-E\Phi(\frac{x}{\sqrt{b}}+Y) \ ,$$

where $Y\sim N(0,(b-1)/b)$. So, we can simplify

$$h_{1,b}(x) = E[I\{Z<\frac{x}{\sqrt{b}}+Y\}] \ ,$$

where $Z\sim N(0,1)$ and Z is independent of Y. Therefore,

$$h_{1,b}(x) = 1-\Phi\left(\frac{x}{\sqrt{2b-1}}\right)$$

and

$$\zeta_{1,b} = Var\left[\Phi\left(\frac{X}{\sqrt{2b-1}}\right)\right] \ .$$

By Theorem 6.1, it follows that, under H_0,

$$\sqrt{n}(\bar{p}_n - \frac{1}{2}) = \frac{b}{\sqrt{n}} \sum_{i=1}^{n} [h_{1,b}(X_i) - \frac{1}{2}] + o_P(1) \tag{S.1}$$

and so

$$\sqrt{n}(\bar{p}_n - \frac{1}{2}) \xrightarrow{d} N(0, b^2\zeta_{1,b}) \ .$$

To calculate the limiting distribution under the sequence of alternatives when the mean is $h/\sqrt{n}$, note that by contiguity, the approximation (S.1) holds as well; that is, the term that goes to 0 in probability under $h = 0$ does so under general h as well. The linear term does not have mean 1/2, but we can calculate by a Taylor expansion argument (and noting that the moments in the error term are bounded) that

$$E_h[h_{1,b}(X)] = 1 - E\left[\Phi\left(\frac{Z + h/\sqrt{n}}{\sqrt{2b-1}}\right)\right] \ ,$$

where $Z \sim N(0,1)$. Then

$$E_h[h_{1,b}(X)] = \frac{1}{2} - \frac{h/\sqrt{n}}{\sqrt{2b-1}} E\left[\phi\left(\frac{Z}{\sqrt{2b-1}}\right)\right] + O(1/n) \ .$$

But, using that the moment generating function of Z^2 is $(1-2t)^{-1/2}$, one can calculate

$$E\left[\phi\left(\frac{Z}{\sqrt{2b-1}}\right)\right] = \frac{1}{\sqrt{2\pi}} \cdot (1 - \frac{1}{2b})^{1/2} \ ,$$

and so

$$E_h[h_{1,b}(X)] = \frac{1}{2} - \frac{h}{\sqrt{4\pi bn}} + O(1/n) \ .$$

Also, under $\mu = h/\sqrt{n}$,

$$Var_h[h_{1,b}(X)] = Var\left[\Phi\left(\frac{Z}{\sqrt{2b-1}} + \frac{h/\sqrt{n}}{\sqrt{2b-1}}\right)\right] = \zeta_{1,b} + o(n^{-1/2}) \ .$$

By (S.1) and these calculations, It follows that, under $h/\sqrt{n}$,

$$\sqrt{n}(\bar{p}_n - \frac{1}{2}) \xrightarrow{d} N(-\frac{\sqrt{b}h}{\sqrt{4\pi}}, b^2\zeta_{1,b}) \ .$$

It now trivially follows that the test that rejects if $\sqrt{n}(\bar{p}_n - \frac{1}{2}) < z_\alpha b\sqrt{\zeta_{1,b}}$ has limiting power or rejection probably under $h/\sqrt{n}$ given by

$$P_h\{\sqrt{n}(\bar{p}_n - \frac{1}{2}) < z_\alpha b\sqrt{\zeta_{1,b}}\} = 1 - \Phi\left(z_{1-\alpha} - \frac{h}{\sqrt{4\pi b\zeta_{1,b}}}\right) \ .$$

We now show $b\zeta_{1,b} \to (4\pi)^{-1}$ as $b \to \infty$. But,

$$b\zeta_{1,b} = bVar\left[\Phi\left(\frac{Z}{\sqrt{2b-1}}\right)\right] = bVar\left[\Phi(0) + \frac{Z}{\sqrt{2b-1}}\phi(0) + r_b\right] \ ,$$

where the error term can be ignored because it has a variance of order $1/b^2$. Hence,

$$b\zeta_{1,b} = b\frac{1}{2\pi(2b-1)} + o(1) \to 1/4\pi \ \ .$$

Thus, as $b \to \infty$, the limiting power tends $1 - \Phi(z_{1-\alpha} - h)$, the same as the UMP test.

In the case $b \to \infty$ at the same time $n \to \infty$, we can just apply Theorem 6.1 along with the same calculations for fixed b. ■

Proof of Remark S.1.2. Define $\sigma_b^2 = \text{var}\left(E(1 - \Phi(\sqrt{b}\bar{X}_b)|X_1)\right)$, the variance of the average of the p-values. Under the sequence of local alternative, $h/\sqrt{n}$,

$$\sqrt{n}\frac{1}{N}\frac{\sum \hat{p}_i - 1/2}{b^2\sigma_b} \to N(h(\phi(0)/\sqrt{2})/(b\sigma_b), 1) \ .$$

Therefore, the limiting power is $1 - \Phi\left(z_{1-\alpha} - h(/\sqrt{4\pi}(b\sigma_b)\right)$. Note that since the distribution of $\sqrt{b}\bar{X}_b$, conditional on X_i is $N(X_1/\sqrt{b}, (b-1)/b)$,

$$E\left(1 - \Phi(\sqrt{b}\bar{X}_b)|X_1\right) = P\left(Z_1 > Z_2\sqrt{(b-1)/b} + X_1/\sqrt{b}|X_1\right) = 1 - \Phi\left(X_1/\sqrt{2b-1}\right) \ ,$$

where Z_1 and Z_2 are independent standard normal random variables. Consequently,

$$b^2\sigma_b^2 = b^2Var\left(1 - \Phi\left(X_1/\sqrt{2b-1}\right)\right) \to 1/4\pi \ .$$

Therefore, as $b \to \infty$, the limiting power tends to that of the UMP test. ■

Proof of Theorem S.1.6. Here we follow the notation of Theorem 6.2 with $k = b$ and

$$h_b(X_1, ..., X_b; t) = I\left\{1 - \Phi(\sqrt{b}\bar{X}_b) > \tilde{\theta}_b + t\right\} \ .$$

Then, $\tilde{\theta}_k$ is the distribution of h_b under $h/\sqrt{n}$, or the median of the distribution of $1 - \Phi(Z + h\sqrt{b/n})$ when Z is standard normal. Thus, a trivial calculation gives $\tilde{\theta}_k = 1 - \Phi(h\sqrt{b/n})$. Then,

$$\phi_{1,b}(x; t) = E[h_b(x, X_2, ..., X_b); t] \ ,$$

and

$$\tilde{\zeta}_{1,b}(t) = \text{Var}[\phi_{1,b}(X; t)] \ .$$

Now,

$$\phi_{1,b}(x; t) = P_h\{1 - \Phi(\sqrt{b}\bar{X}_b) > 1 - \Phi(h\sqrt{b/n}) + t\}$$

$$= P\{\Phi(Y + x/\sqrt{b}) < \Phi(h\sqrt{b/n}) - t\} = P\{Y + x/\sqrt{b} < \Phi^{-1}[\Phi(h\sqrt{b/n}) - t]\} \ ,$$

where Y in normal with mean $(b-1)h/\sqrt{nb}$ and variance $(b-1)/b$. Hence,

$$\phi_{1,b}(x;t) = \Phi\left[\frac{\Phi^{-1}[\Phi(h\sqrt{b/n}) - t] - x/\sqrt{b} - (b-1)h/\sqrt{bn}}{\sqrt{(b-1)/b}}\right]$$

Assume the null hypothesis $h = 0$, in which case $\tilde{\theta}_k = 1/2$. In this case,

$$\phi_{1,n}(x;0) = 1 - \Phi\left(\sqrt{\frac{b}{b-1}}\frac{x}{\sqrt{b}}\right)$$

$$= \frac{1}{2} - \sqrt{\frac{b}{b-1}}\frac{x}{\sqrt{b}}\phi(0) + o\left(1/b\right) \ .$$

and so

$$\frac{\tilde{\zeta}_{1,b}(0)}{(\phi(0))^2/b} \to 1$$

as $b \to \infty$. Similarly, one can show that

$$\zeta_{1,b}(t) = \frac{\phi^2(z_{\frac{1}{2}-t})}{b} + o(1/b) \ ,$$

and so the conditions of Theorem 6.2 are met.

Therefore, we have that, under the null hypothesis

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

Under the sequence of local alternatives, $\mu = h/\sqrt{n}$, the median $\tilde{\theta}_b$ is given by

$$\tilde{\theta}_b = 1 - \Phi\left(h\sqrt{b/n}\right) = 1/2 + \phi(0)h\sqrt{b/n} + o_p(1/\sqrt{n}) \ .$$

By similar arguments, the limiting local power of the test based on the median p-value is

$$1 - \Phi\left(z_{1-\alpha} - h\right) \quad . \ \blacksquare$$